Study on Efficient Prior Control for Realizing Practical Systems of Speech Recognition

(音声認識の実用化のための
効率的な事前制御技術に関する研究)

Graduate School of Information Science and Technology
The University of Tokyo

Satoshi Kobashikawa
小橋川 哲

ABSTRACT

Due to the continual improvements in computer resources on the cloud and smart devices, applications based on speech recognition technologies are becoming more widely used. However, recognition accuracy degraded significantly if the speech is noisy, captured in real environments, or spontaneous, containing ambiguous utterances; both problems are barriers to the practical application of speech recognition. This paper provides useful countermeasures in the form of efficient prior control schemes by leveraging the attributes of practical scenarios.

This research assumes two practical usage targets, a) noise robust speech recognition on tablet devices, and b) spontaneous speech recognition for contact centers and parliament bodies. This work deals with three situations of speech application; i) speech interface for tablet devices, ii) speech mining in contact centers, and iii) speech transcription for parliamentary meetings. We develop, for the first situation, i.e. tablet devices, 1) acoustic model adaptation and normalization using pre-observed noise. For the second situation, i.e. contact center, we develop 2) a fast unsupervised adaptation technique using frame-independent confidence scores, 3) a data selection technique using prior confidence, 4) a recognition time stabilization technique using prior beam width control. We also develop 5) fast acoustic pre-processing for the transcription of parliament meetings, third situation.

To tackle the variation in S/N and convolutional noise expected with tablet devices, we develop acoustic model adaptation and normalization using pre-observed noise. This technique assumes that the background noise is relatively stationary and can be captured. It offers robust speech recognition under a wide range of S/N and convolutional noise; the noise captured prior to speech recognition allows noise reduction through the techniques of spectral subtraction (SS), additive noise adaptation using parallel model combination (PMC), and convolutional noise normalization using cepstral mean normalization (CMN).

To improve accuracy under the constraint of a recognition time limit, often seen in contact centers, we develop a fast unsupervised adaptation technique based on frame-independent confidence scores. This technique leverages the property that the target is stored speech. It uses a limited number of Gaussian mixture models (GMMs) for the target speech in a preliminary step before speech recognition, and then improves accuracy rapidly by gender selection and of the application of maximum likelihood linear regression (MLLR).

For contact centers, we develop a data selection technique with prior confidence estimation to reduce the cost of processing by dropping low confidence speech data; if such data is processed it is likely to disrupt the subsequent text mining functions and will eventually be rejected. The property of this technique is that massive volumes of data are stored and low confidence recognition results

are unnecessary. It estimates prior confidence scores rapidly by using a limited number of GMMs and selects only high prior confidence data for speech recognition.

Furthermore, for contact centers, we develop a prior beam width control technique to reduce the time wasted in processing low quality speech data that should be rejected. This technique also assumes that the target is massive volumes of stored speech. It rapidly estimates prior scores by using a limited number of GMMs, and stabilizes the computation time by controlling the search space spread in decoding.

In addition, we also develop a fast acoustic pre-processing technique that can well handle changes in the recording environment and speaker to realize a parliamentary meeting transcription system. We can leverage the property that pre-processing is available since incoming parliament speech is segmented and temporarily stored in caches. This technique achieves high accuracy, even if computation time constraints are imposed, by combining four fast acoustic pre-processing methods of channel selection, speaker indexing, feature parameter normalization, and unsupervised adaptation.

All five proposed techniques are based on the prior control approach and leverage the properties of practical speech recognition applications; they provide significant benefits over conventional speech recognition schemes.

ABSTRACT IN JAPANESE

近年、高性能のクラウド上の計算機とスマートフォンの普及等により音声検索等の音声認識応用アプリケーションが一般に使われ始めている．しかしながら，背景雑音が混入する環境や曖昧な発声を含む話し言葉音声に対しては，ベースとなる音声認識精度は大幅に劣化するため，実用化のための障壁となっている．そこで，本研究では，実用化のために生じる課題への対応するため，音声認識の利用場面に応じて活用できる前提条件に基づく事前制御技術を導入する事で，通常の音声認識処理では得られない効果を得る事を目的とする．

本研究では，a) タブレット端末利用時における耐雑音音声認識，b) コンタクトセンタや議会における話し言葉音声認識，の 2 つの大きな研究課題を対象とする．音声認識の利用場面としては，i) タブレット端末上での音声インタフェース，ii) コンタクトセンタ音声マイニング，iii) 議会音声書き起こし支援，の 3 つを想定する．ここで，タブレット端末上の音声インタフェースに対しては，高いレスポンスのもとで，雑音耐性の強化が必要になる．また，コンタクトセンタ音声マイニングに対しては，大量の蓄積通話音声データから高い認識精度な音声ドキュメントを収集する事が求められる．議会音声書き起こし支援に対しては，限られた処理時間制約のもので変動する話者・環境に対して高い認識精度が求められる．

まず，タブレット端末利用時では，S/N 比の変動および空間伝達特性の影響に対応するため,「1) 事前観測雑音を用いたモデル適応と正規化」技術を開発した．本技術は，定常的な背景雑音は事前に観測が可能という前提を置き，事前観測した加法性雑音のみを用いて，スペクトルサブトラクション (SS: Spectral Subtraction) による加法性雑音抑圧手法，モデル合成法 (PMC: Parallel Model Combination) による加法性雑音適応手法，ケプストラム正規化 (CMN: Cepstral Mean Normalization) による乗法性雑音正規化手法を融合させて，S/N の変動や空間伝達特性である乗法性雑音 (歪み) に強い音声認識方式である．

次に，コンタクトセンタでは，限られた処理時間要件の下で精度向上に貢献する「2) フレーム独立信頼度を用いた事前高速教師なし適応」技術を開発した．本技術は，低遅延のリアルタイム処理は必須ではない蓄積された音声という前提を置き，対象音声データを音響モデル中の限られた GMM(Gaussian Mixture Model) を用いる事により，最尤線形回帰法 (MLLR: Maximum Likelihood Linear Regression) と性別選択を融合させた高速に精度向上ができる音声認識方式である．

また，コンタクトセンタでは，信頼度が低く後段のテキストマイニング処理に悪影響を及ぼすため棄却される音声データの音声認識処理に過剰な計算コストを大幅に削減する「3) 事前信頼度によるデータ選択」技術を開発した．本技術は，大量音声データが蓄積されている前提を置き，低精度の音声認識結果は不要であることから，音響モデル中の限られた GMM を用いて高速に信頼度を推定して，精度が高い音声のみを音声認識対象とする方式である．

さらに，コンタクトセンタ向けに認識精度が低く後段で棄却すべき低音質の音声データ

に対する音声認識処理の計算コストを削減する「4) 事前ビーム幅制御による認識処理時間安定化」を開発した．本技術は，大量音声データの中に含まれる低精度・低品質音声には計算量が無駄にかけなくて良いという前提を置いて，対象音声データを音響モデル中の限られたGMMを用いて予め高速に走査して得られた事前スコアから，音声認識処理中の探索空間の広がりを制御する事で，音声認識処理時間を安定化させる方式である．

加えて，議会録作成支援のための音声認識システムの実用化に向けて，議会場の収音環境および話者・環境の変動に対応するための「5) 議会録作成支援のための高速事前音響処理」を開発した．本技術は，対象の議会音声が区分化された単位で蓄積されて送られてくる前提を活かし，チャネル選択，話者インデクシング，特徴量正規化，教師なし適応といった複数の事前音響処理を高速に行う事で，限られた計算時間の下で，変動する音声に対して高い音声認識精度を実現する方式である．

以上，5つの方式はいずれも音声認識の利用場面に即して活用できる前提を置くことで導入可能な事前制御に関する処理であり，従来の音声認識方式では得られない効果を得る方式である．

Contents

ABSTRACT **i**

ABSTRACT IN JAPANESE **iii**

CONTENTS **vii**

LIST OF FIGURES **x**

LIST OF TABLES **xi**

1 Introduction **1**
1.1 Background 1
1.2 Goal 2
1.3 Overview 4

2 Proposed Model Adaptation and Normalization Using Pre-Observed Noise **7**
2.1 Introduction 7
2.1.1 Problem of low S/N ratio 8
2.1.2 Problem of changes in S/N ratio 8
2.1.3 Problem of changes in H (transfer characteristics) 9
2.2 Model adaptation and normalization using pre-observed noise 9
2.2.1 Additive noise reduction for the input signal 9
2.2.2 Additive noise adaptation for acoustic model 10
2.2.3 Multi-S/N adaptation 10
2.2.4 Convolutional noise normalization for acoustic model 11
2.2.5 Convolutional noise normalization for input signal 12
2.3 Experiments 13
2.3.1 Experimental condition and task 13
2.3.2 Experimental results 14
2.4 Summary 17

3 Fast Unsupervised Adaptation Based on Efficient Statistics Accumulation Using Frame Independent Confidence within Monophone States **21**
3.1 Introduction 21

3.2 Related work on unsupervised adaptation . . . 22
3.3 Proposed rapid unsupervised adaptation technique . . . 23
3.3.1 Illustration of statistics accumulation . . . 23
3.3.2 Unsupervised adaptation with monophone constraint and power utilization 26
3.3.3 Gender selection with utterance segmentation . . . 28
3.3.4 Framework of proposed system . . . 29
3.4 Experiments . . . 30
3.4.1 Experimental settings . . . 30
3.4.2 Experimental results and discussion . . . 31
3.4.3 Experimental results and discussion for gender selection . . . 31
3.4.4 Experimental results and discussion for unsupervised adaptation . . . 32
3.5 Summary . . . 33

4 Efficient Data Selection for Speech Recognition Based on Prior Confidence Estimation Using Speech and Monophone Models 35
4.1 Introduction . . . 35
4.2 Related work on data selection for speech recognition and its application . . . 36
4.3 Proposed data selection based on prior confidence estimation . . . 37
4.3.1 Formulation of prior confidence estimation . . . 37
4.3.2 Qualitative explanation of prior confidence estimation . . . 39
4.3.3 Procedure of proposed system . . . 40
4.4 Experiments . . . 42
4.4.1 Experimental condition and task . . . 42
4.4.2 Results . . . 44
4.4.3 Discussion . . . 46
4.5 Summary . . . 47

5 Efficient Beam Width Control to Eliminate Excessive Speech Recognition Time Based on Score Range Estimation 49
5.1 Introduction . . . 49
5.2 Related work on decoding parameter control . . . 50
5.3 Proposed beam width control based on score range estimation . . . 51
5.3.1 Framework of proposed approach . . . 51
5.3.2 Formulation of proposed approach . . . 52
5.3.3 Comparison of proposed and existing techniques . . . 57
5.4 Experiments . . . 58
5.4.1 Experimental settings . . . 58
5.4.2 Results and discussions . . . 59
5.5 Summary . . . 61

6 Fast Acoustic Pre-processing against Recording Environment and Speaker Changes for Parliamentary Meeting Transcription **63**
6.1 Introduction 63
6.2 Proposed system 65
6.2.1 Channel selection 65
6.2.2 Speaker indexing 66
6.2.3 Acoustic feature normalization 67
6.2.4 Unsupervised acoustic model adaptation 68
6.3 Experiments 69
6.3.1 Experimental setting 69
6.3.2 Experimental results 70
6.4 Summary 71

7 Conclustion **73**
7.1 Preview of work 73
7.2 Summary 76

ACKNOWLEDGMENTS **77**

ACKNOWLEDGMENTS IN JAPANESE **79**

BIBLIOGRAPHY **81**

LIST OF WORK **91**

List of Figures

1.1 *Positioning study targets in speech recognition* 2
1.2 *Conventional speech recognition framework* 3
1.3 *Proposed speech recognition framework* 3
1.4 *Overview of this work* 6

2.1 *Proposed system.* 10
2.2 *Recording condition.* 18
2.3 *Impulse response at the position of 70 cm.* 19
2.4 *Recognition correct rate versus noise type at the position of 50 cm and 0 degree.* . . 19
2.5 *Recognition correct rate versus the position with PC fan noise.* 20
2.6 *Recognition correct rate versus the position with all noises.* 20

3.1 *Relation between (a) forward-backward algorithm, (b) Viterbi algorithm and (c) proposed frame independent statistics accumulation.* 26
3.2 *State sequence used in gender selection with utterance selection.* 28
3.3 *Framework of proposed system.* 30

4.1 *Relational expression from conventional to proposed prior confidence measure.* . . 40
4.2 *The log-likelihood difference between best monophone (ex. /a/) and speech model.* . 41
4.3 *Proposed system.* 41
4.4 *Prior confidence estimation process.* 42
4.5 *Speech data selection performance in terms of recognition rate.* 45
4.6 *Speech data selection performance in terms of raw Average Precision (AP).* 46

5.1 Schematic diagram of proposed system 52
5.2 *Log-likelihood distribution.* 53
5.3 *Log-likelihood distribution with swapped vertical and horizontal axes.* 54
5.4 *Comparison of log-likelihood distributions of base and target speech.* 55
5.5 *Proposed incremental recognition time control.* 57
5.6 *Computation time of proposed technique.* 60

6.1 *Recording environment.* 64
6.2 *Proposed system.* 64
6.3 *Explanation of channel selection.* 66
6.4 *Process flowchart of speaker indexing.* 67

6.5 *Best state sequence with monophone constraint.* 68

List of Tables

1.1 The required constraint condition in each use case 4
1.2 The issues and properties available in each use case 5

2.1 Speech analysis conditions 14
2.2 Acoustic model conditions 14
2.3 Training data 14
2.4 Evaluation task 15
2.5 Comparative techniques 15

3.1 Comparison of (a) forward-backward algorithm, (b) Viterbi algorithm and (c) proposed frame independent statistics accumulation. 27
3.2 Performance of compared techniques regarding gender selection with utterance segmentation. 31
3.3 Compared techniques regarding unsupervised adaptation. 33
3.4 Performance of unsupervised adaptation without power utilization. 33
3.5 Effect of power utilization in unsupervised adaptation 33

4.1 Spoken document retrieval performance in terms of Mean Average Precision (MAP). 45
4.2 Computation time for confidence estimation. 45
4.3 Effect of proposed speech data selection. 46

5.1 *Recognition rate of proposed technique.* 59
5.2 *Performance of the proposed technique for speech recorded in an actual call center.* 60

6.1 Performance of speech recognition 69
6.2 Computation time of speech recognition 71

7.1 The five issues and property in each use case 75

Chapter 1

Introduction

1.1 Background

Speech information processing is one of the most important topics in the human interface research field. Speech recognition is the key to developing highly effective human interface applications. Speech recognition studies started over fifty years ago [1, 2], and speech recognition techniques have evolved dramatically in the last decade. Recently, due to the enhanced performance of computer resources, broadband wireless networks, and the popularization of smart devices, speech recognition applications are coming into general use such as voice search [3, 4] by Google, Siri virtual personal assistant [5] by Nuance, and Shabette Concier (in japanese) voice-agent application by NTT docomo [6][7]. Other speech recognition technologies for human-aided transcription are also in practical use; examples include a broadcast TV closed-captioning service [8] and a meeting transcription service for the Japanese parliament (Diet) [9].

Speech recognition applications and services are being launched continuously but performance still needs to be improved. Recognition errors are a significant problem when the speech overlays background noise in the distant-talking situation, e.g. smart tablet device. While situations that permit formal and reading-style speech can achieve high accuracy, informal and spontaneous speech generate too many recognition errors; e.g. conversational speech between humans in call centers or parliamentary meetings. These recognition errors become a big barrier to really practical applications. This paper attempts to overcome this barrier and accelerate the spread of speech recognition services.

This work focuses on two practical targets that have wide applicability; a) speech captured by distant-talking microphone (e.g. tablet device) with background noise , and b) natural spontaneous conversational speech between humans (e.g. call center speech or parliamentary discussion). Speech recognition accuracy is degraded significantly in these practical targets; noise and ambiguity in speech are major barriers to the realization of practical speech recognition applications. The research targets are positioned in Figure 1.1. 1st target, i.e., tablet devices, was a far-sighted research goal that predicted the recent popularity of smart devices. The 2nd target is spontaneous speech at the call center or in parliament; there are no truly practical systems since it remains difficult to recognize spontaneous speech accurately.

At that time we focused on tablet devices, the devices' penetration rate at home is definitely not the same at the present days; there were few studies to tackle the home noise by using tablet devices. Furthermore, microphone array techniques are often used to deal with distant-talking speech, and it is more difficult to recognize the distant-talking speech by using internal microphones of the tablet devices.

When we started this work for spontaneous speech, a large-scale national project named entitled "Spontaneous Speech Corpus and Processing Technology" was conducted [10]. Although the project constructed a large-scale spontaneous spontaneous speech corpus, the Corpus of Spontaneous Japanese (CSJ), the corpus consists mainly of monolog speech as presentation [11]. Since the dialog speech in call center is conversational and so more spontaneous than monolog presentation, it is more difficult to recognize the dialog spontaneous speech accurately.

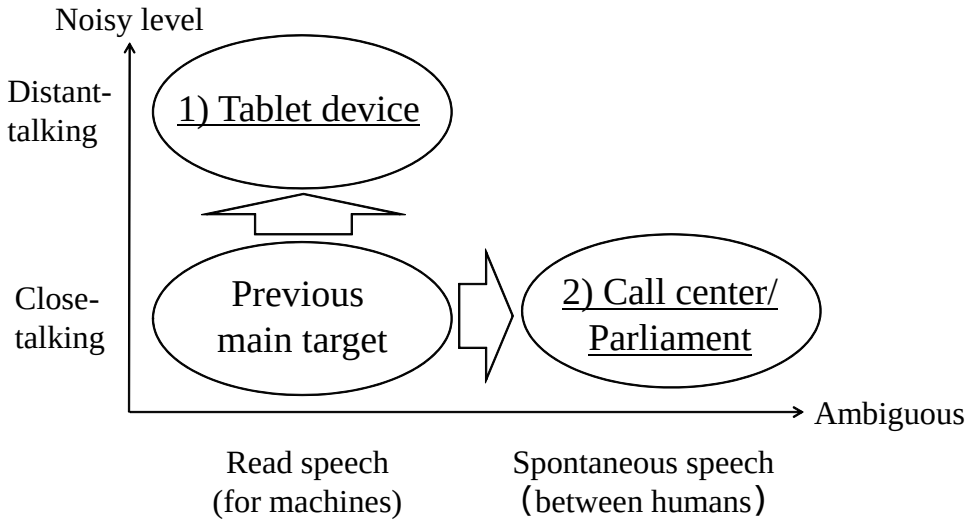

Figure 1.1: *Positioning study targets in speech recognition*

1.2 Goal

One goal of this work is to provide the desired performance, which can not be obtained by the conventional framework, for realizing practical applications based on speech recognition. The approach adopted here is to introduce efficient prior control techniques by leveraging the properties inherent to the applications .

The conventional speech recognition framework is shown in Figure 1.2. This framework consists of 4 components; acoustic analysis, decoder, acoustic model, and language model. Component parameters are basically adapted for each use case by collecting training data. This incurs high production costs since a lot of training data is required to overcome the different problems

encountered in speech recognition; e.g. to achieve sufficient recognition accuracy given limited computer resources.

The framework of this work is shown in Figure 1.3. It offers an additional component, prior control. The component offers an efficient prior process that depends on the available properties of in each practical use case. The prior process alsocontrols subsequent speech recognition components.

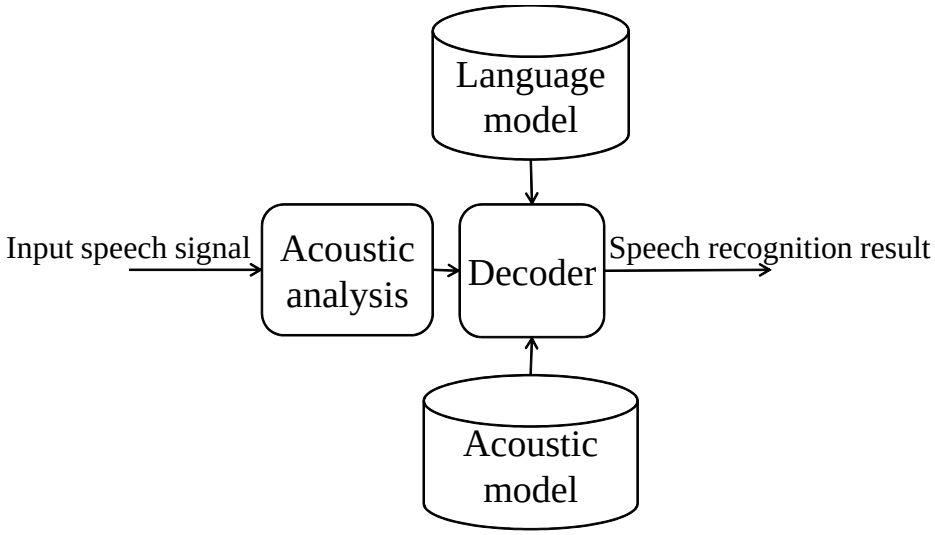

Figure 1.2: *Conventional speech recognition framework*

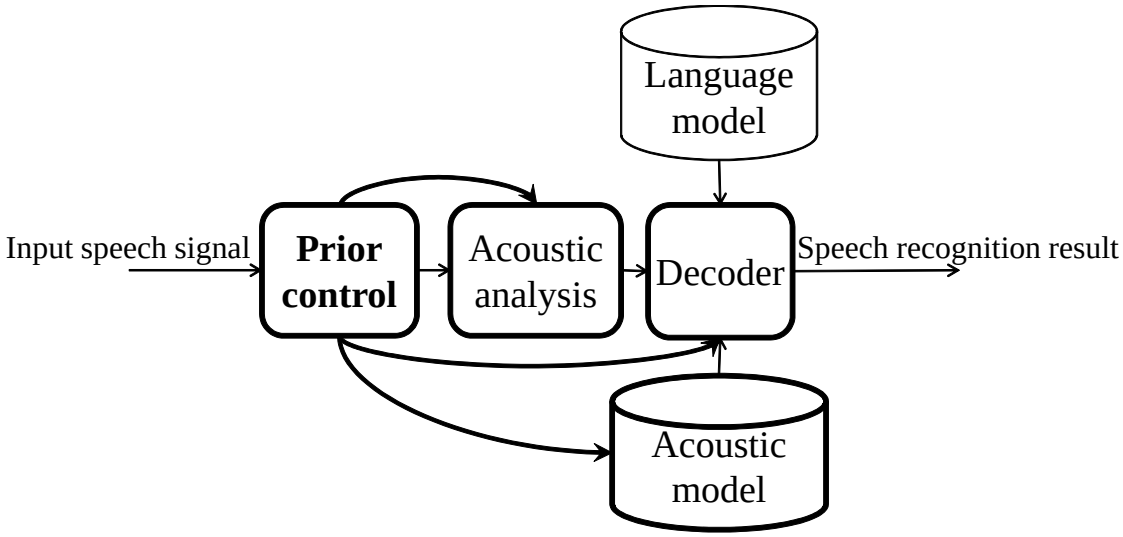

Figure 1.3: *Proposed speech recognition framework*

This work examines use cases of three types; i) speech recognition interface for tablet devices, ii) information extraction from speeches stored in call centers, and 3) transcription support system for parliament speeches. Each use case exhibits a different problem that must be solved if we are to realize practical applications. Table 1.1 shows the required constraint condition in each use case.

Table 1.1: The required constraint condition in each use case

Use case	Response / speed	Offline	Completeness
Tablet device (Speech interface)	High response	Offline for noise	Reject noise data
Call center (Speech mining)	High speed	Offline	Focus on high accurate data
Parliament speech (Speech transcription)	High speed	Offline for short segments	Complete all data

The proposed techniques leverage the available property of each use case. Table 1.2 shows the issues and the the properties focused on in this work. i) Higher accuracy and better response are required to counter background and convolutional noises for tablet device interfaces. ii) Highly accurate spoken documents should be output under limited computer resources from massive volumes of speech data in call centers. iii) High recognition accuracy is required within strictcomputation time limits.

1.3 Overview

This subsection provides an overview of the work in this thesis (Figure 1.4). The proposed prior control techniques leverage the available properties to resolve the issues present in each use case. This thesis, in the following chapters, details the speech recognition experiments conducted to confirm the effectiveness of the proposed prior control techniques .

- Chapter 2 describes the proposed model adaptation and normalization technique that uses the pre-recognition noise for tablet device interfaces. This proposed technique improves the recognition accuracy against convolutional and additive noises by using pre-recognition background noise.

Table 1.2: The issues and properties available in each use case

Use case	Issue	Property
Tablet device	Higher accuracy and better response under noise	Background noise can be pre-observed
Call center	Higher accurate spoken documents under limited computer resources	Low latency is not required for stored speech samples
		Speech samples yielding low recognition quality can be omitted given the massive amounts of data available
		Low quality speech does not need to be recognized carefully
Parliament speech	High accuracy strict time limits	Prior normalization and adaptation are available for the stored speech

The following three chapters are aimed at information extraction from the massive amounts of speech data stored in call centers.

- Chapter 3 describes a fast unsupervised adaptation technique based on efficient statistics accumulation using frame-independent confidence scores within monophone states. This proposed technique improves the recognition accuracy with no increase in computation requirements by using fast prior unsupervised adaptation for the target speech before recognition.
- Chapter 4 describes an efficient data selection technique for speech recognition based on prior confidence estimation using speech and monophone models. This proposed technique estimates prior confidence scores rapidly, and selects high confidence speech data, data that can be expected to yield highly accurate speech recognition results from massive volumes of data.
- Chapter 5 describes efficient beam width control to eliminate excessive speech recognition time through the use of score range estimation. This proposed technique controls search beam width prior to decoding, since low quality speech shouldn't be recognize carefully.
- Chapter 6 describes the proposed fast acoustic pre-processing technique against that can handle changes in the recording environment and the speaker for parliamentary meeting

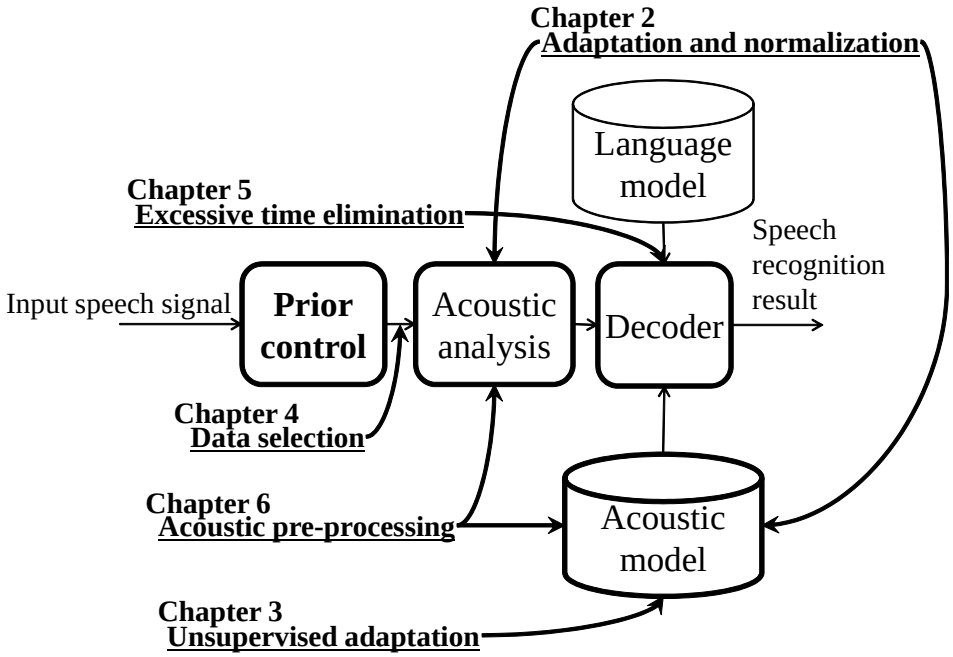

Figure 1.4: *Overview of this work*

transcription. The proposed technique can utilize several pre-processing methods to achieve the desired accuracy within a computation time limit, since the pre-processing approach is available for the cached incoming speech.

Finally, Chapter 7 reviews this thesis and discusses the realization of practical speech recognition systems.

Chapter 2

Proposed Model Adaptation and Normalization Using Pre-Observed Noise

2.1 Introduction

Business is demanding more effective speech dialog systems with the emphasis being placed on rapid response. Early research mainly focused on robustness. Model adaptation techniques [12][13][14][15] to counter noise were developed. Robustness is achieved by the adaptation processing provided by these techniques. Focusing on just accuracy is not sufficient since practical systems also need rapid response to satisfy the user requirements. Higher system response speeds are essential.

Speech recognition accuracy is often degraded by convolutional and additive noises [16]. While the former can be offset by CMN (cepstral mean normalization) [17], HMM-composition, called as NOVO (noise and voice composition) [12] or PMC (parallel model combination) [13], is used to offset the latter. These techniques, both which model the convolutional and additive noise characteristics, are adopted by CMS/PMC [14] and E-CMN/PMC [15]. Recognition accuracy can be increased by model adaptation: the model is optimized to match the instantaneous noise characteristics. These conventional techniques, CMS/PMC and E-CMN/PMC, first adapt the cepstral mean parameters of the acoustic model using the user's utterances against convolutional noise, and so must acquire user speech samples to initialize the speech recognition system. After that, compensation is followed by HMM-composition against the additive noise, which adapts to the observed additive noise but further delays the system's response to the user. Conventional techniques can not achieve adequate response speeds because model compensation and adaptation are performed only after the speech sample is received.

In real situation, the speech recognition system is able to observe the additive noise when user doesn't use the application. Focus on this point, our model generation strategy uses only the additive noise observed before user start to utter. Our technique doesn't have to wait to capture the user's speech sample, therefore can achieve high response. One technique, named NOVO+CMN [18], achieves faster response by this strategy. Model generation, realized by HMM-composition and CMN, is performed intermittently using the additive noise observed by the system. Since noise is effectively constant over short periods in real applications, the results of noise adaptation

are valid and are expected to achieve high performance. This means that after the speaker's speech sample is captured; only CMN need be performed to start recognition processing. In another advance, we create several HMMs to cover the wide S/N range expected in real world applications, because the S/N value can not be known without the user's speech sample. Furthermore, we use an additive noise reduction method like SS (Spectral Subtraction) at the front end of our NOVO+CMN technique to raise the S/N. Simulations show that the technique proposed herein, called SS-NOVO+CMN, achieves better recognition accuracy than the basic methods. The proposed technique is far more practical than either CMS/PMC or E-CMN/PMC since it eliminates the delay imposed by performing model adaptation after the speech sample is received. This paper reports experiments made using the multi-speaker dictation task.

on accuracy is often lost due to convolutional noise and additive noise. There are three main sources of problems with these applications 2.1.1) low Speech/Noise (S/N) ratios, 2.1.2) changes in S/N, and 2.1.3) changes in transfer characteristics (H) between the microphone and the user's mouth.

2.1.1 Problem of low S/N ratio

In the distant talking situation, as the distance between the speaker's mouth and the microphone increases, ambient noise increases and the S/N decreases. The HMM-composition method, NOVO [12] and PMC [13], are well-known noise adaptation methods that can improve speech recognition performance in noisy environments. However, the recognition performance of noise adaptation methods like NOVO, is not sufficient if the S/N is too low, because the speech features are buried in the noise. Methods to raise the S/N of the observed signals by using noise reduction such as the SS method [19] and the Wiener filter (WF) method [20] can be used. These methods, however, are not able to remove noise completely, and they create new problems in that insufficient or excessive reduction processing leads to remaining noise or speech distortion, respectively. One technique used noise reduction to raise the S/N and then noise adaptation to compensate the remaining additive noise [21]. In this paper, we use SS at the front end of this system and adapt the remaining noise by NOVO against the low S/N ratio problem.

2.1.2 Problem of changes in S/N ratio

In the real world, ambient noise level changes occasionally. Even if the ambient noise level remains fixed, the speech level picked up at the microphone depends on the loudness of the speaker's voice, the words uttered, and the position of the speaker in relation to the microphone; thus the S/N ratio changes often and widely. To overcome this problem, we proposed the use of several acoustic models formed under various S/N conditions [21]; speech recognition processing is carried out in parallel using these models and the output of the best performing model, the one with the highest likelihood, is selected. In this paper, we adopt this multi-S/N approach to counter the variation in S/N ratio.

2.1.3 Problem of changes in H (transfer characteristics)

The convolutional noise, created by the space transfer characteristics (H) between the microphone and the user's mouth, changes often in the distant talking situation on a Tablet PC or PDA. CMN [17] is commonly used to counter convolutional noise. Against this problem, we proposed the NOVO+CMN technique [18].

2.2 Model adaptation and normalization using pre-observed noise

Figure 2.1 shows the framework of SS-NOVO+CMN, the system proposed here. The important point of our system is that, we capture only the additive noise (non-utterance) for model generation in offline step. The system puts SS at the front end to counter low S/N values. The system generates several acoustic models for various S/N values against additive noise and normalizes cepstral mean parameters of the noise adapted acoustic models against convolution noise in an offline process. In an online process, it applies the cepstral mean normalization method to the input signal features and selects the recognition result from the acoustic model with highest likelihood S/N value.

Note: An additive noise reduction method and additive noise adaptation method are described in Section 2.2.1 and 2.2.2, respectively. The method of generating several S/N noise adapted models to handle the variation in S/N is described in Section 2.2.3. Section 2.2.4 and 2.2.5 explain convolutional noise normalization for the acoustic model and for the input signal, respectively.

2.2.1 Additive noise reduction for the input signal

In this first step, we use SS [19] to reduce the additive noise and so raise the S/N. In the SS method, $|O|^2$ is the untreated input power spectrum and $|N|^2$ is the observed noise power spectrum. $|\tilde{N}|^2$ is estimated noise power spectrum and is held constant. $|\tilde{N}|^2$ is estimated by the noise observed in non-utterance regions. The estimated noise reduced power spectrum $|\tilde{S}|^2$ and the remaining noise power spectrum $|\tilde{N}_r|^2$ are given by

$$\begin{aligned} |\tilde{S}|^2 &= \max\{|O|^2 - \alpha|\tilde{N}|^2, f|O|^2\} \\ |\tilde{N}_r|^2 &= \max\{|N|^2 - \alpha|\tilde{N}|^2, f|N|^2\} \end{aligned} \tag{2.1}$$

Preliminary experiments showed that the optimum overestimation factor, α, was 1.0, while the spectral flooring parameter, f, was set to 0.7; these values were used in subsequent experiments on SS-NOVO+CMN. Speech distortion by excessive noise reduction degrades speech recognition accuracy and can not restore the distortion with latter additive noise adaptation processing. Therefore we use a low level of noise reduction to prevent the speech distortion, and adapt the remaining noise at the latter processing.

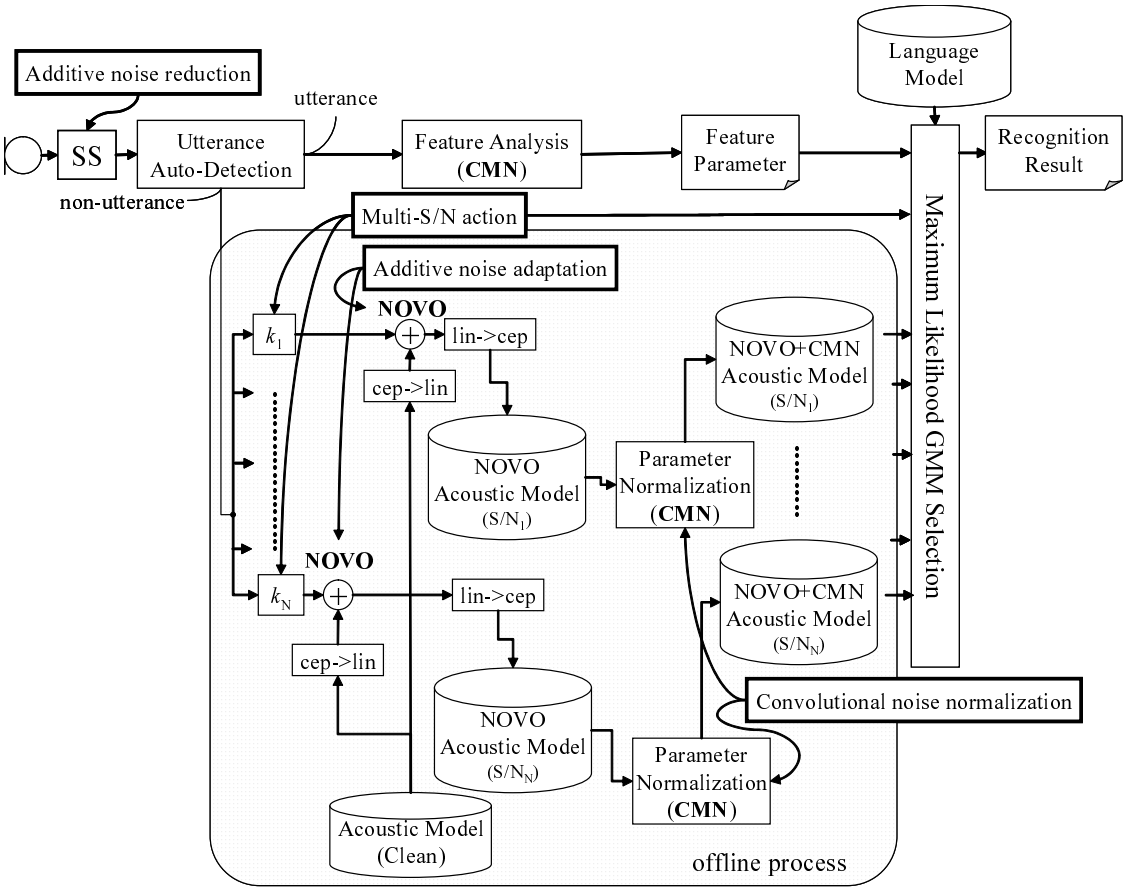

Figure 2.1: *Proposed system.*

2.2.2 Additive noise adaptation for acoustic model

Against the remaining noise, we use NOVO, a noise adaptation technique based on HMM-composition. In this paper, the main parameter of the acoustic model is based on the cepstrum. Accordingly, c_S is the cepstrum of clean speech, c_{Nr} is the cepstrum of the remaining noise, and c_{NOVO} is given by

$$\begin{aligned} c_{NOVO} &= c_{S+N_r} \\ &= F^{-1}(\log[\exp F(c_S)] + k \log[\exp F(c_{N_r}]) \end{aligned} \tag{2.2}$$

F and F^{-1} represent Fourier Transform and Inverse Fourier Transform, respectively.

2.2.3 Multi-S/N adaptation

It is impossible to know the S/N value without an actual speech sample from the speaker in the real world. We use NOVO for noise adaptation to generate several acoustic models for various S/N values in a previous step. Eq. (2.2) makes gain k dependent upon the S/N value. To cope with changes in S/N, we prepare several acoustic models for the various S/N values expected by

changing gain k .We select the recognition result from the estimated S/N acoustic model with maximum likelihood using speech GMMs (Gaussian Mixture Models) generated by multi-S/N HMM-composition.

2.2.4 Convolutional noise normalization for acoustic model

This section explains the cepstral mean normalization method for the noise adapted model generated by HMM-composition. Our technique creates a noise adapted model from the observed noise signal using NOVO as described in [12]. Focusing on the parameters in the spectrum domain converted from the acoustic model parameters, S_S represents the spectrum parameters of the clean speech acoustic model, and the spectrum parameters of the noise adapted acoustic model generated by NOVO, S_{NOVO} , are given by

$$\begin{aligned} S_{NOVO} &= S_{S+N_r} \\ &= S_S + S_{N_r} \end{aligned} \tag{2.3}$$

CMN means that the observed spectrum is normalized by the long-term spectrum in the linear spectrum domain. The long-term spectrum $\overline{S_{NOVO}}$ is given by

$$\begin{aligned} \overline{S_{NOVO}} &= \overline{S_{S+N_r}} \\ &= \overline{S_S + S_{N_r}} \end{aligned} \tag{2.4}$$

where $\overline{X}$ represents the long-term mean of spectrum X . The spectrum parameters of the normalized acoustic model so generated, S^{CMN}_{NOVO}, are given by

$$\begin{aligned} S^{CMN}_{NOVO} &= S^{CMN}_{S+N_r} \\ &= \frac{S_{NOVO}}{\overline{S_{NOVO}}} \\ &= \frac{S_{S+N_r}}{\overline{S_{S+N_r}}} \\ &= \frac{S_S + S_{N_r}}{\overline{S_S + S_{N_r}}} \end{aligned} \tag{2.5}$$

The denominator of Eq. (2.5) is equal to the long-term mean of the noisy speech spectrum. It might be better to use the cepstral mean from the input signal that includes target user speech sample for normalization. From a strategical standpoint, we do not use the input signal because we want to prepare the normalized acoustic model before the user speaks to decrease the response time. Thus our technique uses the parameters of the noise adapted model instead of the input signal.

$\overline{S_{NOVO}}$ represents the long-term mean of the parameters in the linear spectrum domain converted from the noise adapted model parameters in the cepstrum domain; that is the cepstral mean. The noise adapted cepstral mean parameters are approximated to the mean of the average parameters of all distributions in the noise adapted model. This approximation is valid and practical because long-term speech signals contain all phonemes represented by the mixture distributions. The noise adapted cepstral, c^{CMN}_{NOVO}, and cepstral mean, c^{CM}_{NOVO}, are given by,

$$\begin{aligned} c_{NOVO}^{CMN} &= c_{NOVO} - c_{NOVO}^{CM} \\ &= c_{NOVO} - \overline{c_{NOVO}} \\ c_{NOVO}^{CM} &= \overline{c_{NOVO}} \\ &= \overline{c_{S+N_r}} \\ &= \frac{\Sigma_{j=0}^{M-1}\mu_j}{M} \end{aligned} \tag{2.6}$$

where μ_j represents the static cepstral mean parameter of distribution j, and M represents the number of distributions without pause models.

2.2.5 Convolutional noise normalization for input signal

The input signals are subjected to feature analysis processing based on CMN to counter the changes in the transfer characteristics. Focusing on the linear spectrum domain of the input signals, S_S represents the speech spectrum at the sound source, S_{N_r} represents the remaining noise spectrum after noise reduction as observed at the microphone, and H represents the transfer characteristics between the sound source and the microphone; the observed spectrum, S_O, is given by

$$S_O = HS_S + S_N \tag{2.7}$$

Subtracting long-term mean of the feature from the observed feature in the cepstrum domain, CMN, corresponds to division in the linear spectrum domain. The spectrum feature analyzed based on CMN, S_O^{CMN}, is given by

$$\begin{aligned} S_O^{CMN} &= \frac{S_O}{\overline{S_O}} \\ &= \frac{HS_S + S_{N_r}}{\overline{HS_S + S_{N_r}}} \\ &= \frac{S_S + S_{N_r}/H}{\overline{S_S + S_{N_r}/H}} \end{aligned} \tag{2.8}$$

As it assumed that

$$\begin{aligned} S_{N_r} &\cong \frac{S_{N_r}}{H} \\ &= \frac{H^{mic} S_{N_r^{background}}}{H^{mic} H^{space}} \\ &= \frac{S_{N_r^{background}}}{H^{space}} \end{aligned} \tag{2.9}$$

where the microphone characteristics is H^{mic}, the space transfer characteristics is H^{space}, and the remaining background noise after noise reduction is $N_r^{background}$.

This approximation of Eq. (2.9) allows to generate the robust model which adapts the additive noise and normalize the convolutional noise using only the observed the additive noise. From this approximation of Eq. (2.9), S_O^{CMN}, transforms to

$$S_O^{CMN} \cong \frac{S_S + S_{N_r}}{\overline{S_S + S_{N_r}}} \tag{2.10}$$

From Eq. (2.5) and Eq. (2.10), the next equation is obtained.

$$S_O^{CMN} \cong S_{NOVO}^{CMN} \tag{2.11}$$

Normalizing the parameters of the noise adapted acoustic model generated by NOVO by the cepstral mean of the noise adapted model yields the NOVO+CMN acoustic model, S_{NOVO}^{CMN} that matches the input features analyzed with CMN, S_O^{CMN}.

2.3 Experiments

2.3.1 Experimental condition and task

We used artificial hands-free speech data created by convoluting the impulse response and adding noise to 720 clean speech utterances (dry source), each of which consisted of simulated dialogue utterances spoken in different (own) styles. The speech utterances simulated the dialog at call center. Before the recording, we gave the task sheet to the agent and the user. The task sheet included the simulated situation and keywords without reading text. We recorded that spontaneous speech between the agent and the user separately. The subjects were 17 males and 31 females, and each created 15 utterances.

The noises were acquired from a domestic sound database [22], and PC fan noise of a tablet PC was recorded. The noise levels were fixed, and the speech levels were changed by impulse response. The speech power levels were not normalized and varied with the subject and utterance.

The impulse response data was measured with the sound source and the microphone separated by 30, 50 and 70 cm using the TSP (Time-Stretched Pulse) method [23]; a Tablet PC and a speaker were used as shown in Figure 2.2. Figure 2.3 shows the impulse response at the position of 70 cm as an example. The reverberation time is 217 msec measured by square integration method.

At first, we created convolutional noisy speech data by convoluting the impulse response against the clean dry source. We then added the noise samples to the convolutional noisy speech data.

Table 2.1 shows the speech analysis conditions, Table 2.2 shows the acoustic model (HMM) conditions used in the experiments, Table 2.3 shows the training data for the baseline acoustic model, and Table 2.4 shows the evaluation task.

We used a general language model for these experiments, and vocabulary size was 10,000 words. The recognition character correct rate for the dry source were 89.05 %. We utilized a character-based evaluation in the results to eliminate the influence on the length of the Japanese word.

Table 2.5 shows the comparative techniques. SS-CMN use SS and CMN at the front end with CMN acoustic model. Add-matched shows the additive noise matched acoustic model trained using the training data with the matched additive noise. Env-matched shows the environment

matched acoustic model trained using the training data with the matched convolutional and additive noise. These two matched techniques need a lot of time to create the models using large training data against each condition. It is not possible to know the value of S/N beforehand in practice. Our solution is to create several acoustic models with different S/N values, and to perform speech recognition processing in parallel using these models on NOVO-based techniques such as NOVO, SS-NOVO+CMN, and SS-NOVO+CMN(opt.). Here we prepared acoustic models using 3 S/N values: 10 dB, 20 dB, and 30 dB.

SS-NOVO+CMN(opt.) shows the optimum SS-NOVO+CMN. It eliminates the approximation of Eq. (2.9) using S_{N_r}/H instead of S_{N_r} based on the correct impulse response H.

Table 2.1: Speech analysis conditions

Sampling rate	16 [kHz]
Window type	Hamming
Frame width	20 [msec]
Frame shift	10 [msec]
Feature parameter	MFCC(12), ΔMFCC(12), ΔPow

Table 2.2: Acoustic model conditions

HMM	Triphone continuous mixture distribution
# of states	2000
# of mixtures	16
# of phonemes	30

Table 2.3: Training data

Speakers	96 male and 80 female
Size	49 [hour]
# of utterances	47577

2.3.2 Experimental results

Recognition correct rate versus noise type at the position of 50 cm and 0 degree

Figure 2.4 shows average character correct rate versus four noise types at the position of 50 cm and 0 degree. SS-NOVO+CMN shows the best correct rate in most practical situations except for

Table 2.4: Evaluation task

Dry source	15 utterances / speaker of simulated dialogue speech
Topic	Internet Service Provider, PC support, Telecommunication, Mail order, Finance, Local government unit
Speakers	17 male and 31 female
Impulse response	30, 50, 70 [cm] at 0 [degree] and -45, 0, 45 [degree] at 50 [cm]
Noise type	Cleaner, PC fan, sink, ventilation fan
Reverberation time	217 [msec]

Table 2.5: Comparative techniques

ID and name	Acoustic model	Additive noise	Convolutional noise
1. baseline	clean	unknown	unknown
2. NOVO	adapt	known	unknown
3. SS	clean	known	unknown
4. CMN	clean	unknown	unknown
5. SS-CMN	clean	known	unknown
6. SS-NOVO+CMN	adapt	known	unknown
7. SS-NOVO+CMN(opt.)	adapt	known	known
8. Add-matched	train	known	unknown
9. Env-matched	train	known	known

the matched models and optimum models.

Env-matched shows the greatest performance but Env/Add-matched need a lot of training time and they are not practical. Add-matched shows better performance than our proposed SS-NOVO+CMN with cleaner and PC fan noises. Against these noises, NOVO shows worse performance than Add-matched, thus, the approximation accuracy of NOVO is not sufficient. Due to this wide degradation from Add-matched to NOVO, SS-NOVO+CMN can not restore the performance degradation.

In all situations, SS-NOVO+CMN(opt.) achieves better performance than SS-NOVO+CMN. The influence of the approximation of Eq. (2.9) is not so small but the optimum technique needs the correct impulse response.

Recognition correct rate versus position with PC fan noise

Figure 2.5 shows the average character correct rate versus position with PC fan noise. SS-NOVO+CMN shows the best correct rate in average except for the matched and optimum techniques. But SS-NOVO+CMN and SS-CMN don't have statistically significant difference.

At the position of 30 cm, the improvement of Env-matched is quite bigger than other positions. As the distance between microphone and speaker becomes closer, the value of S/N becomes higher and the early reflection from keyboard area becomes excessive. CMN can not eliminate the excessive reflection perfectly. Especially, SS-NOVO+CMN have an approximation in CMN processing of Eq. (2.6) and it reduces the improvement by approximated CMN under the excessive convolutional noise.

At the position of 70 cm, the distance between microphone and speaker becomes more distant, the value of S/N becomes lower. SS achieves the same performance with NOVO. SS and SS-CMN achieve good improvement with sufficient noise reduction. But SS-NOVO+CMN and SS-NOVO+CMN(opt.) can not achieve good performance.

Recognition correct rate versus position with all noises

Figure 2.6 shows the average character correct rate versus position with several noises. SS-NOVO+CMN shows the best correct rate in most and average situations except for the matched and optimum techniques.

At the position of 70 cm, SS-NOVO+CMN can not achieve better performance than SS-CMN as the same reason of Section 2.3.2.

Overall, SS-NOVO+CMN achieve the best performance. This technique increased the average character correct rate by 1.02 % compared to SS-CMN and 11.62 % compared to CMN, which are both statistically significant at $p < 0.0001$ using a matched pair test.

Computational time

The average previous computational time of SS-NOVO+CMN is 0.50 [sec] for 5.0 [sec] noise data, so the RTF (Real Time Factor) is equal to 0.10 on an Intel®Xeon™3.6 GHz processor. It takes 0.015 [sec] to generate the SS parameter for the additive noise reduction, 0.42 [sec] for generating SS-NOVO (S_{NOVO}) models by HMM-composition for the additive noise adaptation, and 0.063 [sec] for generating SS-NOVO+CMN (S^{CMN}_{NOVO}) models for the convolutional noise normalization. If the techniques need additive noise adaptation processing by HMM-composition after user's speech sample is captured, they use not only the waiting time for user's utterance, but also the processing time for HMM-composition to initialize the speech recognition systems. Our proposed technique, SS-NOVO+CMN, use the computational times in the previous step when user doesn't use the application, because it need only the observed additive noise. SS-NOVO+CMN doesn't waste the time to wait user's utterance start, and it is convenient for use in real applications.

The average online RTF is 0.68 for SS-NOVO+CMN, 0.65 for CMN, and 0.65 for SS-CMN. SS-NOVO+CMN doesn't have an excessive computation at the online step compared to CMN and SS-CMN. Therefore, our proposed technique can achieve good response.

2.4 Summary

Conventional noise adaptation techniques counter additive and convolutional noise but fail to achieve rapid response. To rectify this omission, we proposed SS-NOVO+CMN; it normalizes the cepstral mean for the parameters of the noise adapted acoustic models generated by NOVO (HMM-composition) by using just the remaining additive noise after the application of SS (Spectral Subtraction) in a previous step. Furthermore, it generates several S/N acoustic models to handle changes in S/N values in real applications. In an online step, it needs only SS and CMN (Cepstral Mean Normalization) at the front end and S/N selection using GMM. This proposed technique increased the average character correct rate by 11.62 % compared to CMN condition, which is statistically significant at $p < 0.0001$ using a matched pair test, and it is more practical than conventional techniques since it offers short response times.

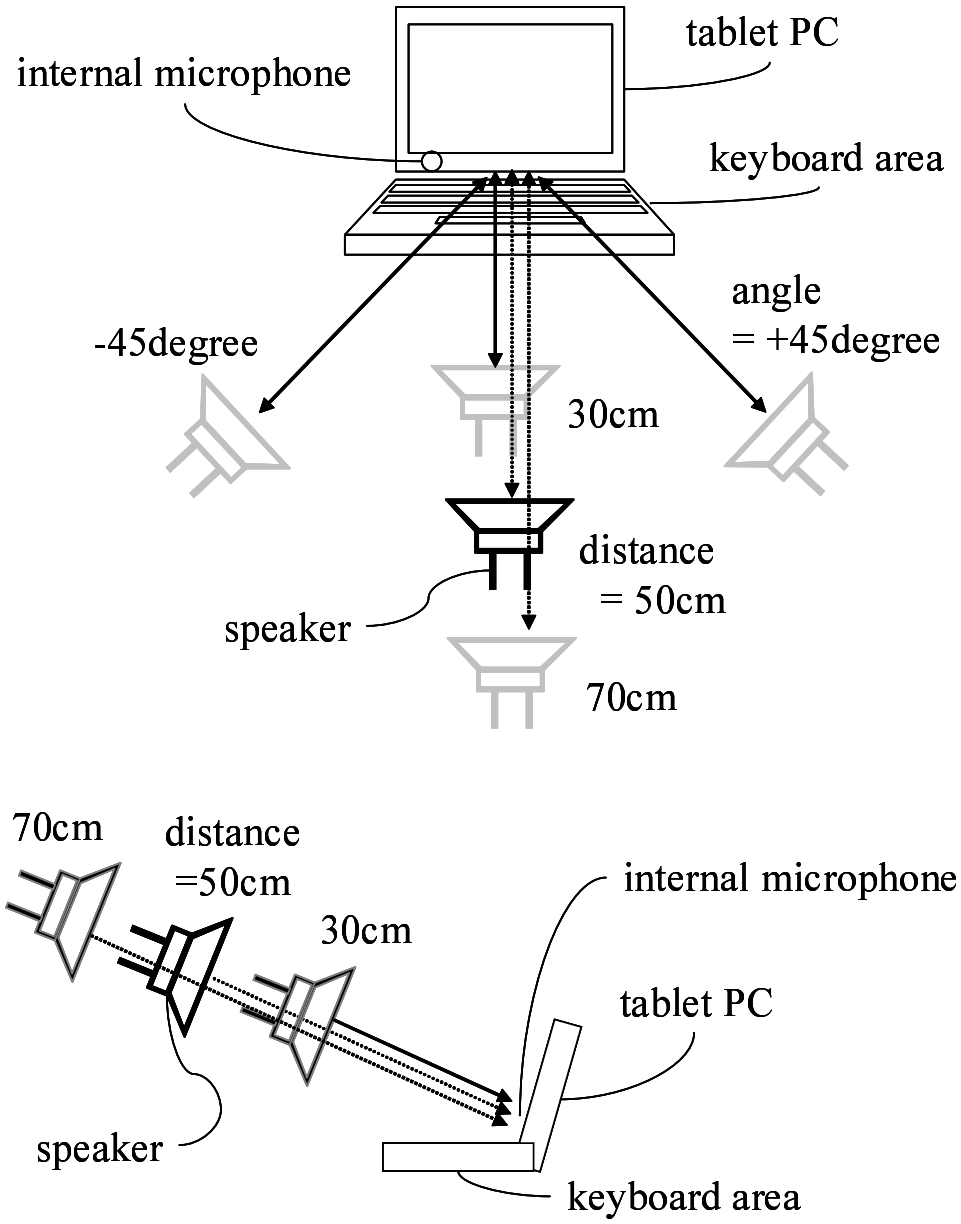

Figure 2.2: *Recording condition.*

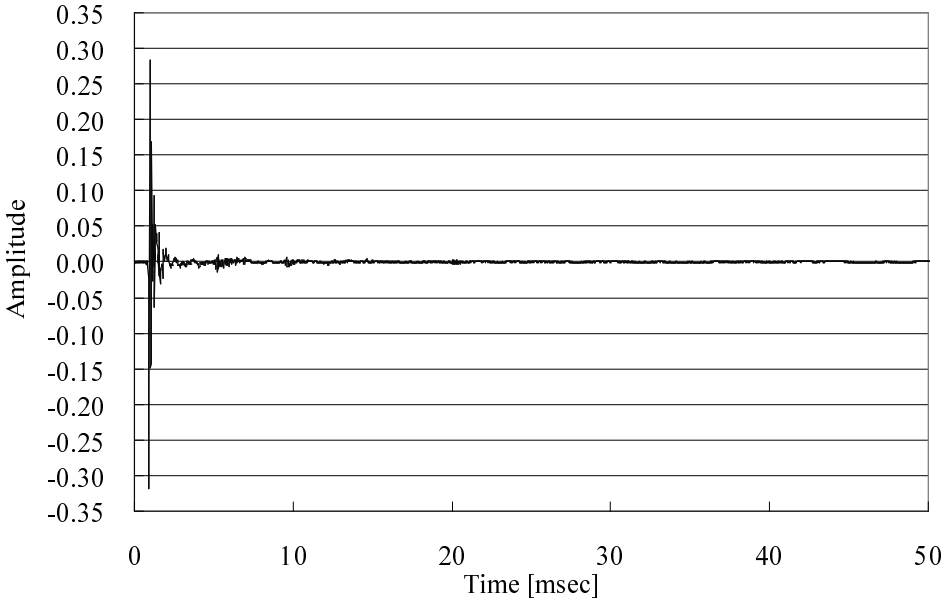

Figure 2.3: *Impulse response at the position of 70 cm.*

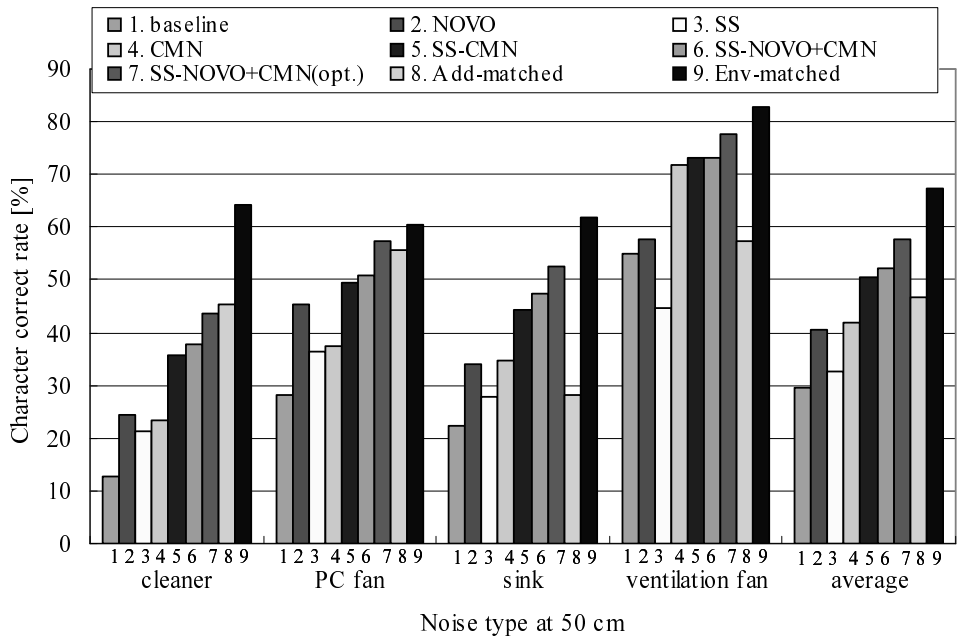

Figure 2.4: *Recognition correct rate versus noise type at the position of 50 cm and 0 degree.*

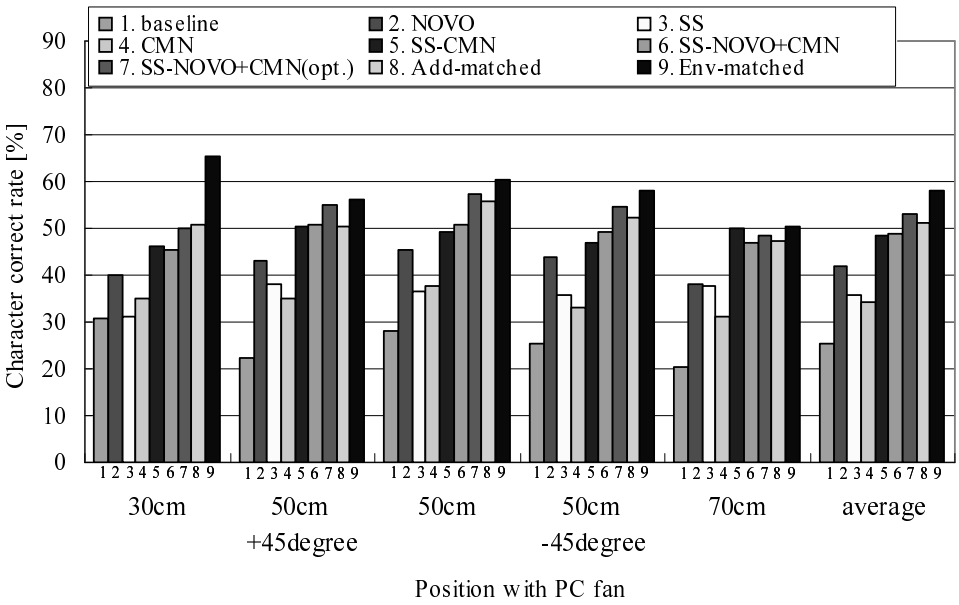

Figure 2.5: *Recognition correct rate versus the position with PC fan noise.*

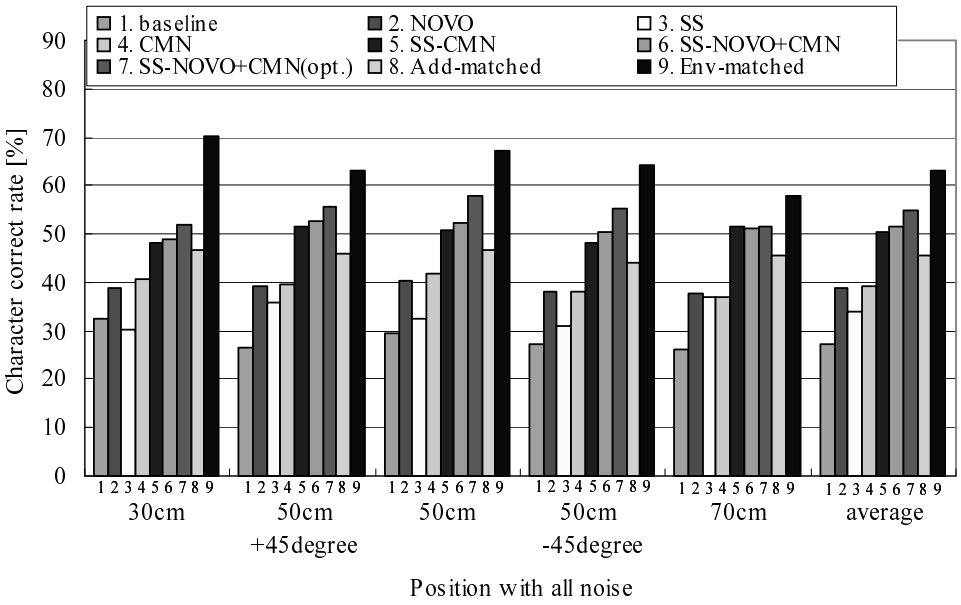

Figure 2.6: *Recognition correct rate versus the position with all noises.*

Chapter 3

Fast Unsupervised Adaptation Based on Efficient Statistics Accumulation Using Frame Independent Confidence within Monophone States

3.1 Introduction

Massive quantities of videos and dialogs are stored every day by servers; typical examples include video sharing services on the Internet and call center services provided by many companies. Speech recognition technologies can automatically transcribe the speech contained in the stored items, and thus make the items searchable via these recognized transcriptions [24]. With respect to call centers, several studies have analyzed customer needs by applying text mining to automatically-transcribed spoken documents [25, 26]. Analysts demand high accuracy from speech recognition systems to enable them to extract the customers' needs efficiently. However managers are reluctant to improve recognition accuracy by increasing computer resources. Even a small increase in computation time generates vast computation costs when related to several tens of thousands of calls per day. Therefore, we must improve recognition accuracy with limited computer resources. This paper aims to achieve higher accuracy within the same computation time required by the baseline system.

Adaptation techniques can improve recognition accuracy significantly when faced with the intrinsic variability of speech such as speaker characteristics and the recording environment [27]. It is difficult to deal with all the variability of speech using supervised adaptation techniques. We can improve recognition accuracy at low cost by adapting these techniques for the acoustic characteristics of the input speech based on unsupervised transcription. Since stored speech does not require the real time processing, a batch-type unsupervised acoustic model adaptation is effective. With an unsupervised acoustic model adaptation it is necessary to generate unsupervised transcriptions and accumulate statistics using this transcription. The conventional accumulation of statistics is based on forward-backward [28] or Viterbi [29] algorithms. These algorithms require state sequences and so need to determine the transcription in advance. Computation time is needed for decoding

and for generating an unsupervised transcription for unsupervised adaptation.

Unsupervised adaptation techniques based on Maximum Likelihood Linear Regression (MLLR) [30] are commonly used to improve accuracy. Constrained MLLR (CMLLR) [31] or feature-space MLLR (fMLLR) often adopts a single global transformation matrix. More complex transformations are sensitive to unsupervised transcription errors in unsupervised adaptation [32]. Based on the premise that it estimates only a single matrix, the existing unsupervised adaptation technique realizes high speed by employing 2 class labels instead of using unsupervised transcription with speech and pause models [33]. However, this approach cannot significantly improve accuracy. We aim to realize fast and highly accurate unsupervised adaptation based on the use of monophone-class labels to estimate a single matrix.

This paper proposes a fast unsupervised adaptation technique for estimating a single transformation matrix similar to CMLLR. Increased speed is achieved simply by using Gaussian Mixture Models (GMMs) belonging to monophones in the acoustic model to generate unsupervised transcriptions and by ignoring the time-wise continuity and adopting the accumulation of frame independent statistics. This approximation did not degrade the accuracy in previous experiments [34] performed under limited restrictions, namely where the gender was known and the quantity of data was small. Experiments show that the proposed technique is fast and highly accurate with a large amount of stored gender-unknown speech. Our technique reduces both the computation time and the number of recognition errors significantly. Since the proposed technique runs at faster than the baseline speed, it does not require a change in the hardware configuration. Therefore, the proposed system avoids the above mentioned criticism regarding computer resources and can be introduced easily into call centers.

3.2 Related work on unsupervised adaptation

Unsupervised adaptation is classified into two types: online and batch (e.g. [35]). The online type described in [36] needs no prior time to generate the unsupervised transcriptions since they are derived from the recognition results of previous utterances. The negative side is that initial utterances receive no adaptation gain. The batch type can be expected to offer improved accuracy for initial utterances. In call-center speech situations such initial utterances often contain important details as the reason for the call [37], and so they are expected to be processed with high accuracy. Therefore, our proposed technique employs batch-type adaptation but the cost is that prior unsupervised transcription is required.

There is an abundance of related work on unsupervised acoustic model adaptation for speech recognition. Most conventional studies are based on three major methods; Maximum a Posteriori (MAP) [38], Eigenvoice [39], and MLLR [30]. In particular, a lot of recently proposed techniques have been based on MLLR. Shift MLLR [40] and eXtended MLLR (XMLLR) [41] change the MLLR transformation formulation for Gaussian mean adaptation techniques, and they use unsupervised transcriptions from the previous stage in multi-pass decoding. Quick fMLLR (Q-fMLLR) [42] reduces the computation time for statistics accumulation, but it also employs the initial unsupervised transcriptions using a speaker independent model for adaptation. Alberti *et al.* needs a

non-negligible computation time to obtain prior unsupervised transcriptions from decoding even with a narrow beam [24]. This approach is computationally expensive since the decoding process uses a large number of Gaussians in the acoustic model, and consumes extra computation time with the exception of the Gaussian output probability calculation [43]. It is essential to eliminate the extra computation time needed for generating unsupervised transcriptions. The proposed technique tackles this time-consuming issue by transcribing using the output probabilities of only a limited number of Gaussians.

Several existing high-speed unsupervised adaptation techniques reduce the computation time by using a small number of class labels for prior unsupervised transcription. Lévy *et al.* has a general GMM as a phoneme-independent model, and phoneme-dependent models are modeled as a transformation of this general GMM. Therefore, only one class GMM has to be adapted and no decoding is needed [44]. Kozat *et al.* uses 2 class labels (speech / pause) before adaptation [33]. Hence they can reduce computation time significantly, and these techniques are effective for word recognition with short data lengths. However, we think that they use too few (1 or 2 classes) labels when transcribing spontaneous conversational speech. Our proposal adopts more classes of labels using monophones (30 classes are used in our system). The proposed technique increases the auto-transcription time, but offers improved accuracy since it realizes unsupervised adaptation from the fine labeling, unlike the coarse transcription approach used in [33].

3.3 Proposed rapid unsupervised adaptation technique

3.3.1 Illustration of statistics accumulation

We compare our proposed statistics accumulation approach with the fundamental forward-backward algorithm [28] and the conventional Viterbi algorithm [29]. We focus on state occurrence probability as the statistic for adaptation.

Conventional statistics accumulation

Acoustic model adaptation methods such as MAP [38] and MLLR [30] accumulate the statistics of posterior probability $\gamma_t(s,m)$ from the m-th mixture component distribution of state s at frame t; $\gamma_t(s,m)$ is calculated from the occurrence probability $\gamma_t(s)$ of state s as follows.

$$\gamma_t(s,m) = \gamma_t(s) \cdot \frac{c_{s,m}\mathcal{N}_{s,m}(\mathbf{O}_t|\boldsymbol{\mu}_{s,m},\boldsymbol{\Sigma}_{s,m})}{\sum_{k=1}^{M_s} c_{s,k}\mathcal{N}_{s,k}(\mathbf{O}_t|\boldsymbol{\mu}_{s,k},\boldsymbol{\Sigma}_{s,k})} \tag{3.1}$$

Here, M_s is the number of distributions belonging to state s, $c_{s,m}$ is the m-th mixture weight and $\mathcal{N}_{s,m}(\cdot)$ is the m-th multidimensional Gaussian distribution function with mean vector $\boldsymbol{\mu}_{s,m}$ and covariance matrix $\boldsymbol{\Sigma}_{s,m}$ of state s feature vector $\mathbf{O}_t$.

$\gamma_t(s)$ is fundamentally estimated using the forward-backward algorithm [28] as follows;

$$\gamma_t(s) = \frac{\alpha_t(s)\beta_t(s)}{\sum_{j=1}^{S} \alpha_t(j)\beta_t(j)} \tag{3.2}$$

where $\alpha_t(s)$ ($\beta_t(s)$) is the forward (backward) probability of state s at frame t, and S is the total number of states used for statistics accumulation. The forward-backward algorithm requires time-series labels as the state sequence; therefore it has to prepare a determined transcription in advance.

The Viterbi algorithm [29] is commonly used in statistics accumulation It approximates $\gamma_t(s)$ as being equal to 1 on the Viterbi path (0 otherwise) as follows;

$$\gamma_t(s) \simeq \begin{cases} 1 & \text{if } s \text{ is on the Viterbi path at } t \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

This Viterbi algorithm also requires the state sequence from the transcription, and estimates the Viterbi path with the state sequence. The Viterbi algorithm requires a pre-determined transcription before undertaking unsupervised adaptation similar to the fundamental forward-backward algorithm. These algorithms must prepare a pre-determined state sequence. Instead of the continuous value ($0 \leq \gamma_t(s) \leq 1$) in the forward-backward algorithm, $\gamma_t(s)$ is a binary value (1 or 0) in the Viterbi algorithm. The Viterbi algorithm estimates $\gamma_t(s)$ with complete confidence on the Viterbi path even though the source label is not reliable in unsupervised adaptation.

Proposed statistic accumulation

In contrast to the forward-backward and Viterbi algorithms, we incorporate a simple reliability expression in the occurrence probability $\gamma_t(s)$ estimation. When reviewing Eq. (3.2) in the forward-backward algorithm, the right side numerator indicates the probability of passing through the target state s at frame t, and the right side denominator indicates the summation of all states' probabilities. $\gamma_t(s)$ can be considered the reliable probability of passing through the target state s within all the state paths at frame t, i.e. state confidence per frame. The Viterbi algorithm ignores this state confidence although it retains a pre-determined state sequence. Instead of the state sequence requirement of this Viterbi algorithm, our proposed technique uses state confidence per frame in statistics accumulation to approximate $\gamma_t(s)$ as follows;

$$\gamma_t(s) \simeq \frac{b_s(\mathbf{O}_t)}{\sum_{j=1}^{S} b_j(\mathbf{O}_t)} \tag{3.4}$$

where $b_s(\mathbf{O}_t)$ is the frame independent output probability of state s for feature vector $\mathbf{O}_t$. This approximate $\gamma_t(s)$ means the posterior probability based on the states' output probabilities. It can

also be considered the state confidence since the posterior probability is often used as a confidence measure [45]. The lattice-based MLLR techniques [46, 47] also use state posterior over an entire lattice instead of our frame independent state confidence.

$\gamma_t(s)$ in the proposed technique is a continuous value similar to the forward-backward algorithm. The lack of a pre-determined state sequence reduces labeling accuracy, but even if the labeling is not accurate in a frame, $\gamma_t(s)$ becomes smaller, so the influence of labeling error is reduced with our frame independent statistics accumulation due to the use of this state confidence per frame. Moreover, to reduce both the detrimental influence of labeling error and the extra computation time, we accumulate only the best state within a frame as shown by the following equation.

$$\gamma_t(s) \simeq \begin{cases} \dfrac{b_s(\mathbf{O}_t)}{\sum_{j=1}^{S} b_j(\mathbf{O}_t)} & \text{if } s \text{ is best state at } t \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

The advantage of this state confidence is investigated by setting $\gamma_t(s)$ equal to 1 at the best state at t within a frame such as the Viterbi algorithm, see Section 3.4. Furthermore, since some decoders ignore the state transition probabilities [48], and MLLR-based adaptation techniques transform the parameters of Gaussian distributions related to output probabilities, we ignore the state transition and use only frame independent output probabilities in statistics accumulation. Thus, the proposed technique can estimate the occurrence probability $\gamma_t(s)$ frame by frame from the frame independent output probability $b_s(\mathbf{O}_t)$ of a state s within frame t, and does not require a determined label from transcription in advance.

Relation between conventional and proposed statistics accumulation

Fig. 3.1 shows the relation between the conventional forward-backward / Viterbi algorithms and the proposed statistics accumulation. The upper part of Fig. 3.1 shows the relation between (a) the forward-backward algorithm, (b) the Viterbi algorithm and (c) the proposed frame independent statistics accumulation with regard to the formulation of the occurrence probability $\gamma_t(s)$ of state s at frame t. The lower left part of the figure shows the state's sequences ($\bullet$, $\leftarrow$ and $\nwarrow$) on the Viterbi path in (b) the Viterbi algorithm. The lower right part shows the best state's sequences ($\bullet$, $\rightarrow$, $\nearrow$, and $\searrow$) at each frame t in (c) the proposed frame independent statistics accumulation. The other states are indicated by $\circ$. T is the total number of frames used for statistics accumulation in (b) the Viterbi algorithm. The states' sequences on the Viterbi path are derived from determined labels by utilizing prior unsupervised labeling. The best states' sequences in (c), namely the proposed technique, are the best states at each frame without the determined label.

Table 3.1 compares (a) the forward-backward algorithm, (b) the Viterbi algorithm and (c) the proposed frame independent statistics accumulation with regard to pre-determined state sequence and state confidence. Conventional statistics accumulation approaches, including both the forward-backward algorithm [28] and the well-known Viterbi training algorithm [29], requires the deter-

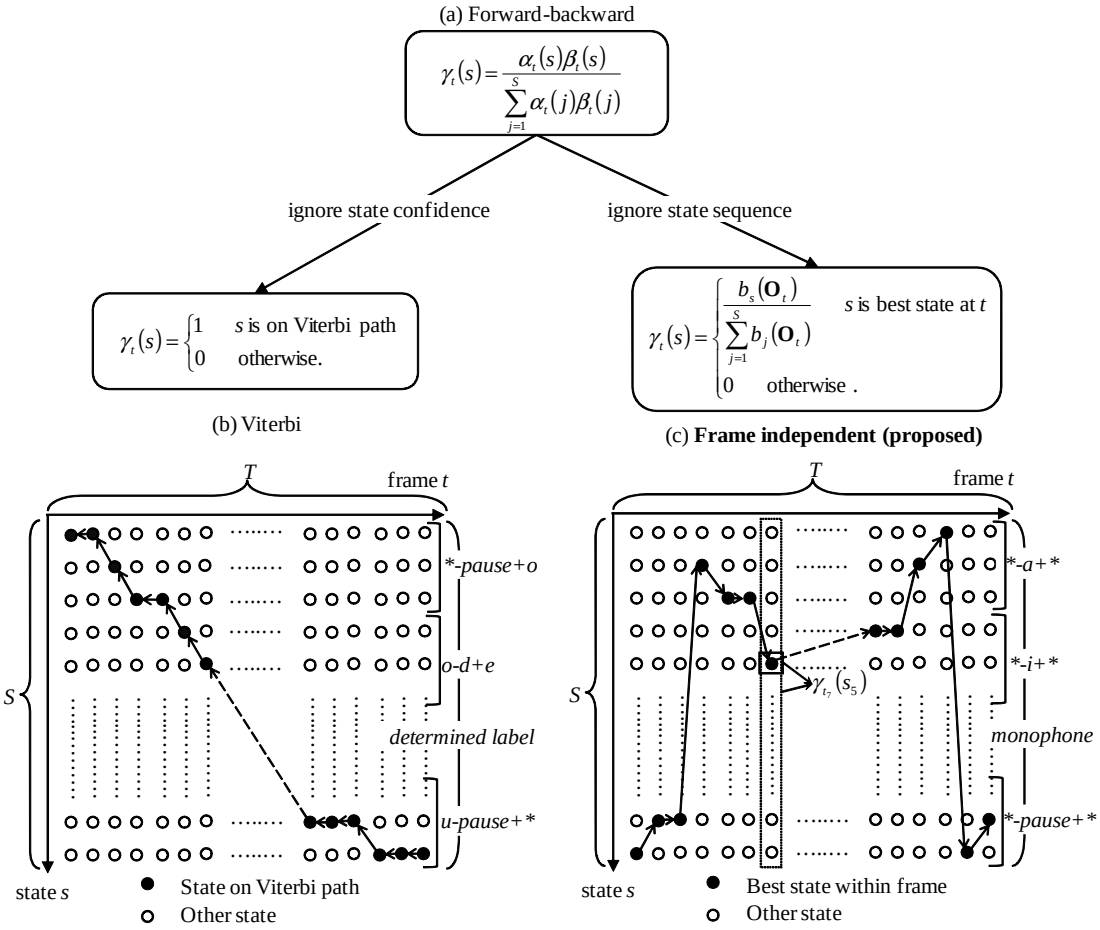

Figure 3.1: *Relation between (a) forward-backward algorithm, (b) Viterbi algorithm and (c) proposed frame independent statistics accumulation.*

mined label (state sequence) to be determined in a prior labeling process. Unlike the conventional prior labeling process, which needs to generate a number of recognition hypotheses to acquire the hypothesis with the maximum likelihood using a large vocabulary word trigram, our frame independent statistics accumulation only has to calculate the output probabilities of the limited states, thus reducing the computation time.

3.3.2 Unsupervised adaptation with monophone constraint and power utilization

We describe the monophone constraint for realizing increased speed and the power utilization for achieving improved accuracy in unsupervised adaptation.

Table 3.1: Comparison of (a) forward-backward algorithm, (b) Viterbi algorithm and (c) proposed frame independent statistics accumulation.

	State sequence	State confidence
(a) Forward-backward algorithm	use (forward / backward path)	use ($0 \leq \gamma_t(s) \leq 1$)
(b) Viterbi algorithm	use (Viterbi path)	ignore ($\gamma_t(s) = 1$ or 0)
(c) **Frame independent**	ignore (frame independent)	use ($0 \leq \gamma_t(s) \leq 1$)

Monophone constraint

Lee *et al.* achieves fast speech recognition by using monophones [49]. The proposed technique also speeds up unsupervised labeling by using only monophones. The assumption is that one monophone is an approximate model of triphones, which have the same central phoneme. Our target is to estimate a single global transformation matrix rapidly using only monophones; i.e. all the Gaussians in the monophones. A single matrix obviously has fewer elements than a multiclass matrix. We consider that the sophisticated labeling provided by triphones is not required to estimate this smaller number of elements. Thus, it is sufficient to use monophones in labeling even if this yields a few errors. Triphones have many more states than monophones, so our monophone constraint can reduce computation time significantly. Coarse state labeling, such as speech / pause labeling, achieves high accuracy [33] but there is little improvement in recognition accuracy. Setting all state loops as unconstrained cannot achieve high unsupervised transcription accuracy, so the improvement in recognition accuracy is also slight. The optimality of our monophones' state constraint is shown by comparison with speech / pause loops and all state loops in Section 3.4.

Posterior probability $\gamma_t(s, m)$ is calculated using the approximate occurrence probability $\gamma_t(s)$ shown in Eq. (3.1). The statistics of mean parameter, $\sum_{t=1}^{T} \gamma_t(s, m) \cdot \mathbf{O}_t$ and $\sum_{t=1}^{T} \gamma_t(s, m)$, are accumulated using posterior probability $\gamma_t(s, m)$ over the total number of frames, T. The single global transformation matrix is generated from these accumulated statistics using the model-space MLLR described in [31]. The mean parameters of all the distributions in the acoustic model (triphones as well as monophones) are transformed by this same matrix.

Power utilization

The speech power changes depending on the positions of the speaker and microphone. Recognition accuracy is degraded if power term is not properly utilized. Accuracy could be improved by with the proper use of a power adapted model. The proposed technique uses the power term an extra feature parameter for speech recognition only after adaptation, not before; the occurrence probability $\gamma_t(s)$ is calculated using the likelihood without power while $\mathcal{N}_{s,m}(\cdot)$ is calculated with power in Eq. (3.1) to generate the power adapted model. Furthermore, the speech power level is normalized utterance by utterance in acoustic model training.

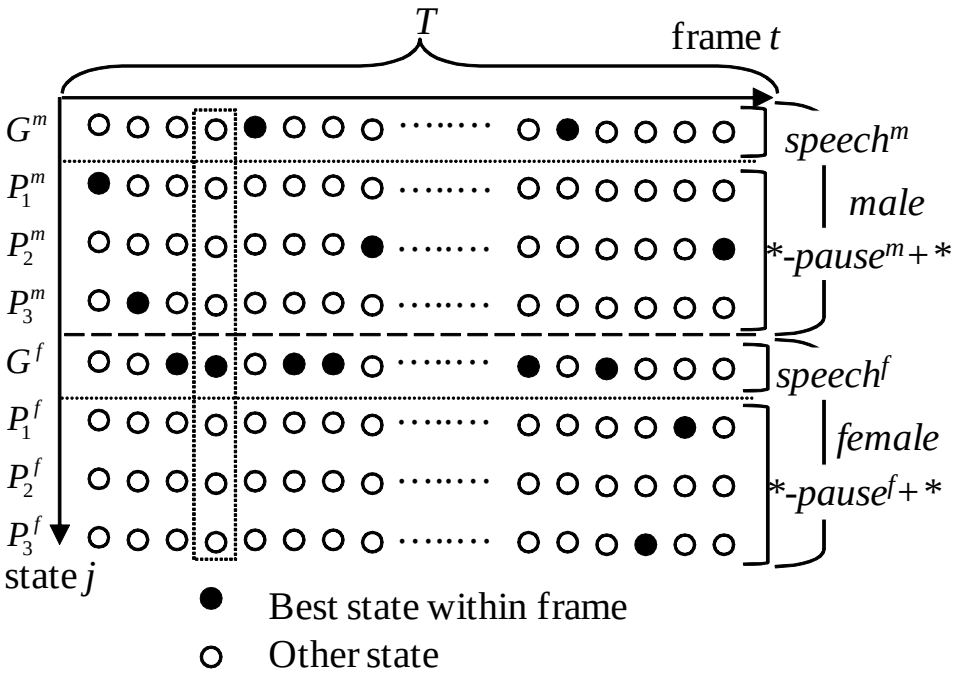

Figure 3.2: *State sequence used in gender selection with utterance selection.*

3.3.3 Gender selection with utterance segmentation

We describe utterance segmentation and gender selection for the selection of an appropriate acoustic model prior to unsupervised adaptation.

Our baseline speech recognition system adopts a conventional parallel decoding technique using dual-gender (male / female) acoustic models. Our baseline decoder shares the search spaces of dual-gender acoustic models and selects gender in the decoding process. This approach is faster than using dual decoders in parallel. Unfortunately, its speed and accuracy are still unacceptable since its search space is larger than that of the 'ideal' single-gender dependent acoustic model.

Imai *et al.* used monophones for 'online' speech detection and speech recognition with dual-gender models [50]. The monophone constraint for gender selection is also effective in reducing computation time in our 'offline' system. The proposed technique investigates the use of only gender dependent speech and pause models as constraints in the gender selection process before speech recognition for a further reduction in computation time. This approach is efficient for selecting the appropriate gender acoustic model for speech recognition, especially with acoustic model adaptation.

Both gender selection and utterance segmentation are performed by using output probabilities from GMMs only in the states belonging to speech (gender) GMMs and pause Hidden Markov Models (HMMs) in dual-gender acoustic models. Fig. 3.2 shows the state sequences used in this process at each frame t. • (○) indicates the best state (the other state) within the frame. T is the

total number of frames used in this process.

Utterance segmentation

Utterance segmentation uses (gender dependent) speech GMMs (G^g) and pause GMMs belonging to pause HMMs (P_1^g, P_2^g, P_3^g; the 3 states of the pause model used in our dual-gender system) in dual gender ($g \in \{m, f\}$; m: male and f: female) dependent acoustic models; since the pause models are not shared by males and females in our system, the pause GMMs are trained dependently by using gender dependent training data and so our pause GMMs are gender dependent. Utterance start-points are detected by using a basic energy-based method with a hangover time. After start-point detection, we calculate the frame independent output probabilities $b_j(\mathbf{O}_t)$ of speech and pause Gaussian mixture models ($j \in G^g, P_1^g, P_2^g, P_3^g$) for feature vector $\mathbf{O}_t$ at frame t. If the speech model ($b_{G^g}(\mathbf{O}_t)$) is the best state (● in Fig. 3.2), frame t is considered to be speech. If not, frame t is considered to be a pause. When the pause frame continues for longer than τ^{pau} (e.g. 0.8 sec), the utterance is segmented as an end-point. Excessive utterance segmentation loses consonant discrimination and degrades accuracy in posterior speech recognition. Therefore, if the interval time between utterances is less than τ^{intvl} (e.g. 1.0 sec), the utterances are concatenated. Whereas Imai *et al.* use monophones for segmentation [50], the proposed technique uses only the output probabilities of speech / pause model so it simplifies implementation and reduces computation time.

Gender selection

The proposed technique selects gender concurrently with utterance segmentation. It determines gender by a majority vote of best state, either male ($b_{G^m}(\mathbf{O}_t)$) or female ($b_{G^f}(\mathbf{O}_t)$) within each speech frame. The above utterance concatenation is performed only between utterances of the same gender. Gender selection only counts the number of best frames against each gender model, and so consumes less computation time.

3.3.4 Framework of proposed system

Fig. 3.3 shows the framework of the proposed system. It consists of two parts; gender selection (Section 3.3.3) and unsupervised adaptation (Section 3.3.1 and 3.3.2). The latter part has the following three component technologies ; frame independent statistics accumulation (Section 3.3.1), monophone constraint (Section 3.3.2) and power utilization (Section 3.3.2). The proposed system performs gender selection utterance by utterance. It then employs fast frame independent statistics accumulation with monophone constraint against the utterances estimated to be from the same gender in the adaptation process, and performs speech recognition with power utilization.

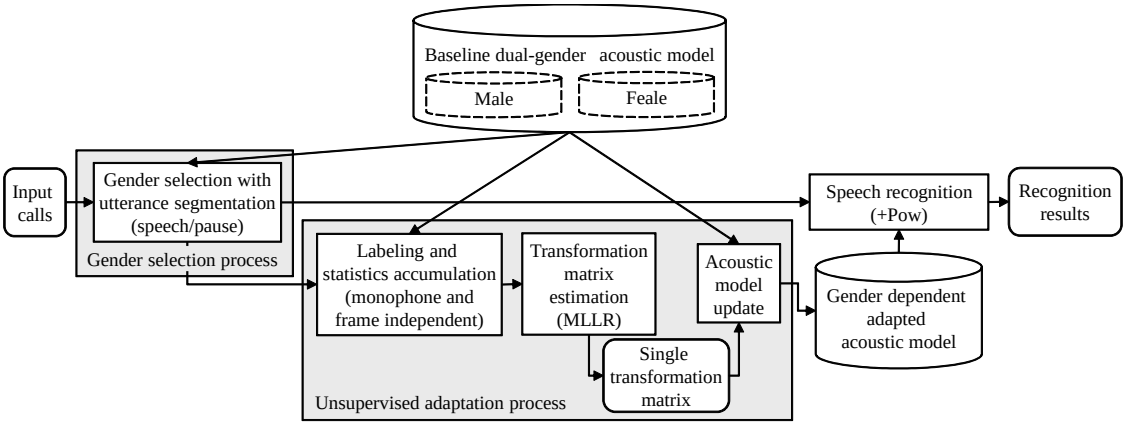

Figure 3.3: *Framework of proposed system.*

3.4 Experiments

We introduce recognition experiments that we performed to investigate the effectiveness of our proposed technique for spontaneous speech; the baseline is no adaptation with a dual-gender acoustic model.

3.4.1 Experimental settings

In this experiment the acoustic analysis condition is as described below; sampling frequency: 16 kHz, 20 msec length Hamming window shifted by 10 msec, and acoustic features: 25 orders (MFCC 12, ΔMFCC 12, Δpower) or 26 orders (power utilization after adaptation). The evaluation task uses 120 calls (9.91 hours and 9,056 utterances) by 24 Japanese speakers (7 males and 17 females), and the speaking style is spontaneous speech in a two-party dialog. We use a dual-gender acoustic model, the number of states is 1,958 in total and 90 monophone states, the distribution number is 26,568 for males and 29,836 for females, and the size of training data is 122.71 hours (109,294 utterances) for males and 113.23 hours (110,792 utterances) for females. The maximum number of mixtures in each state is 16, but some states have fewer mixtures depending on the quantities of training data for each state. The language model is a word trigram developed by using manual transcriptions of dialog speech, and its vocabulary size is 59,676 words. The speech recognition decoder is VoiceRex [51].

The baseline is the technique in the previously adopted speech recognition system; it uses parallel decoding without prior gender selection or unsupervised adaptation. The same beam width is used in all these experiments. The proposed unsupervised adaptation techniques are combined with the proposed gender selection technique.

3.4.2 Experimental results and discussion

We utilized a character-based evaluation to eliminate the influence of Japanese word length; the abbreviations “Cor.” and “Acc.” mean correct rate and accuracy, respectively. The computation time is normalized by the baseline recognition time; “Slct.” is the prior gender selection time with utterance segmentation, “Adpt.” is the prior adaptation time with labeling, and “Sum.” is the total computation time.

3.4.3 Experimental results and discussion for gender selection

We have to select an appropriate gender acoustic model for unsupervised adaptation, and so we first investigate the influence of gender selection. The ID number, the abbreviation (abbr.) and the effect of the compared gender selection techniques are shown in Table 3.2 in relation to gender selection. The effect of gender selection, “GS”, is confirmed by comparing gender-known, “gd”, with -unknown “m/f” (male / female); “gd” uses the ‘ideal’ gender dependent acoustic model. The proposed gender selection with speech / pause models from the dual-gender acoustic model, “GS(s/p)”, is compared to the monophone-based technique, “GS(mo)”, as described in [50].

As the result of gender selection, the baseline technique, “1. m+f: baseline”, exhibited degraded speed and accuracy compared with ‘ideal’ gender dependent technique, “2. gd”, since it expanded the search space and triggered gender selection errors. The prior gender selection techniques, “3. m/f+GS(mo) and “4. m/f+GS(s/p)”, achieved accuracy equivalent to that of the ideal gender dependent technique, “2. gd” and “5. gd+GS(s/p)”, so the adverse impact of our proposed gender selection and utterance segmentation is very small. The proposed speech/pause constraint technique, “4. m/f+GS(s/p)”, is faster than the conventional monophone constraint technique, “3. m/f+GS(mo)”, see [50], with equivalent accuracy. As a result, the proposed gender selection approach, “4. m/f+GS(s/p)”, is equivalent to the ‘ideal’ gender dependent technique, “2. gd”; thus our proposed speech/pause constraint is effective for gender selection.

Table 3.2: Performance of compared techniques regarding gender selection with utterance segmentation.

ID and abbr.	Gender selection	Cor.	Acc.	Sum.	Slct.
1. m/f: baseline	Parallel decoding	79.22	73.79	1.00	-
2. gd	Known	80.78	75.39	.939	-
3. m/f+GS(mo)	Monophone loop	80.72	75.40	.977	.046
4. **m/f+GS(s/p)**	**Speech / pause loop**	**80.71**	**75.34**	**.956**	**.008**
5. gd+GS(s/p)	Known and segments by speech / pause loop	80.72	75.37	.939	.006

3.4.4 Experimental results and discussion for unsupervised adaptation

Next, we investigate the influence of unsupervised adaptation. ID and abbr. are shown in Table 3.3 regarding unsupervised adaptation. The proposed unsupervised adaptation technique, "pUA", which uses state loop auto-transcription and frame independent statistics accumulation with the proposed gender selection "m/f+GS(s/p)", is compared to the conventional unsupervised adaptation technique, "cUA", using forward-backward statistics accumulation with 'ideal' gender-dependent "gd" acoustic model and auto-transcription by three language models; speech / pause loop "cUA(s/p)" using 2 class labels as described in [33], monophone loop "cUA(mo)" and word trigram "cUA(tri)". The proposed technique with monophone constraint auto-transcription, "pUA(mo)", is compared to all state loop (no constraint) "pUA(all)" and speech / pause loop "pUA(s/p)". "pUA(mo)" is also compared to "w/oSC": without state confidence ($\gamma_t(s) = 1$ at best state within frame). The entire unsupervised adaptation process is performed call by call.

Table 3.4 shows the effect of unsupervised adaptation without power utilization. All unsupervised techniques (6-8 or 9-11) provided better accuracy than the unadapted techniques (2 or 4). The conventional unsupervised adaptation technique using word trigrams achieved the best accuracy but its computation time was twice that of the baseline technique, "1. m+f: baseline", since it requires preparation time to generate a unsupervised transcription. Simplifying the language model used in labeling (6→8) reduced not only the computation time but also the accuracy. The conventional speech / pause based technique, "8. cUA(s/p)", which is similar to [33], is faster than "1. m+f: baseline", but it offers the least improvement in accuracy. The proposed unsupervised adaptation technique, "10. pUA(mo)", matched the accuracy of the conventional monophone-based technique, "7. cUA(mo)", and the total computation time was also less than that of the baseline technique, "1. m+f: baseline". The proposed adaptation process contributed to the higher speed since the adaptation effect increased the beam search efficiency. The proposed adaptation technique, "10. pUA(mo)", has better accuracy than either slow "9. pUA(all)" or fast "11. pUA(s/p)", so our proposed monophone constraint is optimal as expected in Section 3.3.2. The advantage of state confidence appears in the difference between the proposed technique, "10. pUA(mo)", and no state confidence technique, "12. 10+w/oSC". There is some improvement due to the effect of considering state confidence and the validity of the approximation of Eq. (3.5).

Finally, the effect of power utilization (Pow: Power) is verified, see Table 3.5. With power utilization, the techniques demonstrate some advantages; the proposed power utilization approach is effective in improving accuracy.

The proposed technique is indicated by "14. pUA(mo)+Pow". This technique reduced the relative error in the correct rate by 13.7 %, which is statistically significant at $p < .0001$ based on the difference between the means of the two binominal distributions, and in the computation time by 17.9 % compared to the baseline technique, "1. m+f: baseline". Moreover, it offers over twice the speed of the conventional trigram-based adaptation technique under the 'ideal' gender-dependent condition with little degradation in accuracy.

Table 3.3: Compared techniques regarding unsupervised adaptation.

ID and abbr.	Statistic accumulation	Labeling
6. cUA(tri)	Forward-backward	Word trigram
7. cUA(mo)	Forward-backward	Monophone loop
8. cUA(s/p)	Forward-backward	Speech/pause loop
9. pUA(all)	Frame independent	All-state loop
10. pUA(mo)	**Frame independent**	**Monophone state loop**
11. pUA(s/p)	Frame independent	Speech/pause loop
12. w/oSC	Frame independent without state confidence	Monophone state loop

Table 3.4: Performance of unsupervised adaptation without power utilization.

ID and name	Cor.	Acc.	Sum.	Slct.	Adpt.
6. cUA(tri)	82.04	76.76	2.00	-	1.08
7. cUA(mo)	81.61	76.26	1.46	-	.512
8. cUA(s/p)	81.32	76.02	.936	-	.022
9. pUA(all)	81.52	76.25	1.36	.008	.456
10. pUA(mo)	**81.63**	**76.38**	**.922**	**.008**	**.020**
11. pUA(s/p)	81.39	76.20	.918	.008	.008
12. 10+w/oSC	81.53	76.24	.924	.008	.019

Table 3.5: Effect of power utilization in unsupervised adaptation

ID and name	Cor.	Acc.	Sum.	Slct.	Adpt.
1. baseline	79.22	73.79	1.00	-	-
6. cUA(tri)	82.04	76.76	2.00	-	1.08
13. cUA(tri)+Pow	82.07	77.15	1.92	-	1.08
10. pUA(mo)	81.63	76.38	.922	.008	.020
14. pUA(mo)+Pow	**82.07**	**77.00**	**.820**	**.008**	**.020**

3.5 Summary

This paper proposes fast unsupervised adaptation by accumulating the acoustic statistics efficiently using the frame independent output probabilities of speech (gender) / pause / monophone models. The proposed technique segments each utterance individually and selects a gender model per ut-

terance simultaneously with less computation time than if only gender and pause models are used. It offers fast unsupervised adaptation using monophone states by accumulating the statistics with the best state's confidence within a monophone per frame against the selected gender dependent acoustic model. After the adaptation, the approach uses a power term in the speech recognition process to improve accuracy. Tests showed that our technique reduced the relative error in the correct rate by 13.7 %, which is statistically significant at $p < .0001$ based on the difference between the means of two binominal distributions, and the computation time by 17.9 % compared with the baseline without prior gender selection and unsupervised adaptation. Furthermore, the proposed technique reduced the computation time by 57.3 % while exhibiting an accuracy equivalent to that of the conventional adaptation technique.

Chapter 4

Efficient Data Selection for Speech Recognition Based on Prior Confidence Estimation Using Speech and Monophone Models

4.1 Introduction

Massive quantities of videos and dialogs are stored every day; typical examples are video sharing services on the Internet and call center services provided by companies. Speech recognition technologies can transcribe the spoken components of these items automatically thus making the items searchable via their transcripts [24]. Several studies have analyzed customer needs by employing text mining [26, 25] and extracting the reasons for the calls [37] from stored conversational spoken documents. A typical call center will store several tens of thousands of calls per day, and we believe that not all calls should be transcribed for the following three reasons. 1) The computation cost involved in transcribing all calls is excessive. 2) An informative analysis can be achieved from a subset of the calls. 3) The quality of the recorded speech samples varies [27], and erroneous speech recognition (due to the poor input) will degrade the efficiency of subsequent spoken document retrieval [52] and analysis.

Several confidence measures have been proposed for identifying "accurate" speech samples [45]. Unfortunately, they require the computationally expensive step of speech recognition processing to obtain confidence scores, which are estimated from the recognition results; they waste considerable computer resources on samples that will eventually be rejected. Most conventional methods target word or utterance verification. A dialog (similar to spoken document) level confidence measure has been proposed [53], but it is also computationally inefficient because it requires several features including speech recognition results to estimate confidence. Several data selection methods have been proposed [54], but their target is to select training data, so they fail to reduce the computation cost significantly.

Our proposal efficiently identifies speech samples that will be well recognized with an extremely low computation cost prior to speech recognition. It can identify those samples that have

high confidence levels from massive numbers of stored speech samples. Prior confidence must be estimated rapidly because speech recognition can only proceed after the estimation results have been received. The proposed estimation technique utilizes the acoustic model used for posterior speech recognition. The proposal uses only context independent (monophone) models and speech models to reduce the computation cost. For even greater efficiency, its confidence estimation step eliminates all processing other than the calculation of acoustic output likelihood from Gaussian Mixture Models (GMMs). The prior confidence is calculated frame by frame from the difference between the output log-likelihoods of the monophone and speech GMMs. This confidence formulation is an approximation of the state level posterior probability with the state occurrence probability. This paper evaluates the actual efficiency of our technique in speech recognition and spoken document retrieval tasks. Experiments show that the proposed technique is significantly faster than the conventional posterior confidence measure based on speech recognition, while maintaining equivalent data selection performance.

The rest of this paper is organized as follows. Related work is outlined in Section 4.2. The proposed technique is described in Section 4.3. Section 4.4 introduces experiments conducted to confirm the effectiveness of the proposed technique. Our conclusion is presented in Section 4.5.

4.2 Related work on data selection for speech recognition and its application

Since there are many factors that cause variability in speech signals [27], the recognition accuracy is strongly dependent on the data. Several data selection methods have been proposed for training [54, 55] and adapting [56] acoustic models for speech recognition. Wu *et al.* also selected data to be transcribed for training by using the confidence score [57]; this technique is called active learning. A great number of confidence measure methods have been proposed [45] and they could also be useful for selecting data during speech recognition processing, since inaccurately recognized data impacts negatively on the subsequent application. Stoyanchev *et al.* detected misrecognized words in spoken dialog systems [58]. Seigel *et al.* estimated a confidence measure at the word/utterance level by using conditional random fields (CRF). Ogawa *et al.* also used CRF directly to estimate the recognition rate rather than the confidence score both per utterance and per lecture at the spoken document level [59]. Asami *et al.* also estimated the spoken document confidence score by using contextual coherence [60]. Senay *et al.* detected low-quality documents by using a confidence measure and semantic consistency based on the latent Dirichlet allocation (LDA) model for spoken document retrieval [61]. Li *et al.* used semantic similarity to estimate a confidence measure for spoken term detection [62]. There are several confidence measure methods at a variety of levels depending on the application.

Conventional confidence measure estimations require speech recognition results; this means that a lot of computation time is required to recognize low-confidence and unuseful data, which should be rejected. Thus, we attempt to reject unuseful data at the document level to prevent harmful effects on the subsequent application prior to speech recognition. In a conventional approach,

Lee *et al.* proposed rejecting data before speech recognition by using noise GMMs [63]. However, this method could reject data at the utterance level and needs to know the noise type beforehand. Chang *et al.* also proposed a pre-rejection algorithm that enhances the robustness of speech recognition by using pitch correlation [64], which allows it reject seriously distorted speech signal during wireless communication. However, it fails to reject slightly distorted speech with pitch continuity. This paper proposes an efficient method for selecting useful data for speech recognition and subsequent spoken document retrieval at the document level, which consists of many utterances before speech recognition. In addition, since our main target speakers, i.e. operators (agents) in call centers, use headset-type close-talk microphones, the recorded speech has high SNR (speech to noise ratio) without distortion. In call center speech, it is more important to tackle spontaneous speech instead of noisy or distorted speech. Thus, we focus on acoustical confidence.

4.3 Proposed data selection based on prior confidence estimation

4.3.1 Formulation of prior confidence estimation

The most common confidence measure is based on the word posterior probability defined as follows;

$$P(\hat{\mathbf{W}}|\boldsymbol{O}) = \frac{P(\hat{\mathbf{W}})P(\boldsymbol{O}|\hat{\mathbf{W}})}{P(\boldsymbol{O})} = \frac{P(\hat{\mathbf{W}})P(\boldsymbol{O}|\hat{\mathbf{W}})}{\sum_{\mathbf{W}} P(\mathbf{W})P(\boldsymbol{O}|\mathbf{W})} \tag{4.1}$$

where $\boldsymbol{O}$ and $\mathbf{W}$ are an acoustic observation feature sequence ($\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T$) and its corresponding word sequence, respectively, $P(\mathbf{W})$ is word occurrence probability as given by the language model, and "$\hat{}$" means the word, state, or sequence with the highest likelihood. The normalization term $P(\boldsymbol{O})$ cannot be easily computed [65], so conventional schemes approximate it using the N-best list from speech recognition results as in [66].

It requires a high computation cost to extract word sequences by using a language model that covers a large vocabulary. To avoid this cost, our strategy dispenses with the language model and instead targets the state sequence $\mathbf{S}$ in Hidden Markov Models (HMMs);

$$P(\hat{\mathbf{S}}|\boldsymbol{O}) = \frac{P(\hat{\mathbf{S}})P(\boldsymbol{O}|\hat{\mathbf{S}})}{\sum_{\mathbf{S}} P(\mathbf{S})P(\boldsymbol{O}|\mathbf{S})} \tag{4.2}$$

Our proposal eliminates all processing steps other than frame-independent acoustic likelihood calculation to further reduce the computation cost; it ignores the transition probability in the same way as several speech recognition decoders [48], and uses only the Gaussian output probability from GMMs to estimate confidence frame by frame (i.e. frame-independent).

That is, the posterior probability $P(\hat{\mathbf{S}}|\boldsymbol{O})$ is approximately calculated from the frame-independent state posterior probability $P(\hat{\mathrm{s}}|\boldsymbol{o}_t)$ with best state $\hat{\mathrm{s}}$ against observed feature $\boldsymbol{o}_t$ at frame t for data length T as shown below;

$$P(\hat{\mathbf{S}}|\boldsymbol{o}) \simeq \prod_{t=1}^{T} P(\hat{\mathbf{s}}|\boldsymbol{o}_t) \tag{4.3}$$

where the frame-independent state posterior probability $P(\hat{\mathbf{s}}|\boldsymbol{o}_t)$, is calculated from the output probability $b_s(\boldsymbol{o}_t)$ of state s frame by frame as follows;

$$P(\hat{\mathbf{s}}|\boldsymbol{o}_t) = \frac{P(\hat{s})b_{\hat{s}}(\boldsymbol{o}_t)}{\sum_s P(s)b_s(\boldsymbol{o}_t)} \tag{4.4}$$

$$b_s(\boldsymbol{o}_t) = \sum_{m=1}^{M_s} w_{s,m} N_{s,m}(\boldsymbol{o}_t|\mu_{s,m}, \Sigma_{s,m}) \tag{4.5}$$

where $\hat{s}$ is the best state in frame t. M_s is the number of distributions belonging to state s, $w_{s,m}$ is the m-th mixture weight and $N_{s,m}(\cdot)$ is the m-th Gaussian distribution function with mean vector $\mu_{s,m}$ and covariance matrix $\Sigma_{s,m}$ of state s.

To increase the processing speed further, the proposed technique uses only monophones when calculating Eq. (4.4); $\hat{s}$ is the best state, i.e. the state with the maximum Gaussian output probability among the states of the monophone HMMs. The assumption is that triphones can be approximated by monophones, and this assumption is often used to improve the speed of speech recognition processing as in [67]. The monophone in our acoustic model is still trained by the acoustic features with triphone-based alignment. Accordingly, this assumption is a very reasonable approach to improving speed.

The denominator of Eq. (4.4), $\sum_s P(s)b_s(\boldsymbol{o}_t)$ is the sum of all states' (all phonemes') output probabilities; it can be approximated by the speech model as follows;

$$\sum_s P(s)b_s(\boldsymbol{o}_t) \sim P(g)b_g(\boldsymbol{o}_t) \tag{4.6}$$

where g is the state of the speech model (GMM) that is trained from the acoustic features of all phonemes i.e. all states. Our speech model has only a single state, so the occurrence probability of the speech model, $P(g)$, must be equal to 1 in speech frames. By assigning 1 to $P(g)$ in Eq. (4.6), we obtain the following.

$$\sum_s P(s)b_s(\boldsymbol{o}_t) \sim b_g(\boldsymbol{o}_t) \tag{4.7}$$

The second term in Eq. (4.7) is similar to the denominator in the second term in Eq. (4.1); thus this approximation is reasonable.

By substituting this expression into Eq. (4.4), the frame-independent posterior probability, $P(\hat{\mathbf{s}}|\boldsymbol{o}_t)$, is approximately calculated as follows;

$$P(\hat{\mathbf{s}}|\boldsymbol{o}_t) \sim \frac{P(\hat{s})b_{\hat{s}}(\boldsymbol{o}_t)}{b_g(\boldsymbol{o}_t)} \tag{4.8}$$

The occurrence probability of state $\hat{s}$, $P(\hat{s})$, can be calculated from the appearance frequency of state s. We herein assume that there is no significant difference between the state appearance frequencies of the acoustic training speech data and the target speech data; in particular, we use only monophone states, and the difference is not significant at the monophone level. Under this assumption, $P(\hat{s})$ is given by the following equation by using total occupancy $\Gamma(s)$, which reflects the appearance frequency of state s in the acoustic model training data.

$$P(\hat{s}) \simeq \frac{\Gamma(\hat{s})}{\sum_{s} \Gamma(s)} \tag{4.9}$$

The frame-independent confidence score, $c(\boldsymbol{o}_t)$, is transformed in the log domain from Eq. (4.8) as follows.

$$c(\boldsymbol{o}_t) = \log(P(\hat{s}) b_{\hat{s}}(\boldsymbol{o}_t)) - \log b_g(\boldsymbol{o}_t) \tag{4.10}$$

If the speech model is adopted as the Universal Background Model (UBM) and we ignore the state occurrence probability $P(\hat{s})$, Eq. (4.10) is similar to the likelihood ratio often used in speaker verification as in [68].

Prior confidence score C is calculated by normalizing the frame-level prior confidence score, $c(\boldsymbol{o}_t)$, by data length T to allow a comparison of speech samples with different lengths as follows;

$$C = \frac{\sum_{t=1}^{T} c(\boldsymbol{o}_t)}{T} \tag{4.11}$$

Fig. 4.1 summarizes the above-mentioned relational expression from a conventional confidence measure based on posterior word probability to the proposed prior confidence measure. Since our proposed prior confidence can be estimated by using the acoustic likelihood from only speech and monophone models, it is significantly faster than a conventional confidence measure.

4.3.2 Qualitative explanation of prior confidence estimation

Fig. 4.2 shows the difference between the log-likelihoods of a monophone and speech model against clear and ambiguous speech; this is a simplified figure just for explanation as each model has only one state and one distribution. The speech model is trained from the acoustic features of all phonemes in speech frames, so the distributions in the speech model have broader variances than the distribution in the monophone model. Accordingly, the speech model provides a comparatively stable log-likelihood regardless of speech quality. If the input speech is clear and similar to the training acoustic data (the expectation is for high accuracy), the input acoustic features are located around the mean of either monophone's distributions. In this case, the log-likelihood of the best monophone is larger than that of the speech model, and the prior confidence becomes higher. In contrast, with ambiguous speech (the expectation is for low accuracy), the input features are located on the side of the distributions and the monophone's log-likelihood becomes smaller, so the

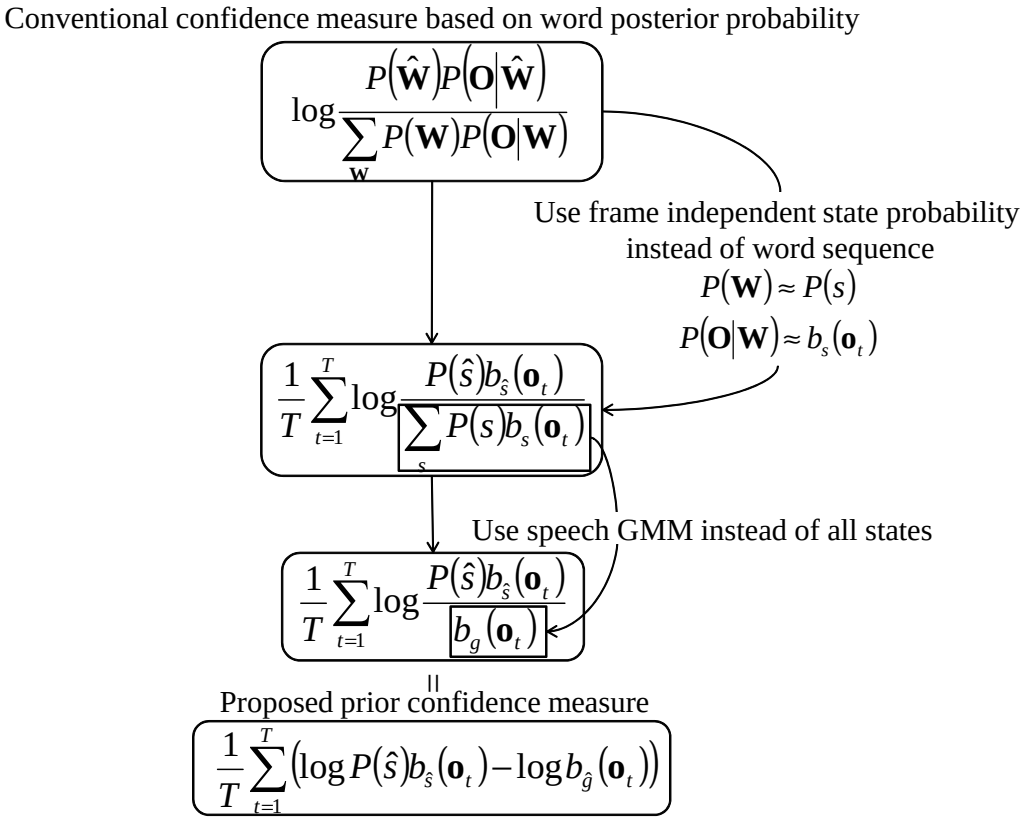

Figure 4.1: *Relational expression from conventional to proposed prior confidence measure.*

prior confidence becomes small. As a result, the difference between the log-likelihoods of the best monophone and the speech model reflects the expected accuracy in posterior speech recognition.

4.3.3 Procedure of proposed system

The procedure of the proposed system is shown in Fig. 4.3. The conventional system subjects all data to speech recognition. In contrast, the proposed system estimates confidence and selects the samples to be passed to speech recognition. In prior data selection, the speech data is ranked by the estimated prior confidence. The proposed system selects those samples that have high confidence scores and then performs speech recognition on the selected samples using triphones in the acoustic and language models. The computation cost falls since the speech recognition step is minimized, and our proposal can efficiently identify well-recognized speech samples.

Fig. 4.4 focuses on the prior confidence estimation process. It calculates the output probability frame by frame from GMMs of monophone HMMs and the speech model in the acoustic model for each complete speech sample. Frame-level prior confidence $c(\boldsymbol{o}_t)$ is estimated from the difference between the log-likelihoods of the best states in the monophone (red filled circle in Fig. 4.4)

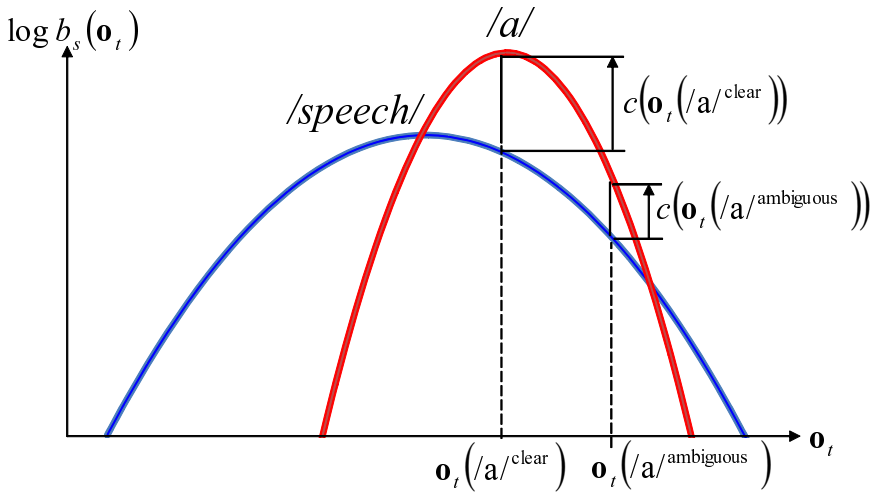

Figure 4.2: *The log-likelihood difference between best monophone (ex. /a/) and speech model.*

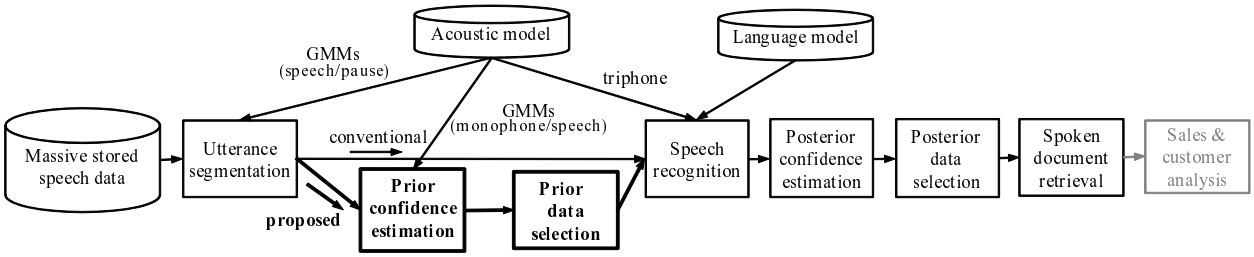

Figure 4.3: *Proposed system.*

and speech models as given by Eq. (4.10). Here, the acoustic features of speech and pause are significantly different. Thus, the proposed technique discriminates speech and pause by using the output probability of GMMs belonging to pause HMMs and the speech model, frame by frame. g is the best state (blue circle in Fig. 4.4) in the speech GMMs and pause HMMs in Eq. (4.10). Our proposed system adopts the utterance segmentation technique described in [69] before confidence estimation. Utterance segmentation uses speech GMMs and pause GMMs belonging to pause HMMs in acoustic model, and when the pause frame continues for longer than $\tau_p au$ (e.g. 0.8 sec), the utterance is segmented as an end-point. The segments include some pause frames by using a hangover scheme [70]. Prior confidence score C is calculated by averaging frame level confidence score as given by Eq. (4.11). We select data to send for speech recognition by using the prior confidence score C.

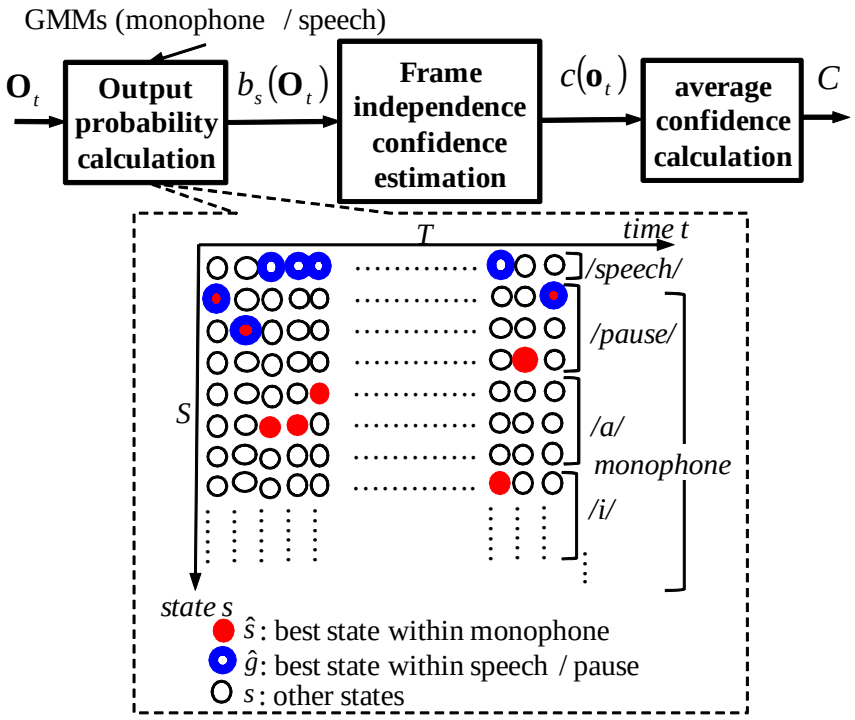

Figure 4.4: *Prior confidence estimation process.*

4.4 Experiments

4.4.1 Experimental condition and task

In this experiment the acoustic analysis condition is as described below; sampling frequency: 16 kHz, 20 msec length Hamming window shifted by 10 msec, and acoustic features: 25 orders (MFCC 12, ΔMFCC 12, Δpower). The evaluation task uses 240 calls (19.81 hours and 17,672 utterances) by 48 Japanese speakers (17 males and 31 females), and the speaking style is spontaneous speech in a two-party dialog, i.e. conversational speech. The recognition rates of the evaluation task are distributed over a wide range; 76.00 % on average, 89.75 % maximum and 47.98 % minimum. The ratio of data with over 80 % recognition rate is 30 % (= 72 /240).

The training data of the acoustic model contains spontaneous and conversational speech in a simulated call center, and the size of training data is 119.51 hours (103,822 utterances) for males and 104.50 hours (97,209 utterances) for females. There are a total of 1,958 states, 90 monophone states, and the number of phonemes is 30, the distribution number is 26,567 for males and 29,836 for females. There are 64 distributions in a single state of the speech model, 1,429 for males and 1,435 for females in monophones, and 26,567 for males and 29,836 for females in total. The maximum number of mixtures in each state is 16, but some states have fewer mixtures depending on the quantities of training data for each state.

The language model is a word trigram developed by using manual transcriptions of dialog speech, and its vocabulary size is 59,676 words. The size of the language model corpus is 44.69 mega words. The perplexity is 111.64, and the OOV (out of vocabulary) rate is 1.65 %. There are 1,500 queries for spoken document retrieval.

The speech recognition decoder is VoiceRex [51]. We used a dual-gender acoustic model and employed our proposed prior gender selection technique [69].

We start by investigating the impact of speech data selection and to this end adopt the metric of the average recognition rate of the selected speech samples. The effect of speech data selection is confirmed by comparing the following 5 conditions; "ideal": selection in descending order of recognition rate, "average": average recognition rate (should simulate random selection), "posterior": selection in descending order of confidence score (estimated by using recognition results after speech recognition based on word trigrams), "mono-loop": selection in descending order of confidence score (estimated by using phoneme recognition results with monophone loop grammar), and "prior": selection in descending order of our proposed prior confidence score. "posterior" adopts our baseline confidence measure that uses the N-best list of recognition results as in [66]. The "mono-loop" confidence measure is calculated by the difference between the acoustic scores of the recognition results of monophone loop grammar and speech / pause loop grammar; it eliminates the effect of the language model used in "posterior". This evaluation is based on character units to eliminate the influence of word length.

Furthermore, we evaluate the data selection performance in the spoken document retrieval task. We retrieve for the spoken documents (calls) by using text queries from 240 calls, the same data used in the previous speech recognition evaluation task. The queries are nouns, adjectives, and verbs. We adopt the mean average precision (MAP) for spoken document retrieval evaluation as in [71], and the data selection performance in spoken document retrieval is evaluated by using the raw average precision (AP) of the selected speech. MAP and raw AP are defined [72] as follows;

$$MAP = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{1}{N_i} \sum_{k=1}^{N_i} \frac{k}{rank_{ik}} \tag{4.12}$$

$$AP = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{N_i}{N} \tag{4.13}$$

where N_q denotes the number of queries, N_i denotes the number of relevant speech samples contained in the N retrieved documents for query q_i and $rank_{ik}$ denotes the rank of the k-th relevant document for query q_i.

The speed of confidence estimation is the time taken to compute the confidence score of the speech. Thus, the "posterior" computation time includes speech recognition processing. Instead of generating the recognition result with maximum likelihood, the proposed technique searches for the best state in terms of monophones frame by frame, hence this computation time comparison is fair with regard to confidence estimation.

4.4.2 Results

The recorded data selection performance is shown in Figs. 4.5 and 4.6. The horizontal axis is the speech data selection rate; it is calculated as the ratio of the number of selected speech samples to the number of all speech samples. The vertical axis in Fig. 4.5 is the average recognition rate of selected speech, and is the raw average precision for spoken document retrieval in Fig. 4.6. The mean average precision is shown in Table 4.1. The confidence estimation time is shown in Table 4.2. The computation time is normalized by that of "posterior" in Table 4.2.

Results in speech recognition

The effect of our data selection proposal is also shown in Fig. 4.5. The proposed "prior" confidence estimation achieved better recognition rates than "average" under all selection rates, so our data selection proposal improved the average recognition rate of the selected speech. It also matched the recognition rate of "posterior" for most selection rates. There was a considerable difference between "prior" and the conventional "posterior" around a selection rate of 10 %. However, we consider this to be due to the effect of the language model used in "posterior" because "mono-loop" also suffered a drop in recognition performance around this selection rate. Besides, recognition rates are degraded even if just a few bits of low-accurate data are selected. The recognition rate increased as the data selection rate decreased; e.g. the correct recognition rate exceeded 80 % at a selection rate of 20 %. Thus, our proposed technique can identify well-recognized speech data before speech recognition.

Results in spoken document retrieval

The evaluation results of our data selection proposal for the spoken document retrieval task are shown in Table 4.1 and Fig. 4.6. The proposed "prior" technique achieved equivalent performance in terms of both mean and raw average precision. In particular, our proposal achieved better than "average" at all selection rates and its improvement of precision increased as the selection rate was reduced as evaluated using the raw average precision. Therefore, our proposed data selection technique can retrieve spoken documents accurately.

Results of computation time

The computation time of confidence estimation is shown in Table 4.2. Our proposal is significantly faster than the conventional "posterior" technique. It is over 50 times faster for confidence estimation. Another important point is that "prior" is also faster than "mono-loop". This reveals that the proposed technique eliminates all processing except for acoustic Gaussian likelihood calculation.

Table 4.3 shows the effect of our speech data selection proposal. It lists the average recognition rate, the raw average precision, and the total computation time with speech recognition processing for several selection rates. Without data selection (selection rate = 100 %), the total computation time is increased due to the overhead processing of the prior confidence estimation, but the increase

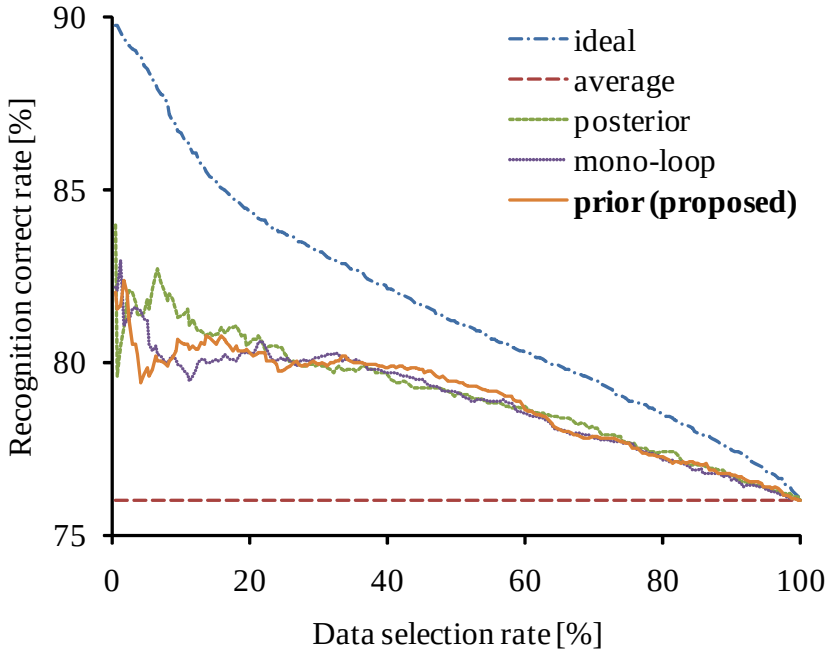

Figure 4.5: *Speech data selection performance in terms of recognition rate.*

Table 4.1: Spoken document retrieval performance in terms of Mean Average Precision (MAP).

ideal	posterior	mono-loop	**prior (proposed)**
58.26	57.63	57.44	**57.43**

Table 4.2: Computation time for confidence estimation.

posterior	mono-loop	**prior (proposed)**
1.00	.0722	**.0184**

is slight. As the selection rate falls, the superiority of the proposed technique strengthens in terms of the recognition rate, raw average precision and processing speed.

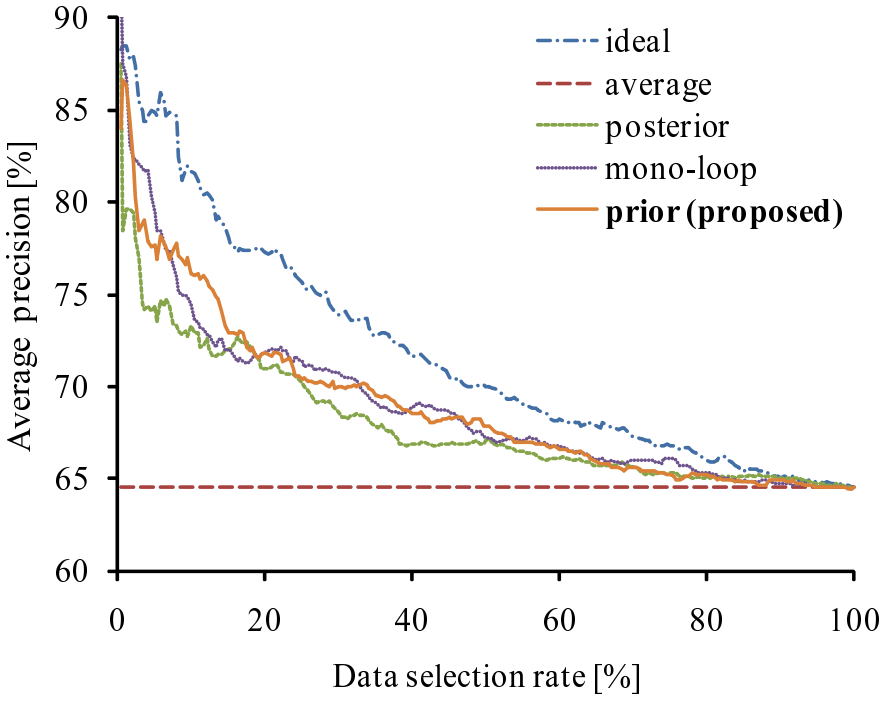

Figure 4.6: *Speech data selection performance in terms of raw Average Precision (AP).*

4.4.3 Discussion

At first, we focus on the difference between the posterior and prior confidence measures. The aim of our proposal is to estimate the overall confidence for each call with a large number of sentences. Fixed phrases (e.g. opening greetings), which are expected to be well recognized, constitutes just one factor in confidence estimation, thus the effect of fixed phrases is not very critical. However, the conventional confidence measure increases with fixed phrases, since fixed phrases provide a higher language score. Instead, our proposed prior confidence estimation approach uses only an acoustic model without the language model, and so the proposed data selection procedure is performed independently of the sentences and these word contexts in a call.

Table 4.3: Effect of proposed speech data selection.

Selection rate [%]	10	20	30	40	50	60	70	80	90	100
Recognition correct rate [%]	80.58	80.31	79.97	79.85	79.44	78.65	77.86	77.27	76.75	76.00
Average precision [%]	76.11	71.84	69.98	68.54	67.83	66.62	65.62	65.19	64.90	64.53
Average recall [%]	13.75	25.89	36.97	45.57	52.91	58.90	64.47	68.16	71.99	74.53
Total computation time	.155	.279	.394	.497	.589	.678	.769	.855	.945	1.02

Second, we discuss the effectiveness of rejecting low-accuracy calls. If an important word is difficult to recognize (e.g. an unpronounceable service name), specific calls containing that important word cannot be retrieved because the prior confidence becomes low owing to the pronunciation problem. Since our call analysis assumes the use of spoken documents with the recognition results, specific calls also cannot be retrieved because of recognition error. It is therefore reasonable to ignore calls with many recognition errors.

Third, we discuss optimization of the data selection rate. Considering the subsequent text mining, we believe that the recognition rate should exceed roughly 80 %, and so the adequate data selection rate is below around 30 % in our experiment. If ten thousand calls arrive every day, 35 ($\sim 19.81\,[hour]\,/240\,[call]*10.000\,[call]\,/24\,[hour/processor]$) speech recognition processes are required to recognize all calls per day. In the case that we can uses only ten recognition processes, we have to select the data at the rate of 30 % ($\sim$ 10 / 35) by the following day. The data volume decreases significantly by around 1/3. However, since many data are stored every day at call centers, the volume can be recovered by increasing the number of days that data are retained.

Finally, we discuss the performance difference between ideal and confidence data selection. There is a large gap between 'ideal' and confidence scores in data selection. This is because the correlation between recognition rate and confidence score is not sufficiently high. Of particular note, since our target task is spontaneous speech which includes ambiguous utterances, the superiority of the recognition result is often not obvious compared to the other recognition candidates; the correlation between the estimated confidence score of the recognition result and the correct recognition rate became small.

4.5 Summary

This paper proposed a rapid prior confidence estimation technique for selecting speech samples that will yield accurate recognition results. It reduces the computation cost since it selects only the best samples for speech recognition based on prior confidence estimation; only a Gaussian acoustic likelihood computation with speech and context independent models is needed. Simulations showed that our confidence estimation technique is over 50 times faster than the conventional posterior confidence measure based on speech recognition, and about 3.9 times faster than the monophone-loop confidence measure based on phoneme recognition. The proposed technique matched the selection performance of the conventional technique, and also improved the precision of the spoken document retrieval task.

Chapter 5

Efficient Beam Width Control to Eliminate Excessive Speech Recognition Time Based on Score Range Estimation

5.1 Introduction

Massive amounts of speech data are stored on a daily basis; typical call centers store several tens of thousands of calls per day. Speech recognition can transcribe these speech data automatically which makes them searchable via the transcripts [24]. Several studies have analyzed customer needs by employing text mining [26, 25] to extract the reason for the call [37] from stored speech documents. To rapidly improve business effectiveness, the analysis results must be extracted by processing this massive dataset on a daily basis; i.e. all calls should be transcribed by the morning of the following day. However, the computation time needed to auto-transcribe a lot of conversational speech can be excessive. In particular, poor quality speech data require a long time to auto-transcribe since they produce hypotheses with no significant score differences, which degrades the pruning efficiency in beam search. Since poor speech data yields erroneous transcripts, the data are of no use in subsequent spoken document processing and should be removed by using confidence measures [52, 60]; recognition time is wasted by producing transcripts that are not useful.

This paper aims to reduce the computation time needed for recogizing the speech data that will yield erroneous transcriptions. To minimize the time taken for speech recognition, several techniques have been proposed that optimize the decoding parameters [73] under a speed constraint [74]. Several adaptive pruning techniques have been proposed [75, 76] to tackle the speech variability problem [27]; they adapt the parameters during decoding. However, both techniques use a development set to determine the parameters. Thus, they fail to fully optimize the parameters if the quality of the target speech is substantially different from that of the development set. To reduce the excessive computation time, histogram pruning [77] is an often-used approach. Since it is not as effective as score beam pruning, pruning criteria are often used in combination [78]. Since histogram pruning is performed frame-wise based on the *instantaneous* score distribution, the beam width should be wide enough to maintain recognition accuracy. Predictive pruning is

also performed based on aspects of the score in the *near* future [79].

On the premise that we are processing stored speech data as in call centers, we target each speech item (i.e. call) directly instead of depending on a development set. The proposed approach controls the score beam width prior to decoding and incrementally in subsequent decoding for each speech item based on a *prolonged* (i.e. not *instantaneous*) score spread. The proposed technique formulates the score range within the beam width, and reduces the speech recognition time by maintaining that range. The score range is rapidly estimated by using only those Gaussians that belong to monophones prior to decoding. There is the possibility that the computation time might be increased by other factors such as confusion at the word and triphone levels, which might not be reflected in the prior monophones' score spread. To better handle the speech data that take inordinately longer than the base computation time, we also restrict the beam width in decoding based on the range by using the processed decoding speed and the time available to process the remainder of the target speech; this yields highly effective timeout control since the proposed method is assured of performing recognition up to the end of the speech segment whereas simple timeout control is likely to terminate before the segment's end which degrades recognition performance.

We evaluate the efficiency of the proposed technique in spontaneous speech recognition tasks with several speech quality levels; i.e. the SNR is varied from 0 dB to ∞ dB. Experiments show that our technique satisfies the speed constraint regardless of the quality of the target speech, and matches the accuracy achieved with ideally optimized parameters. Furthermore, the proposed technique reduces the recognition time needed to transcribe the speech data recorded in an actual call center with no significant drop in accuracy.

The rest of this paper is organized as follows. Related work is outlined in Section 5.2. The proposed technique is described in Section 5.3. Section 5.4 introduces the experiments conducted to confirm the effectiveness of the proposed technique. Our conclusion is presented in Section 5.5.

5.2 Related work on decoding parameter control

The decoding parameters, including beam width, determine speech recognition performance; i.e. accuracy and speed. Decoding parameter optimization is classified into two types: offline and online. Offline-type parameter optimization methods have been proposed that use optimal curve tracking [73] or linear programming [80, 81]; they offer overall optimization with training data. Several methods attempt to improve accuracy by using a response probability model [82] or search error risk minimization [83, 84]; they also require a training phase in the offline step.

Since the decoding speed is dependent on machine and acoustical conditions, namely clean and noisy data [85], online-type parameter optimization methods are essential because the target speech can be used directly to improve performance. This is especially so if there are massive amounts of speech data that vary in quality. Thus, several online pruning methods employ predictive pruning [79], adaptive-beam pruning [86] and dynamic pruning [87] with a confidence measure [88, 89]. These methods don't impose a delay prior to recognizing the initial utterances. Predictive pruning methods aim to improve pruning performance by predicting the future score, but have no mechanism to reduce the excessive computation time for low quality speech since prediction doesn't

necessarily well handle low quality speech. The adaptive and dynamic methods aim to stabilize the surviving number of hypotheses at each frame. It is difficult to reduce the number of hypotheses without search error if instantaneous decisions are made in histogram pruning, and it is difficult to control beam width appropriately.

Most decoding parameter optimization methods require a speech recognition process even for initial utterances. As a result a lot of computation time is required and this reduces the decoding efficiency. The acoustic look ahead method can improve decoding efficiency by using some future time frames [90]. Without using full-scale recognition processing for initial utterances, we attempt to improve efficiency significantly by using a prolonged look-ahead process instead of short time frames with a limited number of Gaussian distributions before decoding. The process can control beam width efficiently since the prolonged frames stabilize the score's statistical distribution.

Bulyko adopted speed constraints to optimize the decoding parameters by using development data [74]. We also introduce a decoding speed constraint to reduce unwarranted and time-consuming processing during decoding.

5.3 Proposed beam width control based on score range estimation

Score beam pruning is the best-known method of controlling the computation time. It retains only those hypotheses whose score is close (within the score beam width) to the best state hypothesis [91]. Since the log-likelihood score distribution of hypotheses varies according to speech quality, the recognition computation time fluctuates with speech quality. The computation time remains stable if the survival rate of hypotheses can be kept at a constant level. The proposed approach estimates the prior score range from prolonged frames of each target speech item by using an acoustic model with a limited number of Gaussians. On the assumption that the prior score spread is proportional to the score spread in subsequent speech recognition events, we stabilize the recognition computation time by reducing the score beam width so as to keep the score range within the beam width just before speech recognition. To further reduce the cost of processing time-wasting speech data, we also reduce the beam width by estimating the required speed-up ratio from the processing speed and the remaining time as indicated by the amount of speech that remains to be processed.

5.3.1 Framework of proposed approach

The framework of the proposed system is shown in Fig. 5.1. The conventional system uses fixed decoding parameters (e.g. $B_{S_{\text{base}}}$) that were optimized against a development set. In contrast, the proposed approach uses the input target speech to estimate the prior score range within the beam width; it then performs speech recognition by using the controlled score beam width, $B_{S_{\text{target}}}$. To estimate the prior score range rapidly, we use only monophones as in [91]. The prior score range is calculated by the average monophone score spread in the log-likelihood scores from the monophone Gaussian Mixture Models (GMMs). Furthermore, the proposed technique offsets the

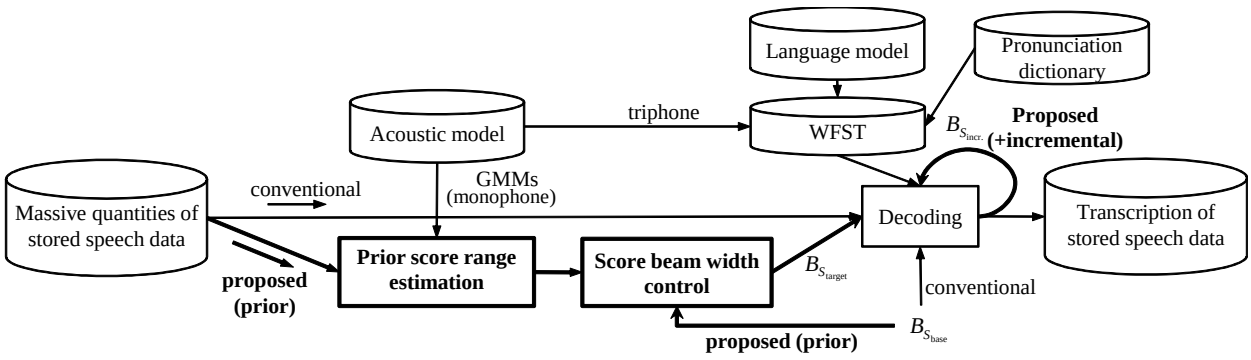

Figure 5.1: Schematic diagram of proposed system

beam width with the score range incrementally during speech recognition by using the speed-up ratio as estimated for each utterance.

5.3.2 Formulation of proposed approach

The proposed technique consists of two parts; prior and incremental beam width control. The former is performed just before decoding, and the latter is performed utterance by utterance during decoding.

Formulation of proposed prior beam width control

Score beam decoding prunes hypotheses whose log-likelihood scores fall under the threshold given by the best log-likelihood score minus the score beam width. If we assume that the hypothesis probability can be simply expressed by a Gaussian distribution, its log-likelihood score distribution, $y = f(x)$, is approximately represented by the logarithm of the Gaussian, namely the quadratic function (parabola) shown in Fig. 5.2. Here variable x indicates different hypotheses at each time frame during decoding, and the x-axis corresponds to a one-dimensional abstraction of the hypothesis space. We assume that the best hypothesis, μ, has the best log-likelihood score y_{best}, which is the vertex of the function at $x = \mu$. The other hypotheses are distributed around μ with log-likelihood y as in Fig. 5.2. Given the score beam width of B_S, the pruning threshold y_{th} equals $y_{best} - B_S$. At this point, the hypothesis survival rate after the pruning process corresponds to the range satisfying $y > y_{th}$, and is correlated with the summation of the log-likelihood score ($\sim$ probability) i.e. score range S: the size of the area within the beam width in Fig. 5.2. The range of log-likelihood scores is spread due to the variance of the Gaussian distribution. By normalizing score range S, we can keep the hypothesis survival rate constant, and therefore keep the computation time constant.

To calculate score range S, we swap the vertical and horizontal axes as in Fig. 5.3, and integrate

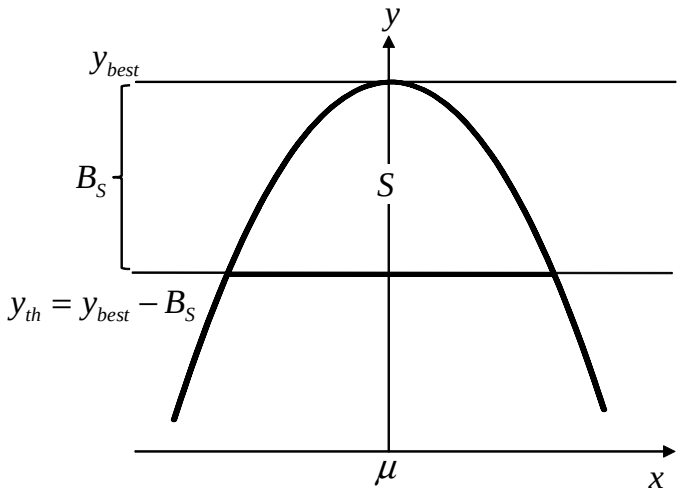

Figure 5.2: *Log-likelihood distribution.*

the inverse function $g(\cdot) = f^{-1}(\cdot)$ as follows;

$$\begin{aligned} S &= 2\int_0^{B_S} f^{-1}(\hat{y})d\hat{y} = 2\int_0^{B_S} g(\hat{y})d\hat{y} \\ &= 2[G(\hat{y})]_0^{B_S} = 2G(B_S) = const. \end{aligned} \tag{5.1}$$

where $\hat{y}$ is the difference score from the best log-likelihood score (i.e. $\hat{y} = y_{max} - y$) and $G(\cdot)$ is the integration function of $g(\cdot)$.

The log-likelihood score is approximated by using the logarithm of the Gaussian distribution $\mathcal{N}(x; \mu, \sigma)$, which is expressed by the following equation, where $-\log\sqrt{2\pi}\sigma$ corresponds to vertex y_{best}.

$$\begin{aligned} y &= \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \\ &= -\log\sqrt{2\pi}\sigma - \frac{(x-\mu)^2}{2\sigma^2} \\ &= y_{best} - \frac{(x-\mu)^2}{2\sigma^2} \end{aligned} \tag{5.2}$$

Thus, difference score $\hat{y}$ corresponds to the second term in Eq. (5.2) and can be converted using the following equation where $\hat{x} = x - \mu$ and $\alpha = \dfrac{1}{2\sigma^2}$.

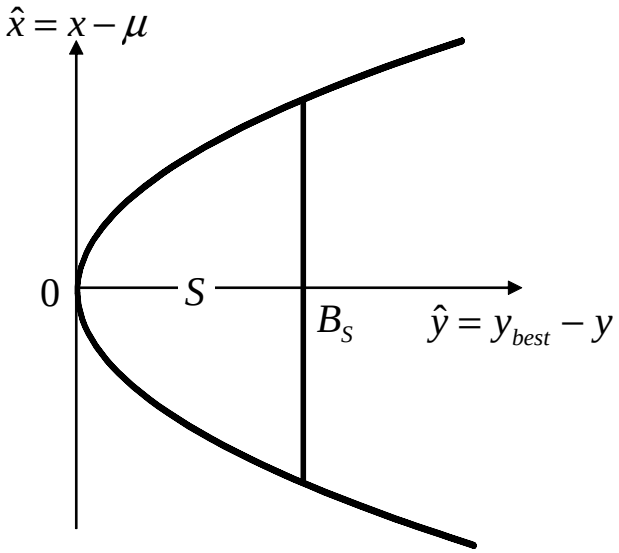

Figure 5.3: *Log-likelihood distribution with swapped vertical and horizontal axes.*

$$\hat{y} = y_{best} - y = \frac{(x-\mu)^2}{2\sigma^2} = \alpha\hat{x}^2 \tag{5.3}$$

Then, by converting Eq. (5.3) into $\hat{x} = \frac{1}{\alpha^{\frac{1}{2}}}\hat{y}^{\frac{1}{2}}$, the inverse function $g(\hat{y})$ becomes

$$g(\hat{y}) = \frac{1}{\alpha^{\frac{1}{2}}}\hat{y}^{\frac{1}{2}}. \tag{5.4}$$

From Eq. (5.4), integral function $G(\hat{y})$ is given by the following equation.

$$G(\hat{y}) = \int g(\hat{y})d\hat{y} = \frac{1}{\left(\frac{1}{2}+1\right)\alpha^{\frac{1}{2}}}\hat{y}^{\left(\frac{1}{2}+1\right)} = \frac{2}{3\alpha^{\frac{1}{2}}}\hat{y}^{\frac{3}{2}} \tag{5.5}$$

By substituting Eq. (5.5) into Eq. (5.1), we obtain the following.

$$S = 2G(B_S) = \frac{4}{3\alpha^{\frac{1}{2}}}B_S^{\frac{3}{2}} = const. \tag{5.6}$$

Therefore, score range S depends on beam width B_S and coefficient α, which is associated with variance σ^2 of the score distribution.

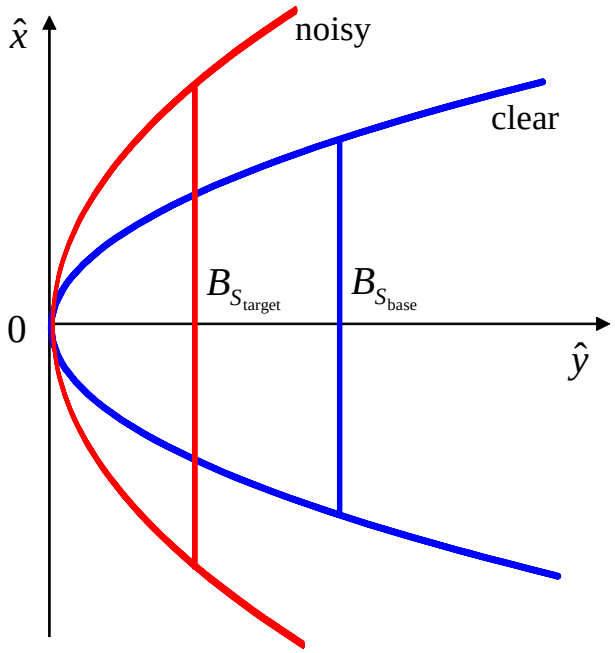

Figure 5.4: *Comparison of log-likelihood distributions of base and target speech.*

With clear speech, the difference score from the best log-likelihood score becomes large, and so the distribution becomes narrow as shown in Fig. 5.4. In contrast, with noisy speech, as also shown in Fig. 5.4, the difference score become small, and thus the distribution becomes wide.

Since the score range S is constant as shown by Eq. (5.6), the beam width relation between the target, $B_{S_{\text{target}}}$, and the base, $B_{S_{\text{base}}}$, is shown by using α_{target} and α_{base} as in the following equation. In Fig. 5.4, the base beam width is fixed by using clear speech; the target speech is noisy.

$$S = \frac{4}{3\alpha_{\text{target}}^{\frac{1}{2}}} B_{S_{\text{target}}}^{\frac{3}{2}} = \frac{4}{3\alpha_{\text{base}}^{\frac{1}{2}}} B_{S_{\text{base}}}^{\frac{3}{2}} \tag{5.7}$$

Target beam width $B_{S_{\text{target}}}$ is calculated from base beam width $B_{S_{\text{base}}}$ as follows;

$$B_{S_{\text{target}}} = \left(\frac{\alpha_{\text{target}}}{\alpha_{\text{base}}} \right)^{\frac{1}{3}} B_{S_{\text{base}}} \tag{5.8}$$

A review of Eq. (5.3) shows that there is a proportional relationship between $\hat{y}$ and α; $\hat{y} \propto \alpha$. $\hat{y}$ indicates the difference from the best log-likelihood score, i.e. the score spread. Instead of using the difference score among usual recognition hypotheses to calculate the score spread, we use

the average of frame-wise difference score $\overline{\hat{y}^{\text{mono}}}$ between the best and worst log-likelihood scores among monophones to reduce the computation time.

$$\frac{\alpha_{\text{target}}}{\alpha_{\text{base}}} = \frac{\overline{\overline{\hat{y}^{\text{mono}}_{\text{target}}}}}{\overline{\overline{\hat{y}^{\text{mono}}_{\text{base}}}}} \tag{5.9}$$

Thus, the target beam width can be calculated from the ratio of the average monophone's score spread and used to normalize the score range within the beam width as follows;

$$B_{S_{\text{target}}} = \left(\frac{\overline{\overline{\hat{y}^{\text{mono}}_{\text{target}}}}}{\overline{\overline{\hat{y}^{\text{mono}}_{\text{base}}}}} \right)^{\frac{1}{3}} B_{S_{\text{base}}} \tag{5.10}$$

Here, we calculate base score spread ($\overline{\hat{y}^{\text{mono}}_{\text{base}}}$) by using the development set beforehand. The beam width, $B_{S_{\text{target}}}$, is kept throughout the duration of the target speech item. Because the purpose of the proposed method is to eliminate unwarranted computation, the beam width is changed when $B_{S_{\text{target}}}$ is smaller than $B_{S_{\text{base}}}$.

We assume that the probabilities of the surviving hypotheses follow an approximately Gaussian distribution, and that their score spread depends on the distance between the observed feature and the acoustic model. Further, that the average score difference in monophone states has a predictable relationship with the subsequent score spread in speech recognition as given by Eq. (5.9), and the score beam width can be optimized using Eq. (5.10).

The proposed method is basically a 2-pass approach. We use GMMs in monophone HMMs to calculate the target beam width from the score spread ($\overline{\hat{y}^{\text{mono}}_{\text{target}}}$) in the 1st pass, and then use triphone HMMs to decode the target speech in the 2nd pass.

Formulation of proposed incremental beam width control

It is possible that prior beam width control cannot adequately minimize computation time since the decoding speed also depends on other factors, such as confusion at the word and triphone level, that are not reflected in the prior monophones' score range. On the assumption that we have stored speech, we can acquire the total data time D_{total} and the processed decoding speed (i.e. real time factor) R_{process} with the data time D_{process} and the computation time T_{process} during speech recognition. The proposed technique incrementally estimates the required decoding speed R_{require} from the remaining data time D_{remain} ($= D_{\text{total}} - D_{\text{process}}$) and the remaining computation time T_{remain} ($= T_{\text{total}} - T_{\text{process}}$) given the limited decoding speed R_{limit} ($= T_{\text{total}}/D_{\text{total}}$, e.g. 1.0), which reflects the speed constraints as in [74], see Fig. 5.5. The required decoding speed is achieved by using the speed-up ratio k ($= R_{\text{require}}/R_{\text{process}}$). Focusing on Eq. (5.6), since the computation time is proportional to score range S, the proposed technique incrementally alters score beam width $B_{S_{\text{incr.}}}$ ($= lB_{S_{\text{prev}}}$; the previous score beam width $B_{S_{\text{prev}}}$), as follows;

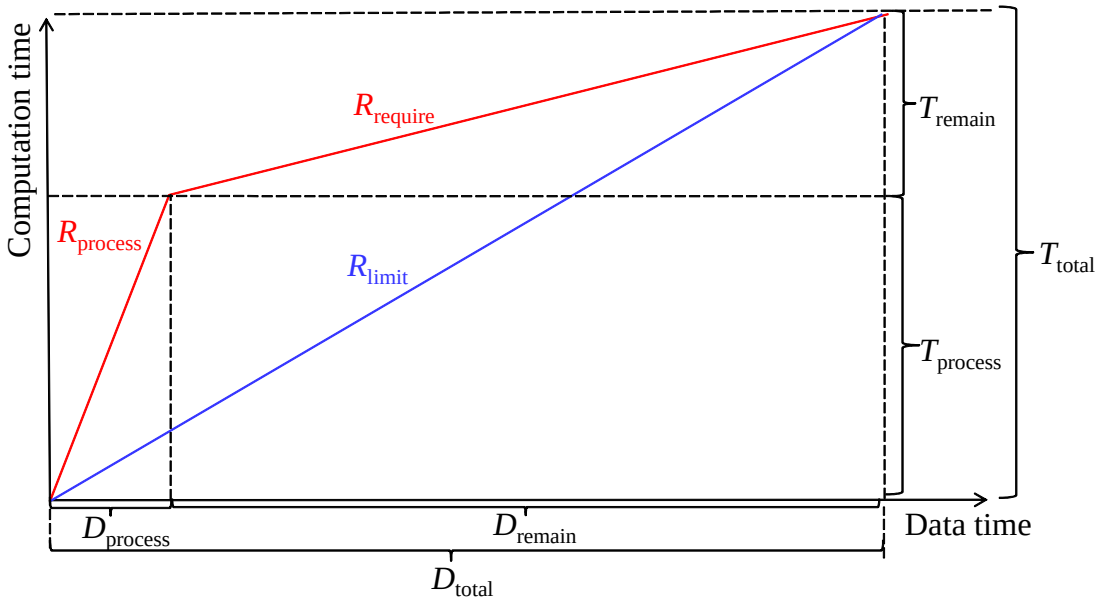

Figure 5.5: *Proposed incremental recognition time control.*

$$
\begin{aligned}
kS &= k\frac{4}{3\alpha^{\frac{1}{2}}}B_{S_{\text{prev}}}^{\frac{3}{2}} = \frac{4}{3\alpha^{\frac{1}{2}}}B_{S_{\text{incr.}}}{}^{\frac{3}{2}} = \frac{4}{3\alpha^{\frac{1}{2}}}\left(lB_{S_{\text{prev}}}\right)^{\frac{3}{2}} \\
kB_{S_{\text{prev}}}^{\frac{3}{2}} &= l^{\frac{3}{2}}B_{S_{\text{prev}}}^{\frac{3}{2}} \\
l &= k^{\frac{2}{3}} = \left(\frac{R_{\text{require}}}{R_{\text{process}}}\right)^{\frac{2}{3}} \\
B_{S_{\text{incr.}}} &= \left(\frac{R_{\text{require}}}{R_{\text{process}}}\right)^{\frac{2}{3}} B_{S_{\text{prev}}}
\end{aligned}
\tag{5.11}
$$

Here, the previous score beam width, $B_{S_{\text{prev}}}$, is equal to $B_{S_{\text{target}}}$ before decoding. Since we apply incremental beam width control utterance by utterance during decoding, the beam width changes frequently. To stabilize the computation time, we use the prolonged progressive average as base beam width $\overline{B_{S_{\text{prev}}}}$ and decoding speed $\overline{R_{\text{process}}}$ as indicated by the following equation. The beam width is adjusted only when the required decoding speed exceeds the speed up to this point.

$$
B_{S_{\text{incr.}}} = \left(\frac{R_{\text{require}}}{\overline{R_{\text{process}}}}\right)^{\frac{2}{3}} \overline{B_{S_{\text{prev}}}} \tag{5.12}
$$

5.3.3 Comparison of proposed and existing techniques

The proposed approach is more efficient than histogram pruning [77]. Histogram pruning restricts the surviving number of hypotheses based on the histogram-based score distribution. However,

pruning is based on instantaneous score histograms and is performed frame by frame, and so a larger beam width is required to maintain equivalent accuracy. The surviving number of hypotheses changes frame by frame, and thus histogram pruning can only work at the frames wherein the surviving number exceeds the beam width after constructing the score histogram. Our proposal for score beam pruning is more efficient since it works immediately if the hypothesis is not close to the best hypothesis. Furthermore, the proposed technique can stably control the score beam width by using the prolonged score spread calculated for all frames of the target speech. It can also adjust the width incrementally during decoding by using the processing speed up to this point and the remaining data length as determined from the remaining amount of unprocessed speech data.

5.4 Experiments

5.4.1 Experimental settings

The acoustic analysis condition in this experiment is as described below; sampling frequency: 16 kHz, 20 msec length Hamming window shifted by 10 msec, and acoustic features: 25 orders (MFCC 12, ΔMFCC 12, Δpower) or 26 orders (power utilization after adaptation). The evaluation task uses 240 speech samples (19.81 hours and 17,672 utterances) $\times$ 10 (SNR) produced by 48 Japanese speakers (17 males and 31 females); the speaking style is spontaneous speech in a two-party dialog. We use a dual-gender acoustic model, the total number of states is 1,958, the number of phonemes is 30, there are 90 monophone states, the distribution number is 26,567 for males and 29,836 for females, the size of the training data is 131.53 hours (114,289 utterances) for males and 123.44 hours (118,219 utterances) for females, and the training is based on differentiated Maximum Mutual Information (dMMI) [92]. The maximum number of mixtures in each state is 16, but some states have fewer mixtures depending on the quantities of training data for each state. The language model, a word trigram developed by using manual transcriptions of dialog speech, has vocabulary size of 59,676 words. The speech recognition decoder is VoiceRex [51, 93]. We used a dual-gender acoustic model and employed the proposed prior gender selection technique [69].

The metrics of the average recognition rate and computation time were used to assess the impact of the beam width optimization proposal on speech samples with several SNR values, from 0 dB to ∞ dB created by adding white noise artificially. The effect of the proposed beam width optimization is confirmed in a four-way comparison; “conventional”: previously-fixed beam width optimized by using development data, “ideal”: the beam width is optimized in a preliminary step by using target SNR speech data with the maximum recognition rate under a speed constraint as in [74], it simulates the ideal condition as the development and target data are the same, “proposed”: proposed beam width optimization; “proposed (prior)” uses only the proposed prior beam width control, and “proposed (prior+incremental)” employs both prior and incremental beam width control. Here, the speed constraint is that we must keep the computation time less than with sno added noise, i.e. SNR $= \infty$ dB. All compared techniques use histogram pruning [77] and number beam width is optimized by referring to development data ; i.e. we use both score and number beam

Table 5.1: *Recognition rate of proposed technique.*

SNR [dB]	conventional	**proposed (prior)**	**proposed (prior+incremental)**	ideal
∞	79.50	**79.50**	**79.47**	79.50
40	78.28	**78.26**	**78.24**	78.27
35	76.40	**76.33**	**76.28**	76.37
30	72.61	**72.39**	**72.35**	72.47
25	63.79	**63.46**	**63.42**	63.51
20	46.60	**45.65**	**45.66**	46.15
15	15.63	**15.84**	**15.71**	15.71
10	8.25	**8.09**	**8.09**	8.42
5	5.57	**5.86**	**5.86**	5.90
0	3.34	**3.32**	**3.32**	3.36

pruning in all techniques. To optimize the basic parameters and thus maximize the recognition accuracy, the development set consisted of mostly clean speech recorded in other call centers (176 calls [32.4 hour]).

We also compare the effectiveness of the "proposed" techniques with the above-mentioned "conventional" technique using speech recorded in an actual call center; 276 speeches (39.25 hours with 17,955 utterances). The remaining conditions are as noted at the head of this section.

5.4.2 Results and discussions

Experiment on speech with additional noise

The recorded beam optimization performance is shown in Table 5.1 and Fig. 5.6. Table 5.1 shows the average recognition rate [%] of speech for a given target SNR. The horizontal axis in Fig. 5.6 is the SNR (speech quality), and the vertical axis is the average computation time normalized by that of the conventional method with ∞ dB.

As shown in Table 5.1, there is no significant difference between the speech recognition rates; the proposed techniques match the recognition rate of the conventional technique regardless of SNR. "Proposed" and "ideal" beam control yield worse recognition rates than "conventional" beam control, since the latter faces no constraint on computation time; fortunately, the recognition rate discrepancy is slight. The effect of our beam optimization proposal is shown in Fig. 5.6. The "conventional" computation time depends on the SNR, and the computation time exceeds that of ∞ dB under several SNR conditions; e.g. the increase is 50 % at 20 dB. In contrast, the "proposed" computation time remains less than that of ∞ dB regardless of SNR; our prior beam width control is effective in reducing excessive the computation time of low quality speech. Since wide regions are buried in noise and are recognized as non-speech periods (pause) at low SNR ($<$ 15 [dB]), both techniques reduce the computation time compared to the 20 dB condition because they immediately pruning the many hypotheses whose scores are lower than the pause hypothesis.

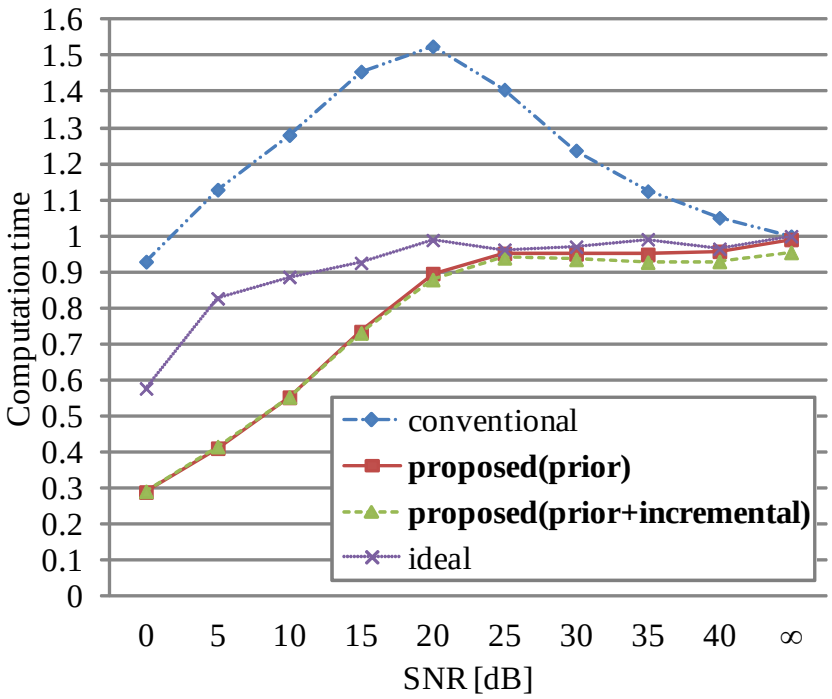

Figure 5.6: *Computation time of proposed technique.*

Table 5.2: *Performance of the proposed technique for speech recorded in an actual call center.*

	Correct	Accuracy	xRT	% of time-consuming process
conventional	88.17	78.74	.632	44.93
proposed (prior)	**88.12**	**78.70**	**.496**	**12.32**
proposed (prior+incremental)	**88.05**	**78.63**	**.471**	**9.42**

However, even at low SNR, the “proposed” techniques provide a significant reduction in computation time with no significant degradation in recognition rate. The proposed techniques achieve a slightly better decoding speed in combination with incremental beam width control at high SNR while providing the same accuracy.

Experiment on speech recorded in actual call center

Table 5.2 shows the performance with speech recorded in an actual call center, namely the average recognition correct rate/accuracy [%] (Correct/Accuracy) and computation time [x RT (Real Time)]. The rate of excessive time-consuming process is also shown in the table. We consider that

time-consuming speech is speech data whose computation time exceeds the average time needed by the “conventional” technique.

The proposed prior beam width control reduces the computation time by 21.5 % while maintaining the accuracy. The proposed technique also reduces the computation time by 25.4 % with no significant degradation in accuracy and the rate of excesive time-consuming process by 21.0 % compared with the “conventional” technique by combing the prior and incremental beam width control. The proposed incremental method can reduce the rate of time-consuming process to under 10 %. This means that the proposed technique can significantly reduce computer resource consumption.

5.5 Summary

This paper proposed an efficient score beam width optimization technique that is performed before and during speech recognition. It formulates the score range within the score beam width, and keeps the range constant by controlling the beam width with the score spread so as to maintain the decoding computation time. The score spread is calculated by using a limited number of GMMs belonging to monophones, and so it offers high speed. On the assumption that there is stored speech, we further control the beam width by estimating, utterance by utterance, the required decoding speed-up ratio for the score range during speech recognition processing to reduce the time by processing les of the intractable speech data. Recognition experiments on spontaneous speech show that the proposed technique maintains the decoding speed regardless of speech quality while matching the recognition accuracy of the conventional approach. Furthermore, the proposed technique recognized speech data recorded in an actual call center while reducing the computation time and the amount of unwarranted data processed significantly with effectively no drop in accuracy.

Chapter 6

Fast Acoustic Pre-processing against Recording Environment and Speaker Changes for Parliamentary Meeting Transcription

6.1 Introduction

Parliamentary meeting records are created daily by teams of stenographers. It would be far more efficient to create the meeting records from the transcriptions generated by automatic speech recognition. Several transcription systems are studied on the subjects of European parliaments [94][95][96] and the Japanese national congress [97]. Our mission is also to develop a Japanese transcription system for actual parliamentary meetings as a rival to [97]. However [97] uses ideal audio data captured by close-talking microphones and manually segmented into speaker-by-speaker. A more practical goal is to tackle actual audio data while keeping the recognition accuracy high. Furthermore, since the manually-corrected meeting records should be fixed quickly after the meeting is over, low latency is required in truly practical speech recognition.

Actual parliamentary meeting speeches pose two significant challenges:

1. the unique audio recording environment,
2. frequent changes of target speaker.

Speech is captured by close and distant microphones and loudspeakers are present in the actual parliamentary meeting room as shown in Fig. 6.1, so the outputs of distant microphones contain reverberation. We should eliminate the harmful effects of the distant microphones to realize accurate recognition. The typical parliamentary meeting consists of several speakers; chairperson, respondent and interpellator. We also have to tackle the issues raised by the intrinsic speech variations [27] like recording environment and speaker characteristics.

Our proposed system employs fast acoustic pre-processing with highly accurate speech recognition; it improves accuracy by offsetting variation in speaker and recording environment rapidly via limited Gaussian computation. First, the proposed method selects the close microphone's

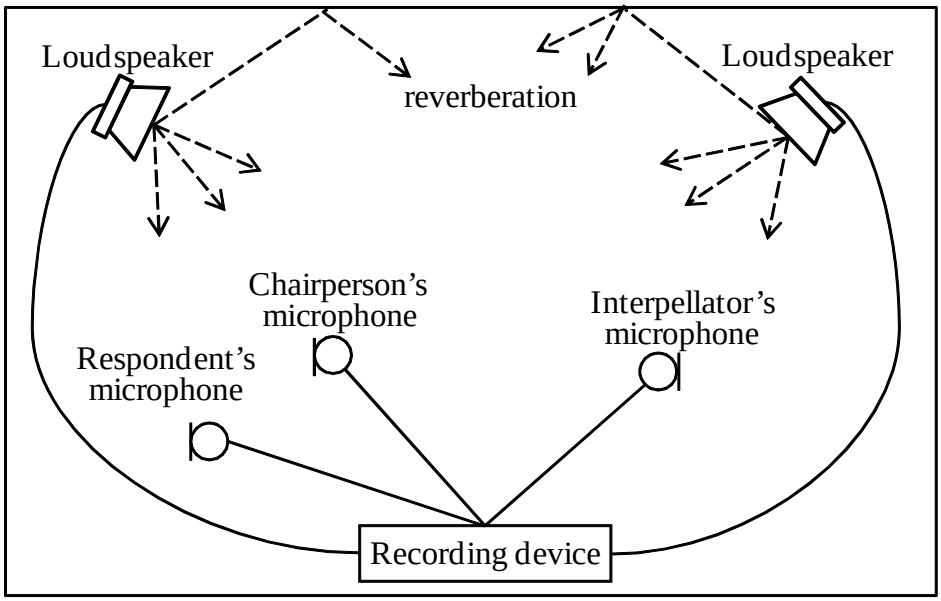

Figure 6.1: *Recording environment.*

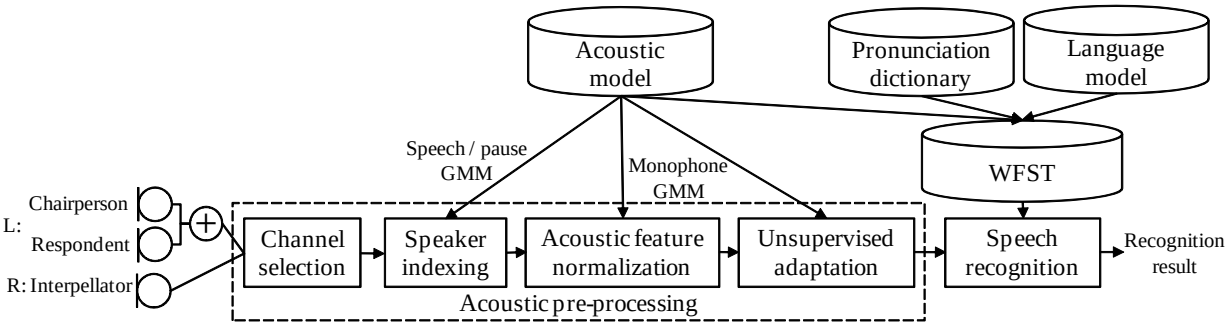

Figure 6.2: *Proposed system.*

signal automatically by comparing the powers, in the frequency domain, of close and distant microphones as in [98]; the selected signal is mostly composed of the target close speaker's speech, so the influence of the reverberation signal is quite small. Second, our method segments the selected signal speaker-by-speaker using clustering the utterances which are split by VAD (Voice Activity Detection) using speech/pause GMMs (Gaussian Mixture Models). Third, the acoustic feature vectors are normalized segment-by-segment using CMN (Cepstral Mean Normalization) [17][99], CVN (Cepstral Variance Normalization) [100] and VTLN (Vocal Tract Length Normalization) [101]; the proposed VTLN method rapidly estimates the frequency warping factor (α) by using a limited number of GMMs from context independent models (monophones). Fourth, we

adopt our proposed fast unsupervised adaptation based on MLLR (Maximum Likelihood Linear Regression) [30] by efficient statistics accumulation through the use of using adapted acoustic model to recognize the normalized acoustic features.

This paper uses actual meeting speeches captured in the parliament to assess the proposed system. Experiments show that the pre-processing proposal is fast enough that speech recognition processing speed is not negatively impacted; the proposed system matches the high recognition accuracy of the ideal recording condition even though its computation time is short.

The rest of this paper is organized as follows; the proposed system is described in Section 6.2. Section 6.3 introduces the experiments conducted to confirm its effectiveness. Our conclusion is drawn in Section 6.4.

6.2 Proposed system

Fig. 6.2 shows the framework of the proposed system. Our proposed acoustic pre-processing technique consists of four methods; 6.2.1) channel selection, 6.2.2) speaker indexing, 6.2.3) acoustic feature normalization, and 6.2.4) unsupervised adaptation. The first method, channel selection, is our countermeasure to the reverberation noise captured by distant microphones. The second method, speaker indexing, deals with the speaker change problem. The latter two methods, acoustic feature normalization and unsupervised adaptation, treat the intrinsic variation issues of recording environment and speaker characteristics. As shown in Fig. 6.2, the latter three methods utilize a limited number of GMMs belonging to the acoustic model. To start creating the meeting records immediately after starting the meeting, the input signal is split at a certain time intervals, and the segments are entered into the proposed system. The system yields the recognition results by application of the proposed acoustic pre-processing and speech recognition processing. Our proposed system employs the speech recognition processing based on WFST (Weighted Finite State Transducer) as in [102].

6.2.1 Channel selection

There are three people in our target parliamentary meeting; chairperson, respondent, and interpellator. Since there is little possibility that the chairperson and respondent speak at the same time, their two speech signals are mixed to yield a mono signal; it becomes the left (L) channel, while the interpellator's signal is set as the right (R) channel. The distant microphone's signal contains the reverberation noise created by the reflection of sound from the walls, thus it is not suitable for speech recognition. The proposed method selects the closer microphone's channel from the stereo right/left signals as the target speech to be recognized.

Our channel selection method is performed based on comparing the power of frequency bins extracted from the stereo signals as in [98]. To reduce the harmful influence of the inherent frequency characteristics of speech, we apply majority voting with the number of the superior bins to determine the candidate channel as shown in Fig. 6.3. The final candidate decision is made with consideration of utterance continuity and breath region. This approach prevents the loss of

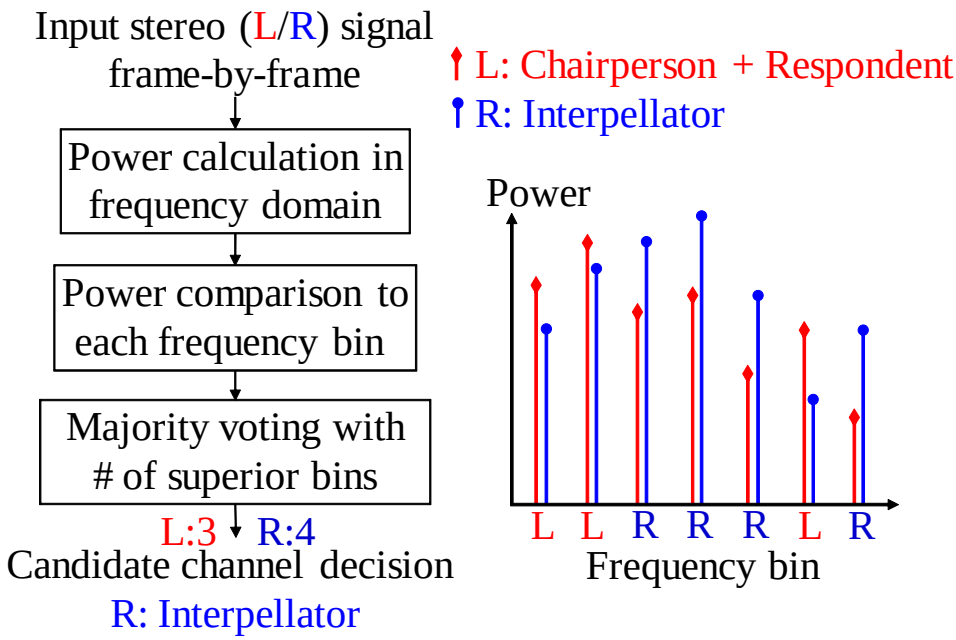

Figure 6.3: *Explanation of channel selection.*

low power consonants with hang-over time, while the ambient noise is suppressed by using noise reduction processing as in [103]. This proposed selection is rapid since it is a simplified method based on power comparison in the frequency domain with no Gaussian computation.

6.2.2 Speaker indexing

To improve recognition accuracy, the acoustic features of each speaker are normalized, and the acoustic model is trained by the per-speaker normalized features. Therefore, speaker indexing is performed to segment the input signal prior to speech recognition. Acoustic feature normalization is performed against the per-speaker segmented signal, and then speech recognition is executed.

Our speaker indexing proposal divides the input signal utterance-by-utterance, and forms speaker segments by clustering the divided utterances as shown in Fig. 6.4. Utterance segmentation is performed by using VAD with speech and pause models as in [104]. If the pause frame is continued over τ (e.g. 0.8 sec), the utterance is segmented. The speech model is composed of GMMs belonging to the context independent models (monophones) in the acoustic model, and the pause model is formed by GMMs belonging to pause HMM (Hidden Markov Model).

Since the input signal is divided at a fixed time interval, the number of speakers in each segment is limited. Conventional speaker indexing methods like variance-BIC [105] have difficulty in controlling the number of speakers, so we adopt speaker clustering as in [106]. However, the CLR (Cross Likelihood Ratio) used in [106] is very computationally expensive since it applies acoustic likelihood computation repeatedly, so we propose a simplified and faster method.

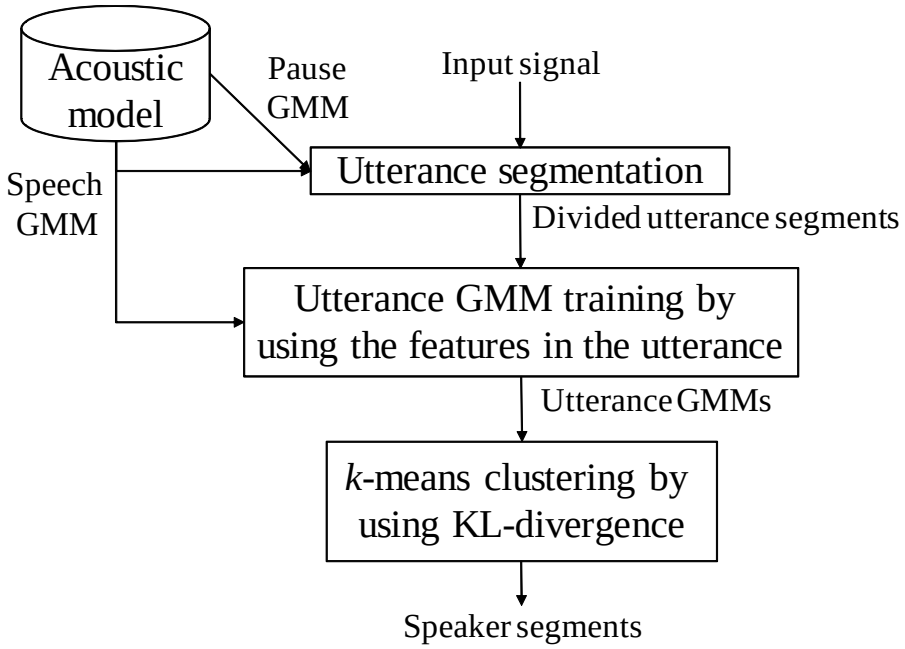

Figure 6.4: *Process flowchart of speaker indexing.*

The proposed method uses GMMs to represent utterance features. Here, the utterance GMM is trained by using the acoustic features in the utterance segment. We perform k-means clustering by using KL (Kullback Leibler) divergence as the distance measure between clusters, and utterance GMM as the initial cluster; the cluster is also represented by GMMs and the cluster GMM is composed from the utterances belonging to the cluster. Short utterance segments (e.g. < 1.0 [sec]) do not offer stable GMM generation, and so are excluded from the clustering. After clustering finishes, short segments are integrated with their neighboring cluster.

6.2.3 Acoustic feature normalization

We adopt the following three acoustic feature normalization methods; CMN [17][99], CVN [100] and VTLN [101]. The acoustic model is trained by the normalized acoustic features, and speaker indexing is performed after CMN/CVN against the entire input signal. The acoustic feature normalization of CMN/CVN/VTLN is run against the speaker segments generated by speaker indexing.

Our proposed VTLN estimates the frequency warping factor (α) with small computation cost. It seeks the best state, $\hat{s}$, with maximum log output probability within the states in the acoustic model against acoustic feature vector $\mathbf{o}_t(\alpha)$ at frame t. As shown in Eq. (6.1), it calculates the sum of log output probabilities in the best state sequence against the candidate warping factors α, and adopts $\hat{\alpha}$ that offers the maximum summation value. We reduce the computation cost by applying the state-wise constraint, which restricts the states for which output probabilities must be

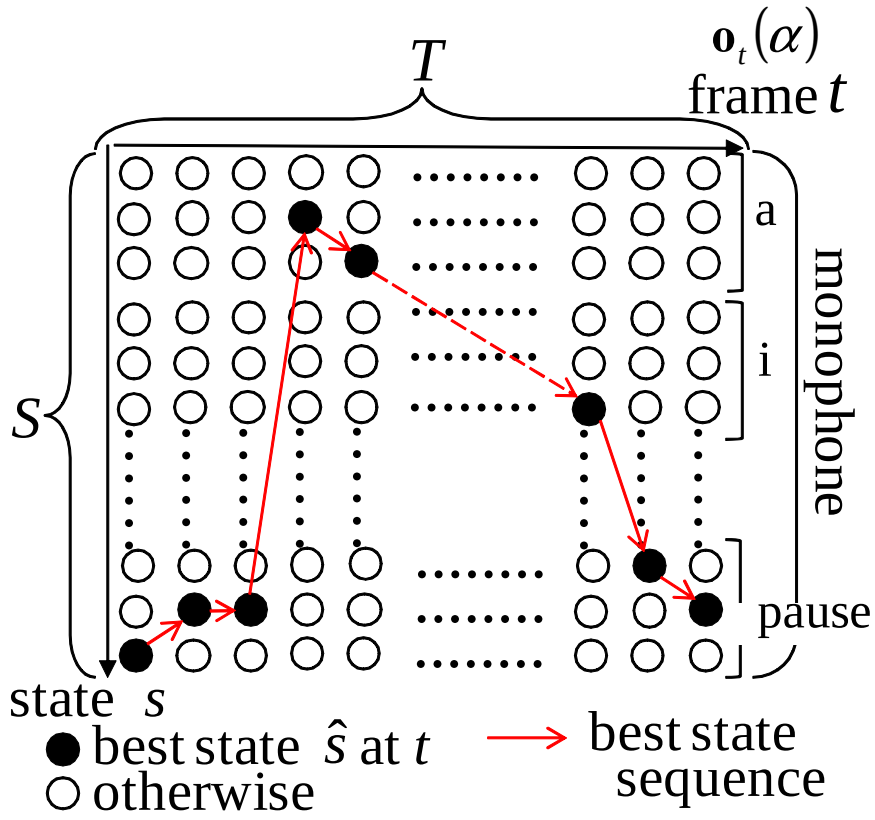

Figure 6.5: *Best state sequence with monophone constraint.*

calculated to monophone states as in Fig. 6.5. We also reduced the number of candidate factors, α, with no significant performance penalty as indicated by preliminary experiments. We further increase speed by adding the frame-wise constraint, which restricts the frame number needed to estimate α.

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \sum_{t=0}^{T-1} \log b_{\hat{s}}(\mathbf{o}_t(\alpha)) \tag{6.1}$$

6.2.4 Unsupervised acoustic model adaptation

For additional accuracy improvement, the unsupervised acoustic model adaptation is conducted against the speaker segments yielded by speaker indexing. Due to the computation time constraints, we employ the proposed fast unsupervised adaptation method as in [107][104]. The proposed unsupervised adaptation is based on the premise of a single-class MLLR [31], and a single global transformation matrix is estimated after accumulating statistics by using the frame independent output probabilities of all states' GMMs that belong to monophones.

Table 6.1: Performance of speech recognition

	Channel selection	Speaker indexing	Unsupervised adaptation	Cor.	Acc.
0.	off	off	off	82.11	75.62
1.	off	on	off	83.65	77.45
2.	on	off	off	85.07	78.59
3.	on	on	off	87.64	81.84
4.	**on**	**on**	**on**	**87.99**	**82.15**
5.	ideal	off	off	85.65	79.35
6.	ideal	on	off	87.70	81.87
7.	ideal	ideal	off	88.73	83.45
8.	ideal	ideal	on	89.93	84.83

Here, to approximate the occurrence probability $\gamma_t(s)$ at state s by Eq. (6.2), the statistics accumulation counts only the best state within frame t by using output probability $b_s(\mathbf{o}_t)$ at state s against the feature vector $\mathbf{o}_t$. The labeling for unsupervised adaptation has to get just the best output probability $b_{\hat{s}}(\mathbf{o}_t)$ by calculating S states belonging to monophones as also shown in Fig. 6.5. The statics $\gamma_t(s,m)$ of distribution m are accumulated by calculating the posterior probability of distribution m on state s based on approximate occurrence probability $\gamma_t(s)$. The single global transformation matrix is generated from these accumulated statistics using the model-space MLLR of [31]. The mean parameters of all distributions in the acoustic model (triphones as well as monophones) are transformed by this matrix.

$$\gamma_t(s) \simeq \begin{cases} \dfrac{b_{\hat{s}}(\mathbf{o}_t)}{\sum_{j=1}^{S} b_j(\mathbf{o}_t)} & \text{if } s \text{ is best state } \hat{s} \text{ at } t \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

6.3 Experiments

6.3.1 Experimental setting

The speech analysis conditions are as follows; 16 kHz sampling frequency, 30 msec Hamming window, 10 msec shift, and the order of feature parameters is 38 (MFCC 12, ΔMFCC 12, ΔΔMFCC 12, Δpower, ΔΔpower). The evaluation material is a 2.82 hour committee meeting with 17 speakers captured in the parliament. The acoustic model is gender and speaker independent following

[108]; it has 3,124 states and 49,984 distributions. The language model is word trigram built on [109], and the vocabulary size is 69,581 words. Speech recognition decoder is VoiceRex [51]. The acoustic model is trained using the normalized acoustic features output by CMN/CVN/VTLN, and these normalizations are used at all conditions because the acoustic model is trained with the normalizations. The input speech signal for this system was divided into segments; the length of each segment is over 300 [sec].

6.3.2 Experimental results

Speech recognition performance is shown in Table 6.1. We use the character-unit speech recognition correct rate (Cor.) and accuracy (Acc.) as the evaluation criteria. The proposed technique uses "on" at all entries with bold font, and the baseline uses "off" at all entries. The proposed channel selection is compared with "ideal": selects target speaker's channel based on hand labeling, and "off": mixes the stereo signals with the monaural signal without channel selection. The speaker indexing is compared with "ideal": speaker-by-speaker segmentation is correctly performed based on hand labeling, and "off" uses the segments divided at fixed time intervals without speaker indexing. The proposed unsupervised acoustic adaptation is compared to "off": does not employ adaptation.

The ideal recording condition, i.e. ideal channel selection and manual speaker indexing (7.), was assumed in previous research [97]; it used close-talking microphones and manually segmented the data into speaker-by-speaker units. Compared to this ideal condition, the proposed system (4.) keeps high accuracy with a slight degradation in recognition correct rate and accuracy; the difference is only around 1 point.

With regard to the effect of channel selection, the proposed method realized a significant recognition error reduction by using channel selection (0. $\rightarrow$ 2. and 1. $\rightarrow$ 3.). Compared to ideal hand-labeled channel selection (3. $\rightarrow$ 6.), it achieved equivalent recognition accuracy. These results demonstrate the effectiveness of our channel selection proposal.

The proposed method improved accuracy by using speaker indexing both with and without channel selection (0. $\rightarrow$ 1. and 2. $\rightarrow$ 3.). It also produced similar improvement under ideal channel selection (5. $\rightarrow$ 6.). These results reflect the effects of acoustic feature normalization (CMN/CVN/VTLN) per speaker.

With regard to unsupervised acoustic model adaptation, the proposed method improved accuracy (3. $\rightarrow$ 4.) with the proposed channel selection and speaker indexing. A similar improvement is also confirmed with ideal channel selection and speaker indexing (7. $\rightarrow$ 8.).

Table 6.2 shows the computation time for each pre-processing step, speech recognition, and the total times for the proposed system (shown as 4.). The computation time is normalized by the time length of the input speech file; i.e. it uses the RTF (Real Time Factor). This table also shows the ratio of each step against the total computation time.

The total computation time is around half the input file duration, so our system is very fast. The acoustic pre-processing proposal occupies about 22 % of the total computation time and its computation cost is considerably smaller than that of speech recognition.

Table 6.2: Computation time of speech recognition

	RTF	Ratio
Sum	.507	1
Acoustic pre-processing	.112	.220
· Channel selection	.026	.051
· Speaker indexing	.004	.009
· Feature normalization	.055	.108
· Unsupervised adaptation	.026	.052
Speech recognition	.395	.780

6.4 Summary

This paper proposed a highly accurate and fast acoustic pre-processing technique for creating records of meeting in the parliament. The first proposal, channel selection, compares power in the frequency domain in different channels to solve the problem of the reverberation noise in signals captured far from the speaker. The second proposal, speaker indexing, tackles the speaker change problem by using utterance clustering. The remaining proposals, acoustic feature normalization and unsupervised adaptation, deal with the variation intrinsic to the recording environment and speaker characteristics by using a limited number of GMMs in the acoustic model. Tests conducted on actual meeting speech recorded in a parliamentary room show that the proposed system basically matches the accuracy achieved with the ideal recording condition at twice the real-time speed.

Chapter 7

Conclustion

7.1 Preview of work

The aim of this thesis was to overcome the barriers that hinder the realization of practical applications based on speech recognition. We focused on three use cases; i) speech interface on tablet devices, ii) information extraction from speech samples stored in call centers, and iii) transcription system for parliament meetings. This work tackled the problems posed by the use cases with five techniques; all leverage the properties of the use cases as shown in Table 7.1.

The problems and available properties of the above-mentioned three use cases addressed by this thesis are described below.

Chapter 2 focused on the issue of improving the recognition accuracy of tablet devices under convolutional and additional noise with rapid response times. This technique leverages the property that background additional noise can be captured in a preliminary step; it allows us to create a noise adapted and normalized model that resolves this issue.

Chapters 3 to 5 introduce three techniques deal with the practical issue of collecting highly accurate spoken documents from massive volumes of spontaneous speech data under the limited computer resources available for information extraction in call centers. These techniques leverage the property that the target data is stored speech, and can be investigated prior to speech recognition.

In chapter 6, the critical issue is achieving high accuracy under processing time limits against changes in the recording environment and speaker for efficient parliamentary meeting transcription. Since parliamentary speech data is segmented and cached before recognition, we can introduce several acoustic pre-processing methods to index, normalize and adapt the target stored speech segments.

The proposed techniques were detailed together with the practical issues and properties of each use case.

Chapter 2 treats the issue that users require speech recognition systems that offer rapid response and high accuracy concurrently. Speech recognition accuracy is degraded by additive noise, imposed by ambient noise, and convolutional noise, created by space transfer characteristics, especially in distant talking situations. Against each type of noise, existing model adaptation techniques

achieve robustness by using HMM-composition and CMN (cepstral mean normalization). Since they need an additive noise sample as well as a user speech sample to generate the models required, they cannot achieve rapid response, though it may be possible to catch just the additive noise in a prior step. In the prior step, the technique proposed herein uses just the additive noise to generate an adapted and normalized model against both types of noise. When the user's speech sample is captured, only online-CMN need be performed to start recognition processing, so the technique offers rapid response. In addition, to cover the unpredictable S/N values possible in real applications, the technique creates several S/N HMMs. Simulations using artificial speech data show that the proposed technique increased the character correct rate by 11.62 % compared to CMN.

Chapter 3 proposed a fast unsupervised acoustic model adaptation technique with efficient statistics accumulation for speech recognition. Conventional adaptation techniques accumulate the acoustic statistics based on a forward-backward algorithm or a Viterbi algorithm. Since both algorithms require a state sequence prior to statistic accumulation, the conventional techniques need time to determine the state sequence by transcribing the target speech in advance. Instead of pre-determining the state sequence, the proposed technique reduces the computation time by accumulating the statistics with state confidence in a monophone per frame basis. It also rapidly selects the appropriate gender acoustic model before adaptation, and further increases the accuracy by employing a power term after adaptation. Recognition experiments using spontaneous speech show that the proposed technique reduces computation time by 57.3 % while providing the same accuracy as the conventional adaptation technique.

Chapter 4 proposed an efficient speech data selection technique that can identify those data that will be well recognized. Conventional confidence measure techniques can also identify well-recognized speech data. However, those techniques require a lot of computation time for speech recognition processing to estimate confidence scores. Speech data with low confidence should not go through the time-consuming recognition process since it will yield erroneous spoken documents that will eventually be rejected. The proposed technique can select the speech data that will be acceptable for speech recognition applications. It rapidly selects speech data with high prior confidence based on acoustic likelihood values with only speech and monophone models. Experiments show that the proposed confidence estimation technique is over fifty times faster than the conventional posterior confidence measure while providing equivalent data selection performance for speech recognition and spoken document retrieval.

Chapter 5 proposed a technique that provides efficient beam width control; it yields practical computation times for the auto-transcription of massive amounts of speech data. We focus on the fact that a lot of time is wasted in recognizing poor quality speech data that will ultimately yield erroneous transcriptions and provide no useful results. To stabilize the time regardless of quality, our proposed technique controls the beam width based on pre-estimated prolonged score spread against the target speech; it formulates the score range within the width and maximizes computation efficiency by regulating the range relevant to the hypotheses' survival rate. The proposed technique can control the width rapidly simply by using monophones prior to decoding. It also restricts the beam width in decoding by using the processing speed and remaining data time to better handle stubborn speech data. Experiments on actual call-center speech data with several SNR val-

ues confirm a reduction in computation time while matching the accuracy of existing techniques.

Chapter 6 proposed a fast acoustic pre-processing technique with automatic speech recognition for a system to transcribe parliamentary meetings. It well handles changes in the actual recording environment and speaker to keep recognition accuracy high with low latency. The proposed technique rapidly adapts the system to the environment and the speaker via limited Gaussian computation; it selects the target speaker's signal by comparing signal powers of close and distant microphones in the frequency domain and then segments the signal, speaker-by-speaker, using utterance clustering for acoustic feature normalization. It also employs fast unsupervised acoustic adaptation based on efficient statistics accumulation through the use of monophones. Experiments conducted on actual meeting speeches show that the proposed technique runs at twice the real-time speed with no significant degradation in accuracy compared to the ideal recording condition.

Table 7.1: The five issues and property in each use case

Use case	Issue	Property and technique	Chapter
Tablet device (Speech interface)	Accuracy improvement and high response under noise	Noise adapted/normalized model is generated by using pre-observed background noise	2
Call center (Speech mining)	Highly accurate speech document from massive data	Fast prior unsupervised adaptation can be performed since low latency is not required in stored speech	3
		Low accuracy speech is identified and not processed data	4
		Low quality speech shouldn't be recognized in full detail as doing so would waster resources.	5
Parliament speech (Speech transcription)	High accuracy under time limits	Fast prior normalization and adaptation are applied since speech samples are cached	6

7.2 Summary

This thesis targeted the goal of removing the barriers to the realization of practical systems based on speech recognition. Robust speech recognition systems are required to handle noise-corrupted speech and spontaneous speech. We developed a acoustic model adaptation and normalization technique for noisy speech recognition with tablet devices by using pre-observed noise. For spontaneous speech recognition, we developed three techniques, fast prior unsupervised adaptation using confidence scores, data selection based on prior confidence estimation, computation time reduction by control beam width before decoding for call center speech. We further developed fast acoustic pre-processing for transcribing parliament meetings.

ACKNOWLEDGMENTS

It gives me great pleasure to be able to receive my doctorate from The University of Tokyo, from whence I received my master's degree eleven years ago. I would like to thank Professor Keikichi Hirose who was my main supervisor, and Professors Tohru Asami, Hitoshi Aida, Hitoshi Iba, Nobuaki Minematsu, and Takeshi Naemura who acted as vice supervisors, for giving me this opportunity, as well as for their generous instruction during my PhD coursework. I particularly want to thank Professor Hirose for noticing my work in its early stages, and for offering me much advice, both general and detailed, as I pursued my research and constructed this thesis.

I started my research career during my time in the Department of Electrics at The University of Tokyo. In the 4th year of my bachelors degree I studied at the Saito-Aida Laboratory where I established my research style of identifying the unique attributes existing in actual field data. For the 2 years for my masters degree I studied at the Hirose-Minematsu Laboratory where I entered the research area of speech technology and studied practical issues and team work as essential factors in successful research. Without coaching from Prof. Hirose and Prof. Minematsu, I would not have been able to proceed with my research life.

This thesis was developed while at NTT Cyber Space and Media intelligence Laboratories, NTT Corporation. My continuous 11-year research on speech information processing has been supported and allowed to continue by Dr. Takeshi Kawabata, Mr. Akihoro Imamura, and Dr. Satoshi Takahashi, all group leaders, thanks to their understanding of my work. In this period, my research has developed greatly through various research discussions and communications within the Spoken Dialog System Group. Members of our speech recognition team, Mr. Osamu Yoshioka, Dr. Hirokazu Masataki, Mr. Yoshikazu Yamaguch, Dr. Katsutoshi Ohtsuki (currently at Microsoft Development Co, Ltd.) Dr. Atsunori Ogawa (currently at NTT Communication Science Laboratories) and Taichi Asami have always provided me with valuable technical knowledge with regard to speech recognition, and Dr. Sumitaka Sakauchi and Mr. Kazunori Kobayashi have given me abundant technical knowledge with regard to acoustic processing.

In addition to these colleagues, members of the technical support staff, Mr. Takeshi Konno, Mr. Kouji Ihigami, Mr. Koji Hattori, Mrs. Shinako Ishijima, Mr. Nobuyuki Tsuruta and Mr. Koudai Fukushi, have engaged in the development of the speech recognition research platform VoiceRex, a basic tool of my research activities. I am extremely grateful to all of them. Each person has provided a different viewpoint on speech recognition, and has also encouraged me along the way.

Mr. Michael Blackburn and Mr. David Meacock who have refined the English of most of my work, including this thesis.

Dr. Masato Miyoshi (currently at Kanazawa University), Dr. Tomohiro Nakatani, Dr. Takaaki

Hori and Dr. Shinji Watanabe (currently at MERL) of NTT Communication Science Laboratories have all given me valuable advice from their vast store of experience in speech recognition research. The part of my work dealing with automatic beam control was suggested by Takaaki Hori as he always pointed out the importance of this topic. To continue such work, I would like to maintain this good relationship between the research and development arms of NTT.

The part of my work dealing with acoustic pre-processing for parliament meeting transcription was based on the data gathered by Kyoto University by Professor Tatsuya Kawahara and the members of his laboratory. I must offer many thanks for their help.

I started the work described in this thesis with Dr. Satoshi Takahashi, Dr. Hirokazu Masataki, Mr. Yoshikazu Yamaguchi and Dr. Atsunori Ogawa, specialists in speech recognition. The hard work during that first research period and their strict teaching has a treasured place in my memory. Their teaching covered many aspects such as research and social postures, technical and business writing, how to conduct research, as well as research activities. They also showed me their own kindness and consideration, which was especially encouraging given that I had just commenced my research life. I truly feel it would have been hard to find another equally fortunate and advantageous learning environment with such knowledgeable supervisors.

Finally I would like to thank all of my friends, my parents, my brother and sister, my wife, my son and my daughter for their wonderful support throughout these years.

ACKNOWLEDGMENTS IN JAPANESE

愛する母校である東京大学で博士学位を取れることをこの上なく幸いに思います. 本学位論文の主査としてこのような機会を与えてくださった東京大学廣瀬啓吉教授ならびに副査の浅見徹教授，相田仁教授，伊庭斉志教授，峯松信明教授，苗村健教授には，その指導も含めて大変感謝しております. 特に廣瀬教授には本学位論文をまとめる上で適切な助言を頂きありがとうございます.

私の研究経歴は東京大学工学部電子工学科時代にはじまります. 学部 4 年の 1 年間所属した東京大学齋藤・相田研究室が，実環境のデータを対象にして独自性を目指すという自分の研究スタイルを形作ってくれたと思っております．齋藤忠夫教授，相田教授等の諸先生方の指導は今でも自分の身に染み付いております.

また，修士過程の 2 年間所属した東京大学廣瀬・峯松研究室が，音声分野との出会いの場であり，実用的な視点やチームワークといった研究には欠かせない要素を学ぶ機会を与えてくれた場であったと考えております．廣瀬啓吉教授，峯松信明教授等の諸先生方の指導がなければ，その後の研究者としての人生を歩むことは出来なかったと考えております.

本研究は日本電信電話株式会社 NTT サイバースペース研究所および NTT メディアインテリジェンス研究所での成果です．入社以来 11 年間首尾一貫して音声認識研究を続けられたのは，川端豪氏，今村明弘氏，高橋敏博士がグループリーダーとして本研究に常に理解を示しサポートしてくれたおかげです. その間，音声対話インタフェースグループや音声言語メディア処理プロジェクトに所属した際の様々な研究及び多くの方々との交流が本成果を生み出したといえます．認識サブグループの吉岡理氏，政瀧浩和博士，山口義和氏，大附克利博士 (現在マイクロソフトデベロップメント)，小川厚徳博士 (現在 NTT コミュニケーション科学基礎研究所)，浅見太一氏は音声認識に関する貴重な専門知識，阪内 澄宇博士，小林 和則氏は音響処理に関する貴重な専門知識を数多く提供してくれました.

上記のメンバーに加えて，今野健氏，石上宏二氏，服部浩司氏，石島姿子氏，鶴田信行氏，福士広大氏，をはじめとする研究補助員のサポートにより日々進化していった音声認識エンジン VoiceRex により本成果は実現できたといえます．彼らの本成果への貢献に大きく感謝します.

また，マイケル・ブラックバーン氏やデービッド・ミーコック氏の本学位論文を含めた英文添削も本研究を大きく助けてくれました.

さらに，三好正人博士 (現在金沢大学)，中谷智広博士，堀貴明博士，渡部慎治博士 (現在 MERL) をはじめとする NTT コミュニケーション科学基礎研究所のメンバーには音声認識の基礎研究の観点から貴重な意見を数々頂きました．特に 5 章の認識処理時間安定化の研究は，堀博士のアドバイスを受けて完成した成果であります．このような研究を今後も進めていくために，実用的研究と基礎研究の緊密な関係を保って生きたいです.

加えて 6 章の議会録作成支援のための事前音響処理の研究は，河原達也教授をはじめとする京都大学で構築されたデータの協力を受けて完成した成果であります．彼らの協力に深く感謝します.

本成果は，高橋敏博士，政瀧浩和博士，山口義和氏，小川厚徳氏博士の指導の下，音声認識の実用化の場面を強く意識した事で実現できたものになります．当初は難航した研究活動も厳しかった指導も今思えば大変貴重な時間でありました．彼らの指導は多岐に渡り研究活動のみならず会社人としての心構え，文章能力，そして何よりも実用的な研究はどうあるべきかを常に自分に提示してくれました. またそれぞれが持つ異なる側面の優しさが，自分を大いに励ましてくれました．当時のこのように贅沢な指導は他にはないのではないかと思っています.

最後に今日まで自分を支えてくれた友人達，両親，兄妹，妻と 2 人の子どもに感謝の意を表します.

Bibliography

[1] Sadaoki Furui. 50 years of progress in speech and speaker ecognition. *SPECOM*, pages 1–9, October 2005.

[2] B. H. Juang and Lawrence R. Rabiner. *Automatic Speech Recognition - A Brief History of the Technology Development*. Elsevier Encyclopedia of Language and Linguistics, 2004.

[3] Johan Schalkwy, Doug Beeferman, Francoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Garret, and Brian Strope. Google search by voice: A case study. In *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, pages 61–90. Springer, 2010.

[4] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. *International Conference on Acoustics, Speech, and Signal Processing*, pages 5149–5152, 2012.

[5] Jérôme R. Bellegarda. Spoken language understanding for natural interaction: the Siri experience. *International Workshop on Spoken Dialogue Systems*, pages 3–14, 2012.

[6] Kosuke Tsujino, Minoru Eto, Yoshinori Eto, and Shinya Iizuka. Speech recognition and natural language interface in industry (in japanese). *Transaction of the Japanese Society for Artificial Intelligence*, (1):75–81, 2013.

[7] NTT docomo. DOCOMO to Launch Shabette Concier Voice-agent Application. http://www.nttdocomo.co.jp/english/info/media_center/pr/2012/001580.html.

[8] Toru Timai, Shinichi Homma, Akio Kobayashi, Shoei Sato, Tohru Takagi, Kyouichi Saitou, and Satoshi Hara. Real-time closed-captioning using speech recognition. *ABU Technical Committee Annual Meeting*, pages 15–19, 2007.

[9] Tatsuya Kawahara. Transcription system using automatic speech recognition for the japanese parliament (diet). *Twenty-Fourth Innovative Appications of Artificial Intelligence Conference*, pages 2224–2228, 2012.

[10] Sadaoki Furui. Selected topics from 40 years of research on speech and speaker ecognition. *INTERSPEECH 2009*, pages 1–8, September 2009.

[11] Sadaoki Furui. Recent progress in corpus-based spontaneous speech recognition. *IEICE Trans. Inf. & Syst.*, E88-D(3):366–375, March 2005.

[12] Franck Martin, Kiyohiro Shikano, and Yasuhiro Minami. Recognition of noisy speech by composition of hidden Markov models. *EUROSPEECH*, pages 1031–1034, September 1993.

[13] Mark J. F. Gales and Steve J. Young. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, 4(5):352–359, September 1996.

[14] Hiroaki Yamamoto, Tetsuo Kosaka, Masayuki Yamada, Yasuhiro Komori, and Minoru Fujita. Fast speech recognition algorithm under noisy environment using modified CMS-PMC and improved IDMM+SQ. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:847–850, April 1997.

[15] Makoto Shozakai, Satoshi Nakamura, and Kiyohiro Shikano. A non-iterative model-adaptive e-cmn/pmc approach for speech recognition in car environments. *EUROSPEECH*, pages 287–290, September 1997.

[16] Kuo-Hwei Yuo and Hsiao-Chuan Wang. Robust features derived from temporal trajectory filtering for speech recognition under the corruption of additive and convolutional noises. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:577–580, May 1998.

[17] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6):1304–1312, June 1974.

[18] KOBASHIKAWA Satoshi, TAKAHASHI Satoshi, YAMAGUCHI Yoshikazu, and OGAWA Atsunori. Rapid response and robust speech recognition by preliminary model adaptation for additive and convolutional noise. *INTERSPEECH 2005 - EUROSPEECH*, pages 965–968, September 2005.

[19] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Audio Processing*, pages 113–120, April 1979.

[20] Jae. S. Lim and Alan. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, Dec 1979.

[21] KOBASHIKAWA Satoshi, SAKAUCHI Sumitaka, YAMAGUCHI Yoshikazu, and TAKAHASHI Satoshi. Robust speech recognition based on hmm composition and modified wiener filter. *INTERSPEECH 2004 - ICSLP*, pages 2053–2056, 2004.

[22] JIS TR S 0001:2002. *A guideline for determining the acoustic properties of auditory signals used in consumer products – A database of domestic sounds*. 2002.

[23] Yoiti Suzuki, Futoshi Asano, Hack-Yoon Kim, and Toshio Sone. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *J. Acoust. Soc. Am.*, 97(2):1119–1123, 1995.

[24] Christopher Albertia, Michiel Bacchiani, Ari Bezman, Ciprian Chelba, Anastassia Drofa, Hank Liao, Pedro Moreno, Arnaud Sahuguet Ted Power, Maria Shugrina, and Olivier Siohan. An audio indexing system for election video material. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4873–4876, Taipei, Taiwan, April 2009.

[25] Martine Garnier-Rizet, Gilles Adda, Jean-Luc Gauvain, Frederik Cailliau, Sylvie Guillemin-Lanne, Claire Waast-Richard, Lori Lamel, Stephan Vanni, and Claire Waast-Richard. CallSurf: Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In *Proc. LREC*, Marrakech, Morocco, May 2008.

[26] L. Venkata Subramaniam, Tanveer A. Faruquie, Shajith Ikbal, Shantanu Godbole, and Mukesh K. Mohania. Business intelligence from voice of customer. In *Proc. IEEE International Conference in Data Engineering*, pages 1391–1402, Shanghai, China, March 2009.

[27] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *Speech Commumication*, 49:763–786, October 2007.

[28] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. IEEE*, volume 77, pages 257–286, February 1989.

[29] Biing-Hwang Juang and L. R. Rabiner. The segmental k-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust. Speech, Signal Process.*, 38(9):1639–1641, September 1990.

[30] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185, 1995.

[31] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[32] R. Wallace, K. Thambiratnam, and F. Seide. Unsupervised speaker adaptation for telephone call transcription. In *Proc. ICASSP*, pages 4393–4396, Taipei, Taiwan, April 2009.

[33] Suleyman S. Kozat, Karthik Visweswariah, and Ramesh Gopinath. Efficient, low latency adaptation for speech recognition. In *Proc. ICASSP 2007*, volume 4, pages 777–780, Honolulu, April 2007.

[34] Satoshi Kobashikawa, Atsunori Ogawa, Yoshikazu Yamaguchi, and Satoshi Takahashi. Rapid unsupervised adaptation using context independent phoneme model. In *Proc. IEEE International Symposium on Consumer Electronics*, pages 209–212, Kyoto, May 2009.

[35] G. Zavaliagkos, R. Schwartz, and J. Makhoul. Batch, incremental and instantaneous adaptation techniques for speech recognition. In *Proc. ICASSP*, pages 676–689, Detroit, Michigan, USA, May 1995.

[36] Yongxin Li, Hakan Erdogan, Yuqing Gao, and Etienne Marcheret. Incremental on-line feature space MLLR adaptation for telephony speech recognition. In *Proc. ICSLP*, pages 1417–1420, Denver, Colorado, USA, September 2002.

[37] Takaaki Fukutomi, Satoshi Kobashikawa, Taichi Asami, Tsubasa Shinozaki, Hirokazu Masataki, and Satoshi Takahashi. Extracting call-reason segments from contact center dialogs by using automatically acquired boundary expressions. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5584–5587, Prague, Czech Republic, May 2011.

[38] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, April 1994.

[39] Roland Kuhn, Jean-Claude Junqua, Patrick Nguyen, and Nancy Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, November 2000.

[40] Jonas Lööf, Christian Gollan, and Hermann Ney. Speaker adaptive training using shift-MLLR. In *Proc. INTERSPEECH*, pages 1701–1704, Brisbane, Australia, September 2008.

[41] Daniel Povey and Hong-Kwang Jeff Kuo. XMLLR for improved speaker adaptation in speech recognition. In *Proc. INTERSPEECH*, pages 1705–1708, Brisbane, Australia, September 2008.

[42] Balakrishnan Varadarajan, Daniel Povey, and Stephen M. Chu. Quick fMLLR for speaker adaptation in speech recognition. In *Proc. ICASSP*, pages 4297–4300, Las Vegas, Nevada, USA, April 2008.

[43] Naveen Parihar, Ralf Schlöter, David Rybach, and Eric A. Hansen. Parallel fast likelihood computation for LVCSR using mixture decomposition. In *Proc. INTERSPEECH*, pages 3047–3050, Brighton, UK, September 2009.

[44] Christophe Lévy, Georges Linarès, and Jean-François Bonastre. Fast adaptation of GMM-based compact models. In *Proc. INTERSPEECH*, pages 286–289, Antwerp, Belgium, August 2007.

[45] Hui Jiang. Confidence measures for speech recognition: A survey. *Speech Commumication*, 45:455–470, March 2005.

[46] Mukund Padmanabhan, George Saon, and Geoffrey Zweig. Lattice-based unsupervised MLLR for speaker adaptation. In *Proc. ISCA Tutorial and Research Workshop Automatic Speech Recognition: Challenges for the new Millenium*, pages 128–132, Paris, France, August 2000.

[47] L. F. Uebel and P. C. Woodland. Speaker adaptation using lattice-based MLLR. In *Proc. ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, pages 57–60, Sophia Antipolis, France, August 2001.

[48] James R. Glass. A probabilistic framework for segmented-based speech recognition. *Computer Speech and Language*, 17:137–152, 2003.

[49] Akinonbu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Gaussian mixture selection using context-independent HMM. In *Proc. ICASSP*, volume 1, pages 69–72, Salt Lake City, Utah, USA, May 2001.

[50] Toru Imai, Shoei Sato, Akio Kobayashi, Kazuo Onoe, and Shinichi Homma. Online speech detection and dual-gender speech recognition for captioning broadcast news. In *Proc. INTERSPEECH - ICSLP*, pages 1602–1605, Pittsburgh, Utah, USA, September 2006.

[51] Hirokazu Masataki, Daisuke Shibata, Yuichi Nakazawa, Satoshi Kobashikawa, Atsunori Ogawa, and Katsutoshi Ohtsuki. VoiceRex – Spontaneous speech recognition technology for contact-center conversations. *NTT Technical Review*, 5(1):22–27, 2007.

[52] Mark Sanderson and Xiao Mang Shou. Search of spoken documents retrieves well recognized transcripts. In *Proc. the 29th European Conference on Information Retrieval*, pages 505–516, Rome, Italy, April 2007.

[53] Diane J. Litman, Marilyn A. Walker, and Michael S. Kearns. Automatic detection of poor speech recognition at the dialogue level. In *Proc. the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 309–316, College Park, Maryland, U.S.A., June 1999.

[54] Yi Wu, Rong Zhang, and Alexander Rudnicky. Data selection for speech recognition. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pages 562–565, Kyoto, Japan, December 2007.

[55] Hui Lin and Jeff Bilmes. How to select a good training-data subset for transcription: Submodular active selection for sequences. In *Proc. INTERSPEECH*, pages 2859–2862, Brighton, U. K., September 2009.

[56] Tobias Cincarek, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Acoustic modeling for spoken dialogue systems based on unsupervised utterance-based selective training. In *Proc. INTERSPEECH*, pages 1722–1755, Pittsburgh, PA, USA, September 2006.

[57] Ji Wu, Zhiyang He, and Ping Lv. An active learning approach to task adaptation. In *Proc. INTERSPEECH*, pages 2597–2601, Florence, Italy, August 2011.

[58] Svetlana Stoyanchev, Philipp Salletmayr, Jingbo Yang, and Julia Hirschberg. Localized detection of speech recognition errors. In *Proc. IEEE Workshop on Spoken Language Technology*, pages 25–30, Miami, FL. U. S. A., December 2012.

[59] Atsunori Ogawa, Takaaki Hori, and Atsushi Nakamura. Recognition rate estimation based on word alignment network and discriminative error type classification. In *Proc. IEEE Workshop on Spoken Language Technology*, pages 113–118, Miami, FL. U. S. A., December 2012.

[60] Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki, and Satoshi Takahashi. Spoken document confidence estimation using contextual coherence. In *Proc. of INTERSPEECH*, pages 1961–1964, Florence, Italy, August 2011.

[61] Gŕegory Senay and Georges Linaŕes. Confidence measure for speech indexing based on latent Dirichlet allocation. In *Proc. INTERSPEECH*, Portland, OR. U. S. A., September 2012.

[62] Haiyang Li, Jiqing Han, Tieran Zheng, and Guibin Zheng. A novel confidence measure based on context consistency for spoken term detection. In *Proc. INTERSPEECH*, Portland, OR. U. S. A., September 2012.

[63] Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, and Kiyohiro Shikano. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. In *Proc. INTERSPEECH*, pages 173–176, Jeju Island, Korea, September 2004.

[64] Dong Jin Seo, Young-Joon Kim, and Nam Soo Kim. Pre-rejection of distorted speech for speech recognition in wireless communication channel. In *Proc. of IEEE Vehicular Technology Conference*, pages 2803–2806, Orlando, Florida, U. S. A., April 2003.

[65] Gang Guo, Chao Huang, Hui Jiang, and Ren-Hua Wang. A comparative study on various confidence measures in large vocabulary speech recognition. In *Proc. ISCSLP*, volume 4, pages 9–12, Hong Kong, China, December 2004.

[66] Bernhard Rueber. Obtaining confidence measures from sentence probabilities. In *Proc. Eurospeech*, pages 739–742, Rhodes, Greece, September 1997.

[67] Akinonbu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Gaussian mixture selection using context-independent HMM. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 69–72, Salt Lake City, U. S. A., May 2001.

[68] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10:19–41, January 2000.

[69] Satoshi Kobashikawa, Atsunori Ogawa, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki, and Satoshi Takahashi. Fast unsupervised adaptation based on efficient statistics accumulation using frame independent confidence within monophone states. *Computer Speech and Language*, 27:369–379, 2013.

[70] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, January 1999.

[71] Jonathan Mamou, David Carmel, and Ron Hoory. Spoken document retrieval from call-center conversations. In *Proc. the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–58, Seattle, Washington, U.S.A., August 2006.

[72] Chia-Hsin Hsieh, Chien-Lin Huang, and Chung-Hsien Wu. Spoken document summarization using topic-related corpus and semantic dependency grammar. In *Proc. ISCSLP*, pages 333–336, Hong Kong, China, December 2004.

[73] Asmaa El Hannani and Thomas Hain. Automatic optimization of speech decoder parameters. *IEEE Signal Processing Letters*, 17(1):95–98, January 2010.

[74] Ivan Bulyko. Speech recognizer optimization under speed constraints. In *Proc. of INTERSPEECH*, pages 1497–1500, Makuhari, Chiba, Japan, September 2010.

[75] Dongbin Zhang and Limin Du. Dynamic beam pruning strategy using adaptive control. In *Proc. of INTERSPEECH*, pages 285–288, Jeju Island, Korea, October 2004.

[76] Tibor Fabian, Robert Lieb, Günther Ruske, and Matthias Thomae. A confidence-guided dynamic pruning approach - utilization of confidence measurement in speech recognition. In *Proc. of INTERSPEECH*, pages 585–588, Lisboa, Portugal, September 2005.

[77] Volker Steinbiss, Bach-Hiep Tran, and Hermann Ney. Improvements in beam search. In *ICSLP*, pages 2143–2146, Yokohama, Japan, September 1994.

[78] Janne Pylkköen. New pruning criteria for efficient decoding. In *Proc. of INTERSPEECH*, pages 581–584, Lisboa, Portugal, September 2005.

[79] Jeff Bilmes and Hui Lin. Online adaptive learning for speech recognition decoding. In *Proc. of INTERSPEECH*, pages 1958–1961, Makuhari, Chiba, Japan, September 2010.

[80] Brian Mak and Tom Ko. Min-max discriminative training of decoding parameters using iterative linear programming. In *Proc. of INTERSPEECH*, pages 915–918, Brisbane, Australia, September 2008.

[81] Brian Mak and Tom Ko. Automatic estimation of decoding parameters using large-margin iterative linear programming. In *Proc. of INTERSPEECH*, pages 1219–1222, Brighton, U.K., September 2009.

[82] Zhanlei Yang, Hao Chao, and Wnju Liu. Response probability based decoding algorithm for large vocabulary continous speech recognition. In *Proc. of INTERSPEECH*, pages 1929–1932, Florence, Italy, August 2011.

[83] Takaaki Hori, Shinji Watanabe, and Atsushi Nakamura. Improvement of search risk minimization in viterbi beam search for speech recognition. In *Proc. of Interspeech*, pages 1962–1965, Makuhari, Chiba, Japan, September 2010.

[84] Takaaki Hori, Shinji Watanabe, and Atsushi Nakamura. Search error risk minimization in viterbi beam search for speech recognition. In *Proc. of IEEE International Conference on Acoustics Speech and Signal Processing*, pages 4934–4937, Dallas, TX, U.S.A., March 2010.

[85] Atsunori Ogawa, Satoshi Takahashi, and Atsushi Nakamura. Machine and acousticalcondition dependency analysis for fast acoustic likelihood calculation techniques. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5156–5159, Prague, Czech Republic, May 2011.

[86] Hugo Van Hamme and Filip Van Aelten. An adaptive-beam pruning technique for continuous speech recognition. In *Proc. of ICSLP*, pages 2083–2086, Philadelphia, PA, USA, October 1996.

[87] XIE Lingyun and DU Limin. Efficient Viterbi beam search algorithm using dynamic pruning. In *Proc. of ICSP*, pages 699–702, Beijing, China, August 2004.

[88] Sherif Abdou and Michael S. Scordilis. Beam search pruning in speech recognition using a posterior probability-based confidence measure. *Speech Communication*, 42:409–428, January 2004.

[89] Tibor Fabian and Günther Ruske. Comparing confidence-guided and adaptive dynamic pruning techniques for speech recognition. In *Proc. of European Signal Processing Conference*, Florence, Italy, September 2006.

[90] D. Nolden, R. Schlüter, and H. Ney. Acoustic look-ahead for more efficient decoding in lvcsr. In *Proc. of INTERSPEECH*, pages 893–896, Florence, Italy, August 2011.

[91] Stefan Ortmannsa and Hermann Ney. Look-ahead techniques for fast beam search. *Computer Speech and Language*, 14:15–32, January 2000.

[92] Erik McDermott, Shinji Watanabe, and Atsushi Nakamura. Discriminative training based on an integrated view of MPE and MMI in margin and error space. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4894–4897, 2010.

[93] Takaaki Hori, Chiori Hori, Yasuhiro Minami, , and Atsushi Nakamura. Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Trans. on Audio, Speech and Lang. Process.*, 15(4):1352–1365, 2007.

[94] Lori Lamel, Jean-Luc Gauvain, Gilles Adda, Claude Barras, Eric Bilinski, Olivier Galibert, Agusti Pujol, Holger Schwenk, and Xuan Zhu. The LIMSI 2006 TC-STAR EPPS transcription systems. In *Proc. ICASSP*, volume 4, pages 997–1000, Honolulu, Hawaii, U.S.A., May 2007.

[95] J. Lööf, M. Bisani, Ch. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schlüter, and H. Ney. The 2006 RWTH parliamentary speeches transcription system. In *Proc. INTERSPEECH - ICSLP*, pages 104–107, Pittsburgh, Pennsylvania, U.S.A., September 2006.

[96] B. Ramabhadran, O. Siohan, L. Mangu, G. Zweig, M. Westphal. H. Schulz, and A. Soneiro. The IBM 2007 speech transcription system for European parliamentary speeches. In *Proc. ASRU*, pages 472–477, Kyoto, Japan, December 2007.

[97] Yuya Akita, Masato Mimura, and Tatsuya Kawahara. Automatic transcription system for meetings of the Japanese national congress. In *Proc. INTERSPEECH*, pages 84–87, Brighton, U. K., September 2009.

[98] Mariko Aoki, Manabu Okamoto, Shigeaki Aoki, Hiroyuki Matsui, Tetsuma Sakurai, and Yutaka Kaneda. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoustical Science and Technology*, 22:149–157, February 2001.

[99] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoustics, Speech and Signal Processing*, 29:254–272, April 1981.

[100] O. Viiki, D. Bye, and K. Laurila. A recursive feature vector normalization approach for robust speech recognition in noise. In *Proc. ICASSP*, volume 2, pages 733–736, Seattle, WA, U. S. A., May 1998.

[101] Li Lee and Richard Rose. A frequency warping approach to speaker normalization. *IEEE Trans. Speech and Audio Processing*, 6(1):49–60, January 1998.

[102] Takaaki Hori, Chiori Hori, Yasuhiro Minami, and Atsushi Nakamura. Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 15:1352–1365, 2007.

[103] Sumitaka Sakauchi, Akira Nakagawa, Yoichi Haneda, and Akitoshi Kataoka. Implementing and evaluating of an audio teleconferencing terminal with noise and echo reduction. In *Proc. International Workshop on Acoustic Echo and Noise Control*, pages 191–194, Kyoto, Japan, September 2003.

[104] Satoshi Kobashikawa, Atsunori Ogawa, Yoshikazu Yamaguchi, and Satoshi Takahashi. Rapid unsupervised adaptation using frame independent output probabilities of gender and context independent phoneme models. In *Proc. INTERSPEECH*, pages 1615–1618, Brighton, U. K., September 2009.

[105] Scott Shaobing Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, Lansdowne, VA, U. S. A., February 1998.

[106] Masafumi Nishida and Tatsuya Kawahara. Speaker indexing and adaptation using speaker clustering based on statistical model selection. In *Proc. ICASSP*, volume 1, pages 353–356, Montreal, Quebec, Canada, May 2004.

[107] Satoshi Kobashikawa, OGAWA Atsunori, YAMAGUCHI Yoshikazu, and TAKAHASHI Satoshi. Rapid unsupervised adaptation using context independent phoneme model. In *ISCE*, pages 209–212, 2009.

[108] Tatsuya Kawahara, Masato Mimura, and Yuya Akita. Language model transformation applied to lightly supervised training of acoustic model for congress meetings. In *Proc. ICASSP*, pages 3853–3856, Taipei, Taiwan, April 2009.

[109] Yuya Akita and Tatsuya Kawahara. Statistical transformation of language and pronunciation models for spontaneous speech recognition. *IEEE Trans. Audio, Speech and Language Processing*, 18(6):1539–1549, 2010.

LIST OF WORK

Journal papers

[J1] Satoshi Kobashikawa and Satoshi Takahashi, "Robust Speech Recognition by Model Adaptation and Normalization Using Pre-Observed Noise,", *IEICE Transactions on Information and Systems*, vol. 91-D-III, no. 3, pp. 422–429, (2008)

[J2] Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Confidence estimation of spoken document recognition using word contextual coherence," (in Japanese), *Journal of the Acoustical Society of Japan*, vol. 68, no. 7, pp.323-330, (2012)

[J3] Satoshi Kobashikawa, Atsunori Ogawa, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Fast unsupervised adaptation based on efficient statistics accumulation using frame independent confidence within monophone states,", *Computer Speech and Language*, vol. 27, no. 1, pp. 369–379, (2013)

[J4] Kenji Imamura, Tomoko Izumi, Kugatsu Sadamitsu, Kuniko Saito, Satoshi Kobashikawa and Hirokazu Masataki, "Morpheme Conversion Using Discriminative Models for Connecting Different. Morphological Systems," (in Japanese), *IEICE Transactions on Information and Systems*, vol. J96-D, no. 1, pp. 239–249, (2013)

[J5] Narichika Nomoto, Satoshi Kobashikawa, Masashi Tamoto, Hirokazu Masataki, Osamu Yoshioka, Satoshi Takahashi, "Estimation of Anger Emotion in Spoken Dialogue Using Conversational Temporal. Relations of Utterances," (in Japanese), *IEICE Transactions on Information and Systems*, vol. J96-D, no. 1, pp. 15–24, (2013)

Masataki, Satoshi Takahashi,

Letters

[L1] Satoshi Kobashikawa, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, 'Efficient data selection for speech recognition based on prior confidence estimation," *Acoustic Science and Technology*. vol. 32, no. 4, pp. 151–153, (2011)

International conferences

[IC1] Nobuyuki Minematsu, Satoshi Kobashikawa, Keikichi Hirose, and Donna Erickson, "Acoustic modeling of sentence stress using differential features between syllables for English rhythm learning system development," In *Proc. International Conference Spoken Language Processing (ICSLP)*, pp. 745-748, (2002).

[IC2] Nobuyuki Minematsu, Satoshi Kobashikawa, Keikichi Hirose, and Donna Erickson, "Acoustic modeling of sentence stress and its detection for learning English rhythm," In *Proc. Conference Integration of Speech Technology into Learning (InSTIL)*, (2002).

[IC3] Satoshi Kobashikawa, Sumitaka Sakauchi, Yoshikazu Yamaguchi, Satoshi Takahashi, "Robust speech recognition based on HMM composition and modified wiener filter," In *Proc. INTERSPEECH*, pp. 2053-2056, (2004).

[IC4] Satoshi Kobashikawa, Satoshi Takahashi, Yoshikazu Yamaguchi and Atsunori Ogawa, "Rapid response and robust speech recognition by preliminary model adaptation for additive and convolutional noise," In *Proc. INTERSPEECH*, pp. 965-968, (2005).

[IC5] Satoshi Kobashikawa, Atsunori Ogawa, Yoshikazu Yamaguchi and Satoshi Takahashi, "Rapid Unsupervised Adaptation Using Frame Independent Output Probabilities of Gender and Context Independent Phoneme Models," In *Proc. INTERSPEECH*, pp. 1615-1618, (2009).

[IC6] Satoshi Kobashikawa, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Efficient Data Selection for Speech Recognition Based on Prior Confidence Estimation Using Speech and Context Independent Models," In *Proc. INTERSPEECH*, pp. 238-241, (2010).

[IC7] Satoshi Kobashikawa, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Efficient Data Selection for Speech Recognition Based on Prior Confidence Estimation Using Speech and Context Independent Models," In *Proc. INTERSPEECH*, pp. 238-241, (2010).

[IC8] Takaaki Fukutomi, Satoshi Kobashikawa, Taichi Asami, Tsubasa Asami, Hirokazu Masataki, and Satoshi Takahashi, "Extracting Call-reason Segments From Contact Center Dialogs By Using Automatically Acquired Boundary Expressions," In *Proc. International Conference on Acoustics, Speech, and Signal Processing* , pp.5584-5587, (2011).

[IC9] Taichi Asami, Narichika Nomoto, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki, and Satoshi Takahashi, "Spoken Document Confidence Estimation Using Contextual Coherence," In *Proc. INTERSPEECH*, pp.1961-1964, (2011).

[IC10] Kenji Imamura, Tomoko Izumi, Kugatsu Sadamitsu1, Kuniko Saito, Satoshi Kobashikawa, and Hirokazu Masataki, "Morpheme Conversion for Connecting Speech Recognizer and Language Analyzers in Unsegmented Languages," In *Proc. INTERSPEECH*, pp.1405-1408, (2011).

[IC11] Satoshi Kobashikawa, Takaaki Hori, Yoshikazu Yamaguchi, Taichi Asami, Hirokazu Masataki and Satoshi Takahashi, "Efficient Beam Width Control to Suppress Excessive Speech Recognition Computation Time Based on Prior Score Range Normalization," In *Proc. INTERSPEECH*, (2012).

[IC12] Taichi Asami, Satoshi Kobashikawa, Hirokazu Masataki, Osamu Yoshioka and Satoshi Takahashi, "Speech Data Clustering Based on Phoneme Error Trend for Unsupervised Acoustic Model Adaptation," In *Proc. INTERSPEECH*, (2012).

International symposiums and workshops

[IW1] Satoshi Kobashikawa, Atsunori Ogawa, Yoshikazu Yamaguchi and Satoshi Takahashi, "Rapid Unsupervised Adaptation Using Context Independent Phoneme Model," In *Proc. The 13th IEEE International Symposium on Consumer Electronics (ISCE2009)*, pp. 2009-212, (2009).

[IW2] Satoshi Kobashikawa, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi, "Efficient Data Selection for Spoken Document Retrieval Based on Prior Confidence Estimation Using Speech and Context Independent Models," In *Proc. IEEE Workshop on Spoken Language Technology (SLT)*, pp. 188-193, (2010).

[IW3] Ryuichiro Higashinaka, Yasuhiro Minami, Hitoshi Nishikawa, Kohji Dohsaka, Toyomi Meguro, Satoshi Kobashikawa, Hirokazu Masataki, Osamu Yoshioka, Satoshi Takahashi, and Genichiro Kikui, "Improving HMM-Based Extractive Summarization for Multi-domain Contact Center Dialogues," In *Proc. IEEE Workshop on Spoken Language Technology (SLT)*, pp. 61–66, (2010).

[IW4] Satoshi Kobashikawa, Takaaki Hori, Yoshikazu Yamaguchi, Taichi Asami, Hirokazu Masataki and Satoshi Takahashi, "Efficient Beam Width Control to Suppress Excessive Speech Recognition Computation Time Based on Prior Score Range Normalization," In *Proc. IEEE Workshop on Spoken Language Technology (SLT)*, pp. 125–130, (2012).

Domestic conferences (in Japanese)

[DC1] Yoshimi Seto, Satoshi Kobashikawa, Yosuke Nishimuro, Hitoshi Aida and Tadao Saito, "A method of synthesizing images of virtual view from stereo images using zooming and estimation of parallax," In *Proc. Spring Meeting of IPSJ 2000*, pp. 203–204, (2000.3)

[DC2] Nobuaki Minematsu, Satoshi Kobashikawa and Keikichi Hirose, "An experimental study on automatic detection of stressed syllables in English sentence utterances," In *Proc. Spring Meeting of ASJ 2001*, 3-6-6, pp. 331–332, (2001.3)

[DC3] Satoshi Kobashikawa, Nobuaki Minematsu, Donna Erickson and Keikichi Hirose, "Automatic detection of stressed syllables in English sentence utterances," In *Proc. Fall Meeting of ASJ 2001*, 2-7-7, pp. 147–148, (2001.10) (recieved the poster award from the ASJ).

[DC4] Satoshi Kobashikawa, Nobuaki Minematsu, Donna Erickson and Keikichi Hirose, "English Rhythm learning based on automatic detection of stressed syllables in sentences," In *Proc. Spring Meeting of ASJ 2002*, 2-5-22, pp. 117–118, (2002.3)

[DC5] Satoshi Kobashikawa, Satoshi Takahashi, Sumitaka Sakauchi and Yoshikazu Yamaguchi, "Robust Speech Recognition for multi-SNR noisy speech based on Noise Reduction and HMM Composition," In *Proc. Fall Meeting of ASJ 2004*, 1-1-12, pp. 23–24, (2004.9)

[DC6] Satoshi Kobashikawa, Satoshi Takahashi, Yoshikazu Yamaguchi and Atsunori Ogawa, "Speech Recognition by Preliminary Acoustic Model Adaptation based on HMM-composition and CMN," In *Proc. Spring Meeting of ASJ 2006*, 3-1-13, pp. 133–134, (2006.3)

[DC7] Satoshi Kobashikawa, Atsunori Ogawa, Hirokazu Masataki and Satoshi Takahashi, "A Study on Accuracy Improvement by Boosting Sufficient Statistics with Keywords," In *Proc. Spring Meeting of ASJ 2008*, 1-Q-23, pp. 133–134, (2008.3)

[DC8] Satoshi Kobashikawa, Atsunori Ogawa, Yoshikazu Yamaguchi and Satoshi Takahashi, "A Study on Rapid Unsupervised Adaptaiton Using Context Independent Phoneme Model," In *Proc. Spring Meeting of ASJ 2009*, 1-P-30, pp. 195–196, (2009.3)

[DC9] Satoshi Kobashikawa, Taichi Asami, Yoshikazu Yamaguchi, Hirokazu Masataki and Satoshi Takahashi "Data Selection for Speech Recognition Based on Prior Confidence Estimation," In *Proc. Spring Meeting of ASJ 2010*, 1-Q-8, pp. 167–168, (2010.3)

[DC10] Takaaki Fukutomi, Satoshi Kobashikawa, Taichi Asami, Tsubasa Shinozaki, Hirokazu Masataki, Osamu Yoshioka and Satoshi Takahashi "Acquisition of Specific Expressions Around Call Reason Phase in Contact Center Dialogue," In *Proc. Fall Meeting of ASJ 2010*, 1-Q-31, pp. 193–194, (2010.9)

[DC11] Takaaki Hori, Atsushi Nakamura, Yoshikazu Yamaguchi, Satoshi Kobashikawa, Taichi Asami, Hirokazu Masataki, Satoshi Takahashi and Tatsuya Kawahara, "Automatic Speech Recognition System for Creation of Meeting Records in the House of Representatives - Acoustic Processing," In *Proc. Spring Meeting of ASJ 2011*, 3-5-8, pp. 89–90, (2011.3)

[DC12] Satoshi Kobashikawa, Taichi Asami, Yoshikazu Yamaguchi, Sumitaka Sakauchi, Atsunori Ogawa, Hirokazu Masataki, Satoshi Takahashi and Tatsuya Kawahara, "Automatic Speech Recognition System for Creation of Meeting Records in the House of Representatives - Acoustic Processing," In *Proc. Spring Meeting of ASJ 2011*, 3-5-9, pp. 91–94, (2011.3)

[DC13] Taichi Asami, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu and Satoshi Takahashi, "Evaluation of confidence estimation using word contextual coherence and acoustic likelihood," In *Proc. Spring Meeting of ASJ 2011*, 3-5-20, pp. 89–94, (2011.3)

[DC14] Satoshi Kobashikawa, Takaaki Hori, Yoshikazu Yamaguchi, Taichi Asami, Hirokazu Masataki and Satoshi Takahashi, "," In *Proc. Spring Meeting of ASJ 2012*, 2-7-7, pp. –, (2012.3)

Domestic workshops in Japanese

[DW1] Satoshi Kobashikawa, Nobuaki Minematsu, Donna Erickson and Keikichi Hirose, "Modeling of Stressed Syllables for their Detection in English Sentences to Develop an English Rhythm Learning System," *Technical Report of IEICE*, vol. 101, no. 520, pp. 99-104, (2001).

[DW2] Taichi Asami, Satoshi Kobashikawa, Yoshikazu Yamaguchi, Hirokazu Masataki, and Satoshi Takahashi, "Confidence Estimation at the Spoken Document Level Using Word Contextual Coherence and Acoustic Likelihood," *Technical Report of IEICE*, vol. 110, no. 143, pp. 43–48, (2010).

[DW3] Takaaki Fukutomi, Satoshi Kobashikawa, Taichi Asami, Tsubasa Shinozaki, Hirokazu Masataki, and Satoshi Takahashi, "Accurate Call-reason Segment Extraction based on Typical Phrase Detection," *Technical Report of IEICE*, vol. 110, no.401 , pp. 43–48, (2011).

Commentary

[C1] Hirokazu Masataki, Daisuke Shibata, Yuichi Nakazawa, Satoshi Kobashikawa, Atsunori Ogawa, and Katsutoshi Ohtsuki, "VoiceRex – Spontaneous Speech Recognition Technology for Contact-center Conversations," *NTT Technical Review*, vol. 5, no. 1 , pp. 22–27, (2007).

Commentary in Japanese

[CJ1] 政瀧 浩和，柴田 大輔，中澤 裕一，小橋川 哲，小川 厚徳，大附 克年, "顧客との自然な会話を聞き取る自由発話音声認識技術「VoiceRex」," *NTT* 技術ジャーナル, vol. 18, no. 11 , pp. 15–18, (2006).

Awards

[A1] The poster award from the ASJ in 2001

[A2] The president award from NTT in 2009

[A3] The president award from NTT in 2011

[A4] The Kiyasu special industrial achievement award from the IPSJ in 2012

[A5] The Maejima Hisoka award from the Teishin Association in 2013

Patents in Japanese

[P1] 小橋川 哲，高橋 敏，”マイク位置決定方法、マイク位置決定装置、マイク位置決定プログラム”，特許 4173462(特開 2005-303898, 特願 2004-120377)

[P2] 小橋川 哲，高橋 敏，山口 義和，今村 明弘，”音声認識方法、その装置およびプログラム、その記録媒体”，特許 4242320(特開 2005-326672, 特願 2004-145334)

[P3] 小橋川 哲，高橋 敏，山口 義和，小川 厚徳，”音声認識方法およびこの方法を実施する装置”，特許 4291728(特開 2005-301097, 特願 2004-119931)

[P4] 小橋川 哲，高橋 敏，山口 義和，小川 厚徳，”音声認識方法、この方法を実施する装置、プログラムおよびその記録媒体”，特許 4464797(特開 2006-145694, 特願 2004-333487)

[P5] 小橋川 哲，大附 克年，小川 厚徳，政瀧 浩和，”音響モデル作成装置、音響モデル作成方法、そのプログラムおよびその記録媒体”，特許 4571922(特開 2007-249051, 特願 2006-075374)

[P6] 小橋川 哲，高橋 敏，山口 義和，今村 明弘，”音響モデル作成装置、音声認識装置、音響モデル作成方法、音声認識方法、音響モデル作成プログラム、音声認識プログラムおよび記録媒体”，特許 4705414(特開 2006-349723, 特願 2005-172122)

[P7] 小橋川 哲，大附 克年，”音響モデル生成装置、方法、プログラム及びその記録媒体”，特許 4705557(特開 2008-129527, 特願 2006-317361)

[P8] 小橋川 哲，浅見 太一，政瀧 浩和，高橋 敏，”音声認識装置とその方法と、プログラムとその記録媒体”，特許 4729078(特開 2009-300716, 特願 2008-154933)

[P9] 小橋川 哲，高橋 敏，小川 厚徳，政瀧 浩和，”音声認識装置、音声認識方法、そのプログラムおよびその記録媒体”，特許 4728791(特開 2007-156364, 特願 2005-355460)

[P10] 小橋川 哲，政瀧 浩和，”学習データ選択装置、学習データ選択方法、プログラムおよび記録媒体、音響モデル作成装置、音響モデル作成方法、プログラムおよび記録媒体”，特許 4829871(特開 2009-128490, 特願 2007-301625)

[P11] 小橋川 哲，政瀧 浩和，高橋 敏，”音響分析パラメータ生成装置とその方法と、それを用いた音声認識装置と、プログラムと記録媒体”，特許 4843646(特開 2009-300837, 特願 2008-156501)

[P12] 小橋川 哲，山口 義和，浅見 太一，政瀧 浩和，高橋 敏，”音声認識装置とその方法と、プログラム”，特許 4852129(特開 2011-13543, 特願 2009-158783)

[P13] 小橋川 哲，小川 厚徳，”音響モデルパラメータ更新処理方法、音響モデルパラメータ更新処理装置、プログラム、記録媒体”，特許 4856526(特開 2008-139747, 特願 2006-328029)

[P14] 小橋川 哲，政瀧 浩和，高橋 敏，”音響モデル作成方法、音響モデル作成装置、そのプログラム、その記録媒体”，特許 4909318(特開 2009-300830, 特願 2008-156458)

[P15] 小橋川 哲，浅見 太一，山口 義和，政瀧 浩和，高橋 敏，”音声認識装置とその方法と、プログラム”，特許 4922377(特開 2011-075973, 特願 2009-229338)

[P16] 小橋川 哲，中澤 裕一，”学習データのラベル誤り候補抽出装置、その方法及びプログラム、その記録媒体”，特許 4981519(特開 2008-292789, 特願 2007-138626)

[P17] 小橋川 哲，山口 義和，浅見 太一，神 明夫，政瀧 浩和，高橋 敏，”音声認識装置とその方法と、プログラムと記録媒体”，特許 4981850(特開 2011-002494, 特願 2009-143173)

[P18] 小橋川 哲，小川 厚徳，”音響モデル生成装置、方法、プログラム及びその記録媒体”，特許 5006768(特開 2009-128496, 特願 2007-301689)

[P19] 小橋川 哲，山口 義和，政瀧 浩和，高橋 敏，”発話区間話者分類装置とその方法と、その装置を用いた音声認識装置とその方法と、プログラムと記録媒体”，特許 5052449(特開 2010-32792, 特願 2008-195136)

[P20] 小橋川 哲，浅見 太一，山口 義和，”音響モデル作成装置、その方法及びプログラム”，特許 5089655(特開 2011-002792, 特願 2009-148089)

[P21] 小橋川 哲，浅見 太一，”音声認識装置及び音響モデル作成装置とそれらの方法と、プログラムと記録媒体”，特許 5089651(特開 2010-286586, 特願 2009-138987)

[P22] 小橋川 哲，山口 義和，”音響分析パラメータ生成方法とその装置と、プログラムと記録媒体”，特許 5166195(特開 2010-96808, 特願 2008-264911)

[P23] 高橋 敏，小橋川 哲，”音声認識方法、その装置およびプログラム、その記録媒体”，特許 4313728(特開 2006-3617, 特願 2004-179723)

[P24] 小川 厚徳，小橋川 哲，高橋 敏，”音響モデル雑音適応化方法およびこの方法を実施する装置”，特許 4510517(特開 2005-338358, 特願 2004-156037)

[P25] 政瀧 浩和，小橋川 哲，”言語モデル作成装置、言語モデル作成方法、そのプログラムおよびその記録媒体”，特許 4537970(特開 2007-249050, 特願 2006-075364)

[P26] 中澤 裕一，小橋川 哲，小川 厚徳，政瀧 浩和，”音響モデル適応装置、音響モデル適応方法、音響モデル適応プログラム及び記録媒体”，特許 4594885(特開 2007-248730, 特願 2006-070961)

[P27] 神 明夫，小橋川 哲，浅見 太一，”通話区間検出装置、その方法、プログラム及び記録媒体”，特許 4825153(特開 2008-216273, 特願 2007-049150)

Bachelor’s thesis in Japanese

[P1] 小橋川 哲，”二眼立体視におけるズーミングによる奥行き感低下の軽減法”，東京大学工学部 電子工学科 卒業論文，(2000.3)

Master thesis in Japanese

[P1] 小橋川 哲，”英語文強勢のモデル化とその発音教育への応用”，東京大学 工学系研究科 情報工学専攻 修士論文，(2002.3)