NAIST-IS-DT9661015

Doctor's Thesis

Statistical Acoustic Model Adaptation for Robust Speech Recognition in Noisy Reverberant Environments

Tetsuya Takiguchi

February 8, 1999

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctor's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Tetsuya Takiguchi

Thesis committee: Kiyohiro Shikano, Professor
Yoh'ichi Tohkura, Professor
Naokazu Yokoya, Professor
Satoshi Nakamura, Associate Professor

Statistical Acoustic Model Adaptation for Robust Speech Recognition in Noisy Reverberant Environments*

Tetsuya Takiguchi

Abstract

This thesis presents a hands-free speech recognition method based on the HMM (Hidden Markov Model) composition and the HMM decomposition for speech which is contaminated not only by additive noise but also by an acoustic transfer function. The method realizes an improved user interface such that a user is not encumbered by microphone equipment in noisy reverberant environments. The HMM composition method has already been proposed for additive noise. In this thesis, the HMM composition method for additive noise is extended to handle convolutional acoustic distortion of the reverberant room, by using an HMM to model the acoustic transfer function. The states of the acoustic transfer function HMM correspond to different sound source positions. This HMM can represent the positions of the sound sources, even if the speaker moves.

This thesis also proposes a new method to estimate HMM parameters of the acoustic transfer function based on the HMM decomposition. The proposed method is obtained as the result of the reverse process of the HMM composition, where the model parameters are estimated by maximizing likelihood of adaptation data uttered from an unknown position. Finally, this thesis describes the performance of the HMM composition and decomposition methods on real distant-talking speech.

Keywords:

robust speech recognition, acoustic model, adaptation, noise, reverberation

*Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9661015, February 8, 1999.

雑音・残響環境下でのロバストな音声認識のためのモデル適応化法*

滝口 哲也

内容梗概

雑音及び残響環境下においてマイクロホンから離れて発話した場合、音声認識精度は劣化してしまう。なぜなら、その音声は周囲の雑音及び残響の影響を受けてしまい、音声モデル作成時の学習データと観測データとの間にミスマッチが生じてしまうためである。それらの影響に対処するために、本論文では従来の音声 HMM(Hidden Markov Model) と雑音 HMM の合成に加えて、音響伝達特性 HMM の合成を試みる。この HMM の各状態を音源位置に対応させることにより、ユーザの自由な場所移動にも対処することが可能になり、ユーザインタフェースの向上が実現される。

音響伝達特性 HMM を作成するためには、認識を行なう前にあらかじめ各場所からの音響伝達特性を測定しておく必要があった。しかしながら実際の環境において、あらかじめ音響伝達特性を測定しておくのは非現実的である。そこで本論文では、更に HMM 分解法による音響伝達特性 HMM の推定方法を提案する。HMM 分解法では、ユーザの場所が既知である必要はなく、任意の場所から発話された観測データを用いて音響伝達特性 HMM の推定が可能である。雑音・残響環境下では、この HMM 分解法が 2 回適用される。まず、周波数領域において雑音 HMM からの分解を最尤推定にもとづいて行ない、更に領域変換を行ないケプストラム領域において、音響伝達特性 HMM を最尤推定にもとづいて分解する。この HMM 分解法により各々の場所からの音響伝達特性を推定し、HMM 合成法により雑音 HMM 及び音声 HMM と組み合わせることにより、対象とする雑音及び残響環境下での音声モデルを作成し、音声認識を行なう。提案手法の評価実験を実際に雑音及び残響環境下にて観測された音声に対して行ない、その有効性を確認した。

キーワード

音声認識、モデル適応、雑音、残響

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 修士論文, NAIST-IS-DT9661015, 1999 年 2 月 8 日.

Acknowledgments

I would like to thank Professor Kiyohiro Shikano. Besides providing helpful suggestions and advice, making this work possible, he also gave me many opportunities in many academic aspects of my life during my graduate studies at Nara Institute of Science and Technology (NAIST). I would also like to thank Professor Yoh'ichi Tohkura for reviewing this thesis and enriching it with his comments and suggestions. I would also like to thank Professor Naokazu Yokoya for many key discussions and helpful suggestions.

I would like to thank Associate Professor Satoshi Nakamura who has given me many thoughtful suggestions. His analysis and comments always gave me more insight into the problem of speech recognition. I learned many things from his guidance and support. This thesis would not have advanced without his support.

I would also like to thank Research Associate Shiro Ise (currently Associate Professor at Kyoto University). Although his name does not appear on the author list of some papers, his contributions and help are highly acknowledged. His comments gave me great insight into the problem of signal processing.

I am very grateful to Research Associate Jinlin Lu (NAIST), Dr. Makoto Shozakai (Asahi Chemical Industry Co.Ltd.), Mike Schuster (ATR interpreting telecommunications research laboratories) and Dr. Harald Singer (ATR interpreting telecommunications research laboratories) for many discussions. I am also indebted to Dr. Erik McDermott (ATR human information processing research laboratories) for reviewing this thesis and giving valuable comments.

There are many colleagues at NAIST speech and acoustics laboratories that I would like to thank. Thanks go to my colleagues, especially to the members of the speech group, Takeshi Yamada, Alexandre Girardi, Eli Yamamoto and Tadashi Yonezaki for helpful discussions.

I would like to thank Dr. Qiang Huo (ATR Interpreting Telecommunications research Laboratories (ATR-ITL)) (currently Professor at University of Hong Kong).

He gave me advice and guidance during my stay at ATR-ITL. I would also like to thank Dr. Yoshinori Sagisaka, head of department 1 of ATR-ITL for his support of the collaboration between NAIST and ATR-ITL.

I am indebted to Mr. Masatoshi Morishima and Mr. Toshihiro Isobe (NTT DATA Corporation) for their support in helping me to complete parts of the experiments reported in Chapter 6 of this thesis. I am also indebted to Dr. Nobuo Koizumi (NTT DATA Corporation) for his support of the collaboration between NAIST and NTT DATA Corporation.

I would like to thank my wife for her friendship, encouragement and love, and my parents for their never ending support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Speech recognition systems have been developed for various applications in the last 30 years. Recently, the accuracy of speaker-independent speech recognition has been remarkably improved by use of stochastic modeling of speech and its training algorithm, e.g. [40, 73, 76, 78]. Efficient search engines have also been developed, e.g. [3, 23, 24, 80]. For example, there are commercial continuous-speech recognition systems which run on a PC: IBM's ViaVoice, Dragon's Naturally Speaking, and others. Some key issues for more widespread use are development of recognition technology capable of handling the following kinds of speech (figure 1.1):

- **Noisy speech**
  Speech recognition systems perform remarkably well in non-noisy environments. However, if a user speaks in noisy environments, the recognition accuracy will seriously degrade because of mismatches between the training and the testing environments. Also, noise will increase the difficulty of the speech-boundary detection, and will cause the Lombard effect [41].

- **Distant-talking speech**
  At present, a user must be equipped with a close-talking microphone (desktop microphone, head-mounted microphone, and so on). A key issue for more widespread use is the development of recognition technology of reverberated speech obtained from a distant microphone.

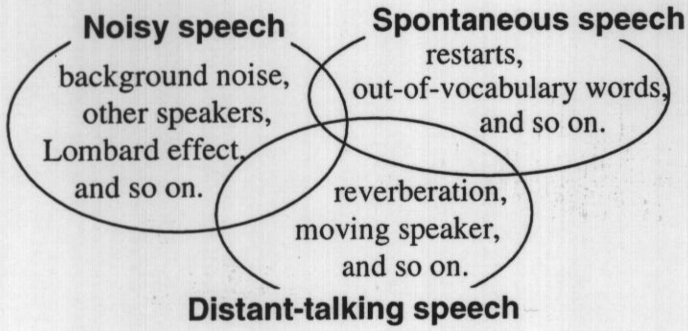

Figure 1.1: Some key issues for more widespread use of speech recognition

- **Spontaneous speech**
  Spontaneous speech is different from carefully produced speech like read speech. It includes restarts, hesitations, and so on. Also, spontaneous speech causes the problem of out-of-vocabulary words.

The most important advantage of the speech interface is to make hands-free speech recognition a reality, where a user is not encumbered with microphone equipment, and a user can speak from a distance while moving. At present, however, to achieve high recognition accuracy, a user must be equipped with a close-talking microphone. If the user speaks from a distance, the recognition accuracy seriously degrades because of the influence of reverberation and environmental noise. Therefore, technology for the distant-talking speech recognition becomes important.

The reverberation is defined by the impulse response (acoustic transfer function). The influence of the reverberation is described by a scalar index of the reverberation time, e.g. [53, 96]. The impulse response will change according to not only the shape of a room but also to temperature and humidity. Figure 1.2 and figure 1.3 show examples of waveform and narrow-band spectrogram for original (clean) speech and reverberated speech. When training data of an acoustic model consists of the clean speech data as shown in figure 1.2, and testing data consists of the reverberated speech as shown in figure 1.3, a serious mismatch between the training data and the test utterances occurs. Present spectral-matching measures have a shortcoming of being easily affected by noise, reverberation, and so on. Those measures are very sensitive to spectral distortion. On the other hand, if the training data consists of speech from

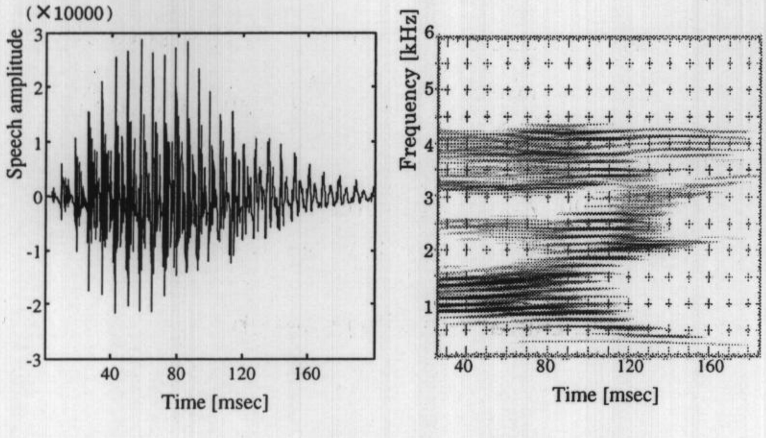

Figure 1.2: Original speech : the speech waveform and narrow-band spectrogram of the Japanese utterance /ai/.

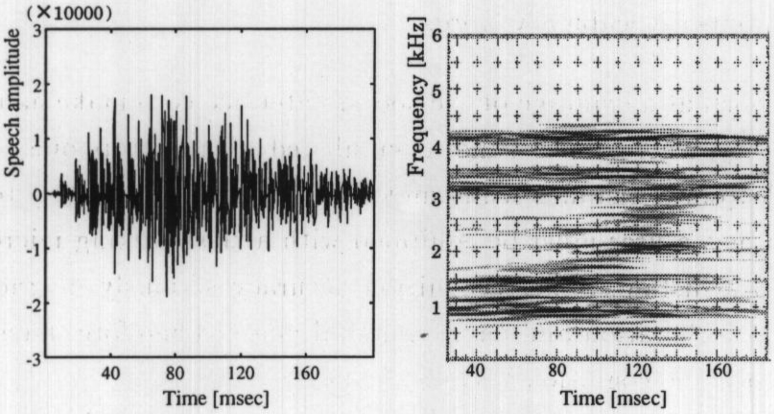

Figure 1.3: Reverberated speech (reverberation time = 0.6 sec) : the speech waveform and narrow-band spectrogram of the Japanese utterance /ai/.

every conceivable combination of signal conditions, the recognition accuracy will not seriously degrade. However, it is not practical to collect a huge set of utterances over every conceivable combination of signal conditions.

Even in the case of a human, adaptability plays a very important role. For example, it is necessary for a human to use a few phrases for adapting to individual speaker differences [45]. Therefore, it is desirable to adapt the acoustic model to the

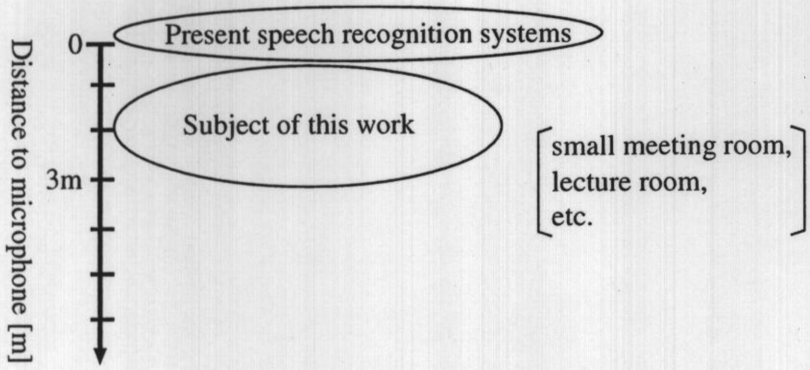

Figure 1.4: Target environments of this work

target environment using a small amount of a user's speech. The aim of the work presented in this thesis is to make automatic speech recognition systems adapt to a new environment.

This thesis details a robust speech recognition technique for the acoustic model adaptation based on the HMM composition and decomposition methods in noisy reverberant environments, where a user speaks from a distance of 0.5 m $\sim$ 3.0 m (figure 1.4). The aim of the HMM composition and decomposition methods is to estimate the model parameters so as to adapt the model to the new environments by using a small amount of a user's speech. The HMM composition algorithm has been proposed for additive noise [16, 55]. In this thesis, the HMM composition algorithm for additive noise is extended to that for the acoustic transfer function of a reverberant room, by using an HMM to model the acoustic transfer function. The states of the acoustic transfer function HMM correspond to different sound source positions. This HMM can represent the positions of the sound sources, even if the speaker moves. This thesis also proposes a new method to estimate HMM parameters of the acoustic transfer function based on the HMM decomposition for hands-free speech recognition. The method is able to estimate the model parameters by using observed speech uttered from an unknown position without measurement of impulse responses. The performance of the HMM composition and decomposition methods is evaluated on real distant-talking speech and telephone speech. It is my hope that this thesis will be useful in human-to-machine communication.

1.2 Literature Review

Much research for robust speech recognition has been done, where the two most important problems to be overcome are

- additive noise,

and

- convolutional distortion.

Additive noise usually consists of background noise, other speakers and so on. Its effect on the speech input is denoted as addition in the wave domain and the linear-spectral domain. Convolutional distortion usually comes from the telephone channel, microphone characteristics, reverberation and so on. Its effect on the speech input is represented as convolution in the wave domain, and is represented as multiplication in the linear-spectral domain.

Many methods have been presented to solve each problem. Those approaches are summarized as follows

- speech enhancement,

and

- model adaptation.

Figure 1.5 shows a robust speech recognition system. We focus on model adaptation using a single microphone in this work. The model adaptation can also be emphasized by using a multi-microphone (microphone array).

The following sections will briefly review some of major approaches to robust speech recognition. Extensive surveys of robust speech recognition can be found in [15, 20, 42, 50].

1.2.1 Speech Enhancement Techniques

This section briefly describes major approaches for robust feature extraction, which reduce the amount of noise or convolutional distortion. The techniques based on microphone arrays are briefly described at the end of this section.

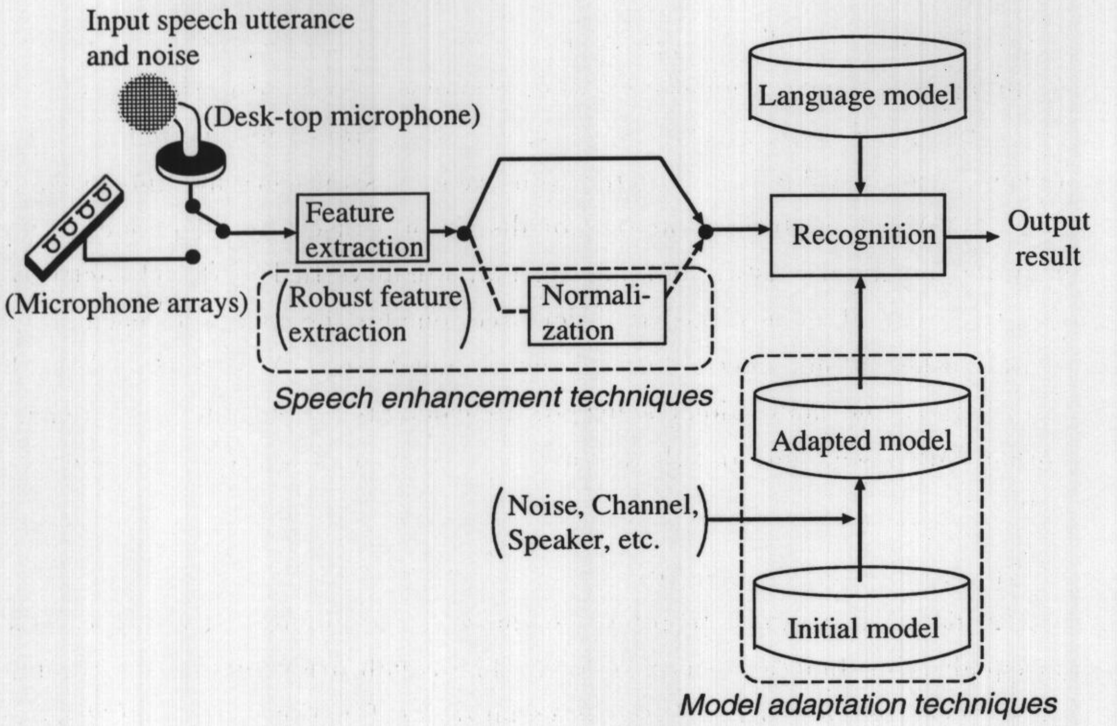

Figure 1.5: Robust speech recognition

For additive noise,

- Spectral subtraction (SS)
  A simple technique is the use of spectral subtraction, where clean speech is estimated by subtracting additive noise from noisy speech in the power-spectral domain [10, 98]. Non-linear spectral subtraction is also proposed in [52].

- Noise robust feature, e.g. [35, 43, 74]
  In [74], the use of formant features (SSC: Spectral Sub-band Centroids) as supplementary features for speech recognition was proposed. Though additive noise affects the speech power spectrum at all the frequencies, the influence is not so much in the higher amplitude (formant) portions of the spectrum. Therefore, the formant features are used as supplementary features for noisy speech in [74].

For convolutional distortion,

- Cepstral mean normalization (CMN)
  A simple bias (convolutional distortion) removal technique is the use of CMN [6]. Here, speech is parameterized using cepstral parameters. To remove a bias, a cepstral mean value is subtracted from each cepstral element. Improved versions of CMN have been proposed for distortion caused by telephone channel, microphone characteristics and speaker individuality. In E-CMN (Exact-CMN) [86], two cepstral mean values are calculated, one for the speech for each speaker and the other for non-speech for each environment. In SCMN (Segmental-CMN) [100, 101], the normalization coefficients are calculated over a sliding finite length normalization segment, and the feature vectors are normalized to zero mean and unit variance.

- Multi-path stochastic equalization
  In [12, 63], a framework to integrate the acoustic model of clean speech and an equalization function was proposed. The equalization function is combined with every possible path, where the equalization function is estimated using a maximum-likelihood framework. In [12], the experiment is conducted on telephone speech data.

For additive noise and convolutional distortion,

- SS and CMN
  Shozakai et al. [87] presented a comparison of some techniques for additive noise such as SS, CSS (Continuous-SS) [67] and MMSE (Minimum Mean Square Error estimation) [14], and proposed a robust speech enhancement approach, E-CMN/CSS, in car environments.

Those techniques have been evaluated on a relatively short impulse response: telephone channel, car environments, and so on. However, they have not been evaluated on noisy reverberant environments.

Speech Enhancement using a Microphone Array

The many techniques using microphone arrays have attempted to enhance speech intelligibility. Robust speech recognition based on microphone arrays has also been investigated recently. The following will briefly review some of the major approaches to robust speech recognition and some of the major approaches to enhance speech using microphone arrays. As for the microphone array techniques, an extensive survey can be found in [71].

For additive noise,

- Beam-forming by a microphone array
  The array signal processing enables high SNR signal-retrieval utilizing information of the differences of speech and noise signal directions. A simple technique is the delay-and-sum beam-former. Adaptive beam-forming techniques have also been proposed [25, 44]. Those techniques are applied to speech recognition, e.g. [22, 29, 72, 103, 104].

For convolutional distortion,

- Inverse filtering of acoustic impulse responses by microphone arrays
  Many techniques to recover reverberated signals, with good intelligibility, in a room have been proposed. For example, in [62], how to calculate the exact inverse of room acoustics by using multiple loudspeakers (or microphones), MINT (Multiple-input/output INverse Theorem) was proposed. In [102], an approach to recover acoustically-reverberated signals using Multi-microphones Sub-Band Envelope Estimation (M-SBEE) was proposed. This technique, using some kind of inverse filtering, is very effective, but a reference signal is required to estimate de-reverberation filters.

For additive noise and convolutional distortion,

- Shields and Campbell [84] reported intelligibility improvements for speech corrupted with noise and reverberation by taking advantage of binaural input channels.

1.2.2 Model Adaptation Techniques

Model adaptation techniques enable robust speech recognition in the acoustic model domain, instead of the parameterization domain. This has the advantage that the observed data are not modified, and front-end processing is not required. Speech recognition systems might be able to have a model of target environments before recognizing observed speech. Model adaptation techniques are also an extension of the technology used in speaker adaptation. The following describes how to deal with additive noise, convolutional distortion, or both. Bayesian adaptive technique is also described briefly at the end of this section.

For additive noise,

- Model (de-)composition of speech and noise
  The observation probability for noisy speech can be calculated from the output of a speech model combined with the output of a noise model. In [99], the output probability is calculated by maximum-approximation in the log-spectral domain, where speech and noise are assumed to be independent. Therefore, their models must be trained in the log-spectral domain. Improved versions have been proposed: PMC (Parallel Model Combination) [16] and HMM composition [54]. In those composition methods, the observed noisy speech is modeled before recognition, and the observation probability is calculated in the cepstral domain. The output probability density function of the speech model and the noise model are converted to the linear-spectral domain, and are composed. Then the composed one is converted back to the cepstral domain. Therefore, the composition method needs the assumption that the sum of two log-normally distributed variables is approximately log-normally distributed (reproduction of distribution) in the linear-spectral domain.

- Parameter generation (Data-driven technique), e.g. [18, 47]
  In [47], an acoustic model parameter estimation method for noisy speech was proposed. The technique is based on cepstral parameter generation from the HMM. The generated sequence of speech and noise from the HMM are combined to yield a noisy speech sequence, and the statistics of the noisy speech sequence are used to obtain the noisy speech model.

- Jacobian adaptation of acoustic model
  In [81], a Jacobian approach to fast adaptation of the acoustic model to noisy environments was proposed. When noise changes slightly, the small changes are approximated by a linear transformation using a Jacobian matrix in the cepstral domain. This technique is based on the idea of adaptation of a model with noise A to a model with noise A'.

For convolutional distortion,

- Maximum-likelihood (ML) approach to stochastic-matching
  In [82, 83], an ML stochastic-matching approach was proposed to decrease the acoustic mismatch between training and testing. The mismatch is reduced by a transformation function that maps the original model to the transformed model that matches better with the testing condition. The transformation function is estimated by using the observed data in an ML framework. In [85], hierarchical structure in the parameter space and an improved version of the ML stochastic-matching are integrated. The experiment for the evaluation is conducted on telephone speech data.

- Linear regression for speaker adaptation
  A popular method for speaker adaptation is maximum-likelihood linear regression (MLLR) [19, 51]. In an MLLR method, the speaker-independent model is adapted to a new speaker by using linear regression transformation. The transformation matrices are calculated in an ML manner.

- Adaptive training for speaker normalization
  In [4, 5, 34], the inter-speaker variability in the training data is reduced. The speaker characteristics are represented as linear transformations of the speaker-independent model. The speaker transformations are calculated in the training phase by an MLLR method [51].

For additive noise and convolutional distortion,

- Model composition and stochastic-matching
  In [61], a technique based on the HMM composition and stochastic-matching was proposed, where additive noise and convolutional distortion are dealt with. The experiment for the evaluation is conducted on simulated data which are passed through a filter.

- Bayesian predictive-classification
  Though prior knowledge about the mismatch mechanism (additive noise or convolutional distortion, etc.) is assumed in many methods, [37, 32, 33] propose a Bayesian predictive-classification approach, where the knowledge of the mismatch mechanism is not assumed.

- Adaptive training

  In [57], the training and testing data are assumed to be recorded with different microphones in a variety of background-noise conditions. The technique consists of combining the noise in testing and training environments. The condition of additive noise becomes the same between training and testing. The new acoustic models are built using the modified noisy speech data.

Bayesian adaptation

The Bayesian learning principle is used to derive maximum a posteriori estimates of the model parameters. The MAP (Maximum A Posteriori) formulation gives a way to combine existing prior knowledge and a small set of newly-acquired task-specific data [49]. The MAP-based techniques have been employed in a number of applications, e.g. [11, 48, 70, 97, 105]. A detailed survey can be found in [49]. For example, [21] proposed the MAP-estimation method for multivariate Gaussian mixture observations of Markov chains. In [58], the use of on-line Bayesian adaptation for speech recognition was proposed. Huo et al. [31] proposed an empirical Bayesian method based on the moment estimates for estimating the parameters of the prior densities.

As previously described, the approaches for robust speech recognition are summarized as speech enhancement and model adaptation. Model adaptation techniques have the advantage that front-end processing is not required, and speech recognition systems might be able to have a model of target environments before recognizing observed speech. The model adaptation technique for robust speech recognition is investigated in this work.

Clearly, many studies that deal with either additive noise or convolutional distortion have been made. However, when both additive noise and convolutional distortion are present, the system's behavior is hard to predict. Some of the studies previously described have dealt with both additive noise and convolutional distortion. But those studies have been done for a relatively short impulse response: telephone channel, car environments, and so on. Those techniques might not be able to deal with the influence of a long impulse response: noisy reverberant environments. Also, techniques using a microphone array to recover reverberated signals, with good intelligibility, in a room have also been proposed. However these techniques require measured impulse responses or a reference signal.

To deal with both noise and reverberation, this thesis proposes a robust speech recognition method based on HMM composition, where utterances are contaminated not only by additive noise but also by an acoustic transfer function. This work focuses on the case that a user speaks from a distance of 0.5 m $\sim$ 3.0 m in a relatively small meeting room, lecture room, etc. This thesis also proposes a new method to estimate HMM parameters of an acoustic transfer function based on HMM decomposition in the model domain. The model parameters are estimated by maximizing the likelihood of adaptation data uttered from an unknown position. The HMM decomposition method does not require measured impulse responses or a reference signal. The proposed method is obtained as the result of the reverse process of the HMM composition.

1.3 Thesis Outline

This thesis is organized as follows. The next chapter, Chapter 2, describes the use of HMMs (Hidden Markov Models) in speech recognition. Chapter 3 describes a robust speech recognition method based on the HMM composition for noisy-distorted speech. Chapter 4 describes a method to estimate HMM parameters of an acoustic transfer function based on the HMM decomposition in the model domain. Chapter 5 describes the performance of the HMM composition and decomposition methods on distant-talking speech. The distant-talking speech is measured in noisy reverberant environments, where a microphone is placed about 2.5 m distant from speakers. Chapter 6 describes performance for the case of a shorter impulse response, telephone speech recognition. The telephone speech data for the evaluation are recorded using 10 kinds of ordinary analog telephone handsets and cordless telephone handsets in a soundproof room, through the public telephone network. Finally, Chapter 7 summarizes this work and suggests future research directions.

Chapter 2

Speech Modeling with HMM

2.1 Stochastic Approach for Speech Recognition

A speech recognition system produces an estimate of the word sequence associated with a given speech waveform. A variety of approaches in speech recognition have been studied, e.g. [77]. In the 1970s, applications of hidden Markov models (HMM) to speech recognition have become a research topic. The models are very rich in mathematical structure. Also, there is the existence of sophisticated and efficient algorithms for training and recognition. Therefore, the models can form the theoretical basis for use in a wide range of applications. This stochastic approach is used in this work. There are more detailed descriptions of the statistical approach in speech recognition [28, 36, 76, 77].

In the stochastic approach, the estimated word sequence $\hat{W}$ is given by

$$\hat{W} = \operatorname*{argmax}_W \Pr(W|O) = \operatorname*{argmax}_W \frac{\Pr(W)\Pr(O|W)}{\Pr(O)}, \tag{2.1}$$

where O is the observed speech data, $\Pr(W)$ is the a-priori probability of the word sequence, $W = w_1, w_2, \ldots, w_I$, and $\Pr(O|W)$ is the probability of the observed speech given the word sequence. Since $\Pr(O)$ is not dependent on the word sequence W, equation (2.1) is rewritten as

$$\hat{W} = \operatorname*{argmax}_W \Pr(W)\Pr(O|W).$$

$\Pr(W)$ is calculated from the language model, where the information about which word sequences are allowable is contained. The probability $\Pr(W)$ is rewritten as

$$\Pr(W) = \Pr(w_1) \prod_{i=2}^{I} \Pr(w_i | w_{i-1}, \ldots, w_1). \tag{2.2}$$

Equation (2.2) is usually approximated by N-grams,

$$\Pr(W) \simeq \Pr(w_1) \prod_{i=2}^{N} \Pr(w_i | w_{i-N+1}, \ldots, w_{i-1}).$$

This means that the probability of the current word is only dependent on the previous $N-1$ words. Typically, N is chosen to be 3 (trigram), 2 (bigram) and 1 (unigram).

$\Pr(O|W)$ is calculated from the acoustic models, HMM. This thesis focuses on the adaptation of the acoustic models to new environments. The acoustic model, HMM, is described in the next section.

2.2 Definition of HMM

An HMM is used as the most widely and successful stochastic approach in speech recognition. The unit of the HMM speech model is usually a phoneme or a word. The phoneme is used as the unit of the speech model in this work. In the case of Japanese, there are about 20 kinds of phonemes: vowels, consonants, fricatives, affricates, nasals and so on. The use of the HMM in speech recognition requires an assumption:

- speech is split into small segments, where each segment is considered to be stationary.

Figure 2.1 shows the speech waveform and wide-band spectrogram of the Japanese utterance /aite/. The speech waveform and wide-band spectrogram change in time. But if the speech is split into small segments (20 $\sim$ 40 msec), each segment can be assumed to be stationary. In speech recognition, the speech spectrum is converted to cepstral parameters which can retain useful speech information. Section 2.5 describes how to analyze the speech.

An example of an HMM is shown in figure 2.2. An HMM can be formally defined by the number of states, the initial state probability density function, the state transition probability matrix and the observation probability density function (PDF). A phoneme HMM has usually three states, and has a simple left-to-right topology. Therefore, the initial state probability is 1.0 for the first state and 0 for all other states. In the example

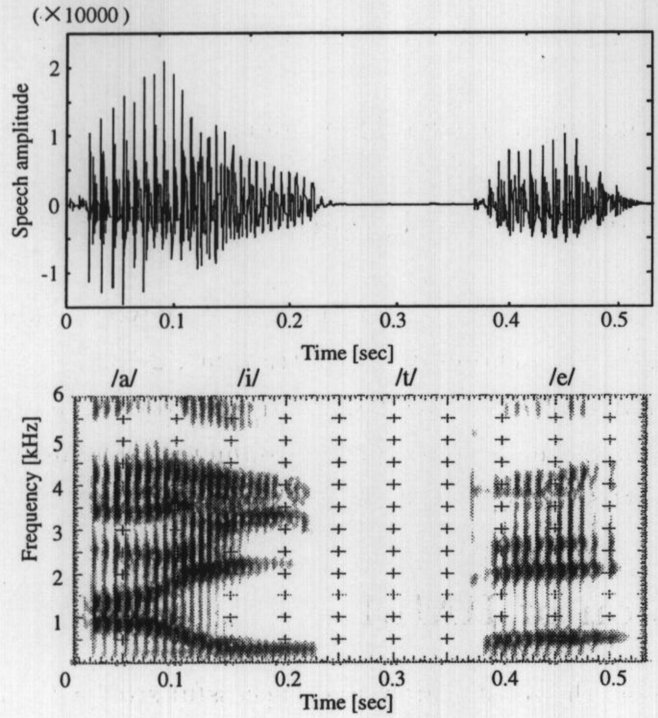

Figure 2.1: Speech waveform and wide-band spectrogram of the Japanese utterance /aite/, where /a/, /i/ and /e/ are vowels, and /t/ is a plosive.

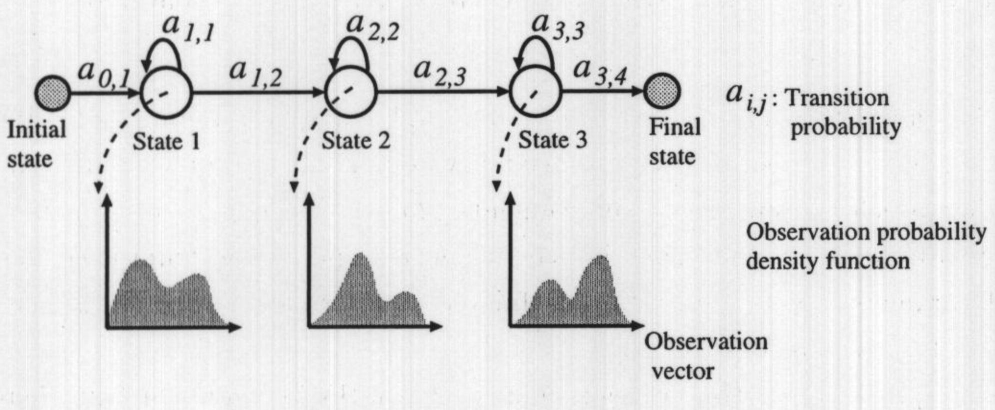

Figure 2.2: 3-state hidden Markov model (HMM) for a left-to-right topology without skips

of figure 2.2, the state transition probability matrix is given by

$$\begin{pmatrix} 0 & a_{0,1} & 0 & 0 & 0 \\ 0 & a_{1,1} & a_{1,2} & 0 & 0 \\ 0 & 0 & a_{2,2} & a_{2,3} & 0 \\ 0 & 0 & 0 & a_{3,3} & a_{3,4} \end{pmatrix},$$

where

$$\sum_j a_{i,j} = 1.$$

The observation probability density function associated with each state is a discrete type or a continuous type. The continuous type is only considered in this work. The observation probability density function for the continuous type is usually modeled by a multivariate Gaussian distribution or a mixture of the multivariate Gaussian distribution. The mixture of the multivariate Gaussian distribution in state j is given by

$$b_j(o_t) = \sum_{k=1}^{K} \omega_{j,k} N(o_t; \mu_{j,k}, \Sigma_{j,k}), \quad \sum_{k=1}^{K} \omega_{j,k} = 1.0,$$

where $N(o_t; \mu_{j,k}, \Sigma_{j,k})$ is a multivariate Gaussian distribution with the mean vector $\mu_{j,k}$ and the covariance matrix $\Sigma_{j,k}$, and K is the total number of multivariate Gaussian distributions.

To improve accuracy of speaker-independent recognition tasks, the total number of mixtures is increased, and context-dependent models are also used, where the current phoneme is dependent on the preceding and following phonemes. However, it is sometimes difficult to obtain sufficient data to accurately estimate all the model parameters. Therefore, it is necessary to tie sets of model parameters together. In a tied-mixture HMM, each model shares the same PDFs which should be representative of the acoustic space. The observation probability density function in state j is given by

$$b_j(o_t) = \sum_{k=1}^{K} \omega_{j,k} N(o_t; \mu_k, \Sigma_k),$$

where the observation probability density function in each state is defined by K mixture weights. Figure 2.3 shows the continuous density function and the tied-mixture density function.

2.3 Recognition Algorithm

The probability of the observation sequence is calculated from the given HMM. Since the state sequence is *hidden* or not observed, the probability for all possible state sequences is given by summing over all possible paths through the HMM

$$\Pr(o|\lambda) = \sum_{\Theta} a_{0,1} \prod_{t=1}^{T} a_{\theta_l(t-1),\theta_l(t)} b_{\theta_l(t)}(o_t), \tag{2.3}$$

where λ is the set of HMMs linked with the word sequence, and Θ is the set of all L possible state sequences of length T in the model λ

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_L\}.$$

It is very expensive to calculate equation (2.3) directly. But an efficient algorithm exists for this calculation. It is called the forward-backward algorithm. The forward-backward algorithm is used to estimate the parameters of the HMM.

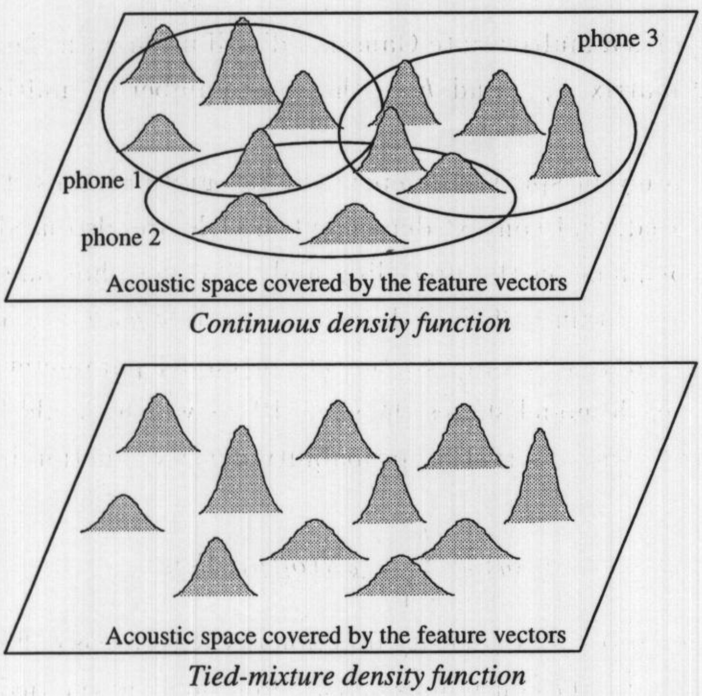

Figure 2.3: Continuous density function and tied-mixture density function

Forward algorithm

- Step 1.

$$\alpha_1(0) = 1,$$

$$\alpha_j(0) = 0, \quad \text{if} \quad j \neq 1.$$

- Step 2. For $t = 1, 2, \ldots, T$,

$$\alpha_j(t) = \left[\sum_{i=1}^{N} \alpha_i(t-1) a_{i,j} \right] b_j(o_t).$$

- Step 3. Terminate with

$$\Pr(o|\lambda) = \sum_{i=1}^{N} \alpha_i(T).$$

Backward algorithm

- Step 1.

$$\beta_j(T) = a_{j,N}, \quad 1 \leq j \leq N.$$

- Step 2. For $t = T-1, T-2, \ldots, 0$,

$$\beta_i(t) = \sum_{j=1}^{N} a_{i,j} b_j(o_{t+1}) \beta_j(t+1).$$

- Step 3. Terminate with

$$\Pr(o|\lambda) = \sum_{i=1}^{N} \beta_i(0).$$

A more efficient algorithm exists for the calculation of equation (2.3). It is called the Viterbi algorithm. The calculation is essentially the same as the forward algorithm except that the summation is replaced by a maximization.

Viterbi algorithm

- Step 1.

$$\alpha_1(0) = 1,$$

$$\alpha_j(0) = 0, \quad \text{if} \quad j \neq 1.$$

- Step 2. For $t = 1, 2, \ldots, T$,

$$\alpha_j(t) = \left[\max_i \alpha_i(t-1)a_{i,j}\right] b_j(o_t).$$

- Step 3. Terminate with

$$\Pr(o|\lambda) = \max_i \alpha_i(T).$$

2.4 Estimation of HMM Parameters

The state sequence cannot be observed directly from a given set of training data. Therefore, a locally-optimal algorithm, the Baum-Welch algorithm [8, 9], based on the Expectation-Maximization (EM) algorithm [13], is usually used. The basic idea is that a good initial estimate of the parameters is first assumed, and the likelihood is optimized iteratively.

The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step (E-step), the auxiliary function is given by

$$\begin{aligned} Q(\lambda, \hat{\lambda}) &= E[\log \Pr(o, \theta|\hat{\lambda}) \mid o, \lambda] \\ &= \sum_{\theta_i \in \Theta} \Pr(o, \theta_i|\lambda) \log \Pr(o, \theta_i|\hat{\lambda}), \end{aligned}$$

where Θ is the set of all possible state sequences. In the second step, called the maximization step (M-step), the estimate of λ, $\hat{\lambda}$ is calculated by maximization of the auxiliary function, $Q(\lambda, \hat{\lambda})$

$$\hat{\lambda} = \operatorname*{argmax}_{\lambda} Q(\lambda, \hat{\lambda}).$$

Iteratively applying the E and M steps guarantees that the likelihood is non-decreasing [8]

$$\Pr(o|\hat{\lambda}) \geq \Pr(o|\lambda).$$

There are many references for the derivation of the model parameter estimation formula (e.g. [28, 36, 76, 77]). Here, only the results for the continuous density HMM are quoted. The estimates for the mean, variances and mixture weights for k-th mixture component in state j are given by

$$\hat{\mu}_{j,k} = \frac{\sum_{t=1}^{T} \gamma_{t,j,k} o_t}{\sum_{t=1}^{T} \gamma_{t,j,k}},$$

$$\hat{\Sigma}_{j,k} = \frac{\sum_{t=1}^{T} \gamma_{t,j,k}(o_t - \hat{\mu}_{j,k})'(o_t - \hat{\mu}_{j,k})}{\sum_{t=1}^{T} \gamma_{t,j,k}},$$

$$\hat{\omega}_{j,k} = \frac{\sum_{t=1}^{T} \gamma_{t,j,k}}{\sum_{t=1}^{T} \gamma_{t,j}},$$

where $'$ denotes the transposition, and γ is given by

$$\gamma_{t,j,k} = \sum_{i} \frac{\alpha_i(t-1)a_{i,j}\omega_{j,k}b_{j,k}(o_t)\beta_j(t)}{\Pr(o|\lambda)}.$$

This maximum-likelihood criterion is considered in this work. Other criteria are also seen in [40, 69].

2.5 Speech Analysis

Cepstral parameters, e.g.[68], are an effective representation to retain useful speech information in speech recognition. At present, many speech recognition systems are based on cepstral parameters. The term *cepstrum* is a word coined from the inverse transform of the spectrum.

Now the speech signal $o(\omega)$ is given by the multiplication of a pseudo-periodic source, $g(\omega)$, and the impulse response of the vocal tract, $v(\omega)$, in the spectral domain as follows:

$$o(\omega) = g(\omega)v(\omega).$$

The cepstrum is given by the inverse Fourier transform of $\log|o(\omega)|$,

$$\mathrm{F}^{-1}\log|o(\omega)| = \mathrm{F}^{-1}\log|g(\omega)| + \mathrm{F}^{-1}\log|v(\omega)|, \tag{2.4}$$

where F and log is the Fourier transform and the logarithm transform, respectively. As shown in the above equation, the cepstral analysis can separate the speech signal into the fine structure (the first term of equation (2.4)) and the spectral envelope (the second term of equation (2.4)). Liftering is the process of weighting in the cepstral domain so as to help separate those two components. Various liftering functions may be used [39, 95]. For this work, the cepstral parameters are weighted according to

$$w(n) = \begin{cases} 1 + \frac{L}{2} \cdot \sin(\frac{n\pi}{L}) & 1 \leq n \leq L \\ 0 & n \leq 0,\ n > L, \end{cases}$$

where L is the liftering coefficient [39]. The form of this filter is shown in figure 2.4.

The human-perceived pitch does not correspond linearly to the physical frequency of the tone. A popular approach to simulate the auditory characteristics more precisely is the use of a mel-scale. For this work, the relationship between frequency and mel-scale is given by

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right),$$

where $\mathrm{Mel}(f)$ is the perceived frequency in mel (Hz) [106]. The mel-scale is shown in figure 2.5. The mapping is approximately linear below 1 kHz and logarithmic above.

The block diagram of cepstral analysis is shown in figure 2.6. The speech waveform is split into a small segment by a window function. Each segment is converted to the linear spectral domain by applying the discrete Fourier transform (DFT). Then the logarithm and inverse discrete Fourier transform (IDFT) are applied, and the cepstral parameters are obtained.

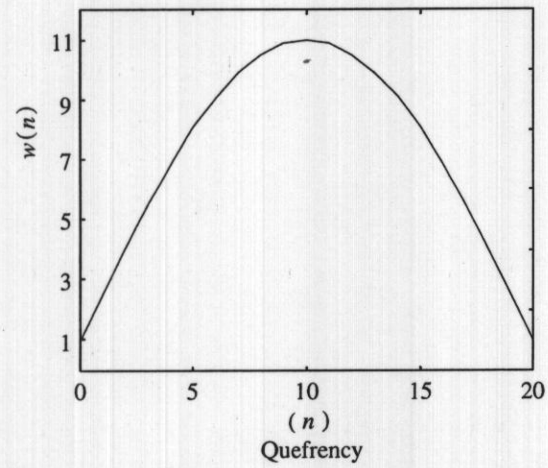

Figure 2.4: An example of low-time lifter. This was used successfully by Juang et al. (1987). Liftering coefficient: $L = 20$.

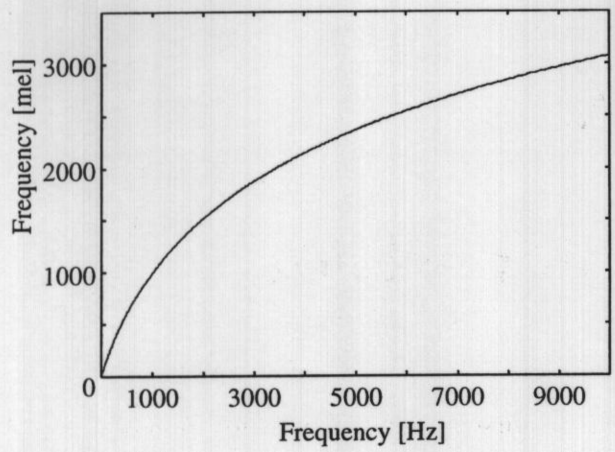

Figure 2.5: An example of mel-scale. The mapping is approximately linear below 1 kHz and logarithmic above.

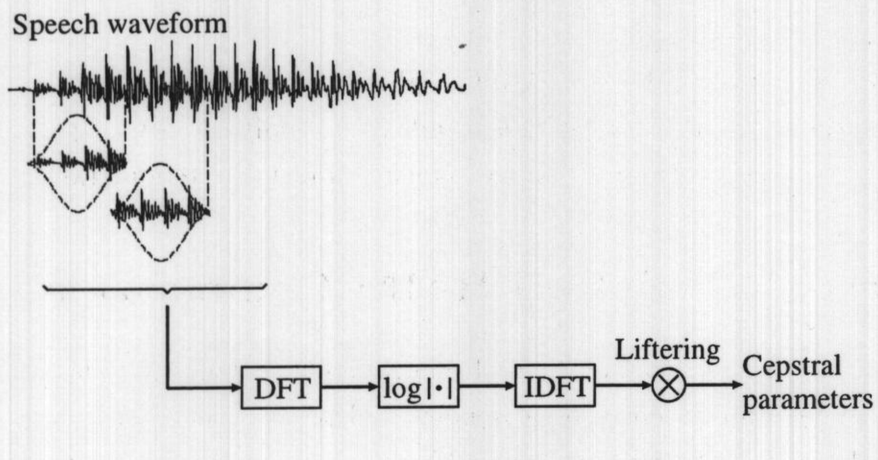

Figure 2.6: Block diagram of cepstral analysis.

Chapter 3

Model Composition

This chapter describes a robust speech recognition method based on HMM composition for noisy and acoustically-distorted speech, where utterances are contaminated not only by additive noise but also by an acoustic transfer function. The method realizes a hands-free user interface such that a user is not encumbered by microphone equipment even in noisy reverberant environments. The HMM composition algorithm has been proposed for additive noise [16, 54, 55]. In this chapter, the HMM composition algorithm for additive noise is extended to that for the acoustic transfer function of a reverberant room [65, 66, 89, 90], by using an HMM to model the acoustic transfer function. The states of the acoustic transfer function HMM correspond to different sound source positions. This HMM can represent the positions of the sound sources, even if the speaker moves.

Section 3.1 describes the basic principle of HMM composition. Section 3.2 describes the HMM composition method for noisy and acoustically-distorted speech in detail. Section 3.3 describes the structure of the acoustic transfer function HMM.

3.1 Basic Principle of HMM Composition

There are many kinds of sounds in real environments. For example, the voices of surrounding people, noisy footsteps, car noise and so on. Then there is also acoustic reflection and reverberation in a room. If a speech recognition system knows the conceivable sounds in the target environment before recognizing observed speech, the system might be able to deal with their influence.

The HMM composition method is based on the idea of "Source Modeling". It can produce speech in the target environment by adapting the model to the target environment, the speaker and so on. Observed speech will be recognized by using the adapted model, even if we do not know about the kinds of noise which contaminates speech.

3.2 HMM Composition for Noisy and Acoustically-Distorted Speech

3.2.1 Structure of Composed HMM

This section describes the HMM composition algorithm for noisy and acoustically-distorted speech. The environment model is shown in figure 3.1, where the speech is contaminated by noise and an acoustic transfer function. An example of this kind of combination is shown in figure 3.2. The structure of the composed HMM is given by the Cartesian product of the component HMMs. The number of states for a noise HMM and an acoustic transfer function HMM are one and three, respectively in this example. Therefore, the parameters of the composed HMM are given by

$$
\begin{aligned}
&\textit{Number of states}\\
&\quad = \mathrm{Num\_states}(\textit{Clean speech HMM}) \times \mathrm{Num\_states}(\textit{Noise HMM})\\
&\qquad\quad \times \mathrm{Num\_states}(\textit{Acoustic transfer function HMM})\\
&\quad = 3 \times 1 \times 3 = 9\,,
\end{aligned}
$$

$$
\begin{aligned}
&\textit{Number of Gaussian mixture components of the output probability}\\
&\quad = \mathrm{Num\_mixtures}(\textit{Clean speech HMM}) \times \mathrm{Num\_mixtures}(\textit{Noise HMM})
\end{aligned}
$$

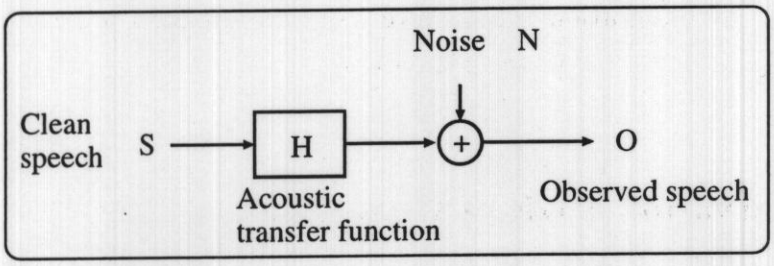

Figure 3.1: The environment model for noisy and acoustically-distorted speech

$\times$ Num_mixtures(*Acoustic transfer function HMM*).

Here, we denote $\omega_{j,k}^{(CleanSpeech)}$ as the weight of the mixture k in the state j of the clean speech HMM, $\omega_{j',k'}^{(Noise)}$ as the weight of the mixture k' in the state j' of the noise HMM, and $\omega_{j'',k''}^{(Acous.t.f.)}$ as the weight of the mixture k'' in the state j'' of the acoustic transfer function HMM. The weights of the composed HMM are given by

$$\begin{aligned}&\textit{Weight of the mixture components}\\ &\quad = \omega_{j,k}^{(CleanSpeech)} \times \omega_{j',k'}^{(Noise)} \times \omega_{j'',k''}^{(Acous.t.f)},\end{aligned}$$

where each state number and mixture number depends on each HMM.

The transition probabilities of the composed HMM are given by

$$a_{i,j}^{(Composed)} = a_{i,j}^{(CleanSpeech)} \times a_{i',j'}^{(Noise)} \times a_{i'',j''}^{(Acous.t.f.)}.$$

For example, the transition probability of the composed HMM in figure 3.2, from the state "A,D,E" to the state "B,D,E", is given by

$$a_{A,B}^{(CleanSpeech)} \times a_{D,D}^{(Noise)} \times a_{E,E}^{(Acous.t.f.)}.$$

The observation probability density function (PDF) of the composed HMM will take the following general form

$$b(o_t) = \int_{C_t} \Pr(s_t, n_t, h_t)$$

$$C_t \equiv \mathcal{F}(s_t, n_t, h_t) = o_t,$$

where the integration is over all triples (s_t, n_t, h_t), and s_t, n_t and h_t are clean speech, noise and an acoustic transfer function at time t, respectively. The function $\mathcal{F}$ denotes the interaction of s_t, n_t and h_t which produces the observation o_t. It is difficult to calculate the above integration. Therefore, some approximation is necessary. The following section describes how to calculate the observation PDF of the composed HMM.

3.2.2 Observation PDF of Composed HMM

First, on the assumption that speech and noise are independent, the observed speech is represented by

$$O(\omega; m) = S(\omega; m) + N(\omega; m).$$

where $O(\omega; m)$, $S(\omega; m)$ and $N(\omega; m)$ are the observed noisy speech, clean speech and noise, respectively. Since this relation is preserved in the linear-spectral domain, we regard $O(\omega; m)$, $S(\omega; m)$, $N(\omega; m)$ as short-time linear spectra at frame m.

The conventional approach estimates noise statistics during a noise period and recognizes an input-noisy speech by using the noise-added reference patterns. The HMM composition executes the addition in the HMM parameter domain instead of the addition in the signal domain. Since the signal level is generally different between training and testing, an adjustment factor k is introduced. Therefore, the observed

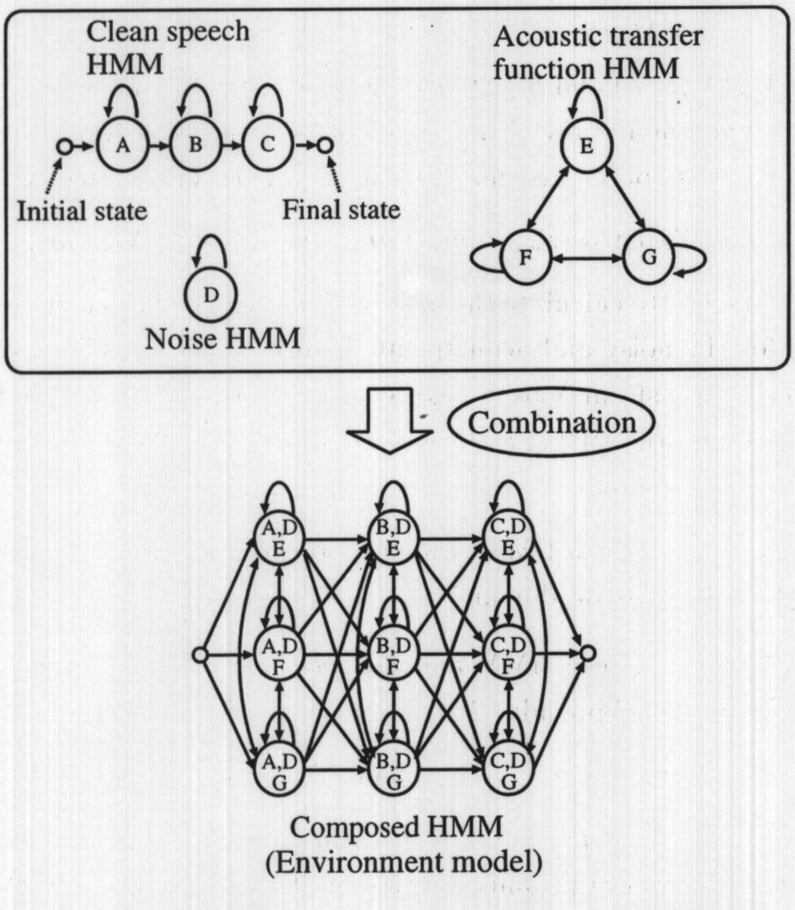

Figure 3.2: An example of a composed HMM

speech is represented by

$$O(\omega; m) = S(\omega; m) + k \cdot N(\omega; m),$$

Generally, model parameters in speech recognition are represented by the cepstrum. The model parameters have to be transformed to the linear domain where the additive property of speech and noise holds [16, 55].

As for a convolutional distortion, the observed spectrum is represented by

$$O(\omega; m) = S(\omega; m) \cdot H(\omega; m),$$

where $H(\omega; m)$ is an acoustic transfer function. If a convolutional distortion is caused by the acoustic transfer function from the sound source to the microphone, $H(\omega; m)$ is a function of frame m, since the sound source may move. The multiplication can be converted to a sum in the cepstral domain as,

$$O_{cep}(t; m) = S_{cep}(t; m) + H_{cep}(t; m),$$

where, $O_{cep}(t; m)$, $H_{cep}(t; m)$ and $S_{cep}(t; m)$ are the cepstra for the observed speech, the acoustic transfer function and the clean speech of quefrency t at frame m, respectively. Therefore, the observed speech, as shown in figure 3.1, is represented by

$$O(t) = \mathrm{Exp}\{\mathrm{Cos}(S_{cep}(t; m) + H_{cep}(t; m))\} + k \cdot N(\omega; m). \tag{3.1}$$

Figure 3.3 shows how to calculate the observation probability density function of the composed PDF for the noisy and acoustically-distorted speech. The cosine transform (Cos), inverse cosine transform (Cos^{-1}), exponential transform (Exp) and log transform (Log) are conducted on the PDFs, as explained in detail in Appendix A.

The procedure is as follows:

1. Estimate the clean speech HMM, the noise HMM and the acoustic transfer function HMM in the cepstral domain.

2. Compose the clean speech HMM and the acoustic transfer function HMM in the cepstral domain (see Appendix A.3)

$$\mu_{(cep\_SH)} = \mu_{(cep\_S)} + \mu_{(cep\_H)} \quad \text{and} \quad \Sigma_{(cep\_SH)} = \Sigma_{(cep\_S)} + \Sigma_{(cep\_H)}.$$

   Here, $\mu_{(cep\_S)}$, $\Sigma_{(cep\_S)}$, $\mu_{(cep\_H)}$, $\Sigma_{(cep\_H)}$, $\mu_{(cep\_SH)}$ and $\Sigma_{(cep\_SH)}$ correspond to a mean vector and a covariance matrix of the clean speech HMM, the acoustic transfer function HMM and the composed HMMs in the cepstral domain, respectively.

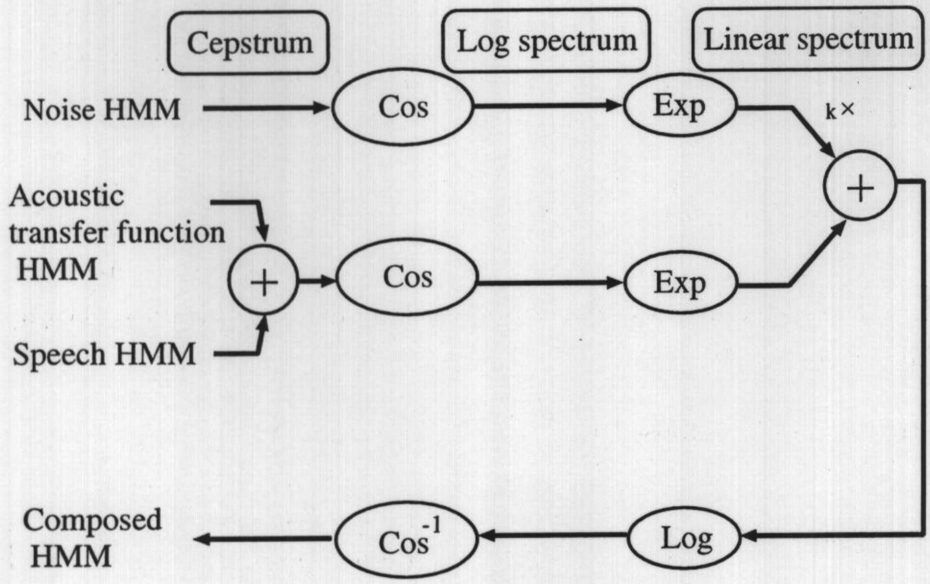

Figure 3.3: Block diagram of the proposed HMM composition

3. Cosine transform of each Gaussian PDF of HMMs (see Appendix A.1)

$$\mu_{(log\_SH)} = \Gamma \cdot \mu_{(cep\_SH)} \quad \text{and} \quad \Sigma_{(log\_SH)} = \Gamma' \cdot \Sigma_{(cep\_SH)} \cdot \Gamma.$$

   Here, Γ is a cosine transform matrix, $\mu_{(log\_SH)}$ and $\Sigma_{(log\_SH)}$ are a mean vector and a covariance matrix of a Gaussian PDF in the log-power spectral domain, respectively.

4. Exponential transform to the linear-spectral domain. (see Appendix A.2)

   The normal random vectors obtained by exponential transform, $Z = \exp(Y)$, has log-normal distribution. The mean and the covariance are given by

$$\mu_{(lin\_SH),i} = \exp\left\{\mu_{(log\_SH),i} + \frac{\sigma^2_{(log\_SH),ii}}{2}\right\}$$

   and

$$\sigma^2_{(lin\_SH),ij} = \mu_{(log\_SH),i} \cdot \mu_{(log\_SH),j} \cdot \{\exp(\sigma^2_{(log\_SH),ij} - 1)\}.$$

   Here, $\mu_{(lin\_SH)}$ and $\Sigma_{(lin\_SH)}$ are the mean vector and the covariance matrix in the linear-power spectral domain.

5. Compose two distributions according to equation (3.1) (see Appendix A.3)

$$\mu_{(lin\_SHN)} = \mu_{(lin\_SH)} + k \cdot \mu_{(lin\_N)}$$

and

$$\Sigma_{(lin\_SHN)} = \Sigma_{(lin\_SH)} + k^2 \cdot \Sigma_{(lin\_N)}.$$

Here, $\mu_{(lin\_N)}$, $\Sigma_{(lin\_N)}$, $\mu_{(lin\_SHN)}$ and $\Sigma_{(lin\_SHN)}$ are the mean vector and the covariance matrix of noise and composed models in the linear-power spectral domain, respectively.

6. Log transform of composed HMMs (see Appendix A.4)

$$\mu_{(log\_SHN),i} = \log \mu_{(lin\_SHN),i} - \frac{1}{2}\left\{\frac{\sigma^2_{(lin\_SHN),ij}}{\mu_{(lin\_SHN),i} \cdot \mu_{(lin\_SHN),i}} + 1\right\}$$

and

$$\sigma^2_{(log\_SHN),ij} = \log\left\{\frac{\sigma^2_{(lin\_SHN),ij}}{\mu_{(lin\_SHN),i} \cdot \mu_{(lin\_SHN),j}} + 1\right\}.$$

7. Inverse cosine transform to the cepstral domain

$$\mu_{(cep\_SHN)} = \Gamma^{-1} \cdot \mu_{(log\_SHN)} \quad \text{and} \quad \Sigma_{(cep\_SHN)} = (\Gamma^{-1})' \cdot \Sigma_{(log\_SHN)} \cdot \Gamma^{-1}.$$

The HMM recognizer decodes observed speech on a trellis diagram according to maximize the log-likelihood. The decoded path will find an optimal combination of speech, noise and the acoustic transfer function.

3.3 Modeling of Acoustic Transfer Function

This section describes the structure of the acoustic transfer function HMM. Figure 3.4 shows the proposed acoustic transfer function HMM in the case of five states. Each state of the acoustic transfer function HMM corresponds to a position of sound sources, and all transitions among states are permitted. Therefore, the proposed acoustic transfer function HMM is able to represent the position of sound sources, even if the speaker moves. Since each state of the acoustic transfer function HMM has Gaussian distributions, it is also possible for the acoustic transfer function HMM to deal with the variation of a user's position or a influence of long impulse response.

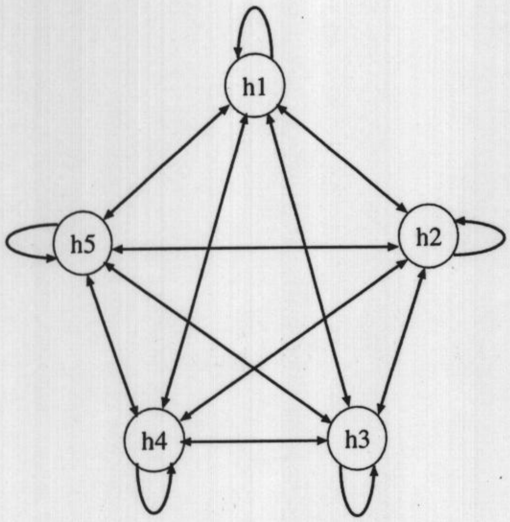

Figure 3.4: An ergodic HMM of acoustic transfer functions

The spectral analysis in speech recognition is based on short time windowing. The multiplication of short time signal spectra and the transfer function is equivalent to the periodic convolution in the time domain. However, actual distorted speech results from the linear convolution. Since the proposed HMM composition of the speech and acoustic transfer function only realizes a periodic convolution, the composed HMM cannot model an actual acoustically-distorted speech accurately. The difference between using periodic and linear convolution will be large according to the length of the impulse response. In this thesis, the covariance matrix of the Gaussian PDF deals with the influence of the long impulse response.

Chapter 4

Model Decomposition

This chapter describes a new method to estimate HMM parameters of an acoustic transfer function based on the HMM decomposition method in the model domain [91]. The model parameters are estimated by maximizing the likelihood of adaptation data. The proposed method is obtained as the result of the reverse process of the HMM composition. The previous chapter described a method which can model observed speech by the composition of HMMs modeling clean speech, noise and the acoustic transfer function. This method, however, requires measurement of impulse responses to train the acoustic transfer function HMM. It is inconvenient and unrealistic to measure impulse responses for a new environment. The new method is able to estimate HMM parameters of the acoustic transfer function from a small amount of adaptation data.

4.1 Basic Principle of HMM Decomposition

Model parameters are estimated in a maximum-likelihood (ML) manner using the expectation-maximization (EM) algorithm, where the likelihood of the observed speech is maximized

$$\hat{\lambda}_H = \underset{\lambda_H}{\operatorname{argmax}} \Pr(O|\lambda_H, \lambda_N, \lambda_S).$$

Here, λ denotes the set of the HMM parameters. S, N and H denote clean speech, noise and the acoustic transfer function.

Now the observed speech is represented by

$$O_{cep}(t;m) = \mathrm{Cos}^{-1}[\mathrm{Log}\{ \mathrm{Exp}(\mathrm{Cos}(S_{cep}(t;m) + H_{cep}(t;m))) + N(\omega;m)\}]. \quad (4.1)$$

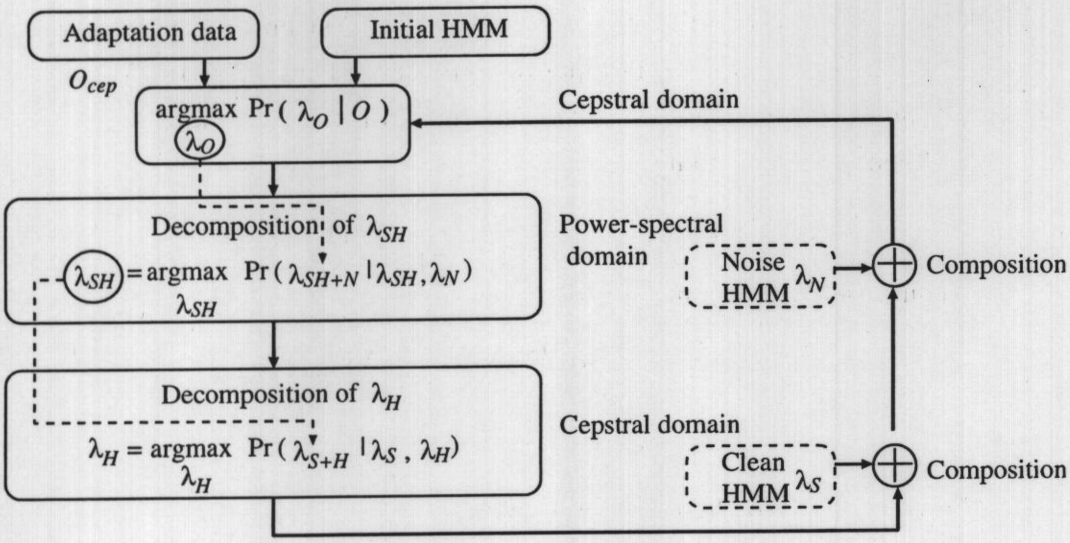

Figure 4.1: Parameter estimation by HMM decomposition

Here, Cos and Cos^{-1} are Fourier (cosine) transform and inverse Fourier (cosine) transform, respectively. $O_{cep}(t;m)$, $S_{cep}(t;m)$ and $H_{cep}(t;m)$ are cepstra for the observed speech, the clean speech and the acoustic transfer function of quefrency t at the m-th frame, and $N(\omega;m)$ is the linear spectrum for noise of frequency ω at the m-th frame. Accordingly, the acoustic transfer function is represented by

$$H_{cep}(t;m) \quad = \quad \text{Cos}^{-1}[\text{Log}\{\ \text{Exp}(\text{Cos}(O_{cep}(t;m))) - N(\omega;m)\}] - S_{cep}(t;m).$$

The estimation equation of the acoustic transfer function HMM is written as

$$\lambda_{H_{cep}} \quad = \quad \text{Cos}^{-1}[\text{Log}\{\ \text{Exp}(\text{Cos}(\lambda_{O_{cep}})) \ominus \lambda_{N_{lin}}\}] \ominus \lambda_{S_{cep}},$$

where the suffixes of *cep* and lin represent the cepstral domain and the linear-spectral domain, respectively. This equation shows that the HMM decomposition is applied twice to the noisy and acoustically-distorted speech. First, the HMM decomposition method is applied in the linear-spectral domain to estimate the distorted speech HMMs by ML estimation. Then, the distorted speech HMMs are converted to the cepstral domain, and the HMM decomposition method is applied again in the cepstral domain to estimate the acoustic transfer function HMM by ML estimation. The procedure is as follows

1. Re-estimate parameters $\lambda_{O_{cep}}$ of composed HMMs using adaptation data with corresponding transcription by ML estimation in the cepstral domain. Next, estimate parameters $\lambda_{N_{cep}}$ of the noise HMM from the signal during noise periods, and then convert $\lambda_{O_{cep}}$ and $\lambda_{N_{cep}}$ to the linear-spectral domain

$$\lambda_{O_{lin}} = \mathrm{Exp}(\mathrm{Cos}(\lambda_{O_{cep}})), \quad \lambda_{N_{lin}} = \mathrm{Exp}(\mathrm{Cos}(\lambda_{N_{cep}})).$$

Decompose $\lambda_{SH_{lin}}$ from $\lambda_{O_{lin}}$ as follows

$$\begin{aligned}\hat{\lambda}_{SH_{lin}} &= \underset{\lambda_{SH_{lin}}}{\mathrm{argmax}} \Pr(\lambda_{SH+N_{lin}} | \lambda_{SH_{lin}}, \lambda_{N_{lin}}) \\ &\triangleq \lambda_{O_{lin}} \ominus \lambda_{N_{lin}}.\end{aligned}$$

2. Convert $\hat{\lambda}_{SH_{lin}}$ to the cepstral domain

$$\lambda_{S+H_{cep}} = \mathrm{Cos}^{-1}(\mathrm{Log}(\lambda_{SH_{lin}})).$$

Then decompose $\lambda_{H_{cep}}$ from $\lambda_{S+H_{cep}}$ as follows

$$\begin{aligned}\hat{\lambda}_{H_{cep}} &= \underset{\lambda_{H_{cep}}}{\mathrm{argmax}} \Pr(\lambda_{S+H_{cep}} | \lambda_{H_{cep}}, \lambda_{S_{cep}}) \\ &\triangleq \lambda_{S+H_{cep}} \ominus \lambda_{S_{cep}}.\end{aligned}$$

The procedure is summarized in figure 4.1. The HMM decomposition method, as shown in figure 4.1, is applied twice to the noisy and acoustically-distorted speech. In the HMM decomposition method, the composed HMM is separated into a known HMM and an unknown HMM by operations on the model parameters based on maximum-likelihood estimation. Figure 4.2 illustrates the HMM decomposition method. The following sections describe the model decomposition in detail. Section 4.2 describes decomposition of a known noise HMM and unknown distorted speech HMMs. Section 4.3 describes decomposition of known speech HMMs and an unknown acoustic transfer function HMM.

4.2 Decomposition of Noise HMM and Distorted Speech HMMs

This section describes the decomposition of a known noise HMM and unknown distorted speech HMMs. Consider tied-mixture HMMs with diagonal covariance matrices,

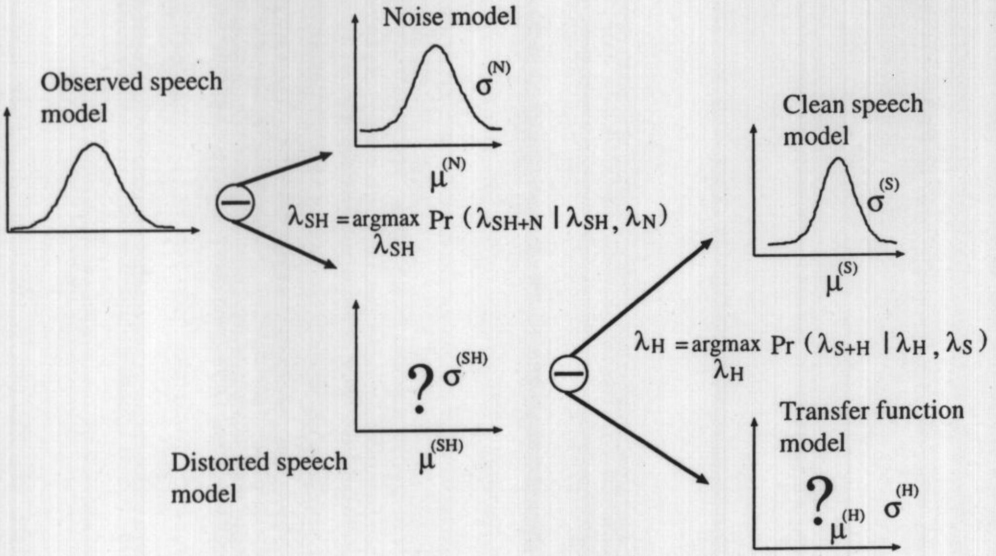

Figure 4.2: HMM decomposition method in noisy reverberant environments

$\lambda = (A, B)$, where $A = [a_{i,j}]$, $i, j = 1, 2, \cdots, J$ is the transition probability matrix, and $B = [b_j]$, $j = 1, 2, \cdots, J$, is the observation probability density function (PDF). J is the number of states. The observation Gaussian PDF $b_j(o_t)$ is given by

$$b_j(o_t) = \sum_{k=1}^{K} \omega_{j,k} N(o_t; \mu_k, \Sigma_k), \tag{4.2}$$

where $\omega_{j,k}$ is the probability of mixture k in state j, and K is the total number of Gaussian PDFs tied by all of the states, and $N(o_t; \mu_k, \Sigma_k)$ is the multivariate Gaussian distribution given by

$$N(o_t; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(o_t - \mu_k)' \Sigma_k{}^{-1} (o_t - \mu_k)\right\},$$

where D is the dimension of the adaptation vector o_t. Next, μ_k and Σ_k are the mean vector and the covariance matrix corresponding to mixture k, respectively, and $'$ denotes transposition.

For an adaptation data sequence o, let s and k be the unobserved state sequence and the unobserved mixture component label, respectively. The probability of observing the state sequence s is simply

$$\Pr(s|A) = \prod_{t=1}^{T} a_{s_{t-1}, s_t},$$

where $a_{s_0,s_1} = 1$. The joint probability for observing the sequences o and s can be evaluated as

$$\Pr(o, s|\lambda) = \prod_{t=1}^{T} a_{s_{t-1},s_t} b_{s_t}(o_t).$$

The joint probability for observing the sequences o, s and k can be evaluated as

$$\Pr(o, s, k|\lambda) = \prod_{t=1}^{T} \{a_{s_{t-1},s_t} \omega_{s_t,k_t} N(o_t; \mu_{k_t}, \Sigma_{k_t}).$$

Therefore, the probability for observing the sequence o is then measured by

$$\Pr(o|\lambda) = \sum_{s} \sum_{k} \Pr(o, s, k|\lambda),$$

where the summations are taken over all possible state sequences and all possible mixture component labels.

Now, the decomposition of distorted speech HMMs ($\lambda_{SH_{lin}}$) is handled in a maximum-likelihood framework

$$\hat{\lambda}_{SH_{lin}} = \operatorname*{argmax}_{\lambda_{SH_{lin}}} \Pr(\lambda_{SH+N_{lin}} | \lambda_{SH_{lin}}, \lambda_{N_{lin}}),$$

where $\lambda_{N_{lin}}$ and $\lambda_{SH+N_{lin}}$ are the model parameter of noise and the model parameter of adaptation data o in the linear-spectral domain, respectively. The above ML parameter estimation can be solved using the EM algorithm. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step (E-step), the following auxiliary function is calculated

$$\begin{aligned} Q(\hat{\lambda}_{SH_{lin}} | \lambda_{SH_{lin}}) &= E[\log \Pr(o, s, k | \hat{\lambda}_{SH_{lin}}, \lambda_{N_{lin}}) | \lambda_{SH_{lin}}, \lambda_{N_{lin}}] \\ &= \sum_{p=1}^{P} \sum_{n=1}^{W_p} \sum_{s^{(p,n)}} \sum_{k^{(p,n)}} \frac{\Pr(o^{(p,n)}, s^{(p,n)}, k^{(p,n)} | \lambda_{SH_{lin}}, \lambda_{N_{lin}})}{\Pr(o^{(p,n)} | \lambda_{SH_{lin}}, \lambda_{N_{lin}})} \\ &\quad \times \log \Pr(o^{(p,n)}, s^{(p,n)}, k^{(p,n)} | \hat{\lambda}_{SH_{lin}}, \lambda_{N_{lin}}) \end{aligned} \tag{4.3}$$

where P is the total number of phonemes, and each phoneme consists of W_p adaptation data. Next, $o^{(p,n)}$ is the n-th observation sequence for a phoneme p, and the length is $T^{(p,n)}$. Finally, $s^{(p,n)}$ and $k^{(p,n)}$ are the unobserved state sequence and the unobserved mixture component labels corresponding to the observation sequence $o^{(p,n)}$.

The joint probability for observing the sequences o, s and k can be evaluated as

$$\Pr(o, s, k | \hat{\lambda}_{SH_{lin}}, \lambda_{N_{lin}}) = \prod_{t=1}^{T} a_{s_{t-1},s_t} \omega_{s_t,k_t} \Pr_o(o_t | \hat{\lambda}_{SH_{lin}}, \lambda_{N_{lin}}), \tag{4.4}$$

where $\Pr(o_t|\hat{\lambda}_{SH_{lin}}, \lambda_{N_{lin}})$ is the probability density function of the random variable o_t.

Let the observation Gaussian PDF in the model $\lambda_{SH_{lin}}$ be the form shown in equation (4.2), and the observation Gaussian PDF in the model $\lambda_{N_{lin}}$ be a single Gaussian. Since the model $\lambda_{SH_{lin}}$ is independent of the model $\lambda_{N_{lin}}$ in the linear-spectral domain, the mean vector and the covariance matrix corresponding to mixture k in the model $\lambda_{O_{lin}}$ are derived by adding the mean vector and the covariance matrix in the model $\hat{\lambda}_{SH_{lin}}$ to the mean vector and the covariance matrix in the model $\lambda_{N_{lin}}$

$$\mu_k^{(O)} = \hat{\mu}_k^{(SH)} + \mu^{(N)} \quad \text{and} \quad \Sigma^{(O)} = \Sigma_k^{(SH)} + \Sigma^{(N)},$$

where $\mu_k^{(SH)}$ and $\Sigma_k^{(SH)}$ are the mean vector and the covariance matrix corresponding to mixture k in the model $\lambda_{SH_{lin}}$. Further, $\mu^{(N)}$ and $\Sigma^{(N)}$ are the mean vector and the covariance matrix in the model $\lambda_{N_{lin}}$. Therefore, equation (4.4) can be written as

$$\Pr(o, s, k|\hat{\lambda}_{SH_{lin}}, \lambda_{N_{lin}}) = \prod_{t=1}^{T}\{a_{s_{t-1},s_t}\omega_{s_t,k_t}N(o_t; \hat{\mu}_{k_t}^{(SH)} + \mu^{(N)}, \ \hat{\Sigma}_{k_t}^{(SH)} + \Sigma^{(N)}).$$

It is straightforward to derive that [38]

$$\begin{aligned}
&Q(\hat{\lambda}_{SH_{lin}}|\lambda_{SH_{lin}}) \\
&= \sum_{p=1}^{P}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{n=1}^{W_p}\sum_{t=1}^{T^{(p,n)}} \Pr(o^{(p,n)}, s_t^{(p,n)} = j, s_{t-1}^{(p,n)} = i|\lambda_{SH_{lin}}, \lambda_{N_{lin}}) \log a_{i,j} \\
&+\sum_{p=1}^{P}\sum_{i=j}^{J}\sum_{k=1}^{K}\sum_{n=1}^{W_p}\sum_{t=1}^{T^{(p,n)}} \Pr(o^{(p,n)}, s_t^{(p,n)} = j, k_t^{(p,n)} = k|\lambda_{SH_{lin}}, \lambda_{N_{lin}}) \log \omega_{j,k} \\
&+\sum_{p=1}^{P}\sum_{k=1}^{K}\sum_{n=1}^{W_p}\sum_{t=1}^{T^{(p,n)}} \Pr(o^{(p,n)}, k_t^{(p,n)} = k|\lambda_{SH_{lin}}, \lambda_{N_{lin}}) \log N(o_t^{(p,n)}; \hat{\mu}_k^{(SH)} + \mu^{(N)}, \hat{\Sigma}_k^{(SH)} + \Sigma^{(N)}).
\end{aligned} \tag{4.5}$$

Here, we focus on only the terms involving ($\hat{\theta}_k = \{\hat{\mu}_k{}^{(SH)}, \hat{\Sigma}_k^{(SH)}\}$). Therefore, equation (4.5) can be written as

$$\begin{aligned}
&Q_{\hat{\theta}_k}(\hat{\lambda}_{SH_{lin}}|\lambda_{SH_{lin}}) \\
&= \sum_{p=1}^{P}\sum_{k=1}^{K}\sum_{n=1}^{W_p}\sum_{t=1}^{T^{(p,n)}} \gamma_{t,k}^{(p,n)} \times \log N(o_t^{(p,n)}; \hat{\mu}_k^{(SH)} + \mu^{(N)}, \ \hat{\Sigma}_k^{(SH)} + \Sigma^{(N)}) \\
&= -\sum_{p=1}^{P}\sum_{k=1}^{K}\sum_{n=1}^{W_p}\sum_{t=1}^{T^{(p,n)}} \gamma_{t,k}^{(p,n)}
\end{aligned}$$

$$\times \left\{ \frac{1}{2}\log(2\pi)^D(\hat{\Sigma}_k^{(SH)} + \Sigma^{(N)}) + \frac{(o_t^{(p,n)} - \hat{\mu}_k^{(SH)} - \mu^{(N)})'(o_t^{(p,n)} - \hat{\mu}_k^{(SH)} - \mu^{(N)})}{2(\hat{\Sigma}_k^{(SH)} + \Sigma^{(N)})} \right\}, \tag{4.6}$$

where

$$\gamma_{t,k}^{(p,n)} = \Pr(o_t^{(p,n)}, k_t^{(p,n)} = k | \lambda_{SH_{lin}}, \lambda_{N_{lin}}),$$

and

$$\gamma_k = \sum_t \gamma_{t,k}.$$

This term, $\gamma_{t,k}$, can be calculated efficiently by using the forward-backward algorithm [9].

The M-step in the EM algorithm maximizes $Q(\hat{\lambda}_{SH_{lin}}|\lambda_{SH_{lin}})$ with respect to $\hat{\lambda}_{SH_{lin}}$

$$\hat{\lambda}_{SH_{lin}} = \underset{\hat{\lambda}_{SH_{lin}}}{\operatorname{argmax}}\, Q(\hat{\lambda}_{SH_{lin}}|\lambda_{SH_{lin}}),$$

which leads to solving $\partial Q_{\hat{\theta}}(\hat{\lambda}_{SH_{lin}}|\lambda_{SH_{lin}})/\partial\hat{\mu}^{(SH)} = 0$ and $\partial Q_{\hat{\theta}}(\hat{\lambda}_{SH_{lin}}|\lambda_{SH_{lin}})/\partial\hat{\Sigma}^{(SH)} = 0$.

Therefore, we get

$$\hat{\mu}_k^{(SH)} = m_k - \mu^{(N)} \quad \text{and} \quad \hat{\sigma}_k^{2(SH)} = v_k - \sigma^{2(N)}, \tag{4.7}$$

where

$$m_k = \sum_p \sum_n \sum_t \gamma_{t,k}^{(p,n)} o_t^{(p,n)} / \gamma_k$$

$$v_k = \sum_p \sum_n \sum_t \gamma_{t,k}^{(p,n)} (o_t^{(p,n)} - m_k)^2 / \gamma_k.$$

Equation (4.7) shows that the HMM decomposition method deals with the model parameter instead of the series of the observed speech.

4.3 Decomposition of Clean Speech HMMs and Acoustic Transfer Function HMM

This section describes the decomposition of known speech HMMs and an unknown acoustic transfer function HMM. The HMM decomposition method is applied in the cepstral domain as shown in figure 4.1. First, the model parameter $\lambda_{SH_{lin}}$, which is estimated in section 4.2, is converted to the cepstral domain. Then, the decomposition of the acoustic transfer function HMM, λ_H, is estimated using maximum-likelihood in the model domain

$$\hat{\lambda}_{H_{cep}} = \underset{\lambda_{H_{cep}}}{\operatorname{argmax}} \Pr(\lambda_{S+H_{cep}} | \lambda_{H_{cep}}, \lambda_{S_{cep}}),$$

where $\lambda_{S+H_{cep}}$ is the model parameter of the distorted speech in the cepstral domain, and λ_S is the model parameter of the clean speech in the cepstral domain. This ML parameter estimation can be solved using the EM algorithm. In [82], an estimation method based on ML is presented, where the estimation of convolutional distortion is implemented in the time domain. On the other hand, we estimate the acoustic transfer function in the model domain. The estimation in the model domain can reduce the amount of computation.

The auxiliary function is defined in a similar way to section 4.2,

$$\begin{aligned} Q(\hat{\lambda}_{H_{cep}} | \lambda_{H_{cep}}) &= E[\log \Pr(o, s, k | \hat{\lambda}_{H_{cep}}, \lambda_{S_{cep}}) | \lambda_{H_{cep}}, \lambda_{S_{cep}}] \\ &= \sum_{p=1}^{P} \sum_{n=1}^{W_p} \sum_{s^{(p,n)}} \sum_{k^{(p,n)}} \frac{\Pr(o^{(p,n)}, s^{(p,n)}, k^{(p,n)} | \lambda_{H_{cep}}, \lambda_{S_{cep}})}{\Pr(o^{(p,n)} | \lambda_{H_{cep}}, \lambda_{S_{cep}})} \\ &\quad \times \log \Pr(o^{(p,n)}, s^{(p,n)}, k^{(p,n)} | \hat{\lambda}_{H_{cep}}, \lambda_{S_{cep}}), \end{aligned} \tag{4.8}$$

where o_{cep} is represented by adding the clean speech data to the acoustic transfer function in the cepstral domain. It is impossible to measure the data o_{cep} practically. However, the HMM decomposition method can deal with the model parameter instead of the series of the data.

Since we focus on only the terms ($\hat{\theta} = \{\hat{\mu}^{(H)}, \hat{\Sigma}^{(H)}\}$), equation (4.8) can be written as

$$\begin{aligned} & Q_{\hat{\theta}}(\hat{\lambda}_{H_{cep}} | \lambda_{H_{cep}}) \\ = & \sum_{p=1}^{P} \sum_{k}^{K} \sum_{n=1}^{W_p} \sum_{t=1}^{T^{(p,n)}} \gamma_{t,k}^{(p,n)} \times \log N(o_t^{(p,n)}; \mu_k^{(S)} + \hat{\mu}^{(H)}, \quad \Sigma_k^{(S)} + \hat{\Sigma}^{(H)}) \\ = & -\sum_{p=1}^{P} \sum_{k}^{K} \sum_{n=1}^{W_p} \sum_{t=1}^{T^{(p,n)}} \gamma_{t,k}^{(p,n)} \\ & \times \left\{ \frac{1}{2} \log(2\pi)^D (\Sigma_k^{(S)} + \hat{\Sigma}^{(H)}) + \frac{(o_t^{(p,n)} - \mu_k^{(S)} - \hat{\mu}^{(H)})'(o_t^{(p,n)} - \mu_k^{(S)} - \hat{\mu}^{(H)})}{2(\Sigma_k^{(S)} + \hat{\Sigma}^{(H)})} \right\}, \end{aligned} \tag{4.9}$$

where $\lambda_{S_{cep}}$ is the tied-mixture HMM, and the total number of Gaussian mixtures is K. Next, $\mu_k^{(S)}$ and $\Sigma_k^{(S)}$ are the mean vector and the covariance matrix corresponding to mixture k, and $\lambda_{H_{cep}}$ is a single Gaussian. We assume that the above $\gamma_{t,k}^{(p,n)}$ is equal to the $\gamma_{t,k}^{(p,n)}$ in section 4.2.

To simplify the equation, the following Δ is defined

$$\hat{\mu}^{(H)} = \mu^{(H)} + \Delta\hat{\mu}^{(H)} \quad \text{and} \quad \hat{\Sigma}^{(H)} = \Sigma^{(H)} + \Delta\hat{\Sigma}^{(H)}.$$

The M-step in the EM algorithm maximizes $Q(\hat{\lambda}_{H_{cep}}|\lambda_{H_{cep}})$ with respect to $\hat{\lambda}_{H_{cep}}$

$$\hat{\lambda}_{H_{cep}} = \underset{\hat{\lambda}_{H_{cep}}}{\operatorname{argmax}}\, Q(\hat{\lambda}_{H_{cep}}|\lambda_{H_{cep}}),$$

which leads to solving $\partial Q_{\hat{\theta}}(\hat{\lambda}_{H_{cep}}|\lambda_{H_{cep}})/\partial\Delta\hat{\mu}^{(H)} = 0$ and $\partial Q_{\hat{\theta}}(\hat{\lambda}_{H_{cep}}|\lambda_{H_{cep}})/\partial\Delta\hat{\Sigma}^{(H)} = 0$. Therefore,

$$\begin{aligned} &\frac{\partial Q_{\hat{\theta}}(\hat{\lambda}_{H_{cep}}|\lambda_{H_{cep}})}{\partial\Delta\hat{\mu}^{(H)}} \\ &= \sum_{p=1}^{P}\sum_{n=1}^{W_p}\sum_{t=1}^{T^{(p,n)}}\sum_{k}^{K}\gamma_{t,k}^{(p,n)}\frac{o_t^{(p,n)} - \mu_k^{(S)} - \mu^{(H)} - \Delta\hat{\mu}^{(H)}}{\Sigma_k^{(S)} + \Sigma^{(H)} + \Delta\hat{\Sigma}^{(H)}} = 0. \end{aligned} \tag{4.10}$$

Since the model parameter $\hat{\lambda}_{SH_{lin}}$ in the linear-spectral domain is calculated by equation (4.7), the model parameter of $o_t^{(p,n)}$, $\hat{\lambda}_{S+H_{cep}}$, is given by

$$\hat{\lambda}_{S+H_{cep}} = \mathrm{Cos}^{-1}\{\ \mathrm{Log}(\ \hat{\lambda}_{SH_{lin}}\ )\ \}.$$

On the other hand, the mean vector and the covariance matrix in the model $\lambda_{S+H_{cep}}$ can also be represented using the term $o_t^{(p,n)}$ as follows

$$\hat{\mu}_k^{(S+H)} = \sum_p\sum_n\sum_t \gamma_{t,k}^{(p,n)} o_t^{(p,n)}/\gamma_k$$

$$\hat{\Sigma}_k^{(S+H)} = \sum_p\sum_n\sum_t \gamma_{t,k}^{(p,n)}(o_t^{(p,n)} - \hat{\mu}_k^{(S+H)})^2/\gamma_k.$$

Then, we get

$$\sum_p\sum_n\sum_t \gamma_{t,k}^{(p,n)} o_t^{(p,n)} = \gamma_k\hat{\mu}_k^{(S+H)}$$

$$\sum_p\sum_n\sum_t \gamma_{t,k}^{(p,n)}(o_t^{(p,n)} - \hat{\mu}_k^{(S+H)})^2 = \gamma_k\hat{\Sigma}_k^{(S+H)}.$$

Therefore, according to equation (4.10), on the assumption that the variance is fixed, the re-estimation formula of $\Delta\hat{\mu}^{(H)}$ is given by

$$\begin{aligned}
\Delta\hat{\mu}^{(H)} &= \frac{\displaystyle\sum_{k=1}^{K} \frac{\sum_p \sum_n \sum_t \gamma_{t,k}^{(p,n)} o_t^{(p,n)} - \gamma_k(\mu_k^{(S)} + \mu^{(H)})}{\Sigma_k^{(S)} + \Sigma^{(H)}}}{\displaystyle\sum_{k=1}^{K} \frac{\gamma_k}{\Sigma_k^{(S)} + \Sigma^{(H)}}} \\
&= \frac{\displaystyle\sum_{k=1}^{K} \gamma_k \frac{\hat{\mu}_k^{(S+H)} - \mu_k^{(S)} - \mu^{(H)}}{\Sigma_k^{(S)} + \Sigma^{(H)}}}{\displaystyle\sum_{k=1}^{K} \frac{\gamma_k}{\Sigma_k^{(S)} + \Sigma^{(H)}}}. \qquad (4.11)
\end{aligned}$$

Equation (4.11) shows that the HMM decomposition method deals with the model parameter $\lambda_{S+H_{cep}}$ instead of the data o_t.

Then, taking the derivative of equation (4.10) with respect to $\Delta\hat{\Sigma}^{(H)}$, and setting to zero, we get

$$\sum_{k}^{K} \gamma_k \frac{\Sigma_k^{(S)} + \Sigma^{(H)} + \Delta\Sigma^{(H)} - \phi_k}{(\Sigma_k^{(S)} + \Sigma^{(H)} + \Delta\Sigma^{(H)})^2} = 0,$$

where

$$\phi_k = \Sigma_k^{(S+H)} + \mu_k^{2(S+H)} + (\mu_k^{(S)} + \hat{\mu}^{(H)})(\mu_k^{(S)} + \hat{\mu}^{(H)} - 2\mu_k^{(S+H)}).$$

There are some approaches to the problem of estimating the covariance matrix $\Sigma^{(H)}$ [82]. In this work, we use a Taylor expansion. Now, define a function F as follows

$$F(\Delta\Sigma^{(H)}) = \frac{\Sigma_k^{(S)} + \Sigma^{(H)} + \Delta\Sigma^{(H)} - \phi_k}{(\Sigma_k^{(S)} + \Sigma^{(H)} + \Delta\Sigma^{(H)})^2}.$$

If F is expanded in a Taylor series through terms of first order, we obtain

$$\begin{aligned}
F(\Delta\Sigma^{(H)}) &\simeq F(0) + \left.\frac{\partial F(\Delta\Sigma^{(H)})}{\partial \Delta\Sigma^{(H)}}\right|_{\Delta\Sigma^{(H)}=0} \times \Delta\Sigma^{(H)} \\
&= \frac{\Sigma_k^{(S)} + \Sigma^{(H)} - \phi_k}{(\Sigma_k^{(S)} + \Sigma^{(H)})^2} - \frac{\Sigma_k^{(S)} + \Sigma^{(H)} - 2\phi_k}{(\Sigma_k^{(S)} + \Sigma^{(H)})^3}\Delta\Sigma^{(H)},
\end{aligned}$$

where $\Delta\Sigma^{(H)}$ converges on 0 by using the EM algorithm. Therefore, the re-estimation formula of $\Delta\hat{\Sigma}^{(H)}$ is given by

$$\Delta\hat{\Sigma}^{(H)} \simeq \frac{\displaystyle\sum_{k}^{K} \gamma_k \left\{ \frac{1}{\Sigma_k^{(S)} + \Sigma^{(H)}} - \frac{\phi_k}{(\Sigma_k^{(S)} + \Sigma^{(H)})^2} \right\}}{\displaystyle\sum_{k}^{K} \gamma_k \left\{ \frac{1}{(\Sigma_k^{(S)} + \Sigma^{(H)})^2} - \frac{2\phi_k}{(\Sigma_k^{(S)} + \Sigma^{(H)})^3} \right\}}. \tag{4.12}$$

Equation (4.12) shows that the HMM decomposition method deals with the model parameter $\lambda_{S+H_{cep}}$ instead of the data o_t.

Chapter 5

Distant-Talking Speech Recognition

This chapter describes the performance of the HMM composition and decomposition methods on real distant-talking speech [93]. We measured the distant-talking speech from four positions. The sound signal is captured by using a single-directional microphone. The HMM decomposition method enables the estimation of parameters of an acoustic transfer function HMM using adaptation speech from an unknown user's location. This chapter also describes the performance of the HMM composition and decomposition methods on speech recognition of a distant moving talker. Speech of the distant moving talker is recognized by using an ergodic-HMM of acoustic transfer

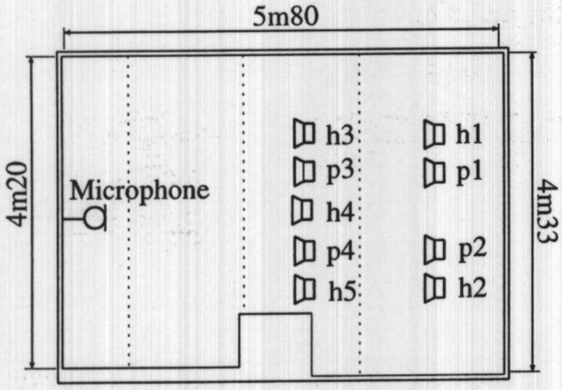

Figure 5.1: Experimental room environment

functions. The experimental results indicate the effectiveness of the HMM composition and decomposition methods. First, section 5.2 describes the performance of the HMM composition method. Section 5.3 describes the performance of the HMM composition and decomposition methods. For speech recognition of the distant moving talker, the performance is described in section 5.4.

5.1 Experimental Conditions

Recognition experiments are conducted to evaluate the effectiveness of the HMM composition and decomposition methods. Figure 5.1 shows a top view of the experimental room. The sound signal is captured by using a single-directional microphone (SONY

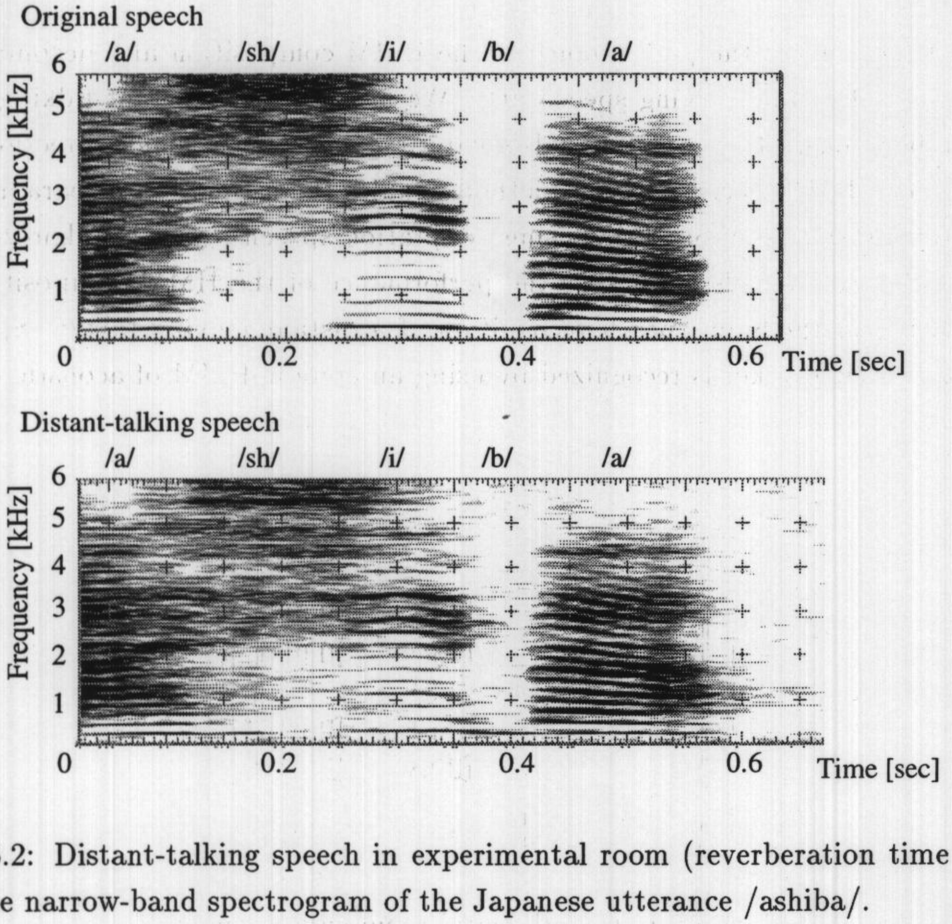

Figure 5.2: Distant-talking speech in experimental room (reverberation time = 0.18 sec) : the narrow-band spectrogram of the Japanese utterance /ashiba/.

C-355). The speech data used for evaluation is the Set-A of the ATR Japanese speech database and the ASJ (Acoustical Society of Japan) continuous speech database.

Real distant-talking speech

To evaluate on real speech, we measured distant-talking speech from four sound source positions, p1, ..., p4. The distant-talking speech is contaminated by computer noise, air conditioner noise and ventilating fan noise, where SNR (Signal to Noise Ratio) is 16.7 dB on average. The SNR is calculated as follows

$$SNR = 10\log_{10}\frac{E[o(t)^2]}{E[n(t)^2]},$$

where $o(t)$ and $n(t)$ denote the observed speech and the noise at time t, respectively. One male speaker is used as the testing speaker in speaker-dependent (SD) experiments. Two male speakers and one female speaker are used as the testing speakers in speaker-independent (SI) experiments. Each testing speaker utters $1 \sim 50$ words ($\times$ 3) as adaptation data which are not used in the training. The related information of the adaptation data used in the following word-recognition experiments are listed in Appendix B. For testing, 500 words which are different from those words in the training are used. The related information of the testing data used in the following word-recognition experiments are listed in Appendix C. Figure 5.2 shows the narrow-band spectrogram for original (clean) speech and distant-talking speech in the experimental room. In section 5.3.2, word-recognition experiments are carried out on the real distant-talking speech.

Simulated distant-talking speech

To evaluate on simulated speech, we measured nine transfer functions corresponding to nine sound source positions, h1, ..., h5 and p1, ..., p4 by using the method reported in [88]. Distant-talking speech is simulated by linear convolution of clean speech and the measured impulse responses. The length of the original impulse response was about 180 msec. The former five positions, h1, ..., h5 are used for the model composition and the latter four positions, p1, ..., p4 are used for the recognition tests. Figure 5.3 shows the measured impulse responses corresponding to four sound source positions, p1, ..., p4. As the distance between the microphone and the sound source position is longer, the delay time is longer. Figure 5.4 shows the cepstral coefficients of the acoustic transfer functions from several training positions. The differences shown will cause degradation of speech recognition. In section 5.2 and 5.3.1, word-recognition experiments are carried out on the simulated distant-talking speech.

The recognition algorithm is based on 256 tied-mixture diagonal covariance HMMs. Each HMM has three states and three self-loops. The models of 54 context-independent phonemes are trained using 2620 words in the ATR database for speaker-dependent HMMs. The other 500 words in the same database are used for testing. The speaker-

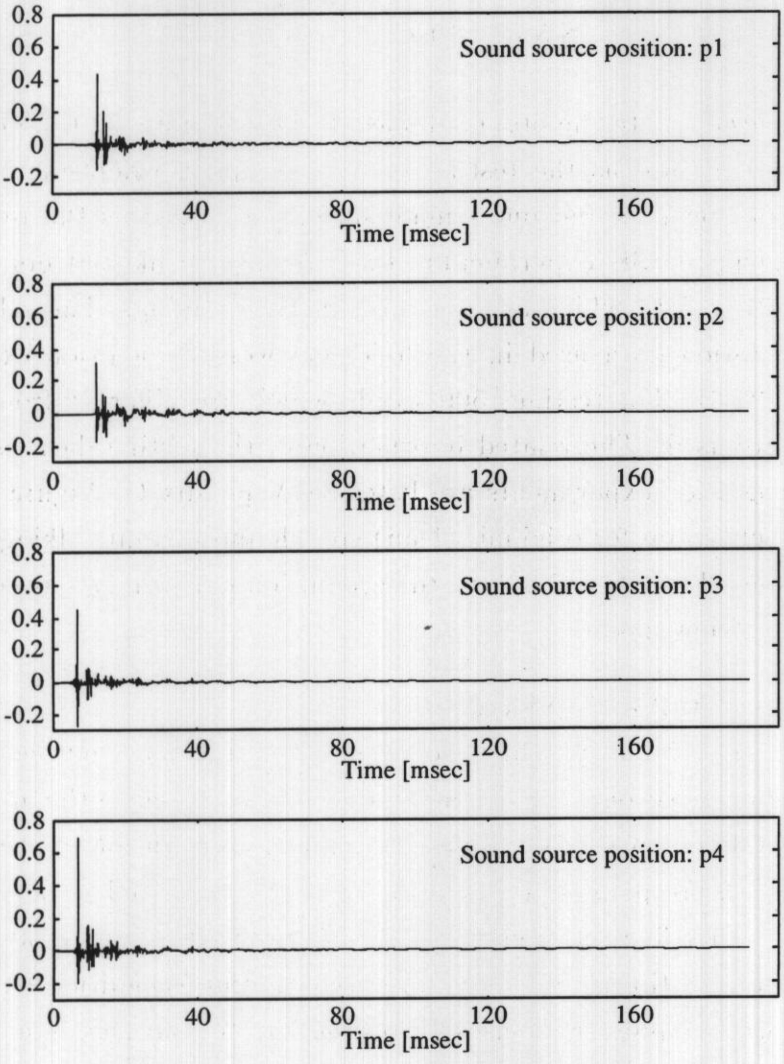

Figure 5.3: Measured impulse responses

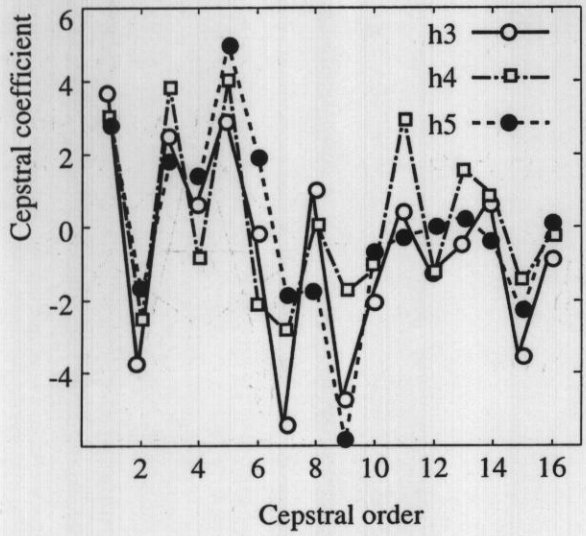

Figure 5.4: Cepstral coefficients for different sound source positions

independent HMMs are trained using about 9600 sentences which are uttered by 64 speakers of the ASJ database.

The speech signal is sampled at 12 kHz and windowed with a 32 msec Hamming window every 8 msec. Then FFT is used to calculate 16-order MFCCs and power. In recognition, the power term is not used, because it is only necessary to adjust the SNR in the HMM composition. The analysis condition is listed in table 5.1.

In section 5.2, we assigned one state for the noise HMM and five states for the acoustic transfer function HMM, and a single Gaussian PDF is used per state. Figure

Table 5.1: Analysis conditions

Sampling freq.	12 kHz
Frame shift	8 msec
Window length	32 msec
Window	Hamming
Pre-emphasis	0.97
Feature parameter	MFCC (order 16)

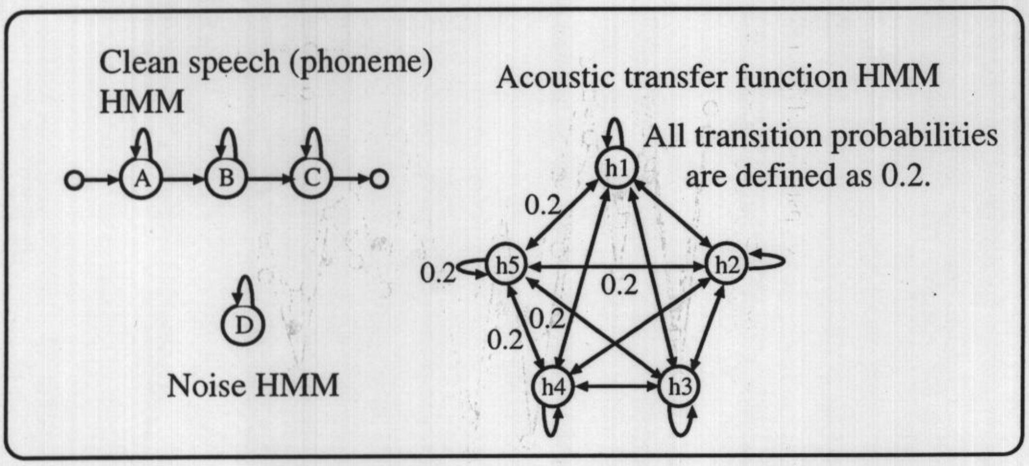

Figure 5.5: Structure of a clean speech HMM, a noise HMM and an acoustic transfer function HMM in experiments

5.5 shows the acoustic transfer function HMM. Each state directly corresponds to one of the training positions, h1, ..., h5. All transitions among states are permitted, and their probabilities are defined as 0.2. In section 5.3, we assigned one state for the noise HMM and one state for the acoustic transfer function HMM.

5.2 Evaluation of HMM Composition

5.2.1 Results for Noisy and Acoustically-Distorted Speech

This section describes performance of the HMM composition method. The points to be investigated are

- performance for the acoustic transfer function HMM,
- improvement of recognition rate for noisy and acoustically-distorted speech,
  - evaluation of speaker-dependent (SD) and speaker-independent (SI) speech recognition performance,

and

- performance for an unknown position of the sound source.

This section also shows the influence of the reverberation time, based on the results of the 500-word recognition experiments, where impulse responses of 100 msec and 32 msec are artificially made from the impulse response of 180 msec. They are made by multiplying the original 180 msec by an exponential function as follows

$$y' = y \cdot e^{-a}, \quad a > 0.$$

Several impulse responses are adjusted by a constant a. The obtained impulse responses are shown in figure 5.6.

Here, the acoustic transfer function in the cepstral domain is obtained by subtracting the cepstrum coefficients of original speech from those of convoluted speech. The mean value of the i-th cepstral coefficient, μ_i, is given by

$$\mu_i = \frac{1}{g}\sum_{j=1}^{g}(s_i'^{(j)} - s_i^{(j)}) = \frac{1}{g}\sum_{j=1}^{g} c_i^{(j)}, \tag{5.1}$$

where g is the total number of frames of the training data, and $s_i'^{(j)}$ is the i-th cepstral coefficient at frame j for the distorted speech which is made by the linear convolution. The clean speech, $s_i^{(j)}$, is the i-th cepstral coefficient at frame j. The covariance σ_{ii} is given by

$$\sigma_{ii} = \frac{1}{g}\sum_{j=1}^{g}(c_i^{(j)} - \mu_i)(c_i^{(j)} - \mu_i), \tag{5.2}$$

$$c_i^{(j)} = s_i'^{(j)} - s_i^{(j)},$$

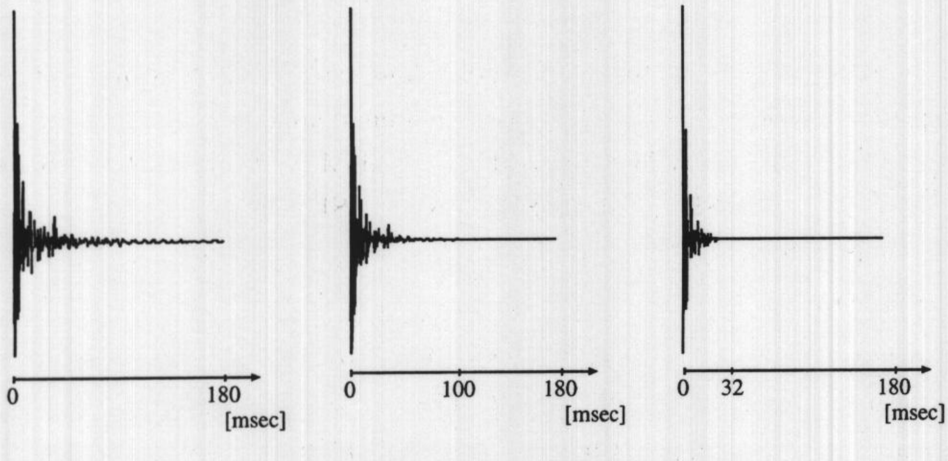

Figure 5.6: Impulse responses (180 msec, 100 msec, and 32 msec)

where we assume that the cepstral coefficients are uncorrelated. In this experiment, 500 words are used for training of the acoustic transfer function HMM. For the speaker-independent model, 125 words are used for each testing speaker (125 words $\times$ 4 testing speakers = 500 words). The structure of the acoustic transfer function HMM is shown in figure 5.5.

Table 5.2 shows the recognition rates for the distorted speech with the speaker-dependent model. The word-recognition rate is shown for the positions, h1, ..., h5, on average. Without any compensation, the word-recognition rate with the clean speech HMMs (HMM-S) is 78.5% in the case of 180 msec impulse response. When only the mean vector are adapted (indicated as HMM-SH(μ)), the word-recognition rate is improved from 78.5% to 87.2%, in the case of the 180 msec impulse response.

Table 5.2: Word-recognition rates [%] for distorted speech with speaker-dependent models

Model	HMM-S	HMM-SH(μ)	HMM-SH(μ, Σ)
Acoustic compensation	×	○	○
180 msec	78.5	87.2	84.0
100 msec	88.0	94.0	92.4
32 msec	88.6	96.2	95.6
0 msec	96.6	-	-

Table 5.3: Word-recognition rates [%] for noisy and acoustically-distorted speech with speaker-dependent models (SD) and speaker-independent models (SI)

Models	HMM-S		HMM-SN		HMM-SHN(μ)		HMM-SHN(μ, Σ)	
	SD	SI	SD	SI	SD	SI	SD	SI
Noise compensation	×	×	○	○	○	○	○	○
Acoustic compensation	×	×	×	×	○	○	○	○
180 msec	4.8	18.7	59.5	53.5	67.2	57.2	55.2	45.4
100 msec	9.9	22.0	76.2	65.7	83.6	66.7	79.7	59.2
32 msec	14.4	21.9	76.5	65.4	87.0	68.4	86.5	65.5

The improvement is also obtained in other cases. In the case of the 100 msec impulse response, the word-recognition rate is improved from 88.0% to 94.0%. In the case of the 32 msec impulse response, the word-recognition rate is improved from 88.6% to 96.2%. In this table, "0 msec" means the clean speech. The word-recognition rate for the clean speech with the clean model is 96.6%.

This table also shows that as the impulse response is longer, the effectiveness of the composed HMM is decreased. The effectiveness of the covariance matrix, Σ, of the acoustic transfer function HMM is not significant, because the variation might be larger than expected and distributed depending on preceding speech characteristics. That result is shown by the label of HMM-SH(μ,Σ), where the mean vector and the covariance matrix are both adapted.

Table 5.3 shows the recognition rates for noisy and acoustically-distorted speech. The recognition rate is shown for the positions, h1, ..., h5, on average. The noise data is collected in a computer room and added to the acoustically-distorted data as the SNR is 15 dB. The recognition rate with the HMM-SN, composed of the HMM-S and the noise HMM, is improved from 4.8% to 59.5% for the SD model. The proposed HMM-SHN(μ), composed of the HMM-SN and the acoustic transfer function HMM, increases the recognition rate by 67.2% for the SD model. On the other hand, the recognition rate with the matched condition is 89.7%, where the phoneme HMM is trained using 2620 words which are simulated by the linear convolution of the speech corpus and the measured transfer function (180 msec impulse response), followed by the addition of the noise data. Comparing this result with that of the composed HMM, HMM-SHN(μ), it shows a difference in performance of 22.5%.

5.2.2 Results for Unknown Positions

The performance of the proposed method is evaluated on unknown positions of the testing speaker. The five positions, h1, ..., h5, are used for the model composition. The other four positions, p1, ..., p4, are used for the recognition tests. Figure 5.7 shows the cepstral distance between the known-training positions and the unknown-testing positions. The cepstral distance, d, is given by

$$d = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\{\mathrm{h}^{(\mathrm{train})}(j) - \mathrm{p}^{(\mathrm{test})}(j)\}^2},$$

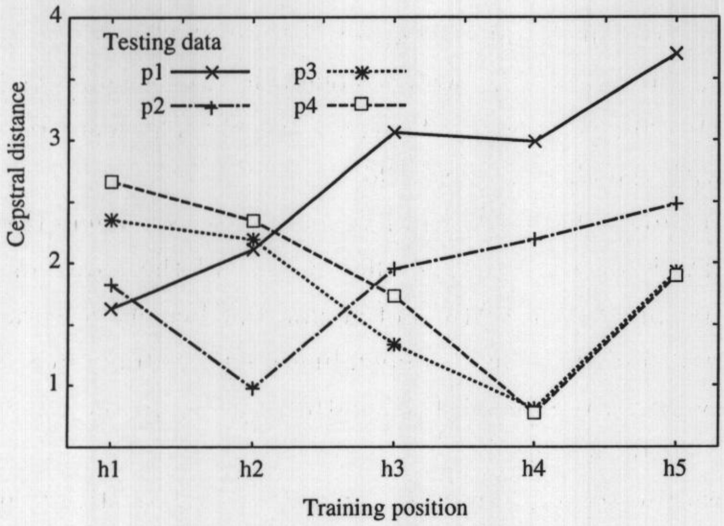

Figure 5.7: Cepstral distance between known-training positions and unknown-testing positions

where J is the cepstral order. Next, $h^{(\text{train})}(j)$ is the j-th cepstral coefficient for the known-training position, and $p^{(\text{test})}(j)$ is the j-th cepstral coefficient for the unknown-testing position. For example, the training position, h2, is the closest position for the testing position, p2, in the cepstral domain.

Table 5.4 shows word-recognition rates with the speaker-dependent model for the known-training positions and the unknown-testing positions on average. The word-recognition rates for the known-training positions are the same rates in table 5.2. The recognition rates with the HMM-SH(μ) for the known-training positions and the unknown-testing positions are 87.2% and 86.2%, respectively. It is confirmed that the degradation between the training sound source positions and the testing sound source positions is relatively small for all composed HMMs. This is because the cepstral distance between an testing position and the closest training position is not so far as shown in figure 5.7.

Figure 5.8 shows the recognition rates for an unknown testing position, p2, by using the acoustic transfer function of each training position, h1, ..., h5, and also shows the cepstral distance between the testing position, p2, and each training position, h1, ..., h5. This figure indicates the closest position results in the best performance, 86.2%. As the cepstral distance is longer, the recognition rate will degrade. In the case of the

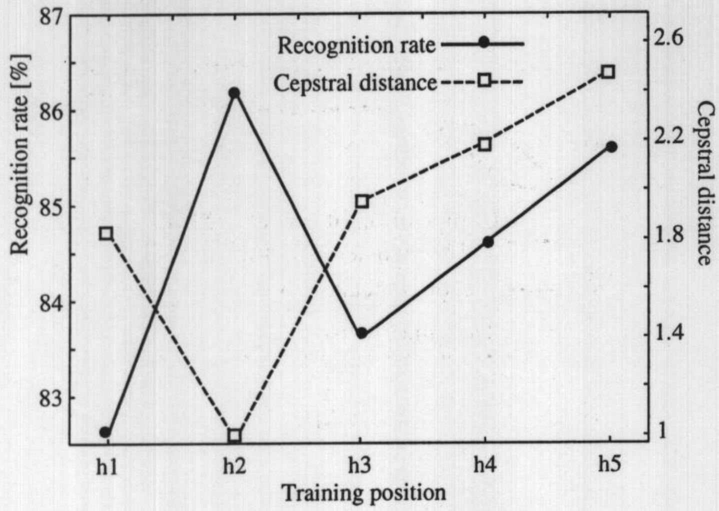

Figure 5.8: Word-recognition rates and cepstral distance for an unknown position (p2)

training position, h1, the recognition rate is decreased to 82.6%.

This figure also shows performance difference between an acoustic transfer function HMM of the closest position and an ergodic HMM of the acoustic transfer function (shown in table 5.4) is quite small. Because the decoded path by using the ergodic HMM finds the optimal combination of the acoustic transfer function HMM and the clean speech HMM.

Table 5.4: Word-recognition rates [%] for known/unknown positions

Model	HMM-S		HMM-SH(μ)		HMM-SH(μ, Σ)	
Acoustic compensation	×	×	○	○	○	○
Distorted speech	known	unknown	known	unknown	known	unknown
	78.5	77.8	87.2	86.2	84.0	83.7

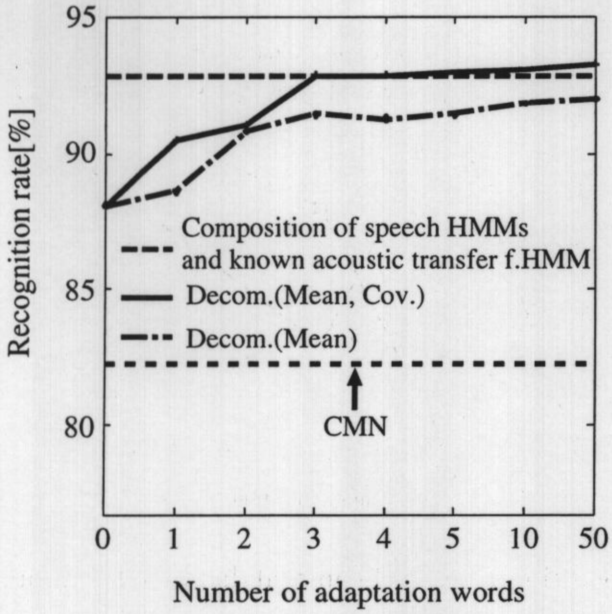

Figure 5.9: Word-recognition rates in reverberant environment

5.3 Evaluation of HMM Decomposition

This section describes the performance of the HMM composition and decomposition methods on distant-talking speech. The model parameters of the acoustic transfer function are estimated by maximizing likelihood of adaptation data uttered from an unknown position.

5.3.1 Results in Simulated Environment

In this section, the speech corpora are processed by linear convolution of clean speech and an impulse response which is measured in an anechoic room to compensate for the influence of the loudspeaker's characteristics. The loudspeaker used in this work is JBL Control 5 Plus. Next, the speech is processed by linear convolution of impulse responses which are measured in figure 5.1.

Figure 5.9 shows the SD experiment results averaged over p1, ..., p4 by using different amounts of adaptation data. The recognition rate with initial HMMs (clean speech HMMs) is 88.1%. By using the HMM composition and decomposition methods, the performance, "Decom.(Mean)", is improved to 91.8% with 10 adaptation words. Finally, applying the HMM decomposition method to both the mean vector and the covariance matrix, "Decom.(Mean,Cov.)" increases the performance by about 1%. These results show the effectiveness of the estimated covariance matrix of the acoustic transfer function.

Figure 5.9 also shows the recognition rate in the case of the known acoustic transfer function, where the model parameters of the acoustic transfer function are estimated according to equation (5.1) and (5.2). The recognition rate in the case of the known acoustic transfer function is 92.8%. These results show that there is essentially no difference between the known acoustic transfer function and the estimated acoustic transfer function.

In the CMN-based testing case, the phoneme HMMs are trained using the CMN-processed clean speech data. By subtracting each cepstral mean value from each testing data, the recognition rate is 80.7%. The experimental results clearly show that the simple CMN technique does not work well. In this simulated experiment, the silence part of the samples is cut off. The length of one word is about 0.6 sec on average. This figure shows, in the case of the 180 msec impulse response, it is difficult to calculate the cepstral mean on a short time.

In the case of the matched condition, the SD recognition rate is 96.6%, where

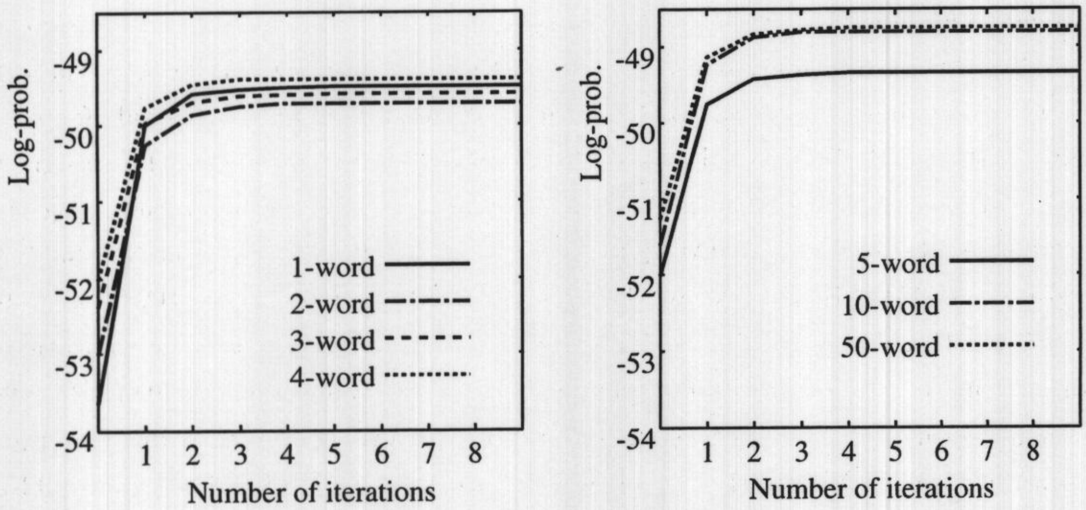

Figure 5.10: Convergence of HMM decomposition training

Table 5.5: Word-recognition rates with 10 adaptation words at various SNRs

Models	HMM-S	HMM-SN	Decom.(Mean)	Matched HMMs
Noise compensation	×	○	○	-
Acoustic compensation	×	×	○	-
0 dB	7.8%	81.6%	82.5%	88.2%
5 dB	35.6%	86.8%	86.8%	
10 dB	59.8%	90.4%	90.8%	
15 dB	76.4%	90.4%	92.5%	
20 dB	82.8%	90.6%	92.7%	96.4%

each phoneme HMM is trained using acoustically-distorted speech. Comparing this result with that of the composed HMM, Decom.(Mean,Cov.), it shows a difference in performance of 3.3%.

Figure 5.10 shows the convergence property of the HMM decomposition method in the SD model. The number of the adaptation words is one, two, three, four, five, ten and fifty in the place, p1. The label of 1-word is /ikioi/, and the label of 2-word is /omoshiroi/, and so on. The other information of the adaptation words are listed in Appendix B. In this figure, the log-likelihood of each adaptation word versus the number of iterations in EM algorithm is plotted. The result shows that three or four iterations seem enough.

Finally, the recognition rates at various SNRs are shown in table 5.5, where the computer-noise signal is added to the acoustically-distorted speech signal, for p1, at various SNRs, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB. Table 5.5 shows the recognition rates with 10 adaptation words at each SNR. In the case of SNR 0 dB, the recognition rate with the clean speech HMMs (HMM-S) is 7.8%. The recognition rate with the composition HMMs (HMM-SN) of the speech HMMs and the noise HMM is 81.6%. Applying the HMM composition and decomposition methods to noisy and acoustically-distorted speech, "Decom.(Mean)", increases the performance by about 1.0%, where the mean vector of the acoustic transfer function HMM is estimated and composed. Also, in the case of SNR 20 dB, the recognition rate is improved from 82.8% to 92.7%. The recognition rate with the matched HMMs is 88.2% at SNR 0 dB, and 96.4% at SNR 20 dB. In comparison with the performance of the matched HMMs, the difference is 5.7% at SNR 0 dB, and 3.7% at SNR 20 dB. The performance at SNR 0 dB is slightly lower than the performance at SNR 20 dB.

5.3.2 Results in Real Environment

Recognition results for real distant-talking speech are shown in figure 5.11 and figure 5.12. The recognition rate with initial HMMs (clean speech HMMs) is 77.2% for the SD model, and 54.4% for the SI model. The recognition rate with composed HMMs of clean speech HMMs and noise HMM is 87.5% for the SD model, and 61.5% for the SI model.

By applying the HMM decomposition method to only the mean vector, "Decom.(Mean)", the recognition rate with 10 adaptation words is improved to 90.5% for the SD model, and 64.9% for the SI model. Then, applying the HMM decomposition method to both the mean vector and the covariance matrix, "Decom.(Mean,Cov.)", increases the performance to 91.2% for the SD model, and 66.2% for the SI model.

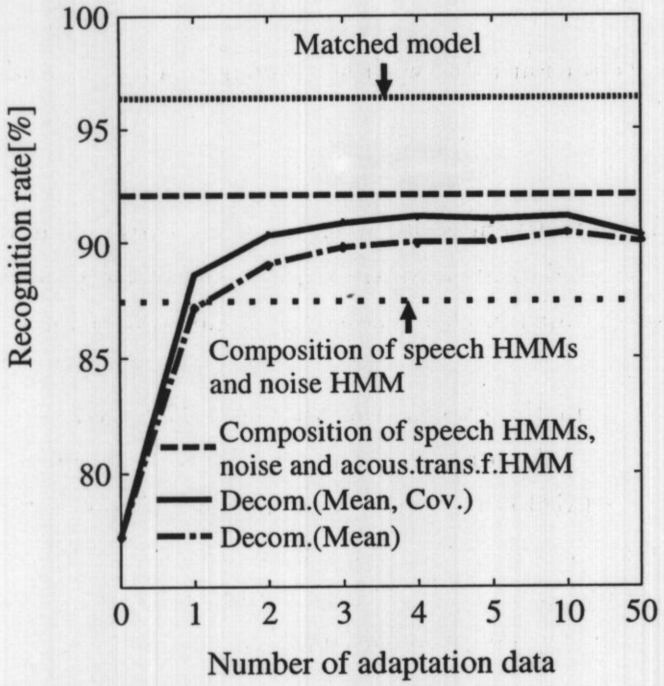

Figure 5.11: Word-recognition rates with speaker-dependent models in real environment

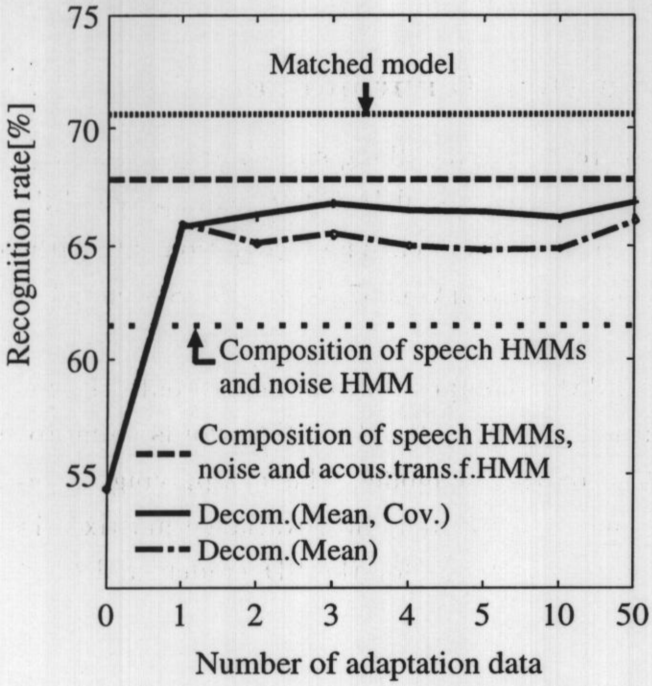

Figure 5.12: Word-recognition rates with speaker-independent models in real environment

These results show the effectiveness of the estimated covariance matrix of the acoustic transfer function. Some adaptation data cause a small decrease in recognition rate. This is because there is a mismatch between the some adaptation data and the testing data.

The recognition rate in the case of the known acoustic transfer function is 92.2% for the SD model, and 67.8% for the SI model. These recognition results show that the performance of the HMM composition and decomposition methods is close to that of the case of the known acoustic transfer function as the number of adaptation data increases. Finally, in the case of the matched condition, the SD and the SI recognition rates are 96.4% and 70.7%, where each phoneme HMM is trained using simulated distant-talking speech.

5.4 Evaluation on Speech Recognition of Distant Moving Talker

This section describes the performance of the HMM composition and decomposition methods on speech recognition of a distant moving talker. Speech of the distant moving talker is recognized by using an ergodic-HMM of acoustic transfer functions. Each state of the ergodic-HMM of acoustic transfer functions corresponds to a position of sound sources, where all transitions among states are permitted. Therefore, the proposed ergodic-HMM of acoustic transfer functions is able to trace the position of sound sources.

5.4.1 Experimental Conditions

Recognition experiments are conducted to evaluate the effectiveness of an ergodic-HMM of acoustic transfer functions on speech recognition of the distant moving talker. Figure 5.13 shows the recording condition of speech of the distant moving talker. One male is walking from "Starting position" shown in figure 5.13. He speaks 31 sentences while moving. One sentence is used for adaptation. Distant-talking speech without moving is also recorded. The position of sound sources is g1, g2 and g3 shown in figure 5.13. Figure 5.14 shows the estimated cepstral coefficients of acoustic transfer functions

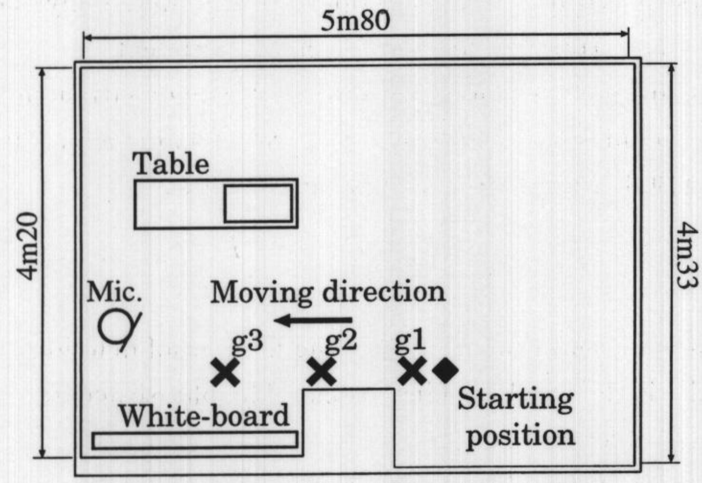

Figure 5.13: Recording condition of speech of a distant moving talker

Figure 5.14: Estimated cepstral coefficients of acoustic transfer functions

at g1, g2 and g3.

We chose 55 context-independent phonemes as clean speech units. Each phoneme is modeled by a single left-to-right 3-state tied-mixture HMM with 3 self-transition loops and without state skipping. Sixteen mel-frequency cepstral coefficients (MFCC) with their first order differentials (ΔMFCC), and first order differentials of normalized logarithmic energy (Δpower) are calculated as an observation vector of each frame. There are 256 Gaussian mixture components with diagonal covariance matrices shared by all of the models for MFCC and ΔMFCC, respectively. There are 128 Gaussian mixture components shared by all of the models for Δpower. Only the mean vector is estimated for an acoustic transfer function in this experiment.

The phrase recognition experiment is carried out using continuous sentence speech, where the sentence includes $6 \sim 7$ phrases on average. This task is 306 phrases with a phrase perplexity of 306. Phrase accuracy is calculated by

$$\text{Accuracy} = \frac{N - D - S - I}{N} \times 100,$$

where N is the total number of phrases, D is the number of deletions, S is the number of substitutions and I is the number of insertions. The phrase accuracy for close-talking speech of the testing talker is 90.4%.

The points to be investigated are the performance of

- parallel models of acoustic transfer functions;

Composed HMMs for each acoustic transfer function are separately set. Likelihood scores for each composed HMMs are calculated, and then composed HMMs having maximum likelihood are selected.

and

- ergodic models of acoustic transfer functions.

5.4.2 Results for Speech of Distant Moving Talker

Table 5.6 shows the average phrase accuracy [%] for distant-talking speech without moving. The phrase accuracy with the clean speech HMMs is 69.5%. Next, we compose the clean speech HMMs and each acoustic transfer function HMM, g1, g2 and g3. The performance of the parallel models, where composed HMMs having maximum likelihood are selected, is 76.5% on average. The performance of the composed ergodic-HMMs (shown in figure 5.15) is 75.5% on average. Comparing this result with that of the parallel model, a difference in performance of 1.0% is shown. This is because all transition probabilities of acoustic transfer functions in the ergodic-HMM are set equally, and a wrong path might be chosen.

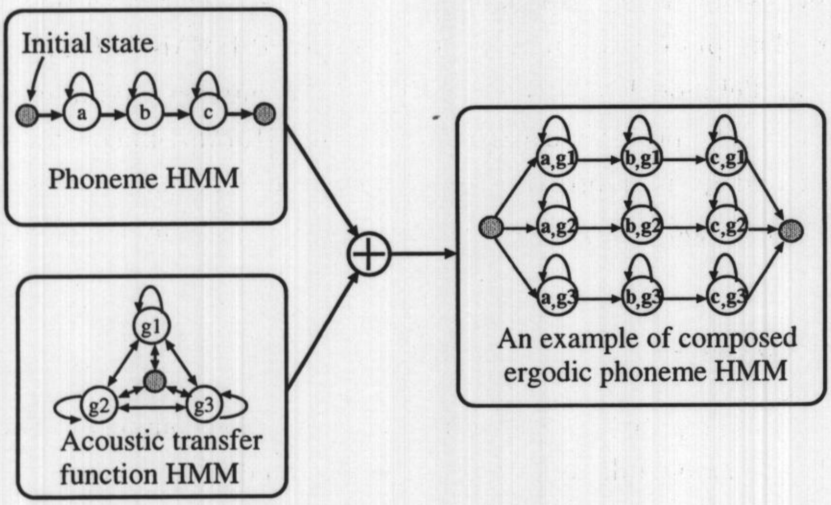

Figure 5.15: An example of a composed HMM in experiments of a distant moving talker

Table 5.6: Phrase accuracy [%] for distant-talking speech without moving

Models	g1	g2	g3	Average
Clean speech HMMs	58.1	72.6	77.7	69.5
Parallel models	67.0	76.3	86.1	76.5
Ergodic-HMMs (g1, g2, g3)	66.1	73.5	87.0	75.5

Table 5.7: Phrase accuracy [%] for speech of a distant moving talker

Models	Phrase accuracy
Clean speech HMMs	63.3
Parallel models	76.7
Ergodic-HMMs (g1, g2, g3)	82.3
Ergodic-HMMs (g1, g2)	78.6
Ergodic-HMMs (g1, g3)	76.3
Ergodic-HMMs (g2, g3)	80.0

Table 5.7 shows the average phrase accuracy [%] for speech recognition of the distant moving talker. The phrase accuracy with clean speech HMMs is 63.3%. The performance of the parallel models, where composed HMMs having maximum likelihood are selected, is 76.7%. The performance with the ergodic-HMMs of acoustic transfer functions at g1, g2 and g3 is improved to 82.3%. These experimental results show the effectiveness of the ergodic-HMMs for speech recognition of the distant moving talker.

5.5 Summary

This chapter has investigated the performance of the HMM composition and decomposition methods on distant-talking speech, where the loudspeaker is set at a distance of about 2.5 m. The HMM decomposition method enables to estimate the parameters of the acoustic transfer function HMM not from one measured impulse responses but by using adaptation speech from an unknown user position. The results are summarized as follows:

- Evaluation on simulated distant-talking speech at various SNRs,
  The proposed method improves the recognition rates for noisy and acoustically-distorted speech at various SNRs, where the computer-noise signal is added to the acoustically-distorted speech signal. In table 5.5, the recognition rates with 10 adaptation words at each SNR are shown for the speaker dependent model. In the case of SNR 0 dB, the recognition rate with the clean speech HMMs is 7.8%. The recognition rate with the composition HMMs of the speech HMMs and the noise HMM is 81.6%. Applying the HMM composition and decomposition methods to noisy and acoustically-distorted speech, increases the performance by about 1.0%, where the mean vector of the acoustic transfer function HMM is estimated and composed.

- Evaluation on real distant-talking speech,
  The proposed method improves the recognition rates for the SD model and the SI model. In figure 5.11 and figure 5.12, recognition results are shown for the speaker dependent (SD) and the speaker independent (SI) model. The recognition rate with clean speech HMMs is 77.2% for the SD model, and 54.4% for the SI model. The recognition rate with composed HMMs of clean speech HMMs and noise HMM is 87.5% for the SD model, and 61.5% for the SI model. Applying the HMM composition and decomposition methods to the real distant-talking speech, the recognition rate with 10 adaptation words is improved to 90.5% for the SD model, and 64.9% for the SI model, where the mean vector of the acoustic transfer function HMM is estimated and composed.

Then,

- applying the HMM decomposition method to both the mean vector and the covariance matrix of the acoustic transfer function HMM, increases the performance to 91.2% for the SD model, and 66.2% for the SI model.

The experimental results show that the proposed method can improve the distant-talking speech recognition performance in comparison with that of using a speech recognizer composed of the clean speech HMMs and the noise HMM (from 87.5% to 91.2% for the SD model, from 61.5% to 66.2% for the SI model). These results also show that the covariance matrix of the acoustic transfer function estimated by the HMM decomposition is effective to compensate for the influence of the long impulse response. However, in the matched condition, the SD and the SI recognition rates are 96.4% and 70.7%, where each phoneme HMM is trained using simulated distant-talking speech. The performance of the proposed method is small in comparison with that of the matched condition. Therefore, the further improvement of the HMM adaptation method would be necessary.

This chapter has also investigated the performance of the HMM composition and decomposition methods on speech recognition of the distant moving talker. Speech of distant moving talker is recognized by using an ergodic-HMM of acoustic transfer functions. Each state of the ergodic-HMM of acoustic transfer functions corresponds to a position of sound sources, where all transitions among states are permitted. The results are summarized as follow:

- The performance of the parallel models, where composed HMMs having maximum likelihood are selected, is 76.7%. On the other hand, the performance with the ergodic-HMMs of acoustic transfer functions is improved to 82.3%. These experimental results show that the ergodic-HMM can improve the speech recognition performance of the distant moving talker.

Chapter 6

Telephone Speech Recognition

There have been many studies that deal with convolutional distortion in telephone speech recognition. For more widespread use of telephone speech recognition, studies to deal with additive noise and convolutional distortion should be made. Recently, since cordless telephone handsets are also used, there is the problem of the difference between ordinary analog telephone handsets and cordless telephone handsets.

The previous chapter describes the performance of the HMM composition and decomposition methods on the real distant-talking speech, where speech is contaminated not only by additive noise but also by an acoustic transfer function. The HMM composition and decomposition methods are able to apply to not only distant-talking speech but also to telephone speech. This chapter explores the case of a shorter impulse response, telephone speech recognition. The recognition experiment shows the problem of cordless telephone handsets, and shows that the HMM composition and decomposition methods is able to improve the performance. Other techniques for telephone speech recognition have been reported in [30, 59, 64, 75, 79].

The telephone speech data for evaluation are recorded using 10 kinds of ordinary analog telephone handsets and cordless telephone handsets in a soundproof room through the public telephone network as shown in figure 6.1.

6.1 Telephone Speech Data

Figure 6.1 shows the recording condition of the telephone speech. Utterances from 60 speakers in the ASJ (Acoustical Society of Japan) continuous speech database are outputted through a mouth simulator, and inputted into 10 kinds of ordinary analog

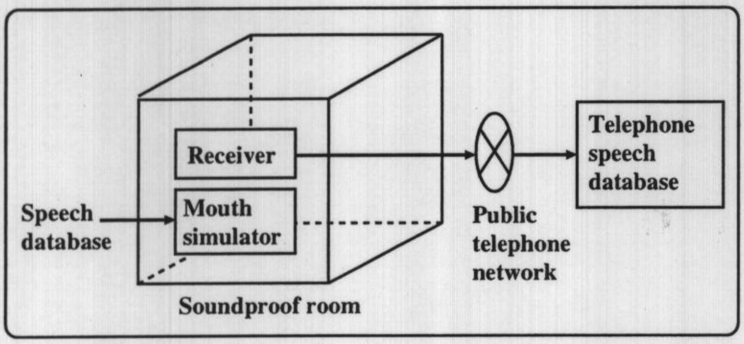

Figure 6.1: Recording condition of telephone speech

telephone handsets and cordless telephone handsets in the soundproof room. Then, their speech is recorded through the public telephone network. Ten kinds of telephone handsets are CANON (CF-H1CL), KENWOOD (IS-W757), NEC (Speax23 CL), NTT (CP-D40), PANASONIC (VE-D67L-K), PIONEER (TF-JP50), SANYO (TEL-L710), SHARP (CJ-H7-B), SONY (SPP-A600) and VICTOR (TN-DJ1-B). Each telephone handset consists of an ordinary analog telephone handset and a cordless telephone handset.

Figure 6.2 shows the log-power spectrum of the clean speech and the telephone speech, which are digitized at an 8 kHz sampling rate. In the case of the speech through cordless telephone handsets, the spectral shape over 3 kHz is distorted. The SNRs of ordinary analog telephone handsets and cordless telephone handsets are 25.1 dB and 20.3 dB, respectively. Their SNRs are calculated by

$$SNR \sim 10\log_{10}\frac{\frac{1}{l}\sum_{t=1}^{l} o(t)^2}{\frac{1}{m}\sum_{t=1}^{m} n(t)^2},$$

where $o(t)$ and $n(t)$ denote the observed speech and the noise at time t, respectively. Next, l and m are the number of total frames of speech data and the number of total frames of noise data, respectively. The problems of cordless telephone handsets are summarized as follows:

- The band-width becomes narrow in comparison with ordinary analog telephone handsets.
- Noise and distortion caused by a wireless system are added to speech.

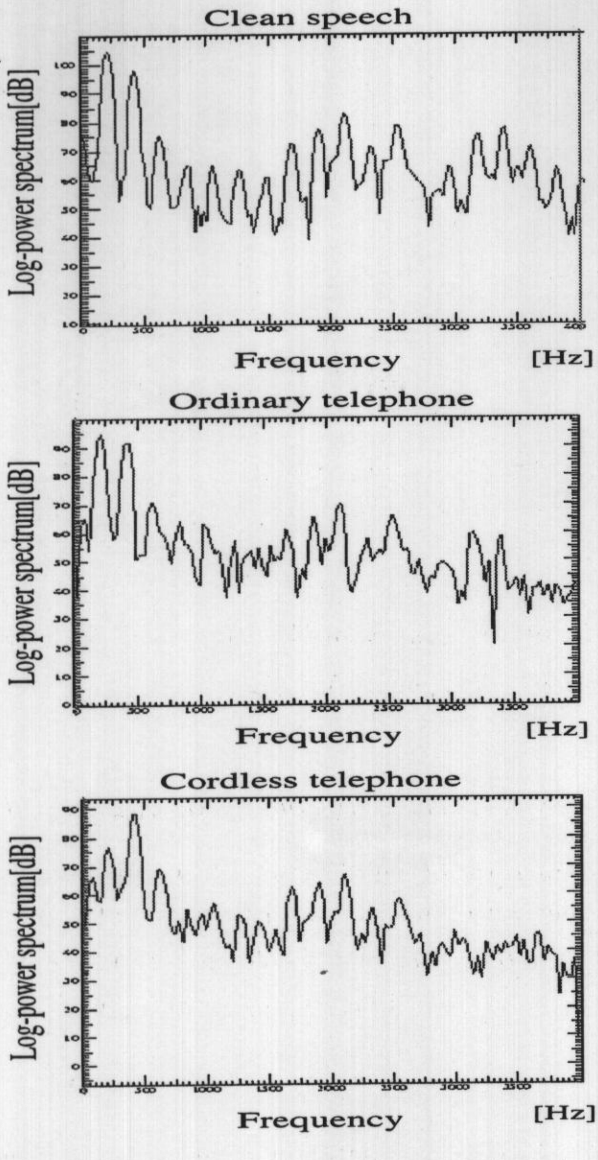

Figure 6.2: Log-power spectrum /u/ of clean speech and telephone speech

- A cordless telephone handset has a scramble-function which causes distortion.

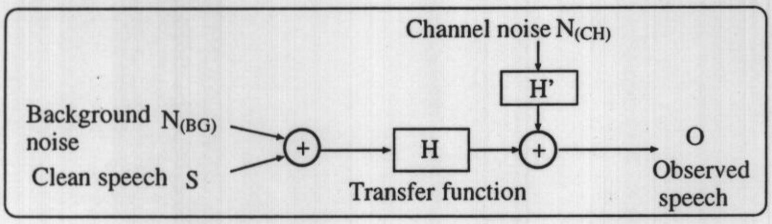

Figure 6.3: Environment model for telephone speech

6.2 HMM Decomposition on Telephone Speech

This section describes the HMM decomposition on telephone speech. Figure 6.3 shows an environment model for the telephone speech. The observed speech $O(\omega; m)$ is represented by

$$\begin{aligned} O(\omega; m) &= \{S(\omega; m) + N_{(BG)}(\omega; m)\} \cdot H(\omega; m) + N_{(CH)}(\omega; m) \cdot H'(\omega; m) \\ &= S(\omega; m) \cdot H(\omega; m) + N(\omega; m), \end{aligned}$$

where

$$N(\omega; m) = N_{(BG)}(\omega; m) \cdot H(\omega; m) + N_{(CH)}(\omega; m) \cdot H'(\omega; m).$$

$S(\omega; m)$, $N_{(BG)}(\omega; m)$, $N_{(CH)}(\omega; m)$, and $N(\omega; m)$ denote the clean speech, the background noise, the channel noise and the observed noise at frame m and frequency ω, respectively. $H(\omega; m)$ and $H'(\omega; m)$ are transfer functions. Accordingly, a composed HMM of the observed speech in the linear-spectral domain is represented by

$$\lambda_{SH+N} = \mathrm{Exp}\{\ \mathrm{Cos}\ (\lambda_{S_{cep}} \oplus \lambda_{H_{cep}})\} \oplus \lambda_{N_{lin}}, \tag{6.1}$$

where λ and $\oplus$ denote a set of model parameters and a model composition procedure, respectively. Exp and Cos are the exponential transform of the distribution function and the cosine transform of the distribution function, respectively. According to equation (6.1), the estimation equation of the transfer function HMM is written in the cepstral domain as follows

$$\lambda_{H_{cep}} = \mathrm{Cos}^{-1}\{\mathrm{Log}(\ \lambda_{SH+N_{lin}} \ominus\ \lambda_{N_{lin}})\} \ominus \lambda_{S_{cep}}, \tag{6.2}$$

where *cep* and *lin* denote the cepstral domain and the linear-spectral domain, respectively. Next, $\ominus$ denotes a model decomposition procedure. Finally, Cos^{-1} and Log are the inverse cosine transform of the distribution function and the logarithm transform of the distribution function, respectively. Equation (6.2) shows that the HMM decomposition method is applied twice in the linear-spectral domain and in the cepstral domain to estimate the transfer function HMM. First, the HMM decomposition method is applied in the linear-spectral domain to estimate the telephone speech HMMs which are free from the influence of noise. The obtained telephone speech HMMs are converted to the cepstral domain. Then, the HMM decomposition method is applied again to estimate the transfer function HMM.

6.3 Experiments and Results

6.3.1 Experimental Conditions

The experiment is conducted on the telephone speech data which we described in section 6.1. About 7500 sentences from 25 males and 25 females are used for the training. Five males and five females for the testing are not used in the training. Each testing speaker utters only one sentence for adaptation for each handset.

We chose 55 context independent phonemes as the clean speech units. Each phoneme is modeled by a single left-to-right 3-state tied-mixture HMM with 3 self-transition loops and without state skipping. Sixteen mel-frequency cepstral coefficients (MFCC) with their first order differentials (ΔMFCC), and the first order differentials for normalized logarithmic energy (Δpower) are calculated as the observation vector for each frame. There are 256 Gaussian mixture components with diagonal covariance matrices shared by all of the models for MFCC and ΔMFCC, respectively. There are 64 Gaus-

Table 6.1: Total number of phrases in testing set

Name of subset	a	b	c	d	e	f	g	h	i	j
Total number of phrases	306	331	327	359	327	366	358	306	292	261
Number of phrases for one sentence	6	7	7	7	7	8	7	6	6	5
Number of phonemes for one phrase	8	7	8	7	7	7	7	7	7	7

Table 6.2: Details of testing set. Five males (m) and five females (f) are used.

Speaker for testing	Subset		
can0001 (m)	a	h	i
tsu0003 (m)	a	b	c
mat0002 (m)	a	i	j
hit0002 (m)	a	d	e
son0001 (m)	a	f	g
ecl1008 (f)	a	e	f
nec1001 (f)	a	b	j
etl1002 (f)	a	c	d
sha1002 (f)	a	g	h
tos1001 (f)	a	h	i

sian mixture components shared by all of the models for Δpower. A single Gaussian is employed to model the noise and the transfer function. Only the mean vector is estimated for the transfer function in this experiment.

The phrase recognition experiment is carried out using continuous sentence speech. Each sentence includes 6 $\sim$ 7 phrases on average. In this task, the ASJ database is divided into 10 subsets. Each subset consists of 50 sentences, except one subset which consists of 53 sentences. One typical subset of this task is 323 phrases with a phrase perplexity of 323 on average. Table 6.1 shows the total number of phrases in detail. Each speaker utters three subsets through one telephone handset as shown in table 6.2.

Phrase accuracy is calculated by

$$\text{Accuracy} = \frac{N - D - S - I}{N} \times 100,$$

where N is the total number of phrases, D is the number of deletions, S is the number of substitutions and I is the number of insertions.

6.3.2 Experimental Results

The points to be investigated are

- improvement of recognition rate by the HMM composition and decomposition methods,

- comparison with cepstral mean normalization (CMN),

and

- comparison with matched condition.

Table 6.3 and table 6.4 show the average phrase accuracy [%] for 10 kinds of ordinary analog telephone handsets and cordless telephone handsets, respectively. The phrase accuracy with the clean HMMs (indicated as HMM-S) is 79.2% for the clean speech. The telephone speech, however, decreases the phrase accuracy to 60.9% for ordinary analog telephone handsets, and 19.6% for cordless telephone handsets.

The phrase accuracy with the HMM-SN, composed of the HMM-S and the noise HMM, is improved to 70.1% for ordinary analog telephone handsets, and 30.3% for cordless telephone handsets. By applying the HMM decomposition method twice in the linear-spectral domain and in the cepstral domain, HMM-SHN, the phrase accuracy is improved from 60.9% to 78.1% for ordinary analog telephone handsets, and from 19.6% to 50.5% for cordless telephone handsets with one adaptation sentence.

Table 6.3 and table 6.4 also include the average phrase accuracy for 10 kinds of the telephone handsets in the matched condition. The phoneme HMMs, HMM-TELE (ordinary tele.) are trained using the speech data through 10 kinds of ordinary analog telephone handsets. The phoneme HMMs, HMM-TELE (cordless tele.), are trained using the speech data through 10 kinds of cordless telephone handsets. The phoneme HMMs, HMM-TELE (ordinary and cordless), are trained using the speech data through 10 kinds of ordinary analog telephone handsets and cordless telephone handsets. The phrase accuracy with the HMM-TELE (ordinary tele.) is 77.7% for ordinary analog telephone handsets. The phrase accuracy with the HMM-TELE (cordless tele.) is 61.0% for cordless telephone handsets. On the other hand, the phrase accuracy with the HMM-TELE (ordinary and cordless) is decreased to 72.7% for ordinary analog telephone handsets, and 60.5% for cordless telephone handsets. This is caused by the mismatched condition between ordinary analog telephone handsets and cordless telephone handsets.

Table 6.5 shows the average phrase accuracy with CMN. In the CMN-based testing case, the phoneme HMMs are trained using the CMN-processed clean speech data.

Table 6.3: Phrase accuracy [%] for 10 ordinary analog telephone handsets

Models	Noise compensation	Channel compensation	Phrase accuracy
HMM-S	×	×	60.9
CMN	×	○	74.7
HMM-SH	×	○	68.6
HMM-SN	○	×	70.1
HMM-SHN	○	○	78.1
HMM-TELE (ordinary tele.)	-	-	77.7
HMM-TELE (ordinary and cordless)	-	-	72.7

Table 6.4: Phrase accuracy [%] for 10 cordless telephone handsets

Models	Noise compensation	Channel compensation	Phrase accuracy
HMM-S	×	×	19.6
CMN	×	○	42.0
HMM-SH	×	○	29.1
HMM-SN	○	×	30.3
HMM-SHN	○	○	50.5
HMM-TELE (cordless tele.)	-	-	61.0
HMM-TELE (ordinary and cordless)	-	-	60.5

By subtracting each cepstral mean value from each testing data, the phrase accuracy is 74.7% for ordinary analog telephone handsets, and 42.0% for cordless telephone handsets. On the other hand, by subtracting the cepstral mean of the same adaptation data to the HMM decomposition from the testing data, the phrase accuracy is dropped

Table 6.5: Comparison with adaptation data in CMN (ordinary/cordless)

Cepstral mean	CMN
Each testing data	74.7% / 42.0%
Adaptation data	72.6% / 38.6%

Table 6.6: Comparison with matched condition for one ordinary analog telephone handset

Input	HMM-S	HMM-SHN	HMM-TELE (matched handset)
matched handset	64.5%	80.1%	86.6%

to 72.6% for ordinary analog telephone handsets, and 38.6% for cordless telephone handsets. This is due to the mismatch of the cepstral mean between adaptation data and each testing data.

Table 6.6 shows the comparison with the matched condition for one ordinary analog telephone handset. In the case of the HMM-TELE (matched handset) which are trained using the speech through only one kind of ordinary analog telephone handset, the performance is 86.6% for the same ordinary analog telephone handset. In the case of the HMM composition and decomposition, the phrase accuracy with the HMM-SHN is 80.1% for the same analog telephone handset with one adaptation sentence. These show a difference in performance of 6.5%. Therefore, further improvement of the HMM adaptation method would be necessary.

6.4 Summary

This chapter has evaluated the performance of the model adaptation based on the previously proposed HMM decomposition method [92] for the telephone speech recognition. The average phrase recognition accuracy with the clean speech HMMs is 60.9% for ordinary analog telephone handsets, and 19.6% for cordless telephone handsets. The average phrase recognition accuracy with the CMN-HMMs is 74.7% for ordinary analog telephone handsets, and 42.0% for cordless telephone handsets. By the HMM decomposition method, the average phrase recognition accuracy is improved to 78.1% for ordinary analog telephone handsets, and 50.5% for cordless telephone handsets. These results show the HMM decomposition method is able to improve the performance. However, in the matched condition, the average phrase recognition accuracy is 77.7% for ordinary analog telephone handsets, and 61.0% for cordless telephone handsets. Therefore, further improvement of the HMM adaptation method is necessary for cordless telephone speech.

Chapter 7

Conclusions

7.1 Summary of Dissertation

The most important advantage of the speech interface is to make hands-free speech recognition a reality, where a user is not encumbered with microphone equipment, and a user can speak from a distance while moving. At present, however, to achieve high recognition accuracy, a user must be equipped with a close-talking microphone. If the user speaks from a distance, the recognition accuracy seriously degrades because of the influence of reverberation and environmental noise. Therefore, technology for the distant-talking speech recognition becomes important.

This thesis has detailed a robust speech recognition technique for acoustic model adaptation based on the HMM composition and decomposition methods in noisy reverberant environments, where a user speaks from a distance of 0.5 m $\sim$ 3.0 m. The aim of the HMM composition and decomposition methods is to estimate the model parameters so as to adapt the model to a target environment by using a small amount of a user's speech in noisy reverberant environments.

In Chapter 3, the HMM composition algorithm for additive noise is extended to model the acoustic transfer function of a reverberant room. In this approach, an HMM attempts to model the acoustic transfer function. The states of the acoustic transfer function HMM correspond to different sound source positions. This HMM can represent the position of sound sources, even if the speaker moves.

This thesis has also proposed, Chapter 4, a new method to estimate HMM parameters of the acoustic transfer function based on the HMM decomposition. This method is able to estimate the model parameters by using observed speech uttered from an unknown position without measurement of impulse responses. The estimated acoustic transfer function, the clean speech HMMs and the noise HMM are composed to recognize noisy and acoustically-distorted speech.

In Chapter 5, speech recognition experiments were carried out to investigate the effectiveness of the HMM composition and decomposition methods on real distant-talking speech, where the loudspeaker is set at a distance of about 2.5 m. The proposed method improves the word-recognition rates for the speaker dependent (SD) model and the speaker independent (SI) model. The word-recognition rate with clean speech HMMs is 77.2% for the SD model, and 54.4% for the SI model. The word-recognition rate with composed HMMs of clean speech HMMs and noise HMM is 87.5% for the SD model, and 61.5% for the SI model. Applying the HMM composition and decomposition methods to the real distant-talking speech, the word-recognition rate with 10 adaptation words is improved to 90.5% for the SD model, and 64.9% for the SI model, where the mean vector of the acoustic transfer function HMM is estimated and composed. Then, applying the HMM decomposition method to both the mean vector and the covariance matrix of the acoustic transfer function HMM, increases the performance to 91.2% for the SD model, and 66.2% for the SI model. It is shown that the covariance matrix of the acoustic transfer function is also effective to compensate for the influence of long impulse responses. However, in the matched condition, the SD and the SI word-recognition rates are 96.4% and 70.7%, where each phoneme HMM is trained using simulated distant-talking speech. The performance of the proposed method is small in comparison with that of the matched condition. Therefore, the further improvement of the HMM adaptation method would be necessary. This chapter has also investigated the performance of the HMM composition and decomposition methods on speech recognition of the distant moving talker. Speech of distant moving talker is recognized by using an ergodic-HMM of acoustic transfer functions. Each state of the ergodic-HMM of acoustic transfer functions corresponds to a position of sound sources, where all transitions among states are permitted. The performance of the parallel models, where composed HMMs having maximum likelihood are selected, is 76.7%. On the other hand, the performance with the proposed ergodic-HMMs of acoustic transfer functions is improved to 82.3%. These experimental results show that the ergodic-HMM can improve the speech recognition performance of the distant moving talker.

Chapter 6 has explored telephone speech recognition. Telephone speech data for evaluation are recorded using 10 kinds of ordinary analog telephone handsets and cordless telephone handsets, in a soundproof room, through the public telephone network [94]. The experimental results show that the HMM decomposition method is able to improve the performance of the telephone speech. However, further improvement would be necessary for cordless telephone speech.

In summary, the HMM composition and decomposition methods are applicable to a wide variety of additive noise and convolutional distortion tasks. We have focused on model adaptation using a single microphone in this paper. The model adaptation, however, can also be emphasized by using a multi-microphone (microphone array). Though the proposed method has currently not achieved distant-talking speech recognition to state-of-the-art level, one can achieve good performance through extensions of the proposed method in real world conditions.

7.2 Future Work

Distant-talking speech recognition is an important research topic with great potential. The technique proposed in this thesis has improved speech recognition performance, where a user speaks from a distance of 0.5 m $\sim$ 3.0 m in noisy reverberant environments. However, there are still some fundamental problems that need to be addressed and carefully studied. For example, the covariance matrix of the acoustic transfer function HMM deals with the influence of the long impulse response in this thesis. Speech recognition performance is improved with this method, but the effect is not sufficient to compensate for the influence completely.

The spectral analysis for speech recognition is based on short-time windowing. The length of the window is smaller than that of the room impulse response. If the window is sufficiently shorter than that of the impulse response, the use of CMN will be effective. However, to make the window longer degrades the speech recognition rate, because the spectra in the window become unstable. In [46], the performance of speech recognition systems is compared with the performance of human listeners on reverberated speech, where some speech enhancement techniques are used: PLP (Perceptual Linear Predictive) [26], log-RASTA-PLP (log-Relative Spectra-PLP) [27], and j-RASTA-PLP [27]. The experiments show that humans are adept at recognizing reverberated speech clearly, while the speech recognition systems are not. Avendano et al. [7] proposes a multi-resolution channel normalization technique for reverberated speech, where the reverberated speech is recovered from narrow-band spectrograms (long-analysis window) and wide-band spectrograms (short-analysis window). When more than one source of distortion exists, e.g., both additive and convolutional distortions, the problem becomes more difficult. When the distortion sources are non-stationary, e.g., when the speaker is moving, some adaptive compensation techniques are needed. Integration of microphone arrays and acoustic model adaptation will be also expected. To enhance the efficacy and the effectiveness of the compensation, those techniques need to better characterize the distribution of possible distortion types, and to use this distribution to choose the appropriate compensation model. We are working along these lines of thought.

Appendix A

Transformation of Probability Distribution

This chapter describes the transformations of a probability distribution which are applied to compose an HMM. The distribution is the output probability distribution of an HMM. The transforms applied in the HMM composition are as follows,

- Cosine transform
- Exponential transform
- Convolution of distribution
- Logarithm transform

A multivariate Gaussian distribution is used in general for the output probability distribution. Therefore, the following transforms should be applied to the multivariate Gaussian distribution. The following refers to [56].

A.1 Cosine Transform

Let a random vector $\boldsymbol{X}$ be a multivariate Gaussian distribution. The cosine transform applied to the random vector $\boldsymbol{X}$ can be described as follows,

$$\boldsymbol{Y} = \Gamma \boldsymbol{X}.$$

Each element of transform matrix is given by

$$c_{ij} = \cos(i(j-0.5)\pi/N), \quad (0 \le i, j \le N), \tag{A.1}$$

where N is the dimension of the distribution.

As the transform is linear, $\boldsymbol{Y}$ has a multivariate Gaussian distribution when $\boldsymbol{X}$ has a multivariate Gaussian distribution.

A transformed mean vector $\boldsymbol{\mu}= (\mu_1, \cdots, \mu_N)'$ and a covariance matrix $\boldsymbol{\Sigma}$ are given by

$$\begin{aligned} \boldsymbol{\mu}^* &= \Gamma\boldsymbol{\mu}, & \text{(A.2)} \\ \boldsymbol{\Sigma}^* &= \Gamma'\boldsymbol{\Sigma}\Gamma, & \text{(A.3)} \end{aligned}$$

where $'$ denotes the transposition.

A.2 Exponential Transform

The exponential transform applied to a random vector $\boldsymbol{X}$ can be described as follows,

$$\boldsymbol{Y} = \exp(\boldsymbol{X}).$$

Note that $\boldsymbol{Y}$ does not have a multivariate Gaussian distribution even if $\boldsymbol{X}$ has a multivariate Gaussian distribution, since the transform is non-linear. Rather, the random vector $\boldsymbol{Y}$ has a log-normal distribution, if the random vector $\boldsymbol{X}$ has a multivariate normal distribution. After applying the exponential transform to $\boldsymbol{X}$, the first moment μ_i^* is obtained as follows.

$$\begin{aligned} \mu_i^* &= E[\exp(x_i)] \\ &= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp(x_i) \exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right\} d\boldsymbol{X} \quad \text{(A.4)} \end{aligned}$$

Now define a variable as follows,

$$\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{\mu}.$$

Inserting this notation into (A.4), μ_i^* is obtained as follows,

$$\begin{aligned} \mu_i^* &= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp(z_i + \mu_i) \exp\left\{-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\right\} d\boldsymbol{X} \\ &= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{z_i + \mu_i - \frac{1}{2}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\right\} d\boldsymbol{X}, \quad \text{(A.5)} \end{aligned}$$

where $z_i \in \mathrm{Z}$. Let's us define e_i to have 1 for the i-th position,

$$e_i \stackrel{\text{def}}{=} (0\,0 \ldots\, 0\,1\,0 \ldots\, 0)'.$$

According to equation (A.5),

$$\begin{aligned}
\mu_i^* &= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{z_i + \mu_i - \frac{1}{2}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\right\} d\boldsymbol{X} \\
&= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{z_i + \mu_i - \frac{1}{2}((\boldsymbol{Z} - \boldsymbol{\Sigma}e_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{\Sigma}e_i) + 2z_i - \sigma_{ii})\right\} d\boldsymbol{X} \\
&= \exp\left(\mu_i + \frac{\sigma_{ii}}{2}\right) \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{Z} - \boldsymbol{\Sigma}e_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{\Sigma}e_i)\right\} d\boldsymbol{X}.
\end{aligned}$$

Therefore, the first moment is given by

$$\mu_i^* = \exp\left(\mu_i + \frac{\sigma_{ii}}{2}\right), \quad (0 \leq i \leq N). \tag{A.6}$$

The second moment σ_{ij}^* can be sought in a similar manner,

$$\sigma_{ij}^* = E[\exp(x_i)\exp(x_j)] - E[\exp(x_i)] \cdot E[\exp(x_j)]. \tag{A.7}$$

This right hand side can be rewritten as follows,

$$\begin{aligned}
&E[\exp(x_i)\exp(x_j)] \\
&= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp(x_i + x_j) \exp\left\{-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right\} d\boldsymbol{X} \\
&= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp(z_i + \mu_i + z_j + \mu_j) \exp\left\{-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\right\} d\boldsymbol{X} \\
&= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{z_i + \mu_i + z_j + \mu_j - \frac{1}{2}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\right\} d\boldsymbol{X} \\
&= \int \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{z_i + \mu_i + z_j + \mu_j \\
&\quad -\frac{1}{2}((\boldsymbol{Z} - \boldsymbol{\Sigma}(e_i + e_j))'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{\Sigma}(e_i + e_j)) + 2z_i - \sigma_{ii} + 2z_j - \sigma_{jj} - 2\sigma_{ij})\Big\} d\boldsymbol{X} \\
&= \exp\left(\mu_i + \frac{\sigma_{ii}}{2}\right) \exp\left(\mu_j + \frac{\sigma_{jj}}{2}\right) \exp(\sigma_{ij}). \qquad (A.8)
\end{aligned}$$

According to equation (A.7) and equation (A.8), the second moment σ_{ij}^* is given by

$$\sigma_{ij}^* = \mu_i^*\mu_j^*(\exp(\sigma_{ij}) - 1), \quad (0 \leq i, j \leq N). \tag{A.9}$$

A.3 Convolution of Probability Distributions

A random vector $\boldsymbol{Y}$ is given by

$$\boldsymbol{Y} = \boldsymbol{X}_1 + \boldsymbol{X}_2,$$

where two random vectors, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent. Therefore, a distribution function of $\boldsymbol{Y}$ is given by

$$\int_{-\infty}^{\infty} F(\boldsymbol{Y} - \boldsymbol{X}_2) dG(\boldsymbol{X}_2) = \int_{-\infty}^{\infty} G(\boldsymbol{Y} - \boldsymbol{X}_1) dF(\boldsymbol{X}_1),$$

where distribution functions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $F(\boldsymbol{X}_1)$ and $G(\boldsymbol{X}_2)$, respectively. This is called convolution of the distribution function F and G. If the random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have a multivariate Gaussian distribution, the addition of them also has a multivariate Gaussian distribution (called reproducibility).

The transformed mean vector μ_i^* and covariance σ_{ij}^* are given by

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \tag{A.10}$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2. \tag{A.11}$$

In fact, however, convolution of the log-normal distributions is executed. An approximation is used to execute it with facility. The approximation is that the sum of log-normally distributed variables has a log-normal distribution. Therefore, an error might be produced.

A.4 Logarithm Transform

The logarithm transform applied to a random vector $\boldsymbol{X}$ can be described as follows,

$$\boldsymbol{Y} = \log(\boldsymbol{X}).$$

The same thing as exponential transform can be said of this transform: it is non-linear. $\boldsymbol{Y}$ must have a multivariate Gaussian distribution. Therefore,

- assume that $\boldsymbol{X}$ is log-normally distributed.

Then, the transformed distribution is a multivariate Gaussian distribution. Its parameters are given by

$$\mu_i^* = \log(\mu_i) - \frac{1}{2}\log\left(\frac{\sigma_{ii}^2}{\mu_i\mu_i} + 1\right), \tag{A.12}$$

$$\sigma_{ij}^* = \log\left(\frac{\sigma_{ij}}{\mu_i\mu_j} + 1\right). \tag{A.13}$$

Appendix B

Lists of Adaptation Data in Word Recognition

In this appendix, the related information of the adaptation data used in the word-recognition experiments are listed, where /Q/ is a double consonant. Each testing speaker utters 1 $\sim$ 50 words ($\times$ 3) as adaptation data which are not used in the training. The speech corpus is the Set-A of the ATR Japanese speech database.

1-word ikioi	2-word omoshiroi	3-word shuukyou	4-word daidokoro	5-word basho	6-word pokeQto	7-word ukeau
8-word seii	9-word teochi	10-word hyakushou	11-word rejaa	12-word ajiwau	13-word aNkeeto	14-word imagoro
15-word umeawaseru	16-word esukareetaa	17-word oiharau	18-word omocha	19-word kimuzukashii	20-word gyuunyuu	21-word kyokutaN
22-word kogiQte	23-word kopii	24-word shuuheN	25-word joukyaku	26-word supiido	27-word soredeha	28-word chouetsu
29-word tenohira	30-word nakanaori	31-word nyoubou	32-word paipu	33-word hikiukeru	34-word byoudou	35-word peeji
36-word misuborashii	37-word myounichi	38-word yuumoa	39-word yotsukado	40-word ryuuchou	41-word reNai	42-word wazawaza
43-word taNoNhyouji	44-word fuseimyaku	45-word kebyou	46-word majo	47-word menyuu	48-word kounyuu	49-word meiryou
50-word iNryoku						

1-word iyoiyo	2-word guai	3-word juNbaN	4-word chaNto	5-word byouiN	6-word boNyari	7-word uchiawase
8-word sobieru	9-word deshabaru	10-word byouniN	11-word wasuremono	12-word atarimae	13-word iede	14-word iraQsharu
15-word uyamau	16-word enerugii	17-word ogosoka	18-word omowazu	19-word gyakutai	20-word kyuuryou	21-word kyozetsu
22-word kokoroyoi	23-word gobusata	24-word shuQse	25-word shoufuda	26-word zeNshuu	27-word zoNzai	28-word tsukekuwaeru
29-word depaato	30-word nyuuiN	31-word neage	32-word hanahada	33-word hiQkurikaesu	34-word hyouhoN	35-word beQdo
36-word misebirakasu	37-word meue	38-word yukizumaru	39-word yoQparai	40-word ryuQkusaQku	41-word rokuoN	42-word wariateru
43-word kaNfuru	44-word karyuu	45-word toQkyo	46-word zahyou	47-word koumyou	48-word saNmyaku	49-word meiryou
50-word gobyuu						

1-word urayamashii	2-word zairyou	3-word suichoku	4-word chuuou	5-word hyoujuN	6-word megane	7-word kareNdaa
8-word chichioya	9-word nisemono	10-word furafura	11-word akachaN	12-word apaato	13-word ichijirushii	14-word udemae
15-word eikyuu	16-word epuroN	17-word oshaberi	18-word gaishutsu	19-word kyuugyou	20-word gyousei	21-word gehiN
22-word kotozute	23-word sakihodo	24-word juNjuNni	25-word shouryaku	26-word zeNtei	27-word daibubuN	28-word dekigoto
29-word toriaezu	30-word nyuujou	31-word nesage	32-word hanabanashii	33-word byousha	34-word puroguramu	35-word poNpu
36-word myaku	37-word mochinushi	38-word yubisasu	39-word ryakusuru	40-word ryougae	41-word roQkaa	42-word waribiki
43-word maehyoubaN	44-word gabyou	45-word soQchoku	46-word koNnyaku	47-word myuujiQku	48-word saNmyaku	49-word tsuikyuu
50-word techou						

Appendix C

Lists of Testing Data in Word Recognition

In this appendix, the related information of the testing data used in the word-recognition experiments are listed. For testing, 500 words which are different from those words in the training are used. The speech corpus is the Set-A of the ATR Japanese speech database.

1. aa	2. aite	3. aoru	4. aki	5. akushu
6. ago	7. ashiba	8. aseru	9. ataru	10. atsui
11. atehamaru	12. apaato	13. amasu	14. ayashimu	15. arasou
16. aru	17. awaseru	18. aNshiN	19. iitsukeru	20. igaku
21. igi	22. ikou	23. iji	24. izeN	25. itamu
26. ichiji	27. ichiryuu	28. iQsou	29. itsuka	30. itonamu
31. ihaN	32. iya	33. iru	34. iroiro	35. iNsotsu
36. ukai	37. uketsuke	38. ushinau	39. usotsuki	40. uchigawa
41. uQtoushii	42. utsuru	43. ubau	44. umeawaseru	45. urameshii
46. ureru	47. uNpaN	48. eiyuu	49. eda	50. eri
51. eNgi	52. eNtotsu	53. oite	54. oufuku	55. ooku
56. okashii	57. okujou	58. okonau	59. oshiire	60. ojiisaN
61. osoreru	62. oQto	63. otoroeru	64. onoono	65. oboreru
66. omote	67. oyobu	68. owari	69. kai	70. kaikei
71. kaijou	72. kaiteki	73. kaihou	74. kaeQte	75. kakaeru
76. kagayaku	77. kaku	78. kakutoku	79. kagu	80. kakeru
81. kago	82. kashikiri	83. kasu	84. kata	85. katamari
86. kaQki	87. katsuyaku	88. kanaeru	89. kane	90. kabuseru
91. kami	92. kayui	93. karini	94. kawaigaru	95. kaN

96. kaNgei	97. kaNshou	98. kaNjiru	99. kaNtaN	100. kaNbeN
101. gaisuru	102. gakusei	103. gaQki	104. gaNjitsu	105. kioN
106. kigaru	107. kikeN	108. kisha	109. kizuku	110. kitai
111. kiQpu	112. kinou	113. kimari	114. kiyaku	115. kyuusho
116. kyoukai	117. kyousou	118. kyoumi	119. kyoneN	120. kiri
121. kire	122. kiNko	123. kiNniku	124. gishiki	125. gyouji
126. ku	127. kuusou	128. kusai	129. kujou	130. kuda
131. kuchou	132. kufuu	133. kumori	134. kurayami	135. kure
136. kuwawaru	137. guNkaN	138. keikaku	139. keishiki	140. keibi
141. keshou	142. keQkyoku	143. keQtei	144. kemui	145. keNka
146. keNjitsu	147. keNmei	148. geshuku	149. geNeki	150. geNshuku
151. geNtei	152. koi	153. koueN	154. koukyuu	155. kougeN
156. kouzaN	157. koujou	158. kousoku	159. koudou	160. kouhyou
161. koumoku	162. koe	163. kokugo	164. kokumiN	165. kokoroyoi
166. kojitsukeru	167. kozou	168. koQkai	169. kotei	170. kotori
171. konogoro	172. komakai	173. koraeru	174. kowai	175. koNshuu
176. koNya	177. goudou	178. gozaimasu	179. sa	180. saisho
181. sainou	182. saeru	183. sakarau	184. saku	185. sakubuN
186. sageru	187. sashisawari	188. sasuru	189. saQkyoku	190. satsubatsu
191. saabisu	192. samui	193. saru	194. saNkaku	195. saNso
196. zaisaN	197. zatsuoN	198. shi	199. shio	200. shikashi
201. shikisai	202. shikujiru	203. shikori	204. shijuu	205. shizeN
206. shitashii	207. shiQso	208. shitsubou	209. shinagire	210. shiharau
211. shihou	212. shimatsu	213. shimekiri	214. shaku	215. shameN
216. shuugou	217. shuuteN	218. shuei	219. shusai	220. shuQsaN
221. shubi	222. shou	223. shoukiN	224. shousuu	225. shoutotsu
226. shoufuda	227. shouri	228. shokuhiN	229. shotoku	230. shirase
231. shirushi	232. shiNka	233. shiNkoku	234. shiNsou	235. shiNpai
236. shiNryaku	237. jiki	238. jigoku	239. jishiN	240. jichou
241. jiQseN	242. jitsuha	243. jimaN	244. juu	245. juutai
246. jugyou	247. juNsa	248. juNbaN	249. joukei	250. joudaN
251. jouriku	252. jiNkeN	253. su	254. suisoku	255. suimiN
256. sukasu	257. sugiru	258. sugoi	259. susumeru	260. suQkari
261. subarashii	262. sumaato	263. suru	264. zuaN	265. se
266. seiki	267. seishiki	268. seizoN	269. seinou	270. seiyou
271. sekiniN	272. seQto	273. senaka	274. sewa	275. seNsaku
276. seNdeN	277. seNryou	278. zero	279. sou	280. sousa
281. souzoushii	282. sokushiN	283. soshi	284. soQkuri	285. sonoue
286. soboku	287. sorezore	288. soN	289. zouri	290. taioN
291. taikou	292. taiji	293. taitou	294. taimaN	295. taeru
296. tagayasu	297. take	298. tatakai	299. tachi	300. taQsei

301. tatoe	302. tanoshimi	303. tabi	304. tamatama	305. tayori
306. taNku	307. taNniN	308. daiji	309. daiyaru	310. daQte
311. daN	312. daNtai	313. chiiki	314. chikayoru	315. chizu
316. chihou	317. chuukai	318. chuusha	319. chuumoN	320. chousho
321. chouwa	322. chirasu	323. tsui	324. tsuuka	325. tsue
326. tsukiataru	327. tsuku	328. tsukeru	329. tsutsushimu	330. tsuneni
331. tsubomi	332. tsume	333. tsurai	334. teate	335. teisha
336. teibou	337. tekisuto	338. tejika	339. tetsudou	340. tema
341. teNkai	342. teNbou	343. deshi	344. deNsha	345. to
346. toushi	347. touchaku	348. touroN	349. tokasu	350. tokushoku
351. toge	352. tojiru	353. totemo	354. tobiagaru	355. tomurau
356. tori	357. toridasu	358. toNdemonai	359. dougu	360. dounika
361. douro	362. dokuseN	363. dorei	364. naizou	365. naka
366. nagame	367. nagedasu	368. nadaraka	369. nanoka	370. nameraka
371. nariyuki	372. naNtonaku	373. nigatsu	374. nikoniko	375. nichiyou
376. nyuusu	377. niNgeN	378. nuku	379. nureru	380. neji
381. netsu	382. neru	383. noumiN	384. nozoku	385. noberu
386. noridasu	387. haaku	388. haichi	389. hakase	390. haku
391. hakobu	392. haji	393. hazumu	394. hada	395. haQkou
396. hatsugeN	397. hanasu	398. hahaoya	399. hayai	400. harigane
401. haNi	402. haNjou	403. haNpa	404. baishuu	405. baketsu
406. barabara	407. hi	408. higai	409. hikidasu	410. hikutsu
411. hijoushiki	412. hiQkurikaesu	413. hitogara	414. hiniku	415. himo
416. hyoujuN	417. hiryou	418. hirogeru	419. byouiN	420. biNbou
421. fuushuu	422. fukisoku	423. fukushuu	424. fukei	425. fusagu
426. fusuma	427. futaN	428. futsuka	429. fuhai	430. fuyu
431. furui	432. fuNgai	433. butai	434. bunaN	435. buNseki
436. buNretsu	437. heisa	438. hedateru	439. heNni	440. beNjo
441. houki	442. houseki	443. houbou	444. hogaraka	445. hoshii
446. hotoke	447. homeru	448. hoNseki	449. boueki	450. bokasu
451. boro	452. mairu	453. maku	454. magokoro	455. masu
456. mataha	457. maQkura	458. mato	459. maneku	460. mamoru
461. mawaru	462. miageru	463. mikata	464. mijikai	465. misoka
466. miQchaku	467. mitodokeru	468. minikui	469. mimai	470. myou
471. miNkaN	472. mukigeN	473. mushi	474. muzukashii	475. munashii
476. mure	477. meisho	478. meirei	479. mezasu	480. memo
481. meNbaa	482. moushiwake	483. mokei	484. mochinushi	485. moto
486. monooki	487. moyasu	488. moroi	489. yaku	490. yakume
491. yasui	492. yaQtsukeru	493. yaburu	494. yu	495. yuugi
496. yuudou	497. yugamu	498. yushutsu	499. yunyuu	500. yurumeru

Bibliography

[1] V. Abrash, A. Sankar, H. Franco, M. Cohen, "Acoustic adaptation using transformations of HMM parameters", *Proc.ICASSP-96*, pp. 729-732, 1996.

[2] A. Acero, "Acoustical and environmental robustness in automatic speech recognition", Ph.D Dissertation, ECE Department, Carnegie Mellon University, Sept. 1990.

[3] F. Alleva, "Search organization in the Whisper continuous speech recognition system", *Proc.IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 295-302, 1997.

[4] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, "A compact model for speaker-adaptive training", *Proc.ICSLP-96*, pp. 1137-1140, 1996.

[5] T. Anastasakos, J. McDonough and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization", *Proc.ICASSP-97*, pp. 1043-1046, 1997.

[6] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *Journal of Acoustical Society of America*, Vol. 55, pp. 1304-1312, 1974.

[7] C. Avendano, S. Tibrewala and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments", *Proc.EUROSPEECH-97*, pp. 1107-1110, 1997.

[8] L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *Annals of Mathematical Statistics*, 41, pp. 164-171, 1970.

[9] L. E. Baum, "An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes", *Inequalities*, 3, pp. 1-8, 1972.

[10] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE,ASSP-27*, No.2, 1979.

[11] J.-T.Chien, C.-H.Lee and H.-C.Wang, "Improved Bayesian learning of hidden Markov models for speaker adaptation", *Proc.ICASSP-97*, pp. 1027-1030, 1997.

[12] L. Delphin-Poulat and C. Mokbel, "Signal bias removal using the multi-path stochastic equalization", *Proc.EUROSPEECH-97*, pp. 2575-2578, 1997.

[13] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39, pp. 1-38, 1977.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 32, No. 6, pp. 1109-1121, 1984.

[15] S. Furui, "Recent advances in robust speech recognition", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 11-20, 1997.

[16] M. J. F. Gales and S. J. Young, "An improved approach to the hidden Markov model decomposition of speech and noise", *Proc.ICASSP-92*, pp. 233-236, 1992.

[17] M. J. F. Gales and S. J. Young, "PMC for speech recognition in additive and convolutional noise", CUED-F-INFENG-TR154, 12, 1993.

[18] M. J. F. Gales and S. J. Young, "A fast and flexible implementation of parallel model combination", *Proc.ICASSP-95*, pp. 133-136, 1995.

[19] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework", Computer Speech and Language, Vol. 10, pp. 249-264, 1996.

[20] M. J. F. Gales, " "NICE" model-based compensation schemes for robust speech recognition", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 55-64, 1997.

[21] J.-L.Gauvain and C.-H.Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, 1994.

[22] D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer, "Experiments of HMM adaptation for hands-free connected digit recognition", *Proc.ICASSP-98*, pp. 473-476, 1998.

[23] P. S. Gopalakrishnan, "A tree search strategy for large vocabulary continuous speech recognition", *Proc.ICASSP-95*, pp. 572-575, 1995.

[24] P. S. Gopalakrishnan and L. R. Bahl, "Fast match techniques", in *Automatic speech and speaker recognition: advanced topics*, C.-H.Lee, F. K. Soong and K. K. Paliwal editors, 1996.

[25] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Transactions on Antennas and Propagation*, Vol. 30, No. 1, pp. 27-34, 1982.

[26] H. Hermansky, "Perceptual linear predictive(PLP) analysis of speech", *Journal of Acoustical Society of America*, Vol. 87, No. 4, pp. 1738-1752, 1990.

[27] H. Hermansky and N. Morgan, "RASTA processing of speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589, 1994.

[28] X. D. Huang, Y. Ariki and M. Jack, "Hidden Markov Models for Speech Recognition", *Edinburgh University Press, Edinburgh*, 1990.

[29] T. B. Hughes, H.-S.Kim, J. H. Dibiase and H. F. Silverman, "Using a real-time, tracking microphone array as input to an HMM speech recognizer", *Proc.ICASSP-98*, pp. 249-252, 1998.

[30] W.-W.Hung and H.-C.Wang, "A comparative analysis of blind channel equalization methods for telephone speech recognition", *Proc.EUROSPEECH-97*, pp. 1515-1518, 1997.

[31] Q. Huo, C. Chan and C.-H.Lee, "Bayesian adaptive learning of the parameters of hidden Markov models for speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 5, pp. 334-345, 1995.

[32] Q. Huo, H. Jiang and C.-H.Lee, "A Bayesian predictive classification approach to robust speech recognition", *Proc.ICASSP-97*, pp. 1547-1550, 1997.

[33] Q. Huo and C.-H.Lee, "Combined on-line model adaptation and Bayesian predictive classification for robust speech recognition", *Proc.EUROSPEECH-97*, pp. 1847-1850, 1997.

[34] J. Ishii and T. Fukada, "Speaker independent acoustic modeling using speaker normalization", *Proc.ICASSP-98*, pp. 97-100, 1998.

[35] N. Iwahashi, H. Pao, H. Honda, K. Minamino and M. Omote, "Stochastic features for noise robust speech recognition", *Proc.ICASSP-98*, pp. 633-636, 1998.

[36] F. Jelinek, "Statistical methods for speech recognition", *The MIT Press*, 1997.

[37] H. Jiang, K. Hirose and Q. Huo, "Robust speech recognition based on Bayesian predictive classification", *Proc.ICASSP-97*, pp. 1551-1554, 1997.

[38] B.-H.Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains.", *AT&T Technical Journal*, 64(6), pp. 1235-1249, 1985.

[39] B. H. Juang, L. R. Rabiner and J. G. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 35, No. 7, pp. 947-954, 1987.

[40] B.-H.Juang, W. Chou and C.-H.Lee, "Statistical and discriminative methods for speech recognition", in *Automatic speech and speaker recognition: advanced topics*, C.-H.Lee, F. K. Soong and K. K. Paliwal editors, 1996.

[41] Jean-Claude Junqua and Yolande Anglade, "Acoustic and perceptual studies of Lombard speech: Application to isolated-words automatic speech recognition", *Proc.ICASSP-90*, pp. 841-844, 1990.

[42] Jean-Claude Junqua and Jean-Paul Haton, "Robustness in automatic speech recognition", *Kluwer Academic Publishers*, 1996.

[43] S. Kajita, K. Takeda and F. Itakura, "Spectral weighting of SBCOR for noise robust speech recognition", *Proc.ICASSP-98*, pp. 621-624, 1998.

[44] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 34, No. 6, pp. 1391-1400, 1986.

[45] K. Kato and K. Kakehi, "Listener adaptability to individual speaker differences in monosyllabic speech perception", *Journal of Acoustical Society of Japan*, Vol. 44, No. 3, pp. 180-186, 1988. (in Japanese)

[46] Brian E. D. Kingsbury and Nelson Morgan, "Recognizing reverberant speech with RASTA-PLP", *Proc.ICASSP-97*, pp. 1259-1262, 1997.

[47] T. Kobayashi, T. Masuko and K. Tokuda, "HMM compensation for noisy speech recognition based on cepstral parameter generation", *Proc.EUROSPEECH-97*, pp. 1583-1586, 1997.

[48] T. Kosaka, H. Yamamoto, M. Yamada and Y. Komori, "Instantaneous environment adaptation techniques based on fast PMC and MAP-CMS methods", *Proc.ICASSP-98*, pp. 789-792, 1998.

[49] C.-H.Lee and J.-L.Gauvain , "Bayesian adaptive learning and MAP estimation of HMM", in *Automatic speech and speaker recognition: advanced topics*, C.-H.Lee, F. K. Soong and K. K. Paliwal editors, 1996.

[50] C.-H.Lee, "On feature and model compensation approach to robust speech recognition", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 45-54, 1997.

[51] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, Vol. 9, pp. 171-185, 1995.

[52] P. Lockwood and J. Boudy, "Experiments with a non-linear spectral subtractor, hidden Markov models and the projection, for robust speech recognition in cars", *Proc.EUROSPEECH-91*, pp. 79-82, 1991.

[53] G. W. Mackenzie, "Acoustics", *Focal Press*, 1964.

[54] F. Martin, K. Shikano, Y. Minami and Y. Okabe, "Recognition of noisy speech by using composition of hidden Markov models", *Proc.ASJ Fall meeting*, 1-7-10, 1992.

[55] F. Martin, K. Shikano and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models", *Proc.EUROSPEECH-93*, pp. 1031-1034, 1993.

[56] F. Martin, "Recognition of noisy speech by composition of hidden Markov models", Master thesis of the course of Electronic Engineering of the University of Tokyo. 1993.

[57] D. Matrouf and J-L. Gauvain, "Model compensation for additive and convolutive noises in training and test data", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 207-210, 1997.

[58] T. Matsuoka and C.-H.Lee, "A study of on-line Bayesian adaptation for HMM-based speech recognition", *Proc.EUROSPEECH-93*, pp. 815-818, 1993.

[59] E. McDermott, E. A. Woudenberg and S. Katagiri, "A telephone-based directory assistance system adaptively trained using minimum classification error/generalized probabilistic descent", *Proc.ICASSP-96*, pp. 3346-3349, 1996.

[60] Y. Minami and S. Furui, "A Maximum likelihood procedure for a universal adaptation method based on HMM composition", *Proc.ICASSP-95*, pp. 129-132, 1995.

[61] Y. Minami and S. Furui, "Adaptation method based on HMM composition and EM algorithm", *Proc.ICASSP-96*, pp. 327-330, 1996.

[62] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 36, No. 2, pp. 145-152, 1988.

[63] C. Mokbel, "MUSE:Multipath stochastic equalization, a theoretical framework to combine equalization and stochastic modeling", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 211-214, 1997.

[64] M. Morishima, T. Isobe and N. Koizumi, "Phonetically balanced cepstrum mean normalization", *Acoust. Soc. America and Acoust. Soc. Japan Third Joint Meeting*, pp. 1105-1108, 1996.

[65] S. Nakamura, T. Yamada, T. Takiguchi and K. Shikano, "Hands free speech recognition by a microphone array and HMM composition", *Proc.International Workshop on Human Interface Technology*, pp. 33-38, 1995.

[66] S. Nakamura, T. Takiguchi and K. Shikano, "Noise and room acoustics distorted speech recognition by HMM composition", *Proc.ICASSP-96*, pp. 69-72, 1996.

[67] J. A. Nolazco-Flores and S. J. Young, "Continuous speech recognition in noise using spectral subtraction and HMM adaptation", *Proc.ICASSP-94*, I-409-412, 1994.

[68] A. M. Noll, "Cepstrum pitch determination", *Journal of Acoustical Society of America*, Vol. 41, pp. 293-309, 1967.

[69] Y. Normandin, "Maximum mutual information estimation of hidden Markov models", in *Automatic speech and speaker recognition: advanced topics*, C.-H.Lee, F. K. Soong and K. K. Paliwal editors, 1996.

[70] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs", *Proc.ICSLP-92*, pp. 369-372, 1992.

[71] M. Omologo, "On the future trends of hands-free ASR: variabilities in the environmental conditions and in the acoustic transduction", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 67-73, 1997.

[72] M. Omologo, M. Matassoni, P. Svaizer and D. Giuliani,"Hands-free speech recognition in a noisy and reverberant environment", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 195-198, 1997.

[73] M. Ostendorf, "From HMMs to segments models: stochastic modeling for CSR", in *Automatic speech and speaker recognition: advanced topics*, C.-H.Lee, F.K.Soong and K. K. Paliwal editors, 1996.

[74] K. K. Paliwal, "Spectral subband centroids as features for speech recognition", *Proc.IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 124-131, 1997.

[75] J.-B.Puel and B. Jacob,"Robust HMM architectures for cellular phone speech recognition", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 215-218, 1997.

[76] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc.IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[77] L. R. Rabiner, B.-H.Juang, "Fundamentals of speech recognition", *PTR Prentice-Hall, Englewood Cliffs,NJ*, 1993.

[78] L. R. Rabiner, B.-H.Juang and C.-H.Lee, "An overview of automatic speech recognition", in *Automatic speech and speaker recognition: advanced topics*, C.-H.Lee, F. K. Soong and K. K. Paliwal editors, 1996.

[79] M. G. Rahim and B.-H.Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 1, pp. 19-30, 1996.

[80] S. Renals and M. Hochberg, "Efficient evaluation of the LVCSR search space using the NOWAY decoder", *Proc.ICASSP-96*, pp. 149-152, 1996.

[81] S. Sagayama, Y. Yamaguchi, S. Takahashi and J. Takahashi, "Jacobian approach to fast acoustic model adaptation", *Proc.ICASSP-97*, pp. 835-838, 1997.

[82] A. Sankar and C.-H.Lee, "Robust speech recognition based on stochastic matching", *Proc.ICASSP-95*, pp. 121-124, 1995.

[83] A. Sankar and C.-H.Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 3, pp. 190-202, 1996.

[84] P. W. Shields and D. R. Campbell, "Intelligibility improvements obtained by an enhancement method applied to speech corrupted by noise and reverberation", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 91-94, 1997.

[85] K. Shinoda and C.-H.Lee, "Unsupervised adaptation using structural Bayes approach", *Proc.ICASSP-98*, pp. 793-796, 1998.

[86] M. Shozakai, S. Nakamura and K. Shikano, "A non-iterative model-adaptive E-CMN/PMC approach for speech recognition in car environments", *Proc. EUROSPEECH-97*, pp. 287-290, 1997.

[87] M. Shozakai, S. Nakamura and K. Shikano, "A speech enhancement approach E-CMN/CSS for speech recognition in car environments", *Proc.IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 450-457, 1997.

[88] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses", *Journal of Acoustical Society of America*, Vol. 97, No. 2, pp.1119-1123, 1995.

[89] T. Takiguchi, S. Nakamura and K. Shikano, "Speech recognition in additive noise and room acoustics distortion by HMM composition", *Technical Report of the Institute of Electronics, Information and Communication Engineers*, SP95-71, pp. 41-46, 1995. (in Japanese)

[90] T. Takiguchi, S. Nakamura and K. Shikano, "Hands-free speech recognition by HMM composition in noisy reverberant environments", *Transactions of the Institute of Electronics, Information and Communication Engineers*, Vol.J79-D-II, No.12, pp. 2047-2053, 1996. (in Japanese)

[91] T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, "Speech recognition by adaptation of model parameters based on HMM decomposition in noisy reverberant environments", *Proc. Acoustical Society of Japan Spring Meeting*, 1-6-17, 1997. (in Japanese)

[92] T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, "Adaptation of model parameters by HMM decomposition in noisy reverberant environments", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 155-158, 1997.

[93] T. Takiguchi, S. Nakamura and K. Shikano, "Model adaptation by HMM decomposition and composition in noisy reverberant environments", *Transactions of the Institute of Electronics, Information and Communication Engineers*, Vol.J81-D-II, No.10, pp. 2231-2238, 1998. (in Japanese)

[94] T. Takiguchi, S. Nakamura and K. Shikano, "Evaluation of model adaptation by HMM decomposition on telephone speech recognition", *Proc.ICSLP-98*, 1998 (to be appeared).

[95] Y. Tohkura, "A weighted cepstral distance measure for speech recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 35, No. 10, pp. 1414-1422, 1987.

[96] M. Tohyama, H. Suzuki and Y. Ando, "The nature and technology of acoustic space", *ACADEMIC PRESS*, 1995.

[97] M. Tonomura, T. Kosaka and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation", *Proc.ICASSP-95*, pp. 688-691, 1995.

[98] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system", *Computer Speech and Language*, Vol.3, No.2, pp. 151-167, 1989.

[99] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise", *Proc.ICASSP-90*, pp. 845-848, 1990.

[100] O. Viikki and K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 107-110, 1997.

[101] O. Viikki, D. Bye and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise", *Proc.ICASSP-98*, pp. 733-736, 1998.

[102] H. Wang and F. Itakura, "An approach of dereverberation using multi-microphone sub-band envelope estimation", *Proc.ICASSP-91*, pp. 953-956, 1991.

[103] T. Yamada, S. Nakamura and K. Shikano, "Robust speech recognition with speaker localization by a microphone array", *Proc.ICSLP*, pp. 1317-1320, 1996.

[104] T. Yamada, S. Nakamura and K. Shikano, "An effect of adaptive beamforming on hands-free speech recognition based on 3-D Viterbi search", *Proc. ICSLP-98*, 1998. (to be appeared)

[105] G. Zavaliagkos, R. Schwartz and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition", *Proc.ICASSP-95*, pp. 676-679, 1995.

[106] E. Zwicker and E. Terhardt, "Analytical expressions for critical bandwidth as a function of frequency", *Journal of Acoustical Society of America*, Vol. 68, No. 5, pp. 1523-1525, 1980.

List of Publications

In the course of this thesis, the following papers have been published:

Journal Papers

- T. Takiguchi, S. Nakamura and K. Shikano, "Model adaptation by HMM decomposition and composition in noisy reverberant environments", *Transactions of the Institute of Electronics, Information and Communication Engineers*, Vol.J81-D-II, No.10, pp. 2231-2238, 1998. (in Japanese)
- T. Takiguchi, S. Nakamura and K. Shikano, "Hands-free speech recognition by HMM composition in noisy reverberant environments", *Transactions of the Institute of Electronics, Information and Communication Engineers*, Vol.J79-D-II, No.12, pp. 2047-2053, 1996. (in Japanese)

International Conference Papers

Selected Conference Papers

- T. Takiguchi, S. Nakamura and K. Shikano, "Evaluation of model adaptation by HMM decomposition on telephone speech recognition", *Proc.Int.Conf. on Spoken Language Processing*, 1998.
- T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, "Model adaptation based on HMM decomposition for reverberant speech recognition", *Proc. IEEE Int.Conf. Acoustics, Speech and Signal Processing*, pp. 827-830, 1997.
- T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, "Adaptation of model parameters by HMM decomposition in noisy reverberant environments", *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 155-158, 1997.

- S. Nakamura, T. Takiguchi and K. Shikano, "Adaptation of model parameters to distorted speech in adverse environment by HMM decomposition", *Proc.ISCIE Int.Symposium on Stochastic System Theory and Its Applications*, 1997.

- S. Nakamura, T. Takiguchi and K. Shikano, "Noise and room acoustics distorted speech recognition by HMM composition", *Proc.IEEE Int.Conf.Acoustics, Speech and Signal Processing*, pp. 69-72, 1996.

Other Conference Papers

- K. Shikano, S. Nakamura, T. Yamada, T. Takiguchi, Eli Yamamoto, M. Inoue, R. Nagai, and T. Aoki, "Hands-free speech recognition and lip-reading/synthesis", *Proc. Int.Workshop on Human Interface Technology*, pp. 47-54, 1997.

- S. Nakamura, T. Yamada, T. Takiguchi, K. Shikano, "Hands-free speech recognition by a microphone array and HMM composition", *Third Joint Meeting of ASA & ASJ*, pp. 1149-1154, 1996.

- S. Nakamura, T. Yamada, T. Takiguchi and K. Shikano, "Hands free speech recognition by a microphone array and HMM composition", *Proc. Int.Workshop on Human Interface Technology*, pp. 33-38, 1995.

Technical Report

- T. Takiguchi, S. Nakamura, K. Shikano, M. Morishima and T. Isobe, "Evaluation of model adaptation by HMM decomposition on telephone speech recognition", *Technical Report of the Institute of Electronics, Information and Communication Engineers*, SP98-28, pp. 39-44, 1998. (in Japanese)

- S. Nakamura , T. Takiguchi and K. Shikano, "A method of reverberation compensation based on short time spectral analysis", *Technical Report of the Institute of Electronics, Information and Communication Engineers*, SP98-25, pp. 17-22, 1998. (in Japanese)

- T. Takiguchi, S. Nakamura and K. Shikano, "Hands-free speech recognition by HMM de-composition in noisy reverberant environments", *Notes of the Information Processing Society of Japan, SIG*, SLP98-20, pp.87-94, 1998. (in Japanese)

- T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, "Speech recognition by adaptation of model parameters based on HMM decomposition in reverberant environments", *Technical Report of the Institute of Electronics, Information and Communication Engineers*, SP96-88, pp. 7-12, 1997. (in Japanese)

- T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, "Adaptation of model parameters by HMM decomposition in reverberant environments", *ATR Technical Report*, TR-IT-0182, 1996.

- T. Takiguchi, S. Nakamura and K. Shikano, "Speech recognition in additive noise and room acoustics distortion by HMM composition", *Technical Report of the Institute of Electronics, Information and Communication Engineers*, SP95-71, pp. 41-46. 1995. (in Japanese)

Meeting of Acoustical Society of Japan (in Japanese)

- S. Nakamura, T. Takiguchi and K. Shikano, "A method of reverberation compensation based on short time spectral analysis", *ASJ spring meeting*, 3-6-11, 1998.

- T. Takiguchi, S. Nakamura and K. Shikano, "Evaluation of model adaptation by HMM decomposition in real environments", *ASJ fall meeting*, 2-Q-27, 1997.

- T. Takiguchi, S. Nakamura, Q. Huo and K. Shikano, "Speech recognition by adaptation of model parameters based on HMM decomposition in noisy reverberant environments", *ASJ spring meeting*, 1-6-17, 1997.

- T. Aoki, T. Yamada, T. Takiguchi, S. Nakamura and K. Shikano, "Speech recognition experiments in real environments using a microphone array and HMM composition", *ASJ fall meeting*, 2-Q-2, 1996.

- T. Takiguchi, S. Nakamura and K. Shikano, "Adaptation of model parameter by HMM composition and decomposition in reverberant environments", *ASJ fall meeting*, 2-Q-9, 1996.

- T. Takiguchi, S. Nakamura and K. Shikano, "Effects of the reverberation time on HMM composition for speech recognition", *ASJ spring meeting*, 1-5-18, 1996.

- T. Takiguchi, S. Nakamura and K. Shikano, "Speech recognition in additive noise and channel distortion by composition of hidden Markov models", *ASJ fall meeting*, 1-2-2, 1995.

Meeting of Information Processing Society of Japan (in Japanese)

- M. Miyamoto, T. Takiguchi, S. Nakamura and K. Shikano, "Effects by noise and handsets in speaker recognition of telephone speech", *IPSJ meeting*, 1998.