Internal Use Only
002

TR-IT-0182

Adaptation of Model Parameters by HMM Decomposition in Reverberant Environments

Tetsuya Takiguchi Qiang Huo

Satoshi Nakamura Kiyohiro Shikano

1996.8.30

Abstract

An HMM composition method has been reported in [5] to cope with the distant-talking speech recognition problem in noisy and/or reverberant environments. In order to estimate the HMM parameters of the acoustic transfer function of the reverberant room, actual impulse responses at different positions of the room have to be measured first. It turns out inconvenient and unrealistic to measure the impulse responses for every possible testing room. In this report, we study a method for estimating the HMM parameters of the acoustic transfer function by using a small amount of adaptation data (possibly derived from actual testing data). The proposed technique is based on HMM decomposition which can be viewed as an inverse process of the HMM composition. We will report some preliminary experimental results to show how it works in a simulated distant-talking speech recognition scenario using a single omni-directional microphone.

Contents

1 Introduction

Recently, the distant-talking speech recognition in a noisy and/or reverberant environment has been becoming increasingly popular due to its potentials in some applications such as hands-free telephony and audio-conferencing. In these applications, the performance of a speech recognizer trained on clean and anechoic speech will degrade drastically due to the mismatch between the training and testing condition caused by multi-path echoes and/or ambient noise interference.

In [5], an HMM composition method has been reported to cope with the distant-talking speech recognition problem in a small but noisy room. The environmental model schematically shown in Figure 1.1 is adopted for the development of the algorithm. More specifically, in [5], the HMM composition framework developed for coping with additive noise in [2, 4] is extended to handle both the additive noise and the convolutional distortion caused by room reverberation. Apart from a set of clean speech HMMs and a noise HMM, a separate HMM is adopted to model the acoustic transfer function of the reverberant room. These three sets of HMMs are then composed to create a set of composed-HMMs to recognize the noisy and acoustically distorted speech. It was shown that this HMM composition technique can improve somehow the performance for the noisy distant-talking speech recognition. However, in [5], in order to estimate the HMM parameters of the acoustic transfer function, actual impulse responses at different positions of the experimental room have to be measured first. It turns out inconvenient and unrealistic to measure the impulse responses for every possible testing room in practice.

In this report, we study a method for estimating the HMM parameters of the acoustic transfer function by using a small amount of adaptation data (possibly derived from actual testing data). The proposed technique is based on HMM decomposition which can be viewed as an inverse process of the HMM composition. As a first step, we will constrain ourselves in this study to the case in which a single omni-directional microphone is used to pick up the speech signal and the background noise is negligible.

The rest of the report is organized as follows. In Section 2, the algorithm to estimate the HMM parameters of the acoustic transfer function is described. In Section 3, some preliminary experimental results are reported. Finally, we conclude the report in Section 4.

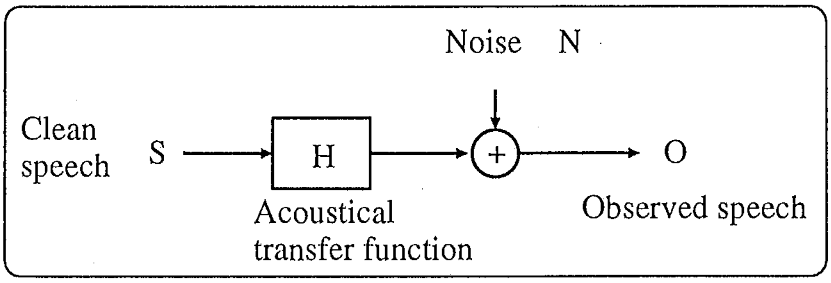

Figure 1.1: An environment model for noisy and acoustically distorted speech

2 Model Adaptation

2.1 Speech recognition method based on HMM composition

In this subsection, we briefly describe the HMM composition method used in [5] for the noisy & distorted speech recognition. The HMM composition method is applicable if two stochastic information sources are additive. Suppose a noisy and acoustically distorted speech observation is characterized by the following equation

$$O_{cep}(t) = \cos^{-1}[\ \log\{\ \exp(\ \cos(S_{cep}(t) + H_{cep}(t))\ ) + N_{\omega}(t)\ \}],$$

where $O_{cep}(t)$, $S_{cep}(t)$, $H_{cep}(t)$ denote the cepstral vectors corresponding to respectively the observed signal, the clean speech signal, the acoustic transfer function; and $N_{\omega}(t)$ denotes the observation of the noise signal in the linear spectral domain. The HMM composition is first carried out in cepstral domain to combine $S_{cep}(t)$ and $H_{cep}(t)$, then the resultant models are further combined with that of $N_{\omega}(t)$ in linear spectral domain, and finally the models are transformed back to the representations in cepstral domain. The number of states of the composed HMM is the product of those numbers of individual compositional models. The same fact applies to the number of the possible state transitions and the number of the mixture components for each state. The state observation probability density function (PDF) of the composed HMM can be obtained theoretically by the convolution of two compositional PDFs, but in practice some approximation methods have to be used [2].

2.2 Estimation of acoustic transfer function based on HMM decomposition

The noise HMM parameters can be estimated by using the signals during the noise periods. Given (1) model structures of a noise HMM and an acoustic transfer function HMM (for simplicity, we call it channel HMM), (2) composed-HMM parameters estimated from some adaptation data collected in the testing environment, (3) the noise HMM parameters, the composite HMM of the clean speech HMM and the channel HMM can be obtained by model decomposition which is an inverse process of the model composition. As a first step, we focus our study here to the case of reverberation distorted speech only. The environmental model is schematically shown in Figure 2.2. We denote the clean speech HMM as M_S, the channel HMM as M_H, the composed HMM from M_S and M_H as M_{SH}. In our experiments, a tied-Gaussian-mixture HMM is used to model each basic speech unit (phone). For simplicity, it is also assumed that M_H has only one state with a single Gaussian PDF. All the involved Gaussian PDFs are assumed to have the diagonal covariance matrices so that all of the following discussions are formulated as a one-dimensional problem. Given some adaptation data, the channel HMM parameters (mean μ and variance σ) can then be estimated as follows:

1. For each speech unit l with adaptation data, starting from $M_S^{(l)}$, re-estimate/adapt its parameters and denote the resultant model as $\hat{M}_{SH}^{(l)}$.

2. Decompose $\hat{M}_H^{(l)}$ from $\hat{M}_{SH}^{(l)}$:
$$\hat{M}_H^{(l)} = \hat{M}_{SH}^{(l)} \ominus M_S^{(l)},$$
where $\ominus$ denotes the decomposition operation which is an inverse process of the composition [5].

3. Compute

$$\bar{\mu}_j^{(l)} = \sum_{k=1}^{K} \lambda_{j,k}^{(l)} \hat{\mu}_{j,k}^{(l)}$$

$$\bar{\sigma}_j^{(l)} = \sum_{k=1}^{K} \lambda_{j,k}^{(l)} \hat{\sigma}_{j,k}^{(l)} + \sum_{k=1}^{K} \lambda_{j,k}^{(l)} (\hat{\mu}_{j,k}^{(l)} - \bar{\mu}_j^{(l)})^2,$$

   where $\hat{\mu}_{j,k}^{(l)}$, $\hat{\sigma}_{j,k}^{(l)}$ are means and variances of kth mixture component in jth state of lth model $\hat{M}_H^{(l)}$; $\lambda_{j,k}^{(l)}$'s are the corresponding mixture coefficients in model $M_S^{(l)}$; K is the total number of Gaussian PDFs shared by all of the states.

4. μ and σ are obtained respectively as an average of the $\bar{\mu}_j^{(l)}$'s and $\bar{\sigma}_j^{(l)}$'s over j and l.

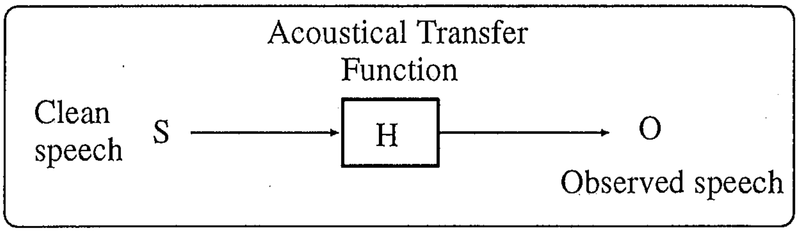

Figure 2.2: An environment model for acoustically distorted speech

3 Experiment

3.1 Experimental setup

To examine the viability and the characteristics of the proposed method, a series of comparative experiments are conducted. Figure 3.3 shows a top view of the experimental room. The sound signal is captured by using a single omni-directional microphone. We measured 4 impulse responses corresponding to 4 sound source positions by using the method reported in [9]. As an example, the impulse response of position p1 is shown in Figure 3.4. The experimental room has a reverberation time of approximately 180msec.

Two speech corpora are used for evaluation. One is the A-set of the ATR Japanese speech database. Another is the ASJ continuous speech database. The former contains word utterances and the latter contains sentence utterances, both spoken by announcers. A set of speaker independent (SI) models are trained by using utterances from 64 speakers in the ASJ database. Another set of speaker dependent (SD) models are trained by using 2620 words of one male speaker from the ATR database. These two sets of models are clean speech models. For testing, we choose 500 words from the same male speaker but different from those words in SD training. The adaptation words are also selected from those 500 words and consequently are excluded from the real testing set. The related information of the adaptation data used in the following experiments are listed in Table 3.1. The test and adaptation data are simulated by linear convolution of the clean speech signal and the measured impulse responses.

We choose 54 context independent phones as the basic speech units. Each speech unit is modeled by a single left-to-right 5-state tied-mixture HMM with 3 self-transition loops and without state skipping. There are in total 256 Gaussian mixture components with diagonal covariance matrices shared by all of the models. Each feature vector used in this study consists of 16 mel-frequency cepstral coefficients (MFCCs). The feature analysis conditions are detailed in Table 3.2. A single Gaussian PDF is used to model the acoustic transfer function for each position.

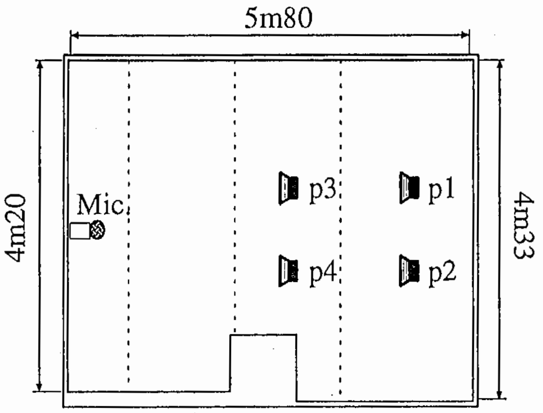

Figure 3.3: A top view of the experimental room

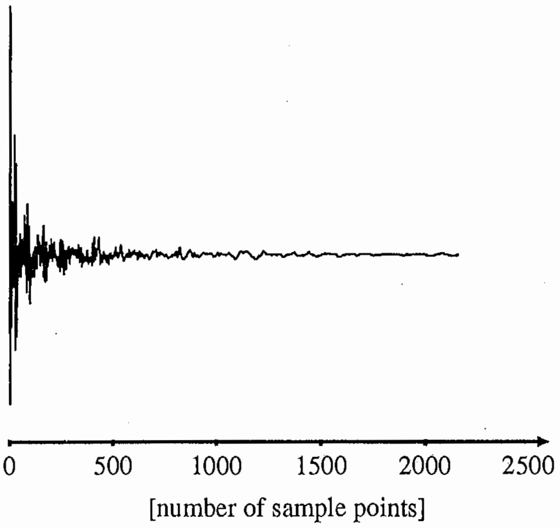

Figure 3.4: Impulse response corresponding to position p1

Table 3.1: Adaptation data information (file name and content)

1-word	MHT10002.LB				
	“ai”				
2-word	MHT10002.LB	MHT12500.LB			
	“ai”	“jouyaku”			
3-word	MHT10002.LB	MHT11660.LB	MHT13320.LB		
	“ai”	“kozeni”	“tsuyoi”		
4-word	MHT10002.LB	MHT11250.LB	MHT13000.LB	MHT13750.LB	
	“ai”	“giNmi”	“tateru”	“nyuuyoku”	
5-word	MHT10002.LB	MHT11000.LB	MHT12000.LB	MHT13000.LB	MHT14000.LB
	“ai”	“gaijiN”	“shiki”	“tateru”	“haN”

Table 3.2: Feature analysis condition

sampling freq.	12kHz
frame shift	8msec
window length	32msec
window type	Hamming
pre-emphasis coef.	0.97
feature vector	16-order MFCC

3.2 Recognition of distorted speech

3.2.1 Results for speaker dependent model

Table 3.3∼3.6 show the testing results (word accuracy in %) at testing positions of p1∼p4 by using different compensation techniques with SD initials and different amount of adaptation data. Only mean vectors are adapted while other model parameters are fixed. "Model-MAP" stands for the case of the MAP (maximum *a posteriori*) adaptation of the model parameters [3] without any HMM composition/decomposition involved. "Decom-ML" and "Decom-MAP" refer to the cases where the proposed acoustic transfer function estimation method is applied and model composition technique is used for recognition. The former corresponds to the case in which ML (maximum likelihood) estimate is used in "step 1" of the channel HMM parameter estimation procedure described in section 2.2, and the latter represents its MAP counterpart.

Without any compensation, the recognition rate with the initial HMMs (clean speech HMMs) is 72.8% at testing position p1. The conventional MAP model adaptation technique is not sufficient when only a small amount of adaptation data are available (e.g., with 5 adaptation words, the performance is improved to 76.2%). By using the proposed techniques (model decomposition for channel parameter estimation and model composition for constructing recognition models), with only 2 adaptation words, the performance is raised to 81.6% in "Decom-ML" case, but this is still far away from the matched-condition (SD HMMs are trained by using the simulated distorted training speech at the testing position) performance of 94.6%. In this case, MAP dose not help so much (no big difference between "Decom-ML" and "Decom-MAP"). The similar facts as the above are also observed in cases corresponding to other three positions of p2, p3 and p4 which have the matched-condition recognition rates of 96.6%, 96.8% and 97.2% respectively.

Due to its heuristic nature of the current formulation, in model decomposition process, we can not guarantee to obtain a variance estimation with positive values. Consequently, in its present implementation, although we formulate the problem at the "model level", i.e., using a single Gaussian to model the acoustic transfer function, we did not apply the decomposition/composition to variance parameters. This will make the model composition technique degenerate to be equivalent to a simple technique of removing a single cepstral bias (i.e., for each testing position, a single bias is estimated) from each frame of the distorted speech (in cepstral domain). The proposed technique provides one of the possibilities to estimate such a bias vector. It will thus become interesting to compare the performance of the proposed technique with that of a simple bias removal technique called cepstral mean subtraction (CMS) [1]. Table 3.7 shows this comparison. In CMS-based testing case (labeled as "CMS"), the SD recognition models are trained by using CMS-processed clean speech data. In "Clean Speech HMM" case, the results are obtained by using the models trained on clean speech (without CMS processing). "Decom-ML" corresponds to the case of using the proposed method with 2 adaptation words. "0msec" refers to the case of testing speech being clean and anechoic. The experimental results clearly show that the simple CMS technique dose not work in reverberation-distorted speech recognition, especially with such a long reverberation time as 180msec in this study. On the other hand, the proposed technique is able to improve the performance somehow.

Table 3.3: Recognition rates (in %) at testing position p1 (SD initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	72.8	72.2	74.0	69.0	74.8	76.2
Decom-ML	72.8	81.6	81.6	78.8	81.2	82.4
Decom-MAP	72.8	79.2	79.6	78.2	80.2	81.8

Table 3.4: Recognition rates (in %) at testing position p2 (SD initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	76.6	75.8	76.8	67.2	78.4	75.6
Decom-ML	76.6	82.8	85.8	81.6	83.2	84.2
Decom-MAP	76.6	82.0	82.8	80.0	82.6	84.0

Table 3.5: Recognition rates (in %) at testing position p3 (SD initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	81.2	81.8	77.0	71.8	80.0	72.2
Decom-ML	81.2	83.6	87.4	86.4	86.0	87.2
Decom-MAP	81.2	85.8	86.8	85.8	85.4	86.0

Table 3.6: Recognition rates (in %) at testing position p4 (SD initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	80.6	79.6	79.2	70.8	79.6	75.0
Decom-ML	80.6	81.6	87.2	84.4	85.2	85.2
Decom-MAP	80.6	83.6	86.0	84.2	85.6	85.4

Table 3.7: Performance comparison (word accuracy in %) between CMS and the proposed method

reverberation	0msec	180msec			
		p1	p2	p3	p4
Clean Speech HMM	96.6	72.8	76.6	81.2	80.6
CMS	95.8	73.8	77.2	75.0	77.0
Decom-MAP	N/A	81.6	85.8	87.4	87.2

3.2.2 Results for speaker independent model

We repeat all of the experiments in SI initial model case. Table 3.8~3.11 show the testing results (word accuracy in %) at testing positions of p1~p4 by using different compensation techniques with SI initials and different amount of adaptation data. Most of the observations are repeated here. But in SI case, overall, "Decom-MAP" achieves a slightly better performance than that of "Decom-ML".

Table 3.8: Recognition rates (in %) at testing position p1 (SI initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	68.6	61.0	63.8	56.4	62.4	56.6
Decom-ML	68.6	60.0	64.0	65.0	66.6	69.8
Decom-MAP	68.6	68.2	67.2	67.8	68.2	71.0

Table 3.9: Recognition rates (in %) at testing position p2 (SI initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	66.6	61.0	64.8	58.8	66.6	62.4
Decom-ML	66.6	62.8	67.6	70.4	72.2	73.2
Decom-MAP	66.6	66.0	67.6	69.8	71.0	71.4

Table 3.10: Recognition rates (in %) at testing position p3 (SI initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	71.2	66.0	69.4	61.4	71.0	60.4
Decom-ML	71.2	67.0	69.2	69.2	68.4	72.8
Decom-MAP	71.2	72.2	72.6	70.8	70.8	73.2

Table 3.11: Recognition rates (in %) at testing position p4 (SI initial)

No. of adapt. data	0-word	1-word	2-word	3-word	4-word	5-word
Model-MAP	71.0	67.6	75.2	66.4	69.0	66.4
Decom-ML	71.0	65.2	68.8	69.8	72.4	74.6
Decom-MAP	71.0	71.6	72.4	70.4	73.4	75.4

4 Conclusion and Discussion

In this report we investigate a new method to cope with the problem of distant-talking speech recognition in reverberant environments. The preliminary experimental results show that the proposed method can improve somehow the reverberation-distorted speech recognition performance in comparison with that of using a speech recognizer trained on clean and anechoic speech. It is also found that the simple CMS technique does not work in this case, especially when the reverberation time is relatively long. As a future work, we will examine the effects of other channel compensation techniques. One possibility is to adopt the model-space stochastic matching technique which is more theoretically sound and flexible to include multiple biases removal, variance compensation, etc. [8]. Another possibility is to use the so-called signal conditioning techniques [6, 7].

Distant talking speech recognition is an important research topic with great potential. The existing techniques in the literature (of course including the one presented in this report) are far from satisfaction. There are many fundamental problems need to be addressed and carefully studied. For example, the acoustic impulse response of a room can be far greater than the analysis interval of 10-30msec used in most of the current speech recognition systems. This suggests the need for feature analysis and model compensation procedures that act over much longer intervals than the traditional assumptions of the short-time stationarity of the speech signal. When more than one sources of distortion exist, e.g., both additive and convolutional distortions, the problem becomes more difficult. Some new theoretical frameworks may be required to directly take into account the nonlinear interaction between different types of distortions. When the distortion sources are nonstationary, e.g., a moving speaker and a nonstationary ambient noise, some adaptive compensation techniques are needed. To enhance the efficacy and the effectiveness of the compensation, those techniques are mostly wanted to be able to better characterize the distribution of the possible distortion types, and use this distribution to choose the appropriate compensation model. We are working along these lines of thoughts.

Acknowledgement

The authors would like to thank Dr. Y. Yamazaki, President, ATR Interpreting Telecommunications Research Laboratories, and Dr. Y. Sagisaka, Head, Department 1 of the ATR-ITL, for their support of this collaborative work.

References

[1] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, Vol. 55, pp.1304-1312, 1974.

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

[3] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. on Signal Processing*, Vol. 39, No. 4, pp.806-814, 1991.

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

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

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

[7] M. G. Rahim, B.-H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Letters*, Vol. 3, No. 4, pp.107-109, 1996.

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

[9] 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," *J. Acoust. Soc. Amer.*, Vol. 97, No. 2, pp.1119-1123, 1995.