社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（20）
2007／12／20

Aspectモデルを用いた話者と環境適応音声認識システムの検討

咸 聖俊† 鈴木 基之† 伊藤 彰則† 牧野 正三†

東北大学大学院工学研究科 〒980-9579 仙台市青葉区荒巻字青葉 6-6-05

E-mail: † {branden65, moto, aito, makimo}@makino.ecei.tohoku.ac.jp

あらまし 適応アルゴリズムで重要な点の１つとして，少ない適応データを利用して多くのパラメタを推定することがある．話者適応では，少量の発話を話者独立システムに組み込んで，その性能を話者依存システムに近づけることを目指す．本研究では，aspect model に基づく音響モデルを用いて話者適応を行うことを目指す．言語モデルに用いられる PLSA と類似の方法を音響モデルについて定式化し，これを用いて話者適応を行った．提案法を代表的な話者適応方法である MAP 推定法と MLLR 法と比較検討し，孤立単語認識での結果を比較分析する．

キーワード 話者適応，潜在モデル，PLSA，話者独立モデル，話者依存モデル

A Study on the Environment and Speaker Adaptation System using Aspect model

Seongjun HAHM† Motoyuki SUZUKI† Akinori ITO† and Shozo MAKINO†

† Graduate School of Engineering, Tohoku University 6-6-05 Aramaki Aza-Aoba, , Aoba-ku, Sendai, 980-9579 Japan

E-mail: † {branden65, moto, aito, makimo}@makino.ecei.tohoku.ac.jp

Abstract One of the key issues for adaptation algorithms is to modify a large number of parameters with only a small amount of adaptation data. Speaker adaptation techniques try to obtain near speaker dependent (SD) performance with only small amounts of specific data and are often based on initial speaker independent (SI) recognition systems. In this paper, we introduce an aspect model into an acoustic model for rapid speaker and environment adaptation. A formulation of probabilistic latent semantic analysis (PLSA) is extended to continuous density HMM. We carried out isolated word recognition experiment, and the results was compared to that of MAP and MLLR.

Keyword Speaker Adaptation, Aspect Models, PLSA, SD Models, SI Models

1. Introduction

When developing a speaker independent (SI) automatic speech recognition system, it is important to account for the wide variability that can be present in any speech waveform. This variability can result from changes in the individual speaker, the speaker's environment, the microphone and channel of the recoding device. Over the last 10-20 years, dramatic improvements in the quality of SI speech recognition technology have been made. With the development and refinement of the Hidden Markov Model (HMM) approach [1], today's speech recognition systems have been shown to work effectively on various large vocabulary, continuous speech, SI tasks. However, despite the high quality of today's SI systems, there can still be a significant gap in performance between these systems and their speaker adaptive (SA) or speaker dependent (SD) counterparts. The reduction in a system's error rate between its SI mode and its speaker dependent mode can be more than 50% [2].

One of the problems to be faced in adaptation is how to adapt a large number of parameters with only a small amount of data. Techniques that only update distributions for which observations occur in the adaptation data, such as those using maximum a posteriori (MAP) adaptation [3], require a relatively large amount of adaptation data to be effective. An alternative approach is to estimate a set of transformations that can be applied to the model parameters. If these transformations can capture general relationships between the original model set and the current speaker or new acoustic environment, they can be effective in adapting all the HMM distributions. One such transformation approach is maximum likelihood linear regression (MLLR) [4] which estimates a set of linear transformations for the mean parameters of a mixture Gaussian HMM system to maximize the likelihood of the adaptation data. It should be noted that while MLLR was initially developed for speaker adaptation, since it reduces the mismatch between a set of models and adaptation data, it can also be used to perform environmental compensation by reducing a mismatch due to channel or additive noise effects.

A method of providing speaker constraint to speech recognition systems that has proven successful is hierarchical speaker clustering [5]. In this approach, similar reference speakers are grouped together into a speaker cluster for which one model is trained.

In this paper, we will examine the Bayesian adaptation method that exploits an aspect model, which is "a mixture of mixture model." We, then, formulate and discuss the potential of the techniques using probabilistic latent semantic analysis (PLSA) [6]. Finally we will show the

experimental results using MAP, MLLR, and the proposed model.

2. Speaker Adaptation Methods

Speaker dependent models usually perform better than speaker independent models. Speaker adaptation refers to the set of techniques that try to modify speaker independent model to approximate speaker dependent models. Here two important Bayesian adaptation methods, Maximum a posteriori Probability (MAP) and maximum likelihood linear regression (MLLR), are briefly explained.

2.1. Bayesian Adaptation

2.1.1. Maximum a Posteriori Estimation

The parameters of most of the speech recognition systems using existing Hidden Markov Models (HMMs) are estimated using Maximum Likelihood Estimation (MLE). In this case, for re-estimation of new speaker utterance all training samples are needed. However, MAP method combines the distributions of adaptation data and existing ones.

The formula for MAP is as follows:

$$\hat{\mu}_{new} = \frac{\mu_{adp\_data} + \tau \times \mu_{ori}}{N_{adp\_data} + \tau} \quad (1)$$

where $\hat{\mu}_{new}$ is updated data, μ_{adp_data} is the mean of adaptation data, μ_{ori} is original mean, N_{adp_data} is the number of available adaptation data and τ is control variable decided empirically. In equation 1, we can see if $\tau \to 0$ the updated mean is dependent on adaptation data. If $\tau \to \infty$, the updated mean keeps the original mean. The MAP method can be seen as finding optimal combination of existing data and adaptation data [3].

2.1.2. Maximum Likelihood Linear Regression

Maximum Likelihood Linear Regression (MLLR) adjusts model parameters using a transformation shared globally or across different units within a class.

Global mean vector scaling, rotation and translation are as follows:

$$\hat{\mu}_{new} = \hat{W}\mu_{ori} + B \quad (2)$$

where $\hat{\mu}_{new}$ is an updated mean, μ_{ori} is an original mean, $\hat{W}$ is a regression matrix, and B is a bias term. The detailed method for calculating regression matrix including a simple example can be found from the paper of J.E. Hamaker [7].

3. Formulation of PLSA in Acoustic Model

3.1. Probabilistic Latent Semantic Analysis

The basic idea of the Latent Semantic Analysis (LSA) is to map high-dimensional count vectors, such as the ones arising in vector space representations of text documents, to a lower dimensional representation in a so-called latent semantic space. The goal of LSA is to find a data mapping which provides information well beyond the lexical level and reveals semantical relations between the entities of interest. PLSA is the probabilistic approach compared to LSA. PLSA is based on a mixture decomposition derived from a latent class model. Figure 2 shows the graphical representation of the PLSA.

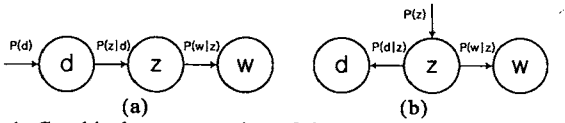


Fig 1. Graphical representation of the aspect model in the asymmetric (a) and symmetric (b) parameterization.

The model of Figure 1(a) is represented by following expression.

$$P(d,w) = p(d)P(w|d),\ P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \quad (3)$$

In this formula, w is a word, d is a document, and z is a latent class. The model of Figure 1(b) is defined by following expression using Bayes' rule.

$$P(z|d) = \frac{P(d|z)P(z)}{P(d)} \quad (4)$$

$$P(d,w) = \sum_{z \in Z} P(z)P(d|z)P(w|z) \quad (5)$$

The standard procedure for maximum likelihood estimation in latent variable models is the Expectation Maximization (EM) algorithm. E-step equation

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')} \quad (6)$$

as well as the following M-step formulae

$$P(w|z) = \frac{\sum_{d} n(d,w)P(z|d,w)}{\sum_{d,w'} n(d,w')P(z|d,w')} \quad (7)$$

$$P(d|z) = \frac{\sum_{w} n(d,w)P(z|d,w)}{\sum_{d',w} n(d',w)P(z|d',w)} \quad (8)$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z|d,w),\ R \equiv \sum_{d,w} n(d,w) \quad (9)$$

3.2. Formulation in Acoustic Model

Figure 2 shows the transition from language model to acoustic model. In language model d means document.

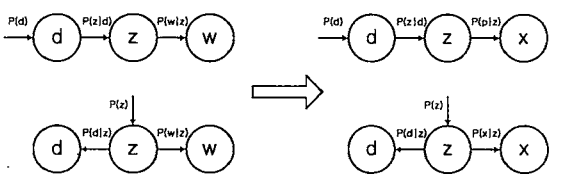


Fig 2. Transition from language models to acoustic models

The d of acoustic model means an environment and x means the speech vector for adaptation. This transition can be represented by following formulae.

$$P(d,x) = p(d)P(x|d),\ P(x|d) = \sum_{z \in Z} P(x|z)P(z|d) \quad (10)$$

$$P(z|d) = \frac{P(d|z)P(z)}{P(d)} \quad (11)$$

$$P(d,x) = \sum_{z \in Z} P(z)P(d \mid z)P(x \mid z) \quad (12)$$

E-step equation for acoustic model is formulated as follows:

$$P(z \mid d,x) = \frac{P(z,d,x)}{\sum_{z'} P(z',d,x)} = \frac{P(z)P(d \mid z)P(x \mid z)}{\sum_{z'} P(z')P(d \mid z')P(x \mid z')} \quad (13)$$

where x is a observed feature parameter for adaptation, d is a certain environment, and z is the a latent class. The $n(d,w)$ of the PLSA is changed into $P_e(x \mid d)$.

$$P_e(x \mid d) = w(x,d)P(x \mid \lambda_d) \quad (14)$$

In this expression $w(x,d)$ means the weighting in the environment, d, and the feature parameter, x. λ_d is the model under the environment, d.

Followings are M-step formulae. In these formulae, x_i^d is the ith feature vector of a phone, x, under the environment, d, and $P_e(\cdot)$ means the probability on empirical models.

$$P(x_i \mid z) = \frac{\sum_d P(z,d,x_i)}{\sum_{d,x'} P(z,d,x')} = \frac{\sum_d P_e(x_i \mid d)P(z \mid d,x_i)}{\int_x \sum_d P_e(x \mid d)P(z \mid d,x)dx} \approx \frac{\sum_d P_e(x_i \mid d)P(z \mid d,x_i)}{\sum_d \sum_i P_e(x_i^d \mid d)P(z \mid d,x_i^d)} \quad (15)$$

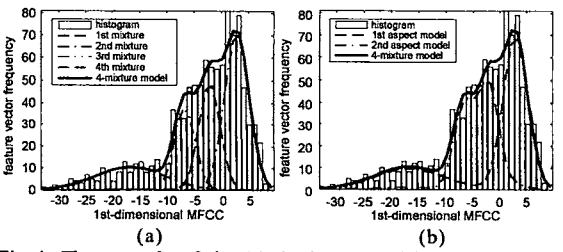
$$P(d \mid z) = \frac{\sum_i P_e(x_i^d \mid d)P(z \mid d,x_i^d)}{\sum_d \sum_i P_e(x_i^d \mid d)P(z \mid d',x_i^d)} \quad (16)$$

$$P(z) = \frac{\sum_d \sum_i P_e(x_i^d \mid d)P(z \mid d,x_i^d)}{\sum_z \sum_d \sum_i P_e(x_i^d \mid d)P(z \mid d',x_i^d)} \quad (17)$$

The formula for speech recognition is as follows:

$$P(x_i \mid \lambda) = \sum_z P_e(x_i \mid d)P(z)P(d \mid z) = \sum_z P_e(x_i \mid d)P(z)\frac{\sum_i P_e(x_i^d \mid d)P(z \mid d,x_i^d)}{\sum_d \sum_i P_e(x_i^d \mid d)P(z \mid d',x_i^d)} \quad (18)$$

This model can be viewed as the combination of the methods, Reference Speaker Weighting (RSW) and Speaker Cluster Weighting (SCW). RSW is an interpolation of models from "reference speakers." SCW is the cluster's mixture models. $P(d \mid z)$ can be thought of as weighting of reference speakers and $P(z)$ could be referred to weighting of speaker clusters if we assume that $P(d \mid z)$ is the speaker cluster.



(a) (b)

Fig 4. The example of the (a) 4-mixture model and (b) aspect models.

Figure 4 shows the example of 4-mixture model and aspect models. PLSA can be viewed as one of the method for mixture decomposition. After decomposition, original distribution is not changed.

After training using PLSA we can get $P(z)$ and $P(d \mid z)$. Figure 5 illustrates the example of calculating aspect models. Firstly, $P(d \mid z)$ is multiplied to each mixture and ten summed. After multiplying $P(z)$, we can obtain each aspect model which can be viewed as a mixture of mixture models.

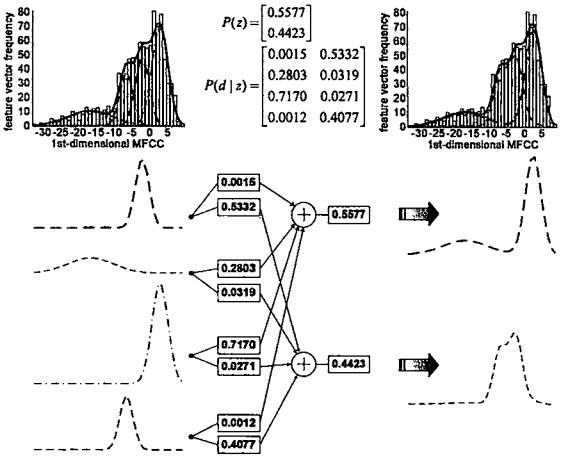


Fig 5. The example of calculating aspect models

Figure 6 shows a comparison of adaptation procedure using mixture and aspect models. Here $P(z)$, the weighting of each aspect model, is the unit for adaptation whereas the unit of each mixture is the unit for adaptation in mixture models. For adaptation of the aspect models for new distributions, EM algorithm is used on the $P(z)$.

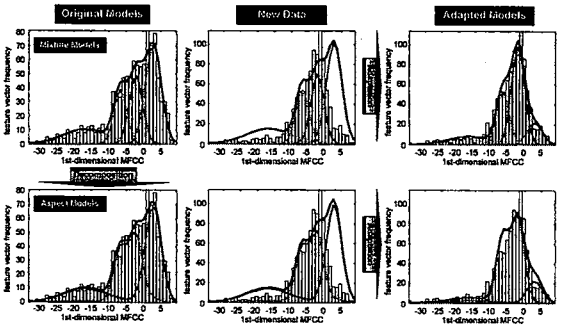


Fig 6. A comparison of adaptation procedure using mixture and aspect models

4. Experimental Results

We used the Korean isolated word databases, such as KLE452 databases. KLE452 consists in recording 70 cooperative speakers (i.e., 38 males only). 35 males are used for training. For testing and adaptation, 3 speakers are used. For adaptation, MAP and MLLR are used in addition to the proposed method. The model of each speaker has 3 states and 1 mixture. Using the model of each speaker we made 35-mixture mono-phone model which is the speaker independent model for speaker adaptation. The experimental results are shown in figure 7.

All models have 3 states and 35 mixtures. Using existing adaptation schemes, MAP and MLLR, we could obtain improved word recognition rate but the proposed method not. In these experiments, the results of MAP are a little bit better than those of MLLR. It seems that the global transformation matrix for all distribution is one of the reasons for the results.

Table 1. Analysis Conditions

Feature extraction method	MFCC
Sampling Rate	16kHz
Pre-emphasis coefficient	0.97
Window	Hamming
Frame length	25ms
Frame Shift	10ms
Cepstral vector dimension	39
Cepstral Mean Normalization	not used

Table 2. Database and Model

Database	KLE452 (38 speakers/ Phone-balanced 452 isolated words
Training data	1st utterances of 35 speakers/ 452 words
Adaptation data	2nd utterances of 3 test speakers/ 1, 5, 10, 50, 100, 150, 200 words
Test data	1st utterances of 3 test speakers/ 452 words
Model type	Mono-phone, 3 states, 35-mixture left-to-right HMM

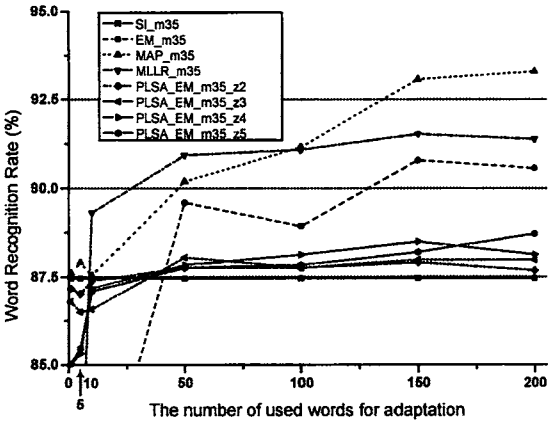


Fig 7. Fixed word recognition rate (%)

5. Conclusion

In this paper, we formulated Probabilistic Latent Semantic Analysis (PLSA) as an acoustic model and evaluated the performance. We expected that PLSA have the power of solving these kinds of problems and modeling effectively. But the word recognition results are not good. It seems that using only weighting values by PLSA based on our formulation is not effective. Future work is expected to perform experiments using Tempered EM algorithm.

Reference

[1] Bahl, L., Jelinek, F., Mercer, R., "A maximum likelihood approach to continuous speech recognition,: IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-5 (2), 179-190, 1983.

[2] T. Hazen, "A comparison of novel techniques for rapid speaker adaptation," Speech Communication, May 2000.

[3] Gauvain, J., Lee, C., "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," IEEE Transactions on Speech and Audio Processing 2 (2), 291-298, 1994.

[4] Leggetter, C., Woodland, P., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language 9 (2), 171-185, 1995.

[5] Furui, S., "Unsupervised speaker adaptation method based on hierarchical spectral clustering," In Proceedings of the ICASSP, 286-289, 1989.

[6] Thomas Hoffman, "Probabilistic latent semantic analysis," in proc. of the 15th Conference on Uncertainty in AI, 1999.

[7] J.E. Hamaker, MLLR: A speaker adaptation technique for LVCSR, Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, November 1999.