Adaptive Stereo-based Stochastic Mapping

Shay Maymon, Pierre Dognin, Xiaodong Cui, and Vaibhava Goel

IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

{maymon, pdognin, cuix, vgoel}@us.ibm.com

Abstract

Stereo-based stochastic mapping (SSM) is a technique based on constructing a Gaussian mixture model for the joint distribution of stereo data. This paper considers the use of SSM for noise robust speech recognition, in which clean and noisy speech features form the stereo data. The Gaussian mixture model, whose parameters are estimated from the observed stereo features during training time, is then used in test time to predict the clean speech from its noisy observation. This paper proposes to leverage the noisy speech observation for updating the model parameters during test time, and thus improve the prediction of the clean speech from its noisy observation. Specifically, an expectation-maximization procedure is developed for adapting the model parameters during test time. This adaptation is especially important when there is a mismatch between the training and testing sets, or when the size of the training set is relatively small, resulting in a poor estimation of the parameters. The proposed method is tested on a noise robustness task and is shown to improve the performance achieved by SSM.

Index Terms: speech recognition, noise robustness, stereo feature, stochastic mapping.

1. Introduction

The performance of speech recognition systems degrades significantly when operated in noisy conditions. Consequently, noise robustness has become an active research area in speech recognition. SSM is a recent development in the family of stereo-based algorithms ([1],[2],[3]) used for noise robustness. Introduced in [4], it uses a maximum a posteriori (MAP) estimate of clean speech given noisy speech observation, and was further extended to a minimum mean square error (MMSE) estimate in [5]. SSM finds applications in a wide variety of areas such as noise robust automatic speech recognition (ASR), automatic bandwidth extension, and voice conversion.

SSM learns the statistical relationship between clean and noisy speech signals by modeling their joint distribution as a Gaussian mixture model (GMM). This GMM is obtained from a multi-conditional training scenario in which the noisy channel of the training data consists of speech signals from various types of noise and signal-to-noise ratios (SNRs).

In this paper, we explore a stochastic mapping approach for noise robust speech recognition based on stereo features under the MMSE criterion. We propose a maximum a posteriori approach to adapt the model parameters during test time, where the estimate of the model parameters obtained during training time serves as a prior information. This approach will be referred to as adaptive SSM (A-SSM). To the extent of our knowledge of prior works, the GMM model is always considered as static at test time.

The rest of the paper is organized as follows: in Section 2 we formulate the mathematical model of stereo features. Section 3 develops the MMSE estimate of the clean speech feature vectors given the noisy speech feature vectors. Maximum likelihood and maximum a posteriori estimation of the model parameters are considered in Section 4 and in Section 5, respectively. Experimental results are discussed in Section 6 and conclusions are given in Section 7.

2. Mathematical Model

A stereo feature $\underline{z}_i$ is defined as the concatenation of the clean speech feature vector $\underline{x}_i$ and the corresponding noisy feature vector $\underline{y}_i$, i.e.,

$$\underline{z}_i = \begin{bmatrix} \underline{x}_i \\ \underline{y}_i \end{bmatrix}. \tag{1}$$

We assume that the stereo feature vectors $\{\underline{z}_i\}$ are independent and identically distributed (i.i.d.). We also assume that each $\underline{z}_i$ is distributed according to a GMM distribution with K mixture components, i.e.,

$$f_{\underline{z}_i}(\underline{z}_i;\underline{\theta}) = \sum_{k=1}^{K} w_k \mathcal{N}(\underline{z}_i;\underline{\mu}_{z,k},\Sigma_{zz,k}). \tag{2}$$

The vector $\underline{\theta}$ denotes the model parameters, which consists of the mixture weights $\{w_k\}_{k=1}^K$ such that $\sum_{k=1}^K w_k = 1$, the mean $\underline{\mu}_{z,k}$ and the covariance $\Sigma_{zz,k}$ of each component, where

$$\underline{\mu}_{z,k} = \begin{bmatrix} \underline{\mu}_{x,k} \\ \underline{\mu}_{y,k} \end{bmatrix}, \tag{3}$$

and

$$\Sigma_{zz,k} = \begin{bmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{bmatrix}. \tag{4}$$

3. MMSE Estimation

This section addresses the problem of estimating the clean speech feature vectors $\{\underline{x}_i\}$ given the observed noisy speech feature vectors $\{\underline{y}_i\}$. The MMSE estimate of $\underline{x}_i$ based on $\underline{y}_i$ is given by the conditional expectation:

$$\hat{\underline{x}}_i = \mathrm{E}(\underline{x}_i|\underline{y}_i;\underline{\theta}), \tag{5}$$

which takes a simple form when the pair $(\underline{x}_i,\underline{y}_i)$ is drawn from a joint GMM distribution. Specifically, using the law of total expectation, we obtain

$$\begin{aligned} \hat{\underline{x}}_i &= \mathrm{E}\left(\mathrm{E}(\underline{x}_i|\underline{y}_i,l_i;\underline{\theta})|\underline{y}_i;\underline{\theta}\right) \\ &= \sum_{k=1}^{K} p(l_i=k|\underline{y}_i;\underline{\theta}) \cdot \underline{\mu}_{\underline{x}_i|\underline{y}_i,k} \end{aligned} \tag{6}$$

where l_i indicates the component density which generated $\underline{z}_i$, and $\underline{\mu}_{\underline{x}_i|\underline{y}_i,k}$ represents the following conditional expectation

$$\underline{\mu}_{\underline{x}_i|\underline{y}_i,k} = \mathrm{E}(\underline{x}_i|\underline{y}_i, l_i = k; \underline{\theta}). \quad (7)$$

Since the conditional distribution of $\underline{z}_i|l_i = k$ is Gaussian, the conditional expectation in (7) takes the following linear form

$$\underline{\mu}_{\underline{x}_i|\underline{y}_i,k} = \underline{\mu}_{x,k} + \Sigma_{xy,k}{\Sigma_{yy,k}}^{-1}\left(\underline{y}_i - \underline{\mu}_{y,k}\right). \quad (8)$$

Also, using Bayes' rule, the posterior probability $p(l_i = k|\underline{y}_i; \underline{\theta})$ in (6) can be computed as

$$p(l_i = k|\underline{y}_i; \underline{\theta}) = \frac{w_k \cdot \mathcal{N}(\underline{\mu}_{y,k}, \Sigma_{yy,k})}{\sum_{l=1}^{K} w_l \cdot \mathcal{N}(\underline{\mu}_{y,l}, \Sigma_{yy,l})}. \quad (9)$$

The mean squared error (MSE) of $\hat{\underline{x}}_i = \mathrm{E}(\underline{x}_i|\underline{y}_i; \underline{\theta})$ can be expressed as

$$\mathrm{E}\left((\underline{x}_i - \hat{\underline{x}}_i)(\underline{x}_i - \hat{\underline{x}}_i)^T; \underline{\theta}\right) = \mathrm{E}\left(\mathrm{Cov}(\underline{x}_i|\underline{y}_i; \underline{\theta}); \underline{\theta}\right) \quad (10)$$

or equivalently as

$$\mathrm{E}\left((\underline{x}_i - \hat{\underline{x}}_i)(\underline{x}_i - \hat{\underline{x}}_i)^T; \underline{\theta}\right) = \mathrm{Cov}\left(\underline{x}_i; \underline{\theta}\right) - \mathrm{Cov}\left(\hat{\underline{x}}_i; \underline{\theta}\right). \quad (11)$$

Since a simple analytic form for the MSE cannot be obtained, in general, we aim to derive a lower bound on it. Denoting by $\tilde{\underline{x}}_i$ the conditional expectation $\mathrm{E}(\underline{x}_i|\underline{y}_i, l_i; \underline{\theta})$, it is straightforward to show that

$$\mathrm{Cov}\left(\tilde{\underline{x}}_i; \underline{\theta}\right) - \mathrm{Cov}(\hat{\underline{x}}_i; \underline{\theta}) = \mathrm{E}\left((\tilde{\underline{x}}_i - \hat{\underline{x}}_i)(\tilde{\underline{x}}_i - \hat{\underline{x}}_i)^T; \underline{\theta}\right) \geq 0.$$

Together with (11), it thus follows that

$$\mathrm{E}\left((\underline{x}_i - \hat{\underline{x}}_i)(\underline{x}_i - \hat{\underline{x}}_i)^T; \underline{\theta}\right) \geq \mathrm{Cov}(\underline{x}_i; \underline{\theta}) - \mathrm{Cov}(\tilde{\underline{x}}_i; \underline{\theta}) \quad (12)$$

or equivalently,

$$\begin{aligned}\mathrm{E}\left((\underline{x}_i - \hat{\underline{x}}_i)(\underline{x}_i - \hat{\underline{x}}_i)^T; \underline{\theta}\right) &\geq \mathrm{E}\left(\mathrm{Cov}(\underline{x}_i|\underline{y}_i, l_i; \underline{\theta}); \underline{\theta}\right) \quad (13)\\ &= \mathrm{E}\left((\underline{x}_i - \tilde{\underline{x}}_i)(\underline{x}_i - \tilde{\underline{x}}_i)^T; \underline{\theta}\right).\end{aligned}$$

This lower bound is intuitively reasonable as we would expect the MMSE estimator $\tilde{\underline{x}}_i = \mathrm{E}(\underline{x}_i|\underline{y}_i, l_i; \underline{\theta})$ of $\underline{x}_i$ to achieve a lower or equal MSE than the estimator $\hat{\underline{x}}_i = \mathrm{E}(\underline{x}_i|\underline{y}_i; \underline{\theta})$. In the trivial case when $K = 1$, i.e., when $\underline{z}_i$ is Gaussian, the lower bound is attained. More generally, the bound is attained if and only if

$$\mathrm{E}\left((\tilde{\underline{x}}_i - \hat{\underline{x}}_i)(\tilde{\underline{x}}_i - \hat{\underline{x}}_i)^T; \underline{\theta}\right) = 0. \quad (14)$$

Explicit computation of the expectation in the expression for the lower bound in (13) yields

$$\mathrm{E}\left(\mathrm{Cov}(\underline{x}_i|\underline{y}_i, l_i; \underline{\theta}); \underline{\theta}\right) = \sum_{k=1}^{K} w_k \cdot \Sigma_{\underline{x}_i|\underline{y}_i,k}. \quad (15)$$

Note that since the conditional distribution of $\underline{z}_i|l_i = k$ is Gaussian, the conditional covariance $\Sigma_{\underline{x}_i|\underline{y}_i,k}$ is not a function of $\underline{y}_i$ and is given by

$$\Sigma_{\underline{x}_i|\underline{y}_i,k} = \Sigma_{xx,k} - \Sigma_{xy,k}{\Sigma_{yy,k}}^{-1}\Sigma_{yx,k}. \quad (16)$$

4. Maximum Likelihood Estimation of the Model Parameters

In the previous section, the model parameters $\underline{\theta}$ were assumed known in estimating the clean speech features from the noisy speech features. This section focuses on maximum likelihood estimation of these parameters using EM [6].

In the case of mixture distributions, the log-likelihood expression of the *incomplete data* is difficult to optimize because it contains the logarithm of a sum. However, if we assume the existence of unobserved labels $\{l_i\}$, whose values specify which component density generated each sample, the likelihood expression is significantly simplified. Specifically, choosing the *complete data* as the set of stereo features $\{\underline{z}_i\}$ together with their corresponding labels $\{l_i\}$, we obtain

$$f_{\underline{z},\underline{l}}\left(\underline{z}, \underline{l}; \underline{\theta}\right) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left(f_{\underline{z}_i|l_i}\left(\underline{z}_i|k; \underline{\theta}\right) \cdot p(l_i = k; \underline{\theta})\right)^{I(l_i=k)}$$

where $I(E)$ is an indicator function of the event E that takes the value 1 when the event E happens and takes the value 0 otherwise. With $p(l_i = k; \underline{\theta}) = w_k$ and $\underline{z}_i|l_i = k \sim \mathcal{N}\left(\underline{\mu}_{z,k}, \Sigma_{zz,k}\right)$, the log-likelihood of the *complete data* takes the following form:

$$\begin{aligned}\log f_{\underline{z},\underline{l}}\left(\underline{z}, \underline{l}; \underline{\theta}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} I(l_i = k)\Bigg(&\log(w_k) - \frac{1}{2}\log|2\pi\Sigma_{zz,k}|\\ &- \frac{1}{2}(\underline{z}_i - \underline{\mu}_{z,k})^T\Sigma_{zz,k}^{-1}(\underline{z}_i - \underline{\mu}_{z,k})\Bigg). \quad (17)\end{aligned}$$

Note that the expression in (17) is much simpler to optimize than the log-likelihood of the *incomplete data*. The problem, of course, is that we do not know the values of the hidden variables which must be estimated from the observed data.

We next discuss two scenarios: training and testing. In the former, both the clean and noisy speech features are observed, i.e., the parameters are estimated using the stereo feature vectors $\{\underline{z}_i\}$. In the latter, only the noisy speech feature vectors $\{\underline{y}_i\}$ are observed.

4.1. Training

Let us denote by $\mathcal{I}_{\text{train}}$ the set of training frames, for which both the clean and noisy features are observed, and by N_{train} the size of this set. Estimating the model parameters $\underline{\theta}$ during training reduces to the problem of estimating the parameters of a GMM distribution, for which EM is the most popular technique used. Although it is a well-studied problem, we will sketch a brief derivation for the sake of completeness.

The first step of the EM algorithm is to compute $Q_z(\underline{\theta}, \underline{\theta}')$, defined as the conditional expectation of the log-likelihood of the *complete data* (17) given the observed data. Specifically,

$$\begin{aligned}Q_z(\underline{\theta}, \underline{\theta}') &= \sum_{i\in\mathcal{I}_{\text{train}}}\mathrm{E}\left(\log f_{\underline{z}_i,\underline{l}_i}\left(\underline{z}_i, \underline{l}_i; \underline{\theta}\right)|\underline{z}_i; \underline{\theta}'\right)\\ &= \sum_{i\in\mathcal{I}_{\text{train}}}\sum_{k=1}^{K} p\left(l_i = k|\underline{z}_i; \underline{\theta}'\right)\Bigg(\log(w_k)\\ &\quad - \frac{1}{2}\log|2\pi\Sigma_{zz,k}| + \frac{1}{2}(\underline{z}_i - \underline{\mu}_{z,k})^T\Sigma_{zz,k}^{-1}(\underline{z}_i - \underline{\mu}_{z,k})\Bigg) \quad (18)\end{aligned}$$

where $\underline{\theta}'$ denotes the current estimate of the model parameters. Similar to (9), the posterior probability $p(l_i = k|\underline{z}_i; \underline{\theta}')$ can be

computed using Bayes' rule, i.e.,

E-Step:

$$p(l_i = k|\underline{z}_i;\underline{\theta}^{'}) = \frac{w_k^{'} \cdot \mathcal{N}(\underline{\mu}_{z,k}^{'}, \Sigma_{zz,k}^{'})}{\sum_{l=1}^{K} w_l^{'} \cdot \mathcal{N}(\underline{\mu}_{z,l}^{'}, \Sigma_{zz,l}^{'})}. \quad (19)$$

The second step of the EM algorithm is to maximize $Q_z(\underline{\theta}, \underline{\theta}^{'})$ computed in (18) with respect to $\underline{\theta}$. This maximization results in the following estimates.

M-Step:

$$w_k^{\text{ML}_{\text{train}}} = \frac{1}{N_{\text{train}}} \sum_{i \in \mathcal{I}_{\text{train}}} p(l_i = k|\underline{z}_i;\underline{\theta}^{'}), \quad (20a)$$

$$\underline{\mu}_{z,k}^{\text{ML}_{\text{train}}} = \frac{\sum_{i \in \mathcal{I}_{\text{train}}} p(l_i = k|\underline{z}_i;\underline{\theta}^{'}) \cdot \underline{z}_i}{\sum_{i \in \mathcal{I}_{\text{train}}} p(l_i = k|\underline{z}_i;\underline{\theta}^{'})}, \quad (20b)$$

$$\Sigma_{zz,k}^{\text{ML}_{\text{train}}} = \frac{\sum_{i \in \mathcal{I}_{\text{train}}} p(l_i = k|\underline{z}_i;\underline{\theta}^{'}) \cdot (\underline{z}_i - \underline{\mu}_{z,k}^{\text{ML}_{\text{train}}})(\underline{z}_i - \underline{\mu}_{z,k}^{\text{ML}_{\text{train}}})^T}{\sum_{i \in \mathcal{I}_{\text{train}}} p(l_i = k|\underline{z}_i;\underline{\theta}^{'})}. \quad (20c)$$

These two steps (19) and (20) are repeated until convergence is achieved, where the newly derived parameters at the current iteration are used as the guess for the next iteration. Each iteration is guaranteed to increase the log-likelihood of the observed data, and the algorithm is guaranteed to converge to a local maximum point of the likelihood function.

4.2. Testing

Let us now denote by $\mathcal{I}_{\text{test}}$ the set of test frames, for which only the noisy speech features $\{\underline{y}_i\}$ are observed, and by N_{test} the size of this set. In this case, the conditional expectation of the log-likelihood of the *complete data* given the observed data takes the following form:

$$Q_y(\underline{\theta}, \underline{\theta}^{'}) = \sum_{i \in \mathcal{I}_{\text{test}}} \text{E}\left(\log f_{\underline{z}_i, l_i}(\underline{z}_i, l_i; \underline{\theta}) \,|\underline{y}_i; \underline{\theta}^{'}\right) \quad (21)$$

$$= \sum_{i \in \mathcal{I}_{\text{test}}} \sum_{k=1}^{K} p(l_i = k|\underline{y}_i;\underline{\theta}^{'}) \cdot \Big(\log(w_k) - \frac{1}{2}\log|2\pi\Sigma_{zz,k}| - \frac{1}{2}\text{E}\left(I(l_i = k) \cdot (\underline{z}_i - \underline{\mu}_{z,k})^T \Sigma_{zz,k}^{-1} (\underline{z}_i - \underline{\mu}_{z,k}) | \underline{y}_i; \underline{\theta}^{'}\right)\Big),$$

where the expectation term in (21) can be simplified by using the law of total expectation, i.e.,

$$\text{E}\left(I(l_i = k) \cdot (\underline{z}_i - \underline{\mu}_{z,k})^T \Sigma_{zz,k}^{-1} (\underline{z}_i - \underline{\mu}_{z,k}) \mid \underline{y}_i; \underline{\theta}^{'}\right) = \quad (22)$$

$$\text{E}\left(I(l_i = k) \cdot \text{E}\left((\underline{z}_i - \underline{\mu}_{z,k})^T \Sigma_{zz,k}^{-1} (\underline{z}_i - \underline{\mu}_{z,k}) | \underline{y}_i, l_i; \underline{\theta}^{'}\right) \mid \underline{y}_i; \underline{\theta}^{'}\right) =$$

$$p(l_i = k|\underline{y}_i;\underline{\theta}^{'}) \cdot \text{E}\left((\underline{z}_i - \underline{\mu}_{z,k})^T \Sigma_{zz,k}^{-1} (\underline{z}_i - \underline{\mu}_{z,k}) \mid \underline{y}_i, k; \underline{\theta}^{'}\right).$$

Eq. (22) can be further simplified by noting that

$$\text{E}\left((\underline{z}_i - \underline{\mu}_{z,k})^T \Sigma_{zz,k}^{-1} (\underline{z}_i - \underline{\mu}_{z,k}) | \underline{y}_i, k; \underline{\theta}^{'}\right) = \text{tr}\left(\Sigma_{zz,k}^{-1} \cdot \Lambda_{i,k}\right) \quad (23)$$

where

$$\Lambda_{i,k} = \text{E}\left((\underline{z}_i - \underline{\mu}_{z,k})(\underline{z}_i - \underline{\mu}_{z,k})^T | \underline{y}_i, k; \underline{\theta}^{'}\right) = \Sigma_{\underline{z}_i|\underline{y}_i,k}^{'} + \left(\underline{\mu}_{\underline{z}_i|\underline{y}_i,k}^{'} - \underline{\mu}_{z,k}\right)\left(\underline{\mu}_{\underline{z}_i|\underline{y}_i,k}^{'} - \underline{\mu}_{z,k}\right)^T \quad (24)$$

and

$$\underline{\mu}_{\underline{z}_i|\underline{y}_i,k}^{'} = \text{E}\left(\underline{z}_i|\underline{y}_i, k; \underline{\theta}^{'}\right) = \begin{bmatrix} \underline{\mu}_{\underline{x}_i|\underline{y}_i,k}^{'} \\ \underline{y}_i \end{bmatrix}, \quad (25a)$$

$$\Sigma_{\underline{z}_i|\underline{y}_i,k}^{'} = \text{Cov}\left(\underline{z}_i|\underline{y}_i, k; \underline{\theta}^{'}\right) = \begin{bmatrix} \Sigma_{\underline{x}_i|\underline{y}_i,k}^{'} & 0 \\ 0 & 0 \end{bmatrix}. \quad (25b)$$

Using (22)-(25) in (21), $Q_y(\underline{\theta}, \underline{\theta}^{'})$ is simplified to

$$Q_y(\underline{\theta}, \underline{\theta}^{'}) = \sum_{i \in \mathcal{I}_{\text{test}}} \sum_{k=1}^{K} p(l_i = k|\underline{y}_i;\underline{\theta}^{'}) \cdot \quad (26)$$

$$\Big(\log(w_k) - \frac{1}{2}\log|2\pi\Sigma_{zz,k}| - \frac{1}{2}\text{tr}\left(\Sigma_{zz,k}^{-1} \cdot \Lambda_{i,k}\right)\Big) =$$

$$\sum_{i \in \mathcal{I}_{\text{test}}} \sum_{k=1}^{K} p(l_i = k|\underline{y}_i;\underline{\theta}^{'}) \cdot \Big(\log(w_k) - \frac{1}{2}\left(\log|2\pi\Sigma_{zz,k}| + \text{tr}\left(\Sigma_{zz,k}^{-1} \cdot \Sigma_{\underline{z}_i|\underline{y}_i,k}^{'}\right) + (\underline{\mu}_{\underline{z}_i|\underline{y}_i,k}^{'} - \underline{\mu}_{z,k})^T \Sigma_{zz,k}^{-1} (\underline{\mu}_{\underline{z}_i|\underline{y}_i,k}^{'} - \underline{\mu}_{z,k})\right)\Big).$$

Maximization of $Q_y(\underline{\theta}, \underline{\theta}^{'})$ w.r.t w_k, $\underline{\mu}_{z,k}$, and $\Sigma_{zz,k}$ then yields:

$$w_k^{\text{ML}_{\text{test}}} = \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test}}} p(l_i = k|\underline{y}_i;\underline{\theta}^{'}), \quad (27a)$$

$$\underline{\mu}_{z,k}^{\text{ML}_{\text{test}}} = \frac{\sum_{i \in \mathcal{I}_{\text{test}}} p(l_i = k|\underline{y}_i;\underline{\theta}^{'}) \cdot \underline{\mu}_{\underline{z}_i|\underline{y}_i,k}^{'}}{\sum_{i \in \mathcal{I}_{\text{test}}} p(l_i = k|\underline{y}_i;\underline{\theta}^{'})}, \quad (27b)$$

$$\Sigma_{zz,k}^{\text{ML}_{\text{test}}} = \frac{\sum_{i \in \mathcal{I}_{\text{test}}} p(l_i = k|\underline{y}_i;\underline{\theta}^{'}) \cdot \Lambda_{i,k}}{\sum_{i \in \mathcal{I}_{\text{test}}} p(l_i = k|\underline{y}_i;\underline{\theta}^{'})}. \quad (27c)$$

Note the similarity between the equations for estimating the parameters in test time, where only $\{\underline{y}_i\}$ are observed, to those equations obtained during training (20). Specifically, $p(l_i = k|\underline{z}_i;\underline{\theta}^{'})$ is replaced with $p(l_i = k|\underline{y}_i;\underline{\theta}^{'})$, $\underline{z}_i$ in (20b) is replaced with $\underline{\mu}_{\underline{z}_i|\underline{y}_i,k}^{'} = \begin{bmatrix} \underline{\mu}_{\underline{x}_i|\underline{y}_i,k}^{'} \\ \underline{y}_i \end{bmatrix}$ in (27b), and $(\underline{z}_i - \underline{\mu}_{z,k})(\underline{z}_i - \underline{\mu}_{z,k})^T$ in (20c) is replaced with $\Lambda_{i,k} = \text{E}\left((\underline{z}_i - \underline{\mu}_{z,k})(\underline{z}_i - \underline{\mu}_{z,k})^T | \underline{y}_i, k; \underline{\theta}^{'}\right)$ in (27c).

5. Maximum a Posteriori Estimation of the Model Parameters

In Section 4 we estimated the model parameters given the observed data using the maximum likelihood approach. Specifically, no prior information on the model parameters $\underline{\theta}$ was assumed in the estimation process. In this section we take a different approach in which the estimates obtained during training are used as priors for the estimation during test-time. As opposed to previous work ([3],[5]) where the model parameters estimated during training were kept fixed at test-time, these parameters are now updated using the maximum a posteriori approach. In developing this approach, we follow a similar derivation to that presented in [7].

Let us denote by $g_{\underline{\theta}}(\underline{\theta}; \underline{\varphi})$ the prior probability density function of $\underline{\theta}$, where $\underline{\varphi}$ are the prior parameters. The MAP estimate $\hat{\underline{\theta}}_{\text{MAP}}$ of $\underline{\theta}$ is then defined as

$$\hat{\underline{\theta}}_{\text{MAP}} = \arg\max_{\underline{\theta}} f(\underline{y}|\underline{\theta}) \cdot g_{\underline{\theta}}(\underline{\theta}; \underline{\varphi}). \quad (28)$$

As noted by Dempster [6], the EM algorithm can also be applied to MAP estimation where instead of maximizing $Q_y(\underline{\theta}, \underline{\theta}')$ as was done in Section 4, we maximize the auxiliary function $R_y(\underline{\theta}, \underline{\theta}') = Q_y(\underline{\theta}, \underline{\theta}') + \log g_{\underline{\theta}}(\underline{\theta}; \underline{\varphi})$.

The prior $g_{\underline{\theta}}(\underline{\theta}; \underline{\varphi})$ is chosen as in [7], i.e.,

$$g_{\underline{\theta}}(\underline{\theta}; \underline{\varphi}) = g_D(\underline{w}; \underline{\nu}) \cdot \prod_{k=1}^{K} g_{NW}(\underline{\mu}_{z,k}, \Sigma_{zz,k}^{-1}; \tau_k, \alpha_k, \underline{\zeta}_k, u_k) \quad (29)$$

where a Dirichlet density $g_D(\underline{w}; \underline{\nu}) \propto \prod_{k=1}^{K} w_k^{\nu_k - 1}$ is chosen to model the prior knowledge about the mixture weights, and a Normal-Wishart density of the form

$$g_{NW}\left(\underline{\mu}_{z,k}, \Sigma_{zz,k}^{-1} | \tau_k, \alpha_k, \underline{\zeta}_k, u_k\right) \propto |\Sigma_{zz,k}^{-1}|^{(\alpha_k - p)/2} \cdot \exp\left(-\frac{\tau_k}{2}(\underline{\mu}_{z,k} - \underline{\zeta}_k)^T \Sigma_{zz,k}^{-1}(\underline{\mu}_{z,k} - \underline{\zeta}_k) - \frac{1}{2}\mathrm{tr}(u_k \Sigma_{zz,k}^{-1})\right)$$

is chosen for the parameters of the individual Gaussian mixture component. $(\{\nu_k\}, \{\tau_k\}, \{\alpha_k\}, \{\underline{\zeta}_k\}, \{u_k\})$ are the prior parameters such that $\nu_k > 0$, $\tau_k > 0$, $\alpha_k > p - 1$, $\underline{\zeta}_k$ is a vector of dimension p, and u_k is a $p \times p$ positive definite matrix where p is the dimension of the stereo feature vector.

Denoting by $\psi(\underline{\theta}, \underline{\theta}') = \exp(R(\underline{\theta}, \underline{\theta}'))$, we can easily show that $\psi(\underline{\theta}, \underline{\theta}') = g_{\underline{\theta}}(\underline{\theta}; \phi) \cdot \exp(Q_y(\underline{\theta}, \underline{\theta}'))$ belongs to the same distribution family as $g(\cdot)$. Maximization of $\psi(\underline{\theta}, \underline{\theta}')$ with respect to $\underline{\theta}$ then yields the following MAP estimates:

$$w_k^{\text{MAP}} = \gamma^w \cdot \left(\frac{\nu_k - 1}{\nu_0}\right) + (1 - \gamma^w) \cdot w_k^{\text{ML}_{\text{test}}}, \quad (30a)$$

$$\underline{\mu}_{z,k}^{\text{MAP}} = \gamma_k^{\mu} \cdot \underline{\zeta}_k + (1 - \gamma_k^{\mu}) \cdot \underline{\mu}_{z,k}^{\text{ML}_{\text{test}}}, \quad (30b)$$

$$\Sigma_{zz,k}^{\text{MAP}} = \gamma_k^{\Sigma} \cdot \left(\frac{u_k}{\alpha_k - p}\right) + (1 - \gamma_k^{\Sigma}) \cdot \Sigma_{zz,k}^{\text{ML}_{\text{test}}} + (1 - \gamma_k^{\Sigma}) \cdot \gamma_k^{\mu} \cdot \left(\underline{\zeta}_k - \underline{\mu}_{z,k}^{\text{ML}_{\text{test}}}\right)\left(\underline{\zeta}_k - \underline{\mu}_{z,k}^{\text{ML}_{\text{test}}}\right)^T, \quad (30c)$$

where $\nu_0 = \sum_{k=1}^{K}(\nu_k - 1)$ and

$$\gamma^w = \frac{\nu_0}{\nu_0 + N_{\text{test}}}, \quad (31a)$$

$$\gamma_k^{\mu} = \frac{\tau_k / N_{\text{test}}}{\tau_k / N_{\text{test}} + w_k^{\text{ML}_{\text{test}}}}, \quad (31b)$$

$$\gamma_k^{\Sigma} = \frac{(\alpha_k - p)/N_{\text{test}}}{(\alpha_k - p)/N_{\text{test}} + w_k^{\text{ML}_{\text{test}}}}. \quad (31c)$$

Note that $(\nu_k - 1)/\nu_0$, $\underline{\zeta}_k$, and $u_k/(\alpha_k - p)$ are the modes of the Dirichlet and Normal-Wishart densities. As such, we can estimate them with $w_k^{\text{ML}_{\text{train}}}$, $\underline{\mu}_{z,k}^{\text{ML}_{\text{train}}}$, and $\Sigma_{zz,k}^{\text{ML}_{\text{train}}}$, respectively. A more detailed discussion on how to estimate the parameters $\underline{\varphi}$ of the prior distribution can be found in [7].

6. Experimental Results

The experimental setup is similar to the one detailed in [8]. All experiments are conducted on English LVCSR with multi-conditional SSM training. Our acoustic model is trained on 280 hours of clean speech signals. It has 5K quinphone states and 100K Gaussians. The tri-gram language model has 330K n-grams and was built on a vocabulary of 45K words and 56K pronunciations. The feature space is constructed by splicing 9 frames of 24-dim PLP features, projecting them down to a 40-dim linear discriminant analysis (LDA) space, and then using a global semi-tied covariance (STC) transformation. The acoustic model is built with feature and model space discriminative training (FMMI and BMMI) [9].

The test set consists of the real conditions scenario referred to as Set C in [8]. This test data is composed of 7 speakers (1421 sentences, 10K words, 1.9 hours total), and is made of real world speech signals recordings in humvee-tank noise condition with SNRs estimated between 5dB and 8dB.

For SSM model training, our 60 hours stereo training data is composed of a clean channel of continuous speech, and a noisy channel generated by artificially corrupting the clean channel. We use 10 types of background noise including M109, Buccaneer, Leopard, wheel carrier, destroyer operation room, HF radio, babble, factory, car and white noise to generate the noisy channel from the clean channel. For each utterance, the noisy channel SNR is randomly chosen between 10dB to 25dB. This 60 hours stereo-data set is only used to train a full covariance (FC) GMM model in the joint channel space. The FC GMM has 512 Gaussian components, and estimates of its parameters are iteratively updated for 45 iterations of EM. All the noise samples used to generate the noisy speech signals in this multi-conditional training are from the NOISEX-92 dataset.

At test time, SSM is carried out in the 40-dim FMMI space for each utterance. For A-SSM, the trained FC-GMM model is used as our initial model for each speaker. The model is then updated for each test utterance from the current speaker. If a speaker turn occurs, we reset our model to the trained FC GMM. All A-SSM results are obtained using MAP updates of the SSM model as described in the previous section. MAP updates enable to have stable estimates for the SSM model parameters even for small test utterances. This is particularly important for the 80×80 GMM covariances.

Table 1 summarizes the performance, measured as word error rates (WER) of SSM and A-SSM for noise compensation compared to our baseline on the real world test data (Set C). A-

Technique/WER	w/o fMLLR	w/ fMLLR
baseline	38.24	34.24
SSM	31.08	28.45
A-SSM	29.17	27.37

Table 1: WERs (in %) for baseline, SSM, and adaptive SSM (A-SSM) noise compensation on the real world test data.

SSM gives 29.17% WER, a 6.15% relative improvement over SSM at 31.08% WER, and 23.72% relative improvement over baseline at 38.24% WER. With fMLLR, commonly compared to noise robustness techniques, we have a 4% absolute gain over baseline, 2.63% absolute gain over SSM, and 1.8% absolute gain over A-SSM, which leads to our best WER at 27.37%. These results show that A-SSM cleans noisy features in a way that fMLLR cannot entirely capture, and is somewhat additive to fMLLR's gain.

7. Conclusions

This paper extends the stochastic mapping approach for noise robust ASR under the MMSE criterion by proposing an EM procedure to perform MAP adaptation of our model parameters at test time. This approach of adaptive SSM not only improves upon SSM by leveraging test time information, but also gives additional gains when used in parallel with fMLLR.

8. References

[1] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, 1993.

[2] J. Droppo, A. Acero, and L. Deng, "Evaluation of the splice algorithm on the aurora 2 database," in *Eurospeech*, 2001, pp. 217–220.

[3] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 17, no. 7, pp. 1325–1334, 2009.

[4] ——, "Stereo-based stochastic mapping for robust speech recognition," in *ICASSP*, 2007, pp. 377–380.

[5] X. Cui, M. Afify, and Y. Gao, "MMSE-based stereo feature stochastic mapping for noise robust speech recognition," in *ICASSP*, 2008, pp. 4077–4080.

[6] 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*, vol. 39, no. 1, pp. 1–38, 1977.

[7] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[8] X. Cui, M. Afify, G. Saon, and V. Goel, "Sparse bayesian factor analysis for stereo-based stochastic mapping," in *INTERSPEECH*, September 2012.

[9] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *ICASSP*, 2008, pp. 4057–4060.