Environmental Compensation Using ASR Model Adaptation by A Bayesian Parametric Representation Method

Xuechuan Wang  Douglas O'Shaughnessy

INRS-EMT, University of Quebec
800 de la Gauchetiere West
Montreal, Quebec, H5A 1K6, Canada
wwang, dougo@inrs-emt.uquebec.ca

Abstract
The mismatch between system training and operating conditions can seriously deteriorate the performance of ASR systems. The maximum a posteriori (MAP) estimation is used for the adaptation of HMM-based multivariate Gaussian mixture models (GMMs). In this paper, we propose an environment independent ASR model parameter adaptation approach based on Bayesian parametric representation (BPR). Compared to the MAP method, the BPR adaptation method has better performance with limited adaptation data. The performances of the two methods are investigated in the experiments designed on the AURORA II noisy speech database.

1. Introduction
Stochastic model-based automatic speech recognition (ASR) systems, such as HMMs, nowadays are able to reach satisfactory rates in environments similar to the acoustic model training ones [1, 2]. In a real-life environment, speech signals are varied by environmental factors (EFs), e.g., noise and channel distortion. The corresponding distribution of speech in feature spaces shifts. When the shifted speech features are beyond the scope of acoustic models, ASR systems will fail to recognize the speech. For an increase of ASR system robustness to the environmental variations, the effects of environmental factors must be eliminated from the system.

It is often difficult and costly to directly remove various EFs from speech features. Statistical ASR model adaptation techniques adapt ASR models to environmental changes by using a significant difference between EFs and speech in stochastic properties, i.e., most EFs are time-invariant, while speech is a time-varying signal. The maximum a posteriori estimation method provides a framework for the adaptation of HMM-based multivariate Gaussian mixture models [8]. MAP adaptation, however, has low efficiency when the adaptation data is limited. This paper proposes a statistical framework based on a Bayesian parametric representation approach: given a measurable probability space, if an a priori distribution on a subspace is defined and an a posteriori distribution is constructed from it, the parametric representation of a target process on a disjoint subspace can be estimated corresponding to conditional expectations under this a posteriori distribution [3]. Starting from the basic assumption of HMM, which assumes that speech signals are generated by a Markov process of finite state random sources $S_i, i = 1, 2, \ldots, L$ [4], we further extend it as that any speech process is a Markov process constructed on a subspace of a measurable probability space $(\Omega, \Phi, \mathcal{P})$, where $\Omega$ is a nonempty random number set, $\Phi$ is a set of $\sigma$-algebras closed under all countable set operations and $0 \leq P \leq 1$ is a probability measurement defined on $\Phi$. Given two speech processes, $X = \{X_t, t \in T\}$ and $Y = \{Y_t, t \in T\}$, generated from subspaces defined by $\Lambda$ and $\Theta$, respectively, where $\Lambda$ and $\Theta \subset \Phi$ are two sub $\sigma$-algebras. From the Radon-Nikodym theorem [5, 6], there exists a Borel mapping function $F$ from $\Lambda$ to $\Theta$, such that

$$\hat{Y}_t = F(X_t) = E(Y|X)_t, \quad t \in T, \quad (1)$$

is a $\Lambda$ measurable process, where $\hat{Y}$ is an expectation or representation of $Y$ given $X$ and $E(Y|X)$ is a conditional expectation. $\hat{Y}$ has similar statistical characteristics as $Y$. Speech signals/features or ASR model parameters obtained in one environment can be represented in different ones if a map $F$ is found. If an ASR system is robust on $X$ in an environment defined by $\Lambda$, it is then able to regain robustness on $\hat{Y}$ in a new application environment defined by $\Theta$ through constructing a $F$ and representing ASR model parameters in the new environment. The minimum mean square error (MMSE) estimate of $F$ results in the conditional expectation $E(Y|X)$ [7]. Parametric representation through $F$ by MMSE estimates is described in Section 2 together with a MAP adaptation formulation. In Section 3, the ASR experimental results on AURORA II [9] databases are presented. The performances of the MAP and the BPR model adaptation methods using different lengths of adaptation data are investigated and compared.
2. Statistical ASR Model Adaptation

2.1. Model Adaptation by the MAP Estimate

The MAP estimate is defined as the mode of the posterior probability density function of model \( \theta \),

\[
\hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta | X)p(\theta),
\]

where \( p(\theta) \) is the \textit{a priori} density. Let \( m_i \) and \( r_i \) be the \( i \)-th mean and precision matrix of a Gaussian mixture model (GMM), respectively, \( i = 1, \ldots, N \). The joint conjugate \textit{a priori} density of \((m_i, r_i)\) is a normal-Wishart density \([2, 8]\):

\[
p(m_i, r_i | \tau_i, \mu_i, \alpha_i, u_i) \propto r_i^{(\alpha_i - 1)/2} \exp\left[-\frac{1}{2} \text{tr}(u_i r_i)\right] \exp\left[-\frac{\tau_i}{2} (m_i - \mu_i)^T r_i^{-1} (m_i - \mu_i)\right],
\]

where \((\tau_i, \mu_i, \alpha_i, u_i)\) is the priori density parameter set.

Eq. 4 yields the MAP estimation at an expectation-maximization (EM) iteration step:

\[
\hat{\mu}_i = \frac{\tau_i \mu_i + \sum_{j=1}^{N} \omega_{ij} \hat{\mu}_{ij}}{\tau_i + \sum_{j=1}^{N} \omega_{ij}}
\]

\[
\hat{r}_i = \frac{\omega_i + \sum_{j=1}^{N} r_{ij} + \tau_i (\hat{\mu}_i - \hat{\mu}_i) (\hat{\mu}_i - \hat{\mu}_i)^T}{\tau_i + \sum_{j=1}^{N} \omega_{ij}},
\]

where \( \omega_{ij} = \omega_{ij}N(\hat{\mu}_i, \hat{\mu}_i) / \sum_{j=1}^{N} \omega_{ij} \) and \( S_i \) is the empirical variance of adaptation data.

2.2. ASR Model Mean Compensation by the Bayesian Parametric Representations

Suppose that two speech sources, \( \Lambda \) and \( \Theta \), are the realizations of a pure speech source \( S \) in two speech environments \( A \) and \( B \), respectively, where \( S \), \( \Lambda \) and \( \Theta \subset \Phi \) are sub-\( \sigma \)-algebras on a probability space \( \{\Omega, \Phi, P\} \). \( X = \{X_t, t = 1, \ldots, T\} \) is a sequence of prior measurements of a random process generated by \( \Lambda \) and \( Y = \{Y_t, t = 1, \ldots, T\} \) by \( \Theta \). It is then possible to find an estimate of \( Y \) based on the value of \( X \) and vice versa. Let \( \hat{Y} = g(X) = \{g(X_t)\} \) be an estimate of \( Y \) given \( X \). If \( X \) is complete and sufficient for \( \Lambda \), then \( g(X) \) is an unbiased estimation of \( Y \) when it satisfies the MMSE criterion. The mean square error (MSE) of an estimate is defined as

\[
\text{MSE}(Y) = E(||Y - g(X)||^2 | X).
\]

According to estimation theory \([7]\), the conditional expectation \( E(Y | X) \) is a unique unbiased estimate of \( Y \), where

\[
E(Y | X) = \int_{\theta} Y \cdot p(Y | X)dY
= \int_{\theta} Y_i \cdot p(Y_i | X_i)dY_i, \quad t \in T.
\]

In the case that both \( X \) and \( Y \) are discrete, \( E(Y | X) \) becomes:

\[
E(Y | X) = \sum_{i=1}^{T} Y_i P(Y = Y_i | X = X_i).
\]

The MMSE estimation in Eqs. (6) and (7), yet, is not immediately applicable, since the prior measurements \( \{Y_t\} \) are often unavailable in most situations. This leaves both \( Y_t \) and \( p(Y_t | X_t) \) undetermined. If, however, the statistical parametric forms of speech sources \( \Lambda \) and \( \Theta \) are known, it is possible to find \( E(Y | X) \) through Bayesian parametric estimation using parameters of \( \Lambda \) and \( \Theta \).

Suppose \( \Lambda \) is partitioned by \( N \) disjoint events or states \( \{\lambda_i, i = 1, 2, \ldots, N\} \), \( \Lambda = \cup_{i=1}^{N} \lambda_i \), and \( \Theta \) by \( \{\theta_j, j = 1, 2, \ldots, M\} \), \( \Theta = \cup_{j=1}^{M} \theta_j \), then the \textit{a priori} densities \( p(X_t) \) and \( p(Y_t) \) can be explicitly expressed as

\[
p(X_t) = p(X_t, \Lambda) = \sum_{i=1}^{N} p(X_t | \lambda_i) p(\lambda_i)
\]

\[
p(Y_t) = p(Y_t, \Theta) = \sum_{j=1}^{M} p(Y_t | \theta_j) p(\theta_j),
\]

where \( \lambda_1, \ldots, \lambda_N \) are disjoint events of a partition of \( \Lambda \), \( \theta_1, \ldots, \theta_M \) are those of \( \Theta, p(\lambda_i) \) is the \textit{a priori} probability of \( \lambda_i \) subject to \( \sum_{i=1}^{N} p(\lambda_i) = 1 \) and \( p(\theta_j) \) is that of \( \theta_j \) subject to \( \sum_{j=1}^{M} p(\theta_j) = 1 \).

The conditional probability density at time \( t \), \( p(Y_t | X_t) \), can be expressed by Bayes rule as:

\[
p(Y_t | X_t) = \frac{p(X_t, Y_t)}{p(X_t)}.
\]

The joint probability density function \( p(X_t, Y_t) \) is then

\[
p(X_t, Y_t) = p(Y_t, \Theta | X_t, \Lambda)p(X_t | \Lambda)
\]

\[
= p(Y_t | \Theta)p(\Theta | X_t)p(X_t | \Lambda)p(\Lambda)
\]

\[
= \sum_{i=1}^{N} \sum_{j=1}^{M} p(Y_t | \theta_j)p(\theta_j | X_t)p(X_t | \lambda_i)p(\lambda_i)
\]

\[
= \sum_{i=1}^{N} \sum_{j=1}^{M} p(Y_t | \theta_j)p(\theta_j)p(X_t | \lambda_i)p(\lambda_i),
\]

where \( p(\theta_j | \lambda_i) \) is the cross-correlation probability between \( \theta_j \) and \( \lambda_i \) and \( \alpha_{ij} = p(\theta_j | \lambda_i) \).

Substitute Eq. (8), (9) and (10) back into Eq. (6), and \( \hat{Y}_t \) becomes:

\[
\hat{Y}_t = E(Y_t | X_t) = \int_{\theta} Y_t \cdot p(Y_t | X_t)dY_t
\]

\[
= \sum_{i=1}^{N} \sum_{j=1}^{M} \int_{\theta} Y_t p(Y_t | \theta_j)dY_t \alpha_{ij} p(X_t | \lambda_i)p(\lambda_i)
\]

\[
= \sum_{i=1}^{N} \sum_{j=1}^{M} \mu_{ij} \mu_{ij} p(X_t | \lambda_i)p(\lambda_i),
\]

where \( \mu_{ij} = E(Y_t | \theta_j) = \int_{\theta} Y_t p(Y_t | \theta_j)dY_t \) is the mean of \( Y_t \) that falls into the area of \( \theta_j \).

Observation vectors of \( X_t \) and \( Y_t \) often have varied amplitude levels in different cases. The proportion of amplitudes of any relating pairs \( (X_t, Y_t) \), however, is always consistent. The amplitude of estimates made by Eq. (11), \( \hat{Y}_t \), is determined by the amplitude of \( \mu_{ij}, j = 1, \ldots, M \), while as the prior knowledge, the amplitude level of \( \mu_{ij} \) is unable to vary by \( X_t \). This problem may cause a complete failure of Eq. (11) to make proper estimation in some applications. The influences of amplitude level can
be eliminated through amplitude normalization, which can be realized by the following procedure.

First, express $X_t, \mu_j$ in an amplitude form as follows:

$$X_t = |A_{X_t}| \sigma_{X_t},$$
$$\mu_j = |A_{\mu_j}| \sigma_{\mu_j},$$  \hspace{1cm} (12)

where $|\sigma_{X_t}| = 1$ and $|\sigma_{\mu_j}| = 1$ are unit vectors. Estimate the unit vector of $Y_t$ in the following step using Eq. (11):

$$\sigma_{Y_t} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \mu_j \rho \sigma_{X_t}|\lambda_i\sigma_{Y_t}(\lambda_i)}}{\sum_{i=1}^{N} \rho(\sigma_{X_t}|\lambda_i)p(\lambda_i)}.$$  \hspace{1cm} (13)

Define an amplitude ratio with respect to $\theta_j$ as

$$\varepsilon_{\theta_j} = \frac{|A_{Y_t}|}{|A_{X_t}| \rho(\theta_j)}.$$  \hspace{1cm} (14)

where $|A_{Y_t}|$ is the amplitude of $Y_t$. The amplitude of $Y_t$ can be estimated by

$$|A_{Y_t}| = |A_{X_t}| \sum_{j=1}^{M} \varepsilon_{\theta_j} p(\theta_j) X_t, Y_t.$$  \hspace{1cm} (15)

Merge Eq. (13) and (15) in the final step to give the estimate of $Y_t$:

$$\hat{Y}_t = |A_{Y_t}| \sigma_{Y_t} = |A_{X_t}| \sum_{j=1}^{M} \sum_{i=1}^{N} \mu_j \rho \sigma_{X_t}|\lambda_i\sigma_{Y_t}(\lambda_i)|$$

$$\sum_{i=1}^{N} \rho(\sigma_{X_t}|\lambda_i)p(\lambda_i).$$  \hspace{1cm} (16)

3. Experiments and Results

3.1. Database and ASR System

3.1.1. Aurora 2 Task

The Aurora 2 database [9] corresponds to TI-DIGITS training data down-sampled to 8 kHz and filtered with a G.712 characteristic. It provides a baseline system performance, which includes clean and multi-condition training sets. The baseline systems use 13 MFCC coefficients with energy extracted from the amplitude of the power spectrum, along with delta and acceleration coefficients. Eleven whole word HMMs are employed. Each has 16 states and 3 diagonal Gaussian mixtures per state. The noisy utterances have SNR conditions from -5 dB to 20 dB.

3.1.2. ASR Systems

The HTK-based speech recognition system [10] is used throughout the experiments. HTK is a Hidden Markov Model (HMM)-based speech recognition system and is designed for both isolated and continuous speech recognition. A continuous whole-word-based speech recognition system is built for AURORA 2 experiments.

3.2. Experimental Results

The ASR models are trained on the clean data from the training set and tested on the noisy speech test set. The noisy speech data in the training set are used as adaptation data. Different durations of adaptation data are selected to investigate the performances of both the MAP and the BPR methods. The results on Test A of the AURORA 2 database are given in Tables 1 to 9. Noisy speech in Test A includes four noise types (subway, babble, car and exhibition) and six SNR levels (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB). The baseline test results are shown in Table 1.

Table 1: Word Recognition Rate (percentage) on Test A set.

<table>
<thead>
<tr>
<th>SNR</th>
<th>20 dB</th>
<th>15 dB</th>
<th>10 dB</th>
<th>5 dB</th>
<th>0 dB</th>
<th>-5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subway</td>
<td>97.85</td>
<td>93.83</td>
<td>82.19</td>
<td>61.77</td>
<td>32.48</td>
<td>13.69</td>
</tr>
<tr>
<td>Babble</td>
<td>97.64</td>
<td>93.56</td>
<td>79.96</td>
<td>56.77</td>
<td>30.14</td>
<td>13.88</td>
</tr>
<tr>
<td>Car</td>
<td>98.06</td>
<td>93.71</td>
<td>75.57</td>
<td>47.18</td>
<td>21.38</td>
<td>9.36</td>
</tr>
<tr>
<td>Exhibition</td>
<td>97.07</td>
<td>92.19</td>
<td>79.02</td>
<td>49.86</td>
<td>20.33</td>
<td>9.94</td>
</tr>
</tbody>
</table>

Tables 2 - 5 show the improvements on word recognition rate after MAP adaptation.

Table 2: Improvements on Word Recognition Rate After MAP Adaptation on Test A (Subway Noise Speech Set).

<table>
<thead>
<tr>
<th>Duration</th>
<th>SNR20</th>
<th>SNR15</th>
<th>SNR10</th>
<th>SNR5</th>
<th>SNR0</th>
<th>SNR-5</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 min</td>
<td>0.00</td>
<td>0.00</td>
<td>0.14</td>
<td>0.21</td>
<td>0.28</td>
<td>0.25</td>
<td>0.15</td>
</tr>
<tr>
<td>4 min</td>
<td>0.02</td>
<td>0.09</td>
<td>1.43</td>
<td>1.97</td>
<td>1.43</td>
<td>0.80</td>
<td>0.96</td>
</tr>
<tr>
<td>6 min</td>
<td>-0.03</td>
<td>0.54</td>
<td>2.45</td>
<td>1.68</td>
<td>1.95</td>
<td>1.40</td>
<td>1.33</td>
</tr>
<tr>
<td>8 min</td>
<td>0.06</td>
<td>1.02</td>
<td>3.04</td>
<td>1.99</td>
<td>2.17</td>
<td>1.40</td>
<td>1.61</td>
</tr>
<tr>
<td>10 min</td>
<td>0.06</td>
<td>1.07</td>
<td>3.69</td>
<td>2.17</td>
<td>2.39</td>
<td>1.40</td>
<td>1.80</td>
</tr>
<tr>
<td>12 min</td>
<td>0.06</td>
<td>1.11</td>
<td>4.46</td>
<td>1.89</td>
<td>2.54</td>
<td>1.17</td>
<td>1.87</td>
</tr>
</tbody>
</table>

Table 3: Improvements on Word Recognition Rate After MAP Adaptation on Test A (Babble Noise Speech Set).

<table>
<thead>
<tr>
<th>Duration</th>
<th>SNR20</th>
<th>SNR15</th>
<th>SNR10</th>
<th>SNR5</th>
<th>SNR0</th>
<th>SNR-5</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 min</td>
<td>0.00</td>
<td>0.02</td>
<td>2.38</td>
<td>0.77</td>
<td>2.18</td>
<td>0.44</td>
<td>0.97</td>
</tr>
<tr>
<td>4 min</td>
<td>0.01</td>
<td>0.93</td>
<td>4.46</td>
<td>1.66</td>
<td>5.62</td>
<td>1.30</td>
<td>2.33</td>
</tr>
<tr>
<td>6 min</td>
<td>0.15</td>
<td>0.81</td>
<td>6.27</td>
<td>2.15</td>
<td>3.34</td>
<td>1.15</td>
<td>2.47</td>
</tr>
<tr>
<td>8 min</td>
<td>0.29</td>
<td>1.05</td>
<td>5.98</td>
<td>2.07</td>
<td>6.10</td>
<td>1.45</td>
<td>2.82</td>
</tr>
<tr>
<td>10 min</td>
<td>0.20</td>
<td>1.18</td>
<td>6.49</td>
<td>2.27</td>
<td>6.39</td>
<td>1.45</td>
<td>3.00</td>
</tr>
<tr>
<td>12 min</td>
<td>0.22</td>
<td>1.85</td>
<td>7.16</td>
<td>2.98</td>
<td>6.62</td>
<td>1.58</td>
<td>3.40</td>
</tr>
</tbody>
</table>

Table 4: Improvements on Word Recognition Rate After MAP Adaptation on Test A (Car Noise Speech Set).

<table>
<thead>
<tr>
<th>Duration</th>
<th>SNR20</th>
<th>SNR15</th>
<th>SNR10</th>
<th>SNR5</th>
<th>SNR0</th>
<th>SNR-5</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 min</td>
<td>0.00</td>
<td>0.13</td>
<td>1.06</td>
<td>0.35</td>
<td>0.48</td>
<td>0.11</td>
<td>0.36</td>
</tr>
<tr>
<td>4 min</td>
<td>-0.03</td>
<td>0.56</td>
<td>4.83</td>
<td>1.45</td>
<td>1.02</td>
<td>0.87</td>
<td>1.45</td>
</tr>
<tr>
<td>6 min</td>
<td>0.12</td>
<td>1.13</td>
<td>8.26</td>
<td>1.57</td>
<td>2.75</td>
<td>1.32</td>
<td>2.53</td>
</tr>
<tr>
<td>8 min</td>
<td>-0.18</td>
<td>1.05</td>
<td>8.78</td>
<td>1.60</td>
<td>3.08</td>
<td>1.48</td>
<td>2.64</td>
</tr>
<tr>
<td>10 min</td>
<td>0.05</td>
<td>1.31</td>
<td>8.70</td>
<td>1.56</td>
<td>3.25</td>
<td>1.48</td>
<td>2.73</td>
</tr>
<tr>
<td>12 min</td>
<td>0.03</td>
<td>1.58</td>
<td>8.54</td>
<td>1.56</td>
<td>3.50</td>
<td>1.48</td>
<td>2.78</td>
</tr>
</tbody>
</table>
Tables 6 to 9 show the improvements on word recognition rate after MAP adaptation. The results show that the performance of BPR adaptation is significantly better than the MAP adaptation when the duration of adaptation data is low (2 min and 6 min). The overall performance of BPR adaptation is better as well.

Table 9: Improvements on Word Recognition Rate After BPR Adaptation on Test A (Exhibition Noise Speech Set).

<table>
<thead>
<tr>
<th>Duration</th>
<th>SNR20</th>
<th>SNR15</th>
<th>SNR10</th>
<th>SNR5</th>
<th>SNR0</th>
<th>SNR-5</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 min</td>
<td>0.12</td>
<td>2.41</td>
<td>9.50</td>
<td>10.65</td>
<td>2.56</td>
<td>3.96</td>
<td>4.87</td>
</tr>
<tr>
<td>4 min</td>
<td>0.52</td>
<td>4.11</td>
<td>12.03</td>
<td>17.13</td>
<td>7.10</td>
<td>3.52</td>
<td>7.40</td>
</tr>
<tr>
<td>6 min</td>
<td>0.65</td>
<td>4.32</td>
<td>11.45</td>
<td>17.96</td>
<td>6.42</td>
<td>3.09</td>
<td>7.32</td>
</tr>
<tr>
<td>8 min</td>
<td>0.52</td>
<td>4.01</td>
<td>11.85</td>
<td>18.54</td>
<td>8.27</td>
<td>4.14</td>
<td>7.89</td>
</tr>
<tr>
<td>10 min</td>
<td>0.99</td>
<td>3.14</td>
<td>12.28</td>
<td>18.05</td>
<td>8.53</td>
<td>4.32</td>
<td>8.05</td>
</tr>
<tr>
<td>12 min</td>
<td>0.86</td>
<td>3.86</td>
<td>12.62</td>
<td>18.08</td>
<td>7.85</td>
<td>4.92</td>
<td>8.05</td>
</tr>
</tbody>
</table>

4. Conclusions

In this paper, we propose a statistical ASR model adaptation approach based on Bayesian parametric representation. Compared to the performance of the MAP adaptation method, the overall performance of the BPR method is better. The BPR method is especially more stable than the MAP when adaptation data is limited.

5. References


