HMM ADAPTATION USING SPARSE PROBABILISTIC SPACE MAPPING FOR NOISY SPEECH

Kaustubh Kalgaonkar and Mark A. Clements

Center for Signal and Image Processing
School of Electrical and Computer Engineering
Georgia Institute of Technology.
Atlanta, Georgia 30339
{kaustubh, clements}@ece.gatech.edu

ABSTRACT

This paper presents an extension of Probabilistic Space Maps (PS-MAPS) to adapt clean acoustic models to a noisy environment. In the presence of noise, the relationship between noisy and clean speech features (MFCC’s, HDLA, etc.) is either nonlinear or unknown. Given the relationship between features, traditional methods try to linearize it using approximations. These methods cannot be used for systems where the mapping model for clean and noisy features is missing.

Given sufficient training data, PS-MAPS [1] provides an excellent framework for extracting and modeling this relationship. The PS-MAP based approach to model adaptation is completely data driven. Experiments were performed on Aurora 2 dataset to evaluate the effectiveness of the algorithm.

Index Terms— Probabilistic Maps, Speech Enhancement, Bayesian Estimation, EM.

1. INTRODUCTION

Noise is a major culprit in the poor performance of automatic speech recognition systems. Speech recognizers often fail due to the mismatch in the training and deployment conditions (channel and noise effects). Although speech enhancement/model adaptation has been an area of active research for many years, improvement is still possible.

Techniques that combat noise in ASR can be broadly classified into three categories: Waveform Enhancement: where the noisy speech is cleaned up and then passed to a recognizer trained on clean data; Feature Enhancement: where the front-end feature extraction algorithm specifically models the new condition; and Model Compensation: where the model trained on clean speech is adapted to the current environment based on the estimates of noise.

Model compensation techniques can be data driven e.g., MLLR [2], or use the actual nonlinear relationship between noisy speech, noise and clean speech e.g., VTS [3].

Several methods for handling the nonlinearity have been proposed. In Vector Taylor Series (VTS) [3] adaptation, the nonlinear function is linearized around expansion points defined by the speech and noise models. In [4], an Unscented Transform was used to estimate to noisy speech distribution using a small number of speech and noise sample points. In [5] a linear spline is used to map the nonlinearity and the phase variations around the mode are modeled with a segmental variance for each spline section.

In this paper, we propose a novel method based on Probabilistic Space Maps (PS-MAPS)[1, 6] to adapt the Gaussians of HMM’s to the noisy environment. This method previously was successfully applied as a front-end feature enhancement scheme to improve the perception of speech in noisy environment [6].

There are several differences between the existing adaptation schemes and the proposed method. First the proposed algorithm automatically extracts the nonlinear relationship between the parameters (clean speech, noisy speech and noise) using the training data. Further the proposed PS-MAP based model adaptation scheme does not approximate the nonlinear relationship between dynamic coefficients as is commonly done in existing adaptation schemes.

Two major sources of uncertainty are present when clean speech models are transformed to noisy speech models. The first is the actual noise and the second is the truncated DCT operation that converts log Mel coefficients to MFCC’s. PS-MAP captures this uncertainty using a many-to-one mapping matrix $A$, that probabilistically models the nonlinear mapping between clean and noisy speech features.

2. PROBABILISTIC STATE MAPPING

Many signal processing applications primarily perform the task of estimating $q \in Q$ from $p \in P$ where $P$ and $Q$ might be the same or different subspaces. Often the transformation $T$ from subspace $P$ to $Q$ is nonlinear and is difficult to invert. As we will discuss in the subsequent section, the relationship between MFCC’s (Mel Frequency Cepstral Coefficients) of clean and noisy speech is nonlinear and is not easily invertible.

Conventional algorithms represent each subspace $(P, Q)$ with a fixed number of bases and estimate the transform ($T$) that best explains the relationship $bases(Q) = T \cdot bases(P)$. Probabilistic space map provides an equivalent probabilistic interpretation of same model. Instead of mapping a subspace with bases, we capture the probabilities of the locations in the subspace. This is done by representing each subspace with a number of latent states. We model each of the $N$ latent states of subspace $P$ with Gaussians $N(\mu^*_n, \sigma^*_n)$, where $n = 1, 2, \ldots, N$ and each of the $M$ latent states of subspace $Q$ with Gaussians $N(\mu^m_n, \sigma^m_n)$ where $m = 1, 2, \ldots, M$.

The mapping between the subspace $P$ and $Q$ is captured in discrete probability matrix $A$ such that $\sum_n a_{m,n} = 1$. Each element of the matrix $A$, $a_{m,n} = p(\gamma_m \mid \pi_n)$ is the belief that the observation $q_i$ is generated using the bases $\gamma_m$ when the input observation $p_i$ was generated using bases $\pi_n$. The relationship between the states
of \( P \) and \( Q \) is many-to-one. Given sufficient training data the parameters of the model \( M = \{ \Pi, \Gamma, A \} \) can be easily estimated using Expectation Maximization [7]

2.1. Parameter Estimation

The current configuration of model permits each state of \( P \) to be mapped to every state of subspace \( Q \), but this situation rarely occurs in real world problems. Due to the limitations of the EM algorithm we will not be able to impose any restrictions on columns of \( A \). Further using a full rank mapping matrix forces the use of a large number of bases to map the entire subspace efficiently. Imposing sparsity on the columns of the mapping matrix \( A \) helps mitigate these problems.

Various metrics have been applied to measure and impose sparsity. \( L_p \) norms are one of the most popular measures of sparsity.

This paper imposes the sparsity using an entropic prior. Given a probability distribution \( \theta \), we can write the entropic prior for the distribution as \( P_1(\theta) \propto \exp(-\beta H(\theta)) \) [8]. For the problem at hand distribution \( \theta \) corresponds to the \( p(\pi) \) and \( p(\gamma|\pi_n) \) for \( n = 1, 2, \ldots, N \). Entropic priors will be used to impose sparsity on the columns of the transition matrix \( A \).

Parameters of the Gaussians of subspace \( P \) and \( Q \) are estimated using the traditional EM algorithm, \( p(\pi) \) and \( p(\gamma|\pi) \) are estimated using maximum a posteriori (MAP) estimation with entropic priors.

A posteriori probability is computed for the E-step:

\[
p(\pi, \gamma_m|p_t, q_t) = \frac{p(p_t|q_t, \pi_n, \gamma_m)}{\sum_{n=1}^{M} \sum_{m=1}^{N} p(p_t|q_t, \pi_n, \gamma_m)}
\]

where the joint probability can be written as:

\[
p(p_t|q_t, \pi_n, \gamma_m) = p(q_t|\gamma_m)p(\gamma_m|\pi_n)p(\pi_t)p(\pi_n)
\]

In the M-step, complete data likelihood \( \mathcal{L} \) is maximized:

\[
\mathcal{L} = E_{\gamma, \pi|p_t, q_t, A}\{\log p(p_t, q_t, \pi, \gamma_m)\}
\]

Solving the Equation (3) for the parameters of the Gaussians (means and variances) in subspaces \( P \) yields:

\[
\mu = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M} p(\pi_m, \gamma_m|p_t, q_t)p_t}{\sum_{t=1}^{T} \sum_{m=1}^{M} p(\pi_m, \gamma_m|p_t, q_t)}
\]

\[\sigma^2 = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M} (p(\pi_m, \gamma_m|p_t, q_t) (p_t - \mu)^2}{\sum_{t=1}^{T} \sum_{m=1}^{M} p(\pi_m, \gamma_m|p_t, q_t)}
\]

The parameters of the Gaussians of subspace \( Q \) can be written similarly to those of subspace \( P \) and are omitted here due to space constraints.

Entropic estimation of \( p(A) \) is performed by maximizing the new augmented likelihood \( \mathcal{R} \)

\[
\mathcal{R} = \mathcal{L} + \tau \left( \sum_{m} p(\gamma_m|\pi_n) - 1 \right) + \delta \sum_{m} p(\gamma_m|\pi_n) \log p(\gamma_m|\pi_n)
\]

where \( \tau \) is the Lagrange multiplier and \( \delta \) is the parameter that controls the sparsity. In the interest of space the details about the solution of Equation (6) are omitted in this paper, the complete derivation and details about the the sparsity parameter can be found in Kalganokar and Clements [1].

Given the trained model \( M = \{ \Pi, \Gamma, A \} \) and \( p_u \), the minimum mean square error (MMSE) estimate of \( q_t \) is given by Equation (7)

\[
\hat{q}_t = E_{q|p} (q|p) = \int q \cdot p(q|p) \, dq
\]

The conditional probability \( p(q|p) \) can be expressed as the marginal of the joint probability. Using Equations (2) and (7) the MMSE estimate of \( q_t \) can be written as

\[
\hat{q}_t = \sum_{m=1}^{M} \sum_{n=1}^{N} \mu_n^u \left( \sum_{n=1}^{N} p(p_t|\gamma_m)p(\gamma_m|\pi_n)p(\pi_n) \right)
\]

3. NONLINEAR DISTORTION AND MODEL ADAPTATION

3.1. Nonlinear Distortion of Speech Cepstra

The power spectrum of noisy speech (\( Y \)) can be expressed as the function of the power spectrum of clean speech (\( X \)) and noise (\( N \)) as shown in the equation below:

\[
|Y|^2 = |X|^2 + |N|^2 + 2|X|\cdot|N|\cos(\phi)
\]

where \( \phi \) is the relative phase between the speech and the noise. This relationship can be easily extend to MFCC’s. MFCC’s are extracted from the power spectrum by computing the DCT of the log of the output of each Mel filter bank. Let \( y, x \) and \( n \) represent the cepstra of noisy speech, clean speech, and noise respectively. The relationship between cepstra of noisy speech and clean speech can be written as

\[
y = n + C \log(1 + e^{D(x-n)} + 2ae^{D(x-n)/2})
\]

where \( C \) is the DCT matrix, \( D \) is the pseudo inverse of the DCT matrix, and \( a \) represents the contribution of the phase term \( \phi \) [9].

A modified form of the Equation (10) can help provide a better insight into the nonlinearity and its application for model adaptation. Using the conventions suggested by McAulay and Malpass let ‘\( u \)' be the a priori SNR and ‘\( v \)' be a posteriori SNR, we can rewrite the Equation (10) as

\[
v = C \log(1 + e^{Du} + 2ae^{Du/2})
\]

\[
v = C \log(1 + e^{Du}) \quad \text{if} \quad a = 0
\]

For reduced complexity, most of the model adaptation algorithms ignore the phase term (\( a = 0 \)) in Equation (11) and use the simplified version (12). Figure 1 shows data represented by the two models. The solid line represents the model as captured by the Equation (12) where the red scatter plot shows the true relationship between the a priori and a posteriori SNR.

3.2. Model Adaptation using Probabilistic Space Maps

Probabilistic space maps provide an ideal framework for performing model adaptations. The subspace \( P \) represents \( u = x - n \) and the subspace \( Q \) represents \( v = y - n \). The experiments presented in this paper use Aurora 2 [10] speech database, which has clean and multi-condition training data. Using this data we compute the cepstra for speech (noisy and clean) and noise and use these to train the model
Given the trained model HMM are adapted using the MMSE estimator (Equation (8)) on the sigma points \([11]\) of the distribution of \(u\) to estimate the parameters for the distribution of \(v\). The sigma points is a set of \((N_u + 1)\) points and associated weights \((w_i)\) deterministically chosen from a distribution (e.g., \(p(u)\)) such that this set of points completely characterizes the said distribution. For a Gaussian distribution these points must completely capture the mean and covariance. The weights can be positive or negative but must sum to unity, and \(N_u\) is the dimension of the vector in subspace \(Q\).

The sigma points for a Gaussian with mean \(\mu_u\) and Covariance \(\Sigma_u\) are given by Equation (13).

\[
\begin{align*}
    u^{(0)} &= \mu_u \\
    w^{(0)} &= 1 - \frac{N_u}{3} \\
    u^{(k)} &= \mu_u \pm \sqrt{\frac{N_u}{1-w^{(0)}}} \Sigma_u \\
    w^{(k)} &= \frac{1 - w^{(0)}}{2N_u}
\end{align*}
\]

In the presence of noise the Gaussians of the clean HMM model are adapted using the statistics of the noise. The new HMM is used for recognition. Estimation of the mean and the variance of the distribution of \(p(y)\) is performed using Algorithm 1.

### Algorithm 1: Adapting the Gaussians in HMM for Noisy Conditions

1. Estimate the noise statistics \([\mu_n, \sigma_n^2]\) for an utterance.
2. For Each Gaussian in the model trained on clean speech do
   3. Estimate the parameters of distribution of \(u\) using \(\mu_u = \mu_x - \mu_n\) and \(\sigma_u^2 = \sigma_n^2 + \sigma_u^2\).
4. Compute the sigma points for distribution \(p(u)\) using Equation (13).
5. Estimate \(u^{(k)}\) for each sigma point \((u^{(k)})\) using Equation (8).
6. Estimate the mean and variance for \(p(v)\)
   - \(\mu_v = \sum_{k=1}^{N_u+1} w^k u^k\)
   - \(\sigma_v^2 = \sum_{k=1}^{N_u+1} w^k (u^k - \mu_v)(u^k - \mu_v)^T\)
7. Estimate the mean and variance of \(p(y)\) using \(\mu_y = \mu_v + \mu_n\) and \(\sigma_y^2 = \sigma_v^2 + \sigma_n^2\).
8. End for

3.3. Adapting Dynamic Coefficients

Many of the traditional adaptation techniques \([3, 5, 4, 12]\) use a continuous time approximation for derivatives and use the static coefficients nonlinearity to adapt the parameters of dynamic coefficients. This gives some benefits and improves the recognition accuracy. A closer look at the actual scatter data (Figure 2) for the dynamic coefficients reveals that the dynamic coefficients do not obey the same nonlinearity.

Probabilistic space maps provide a good alternative in the absence of an exact functional mapping between the dynamic coefficients. The PS-MAPS extracts the relationship between the \(\Delta\)'s and \(\Delta\Delta\)'s of the \(a\) priori and \(a\) posteriori SNR using the training data without making any assumptions about its nature.

During the training phase new PS-MAPS are generated for the delta’s and delta-deltas following the process similar to the static coefficient. Algorithm 1 is used on these new models to estimate the parameters of distributions of \(p(\Delta y)\) and \(p(\Delta\Delta y)\).

4. EXPERIMENTS AND RESULTS

The effectiveness of the new algorithm was evaluated by conducting a series of experiments on the Aurora 2 connected digit recognition corpora using the HTK speech recognition system \([13]\). Aurora 2 consists of artificially degraded data. Eight different kinds of noise with SNR varying from 0 dB to 20 dB are added to clean speech to generate the test and training sets.

The acoustic models were trained using the scripts distributed with Aurora 2. Standard 39-dimensional MFCC features consisting of 13 cepstral features plus delta and delta-delta features were used in the experiments. The cepstral coefficient of order zero (C0) is used instead of log energy. Noise is assumed to be stationary and Gaussian with diagonal covariance. The first and last twenty frames of each utterance are used to estimate the mean and covariance of the noise for that utterance. The PS-MAPS were trained with 20 Gaussians in each subspace.

Table 1 shows the % word accuracy results for the adapted models under different noise and channel conditions. The models used in these experiments do not compensate for the channel effect, adaptation only accounts for the additive noise.

4.1. Discussion of Results

The PS-MAP adapted models show an improvement in recognition over the clean models. This indicates that the PS-MAPS are definitely moving the clean HMM’s in the right noisy space. Further the performance of the adapted models is better for set B as compared...
### Table 1. Average Recognition Accuracy comparisons for the three adaptation schemes

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subway</td>
<td>75.39</td>
<td>79.54</td>
<td>84.48</td>
<td>81.95</td>
</tr>
<tr>
<td>Babble</td>
<td>75.03</td>
<td>75.88</td>
<td>79.73</td>
<td>78.18</td>
</tr>
<tr>
<td>Car</td>
<td>63.91</td>
<td>75.65</td>
<td>81.92</td>
<td>77.95</td>
</tr>
<tr>
<td>Exhibition</td>
<td>68.47</td>
<td>75.39</td>
<td>85.89</td>
<td>81.10</td>
</tr>
<tr>
<td>Avg</td>
<td>75.39</td>
<td>79.54</td>
<td>84.48</td>
<td>81.95</td>
</tr>
<tr>
<td>Street</td>
<td>55.36</td>
<td>59.93</td>
<td>63.91</td>
<td>65.20</td>
</tr>
<tr>
<td>Airport</td>
<td>65.97</td>
<td>65.97</td>
<td>72.96</td>
<td>67.20</td>
</tr>
<tr>
<td>Station</td>
<td>71.25</td>
<td>75.39</td>
<td>80.70</td>
<td>78.18</td>
</tr>
<tr>
<td>Avg</td>
<td>68.48</td>
<td>72.96</td>
<td>78.18</td>
<td>76.42</td>
</tr>
<tr>
<td>Avg</td>
<td>75.39</td>
<td>79.51</td>
<td>83.89</td>
<td>81.10</td>
</tr>
</tbody>
</table>

Fig. 3. Std. Deviation of $v$ as a function of mean of $u$

Preliminary results presented in this paper show that PS-MAPS are viable alternatives both as a front-end (feature enhancement) and back-end (model adaptation) modules. The results presented in this paper can be improved with the use of distance measure in probability spaces and warrant further investigation.

### 6. REFERENCES


