INCREMENTAL LARGEST MARGIN LINEAR REGRESSION AND MAP ADAPTATION FOR SPEECH SEPARATION IN TELMEDECINE APPLICATIONS

Rusheng Hu, Jian Xue and Yunxin Zhao

Department of Computer Science
University of Missouri, Columbia, MO 65211, USA
{rhe02,jxwr7}@mizzou.edu, zhaoy@missouri.edu

1. Introduction

Telemedicine or telehealth is becoming an important means of providing quality health care to rural areas and elderly people in the United States [1], where an important application of spoken language processing is to provide voice-driven automatic captioning system to hearing impaired users. In such a scenario, a speech conversation carried on between a doctor and a patient is acquired in offices of both sites by desktop microphones, and the acquired speech signals are compressed and transmitted over a dedicated network. At the doctor’s site, the desktop microphone not only captures doctor’s speech, but it also picks up patient’s speech that is interleaved speech signals of doctor and patient, and hence a separation step needs to be taken to break up doctor and patient’s speech into two separate speech streams. The separation procedure consists of segmenting recorded signal into speech and background segments, labeling speech segments as doctor or patient, and sorting the labeled speech streams when judged by human listeners.

Gaussian mixture modeling has been shown to be effective in speaker identification [2]. In an online system, incremental speaker model adaptation is often adopted to cope with mismatch of training and test conditions as well as environmental changes, among which MAP [3] and MLLR [4] are most commonly used. Discriminative adaptation techniques such as maximum mutual information (MMI) [5] and minimum classification error (MCE) [6] are supervised in nature, and hence they are rarely used in online adaptation. In this paper, we propose an unsupervised discriminative linear regression approach for online adaptation. This algorithm aims at maximizing an approximated class separation margin and is easily implemented by a simple modification of the standard EM algorithm. It has the similar properties as those of EM and converges fast.

In the current work, both doctor and patient’s speech models are adapted online by combining MAP parameter estimation with LMLR model transformation. The online adaptation is unsupervised and incremental, using speech streams as the unit for sufficient statistics update. In addition, a discount technique is proposed for LMLR to reduce the effect of unreliable initial models on segmental adaptation. This new technique has significantly improved performance of the speech stream separation system, leading to nearly perfectly separated speech streams when judged by human listeners.

The rest of this paper is organized as follows. The proposed LMLR is introduced in section 2. The online incremental adaptation algorithms are described in section 3. The doctor-patient statistical classification method is presented in section 4. Experimental results are summarized in section 5, and a conclusion is made in section 6.

2. Largest Margin Linear Regression

In discriminative training, given any observation \( x_i \), the classifier outputs the class ID \( y_i \), based on the MAP decision rule as follows:

\[
y_i = \arg \max_\theta P(G \mid x_i) = \arg \max_\theta P(x_i \mid \theta) P(G)
\]

where \( \theta_0 \) denotes the model parameters of class \( G \) and \( F(x \mid \theta_0) \) is the discriminant function. Discriminative training methods such as MMI and MCE attempt to minimize the classification error rate by utilizing both correct and confusion information in the training phase. Recently, a largest margin algorithm was proposed for discriminative training [7]. Assume that the model space has \( M \) classes and \( \theta = [\theta_0, \theta_1, ..., \theta_M] \) denote the parameter set for the \( M \) models. Given the labeled training data \( \{ x_i, y_i \} \) with \( y_i \) the correct label of feature vector \( x_i \), the class separate margin for \( x_i \) is defined as

\[
d(x_i) = F(x_i \mid \theta_0) - \max_{\gamma \neq y_i} F(x_i \mid \theta_\gamma)
\]

or alternatively,
\[ d(x_i) = \log F(x_i \mid \theta_c) - \max_{\theta \neq \theta_c} \log F(x_i \mid \theta) \]  
(3)
and the support region is defined by the subset of observations which are correctly classified but are relatively close to the classification boundary, i.e.,
\[ S = \{ x_i \mid 0 \leq d(x_i) \leq \gamma \} \]  
(4)
where \( \gamma > 0 \) is a pre-set upper bound.

In unsupervised training, the correct label \( y_i \) is unknown and the above methods can not be applied. Therefore, a new training criterion is proposed. Consider a classifier with possible error rate \( \varepsilon \) with \( \varepsilon \) small. Denoting the classified label for \( x_i \) by \( \hat{y}^C \), the total class separation margin can be approximated by
\[ D = \sum_{i=1}^{n \times n} \left( \log F(x_i \mid \theta_c) - \max_{\theta \neq \theta_c} \log F(x_i \mid \theta) \right) \]  
(5)

The idea is to adapt the Gaussian mixture models \( \theta = \{ \theta_1, \theta_2, \ldots, \theta_n \} \) by maximizing the following objective function
\[ \hat{\theta} = \arg \max_{\theta} \left( \sum \log F(x_i \mid \theta) - \lambda \sum_{i=1}^{n \times n} \max_{\theta \neq \theta_c} \log F(x_i \mid \theta) \right) \]  
(6)
where \( 0 < \lambda < 1 \) is a tuning factor that controls the rate of discrimination and is determined empirically. However, this kind of objective function might be unbounded and constraints must be introduced. In [7], a batch algorithm called iterative localized optimization algorithm was introduced for constrained optimization. In the current work, we extend the batch algorithm into an incremental localized optimization algorithm as the Algorithm 1.

**Algorithm 1.**

1) Initialize model set \( \Theta \), and set \( k=1 \);
2) Classify incoming observation \( x_k \), and denote the classified label by \( \hat{y}^C \);
3) Estimate model \( \hat{\theta}^{(k)} \) by maximizing the class separation margin using its accumulated discriminative statistics.
4) Set \( k = k+1 \), and go to step 2) for a new observation.

In the separation of doctor and patient’s speech, the online acquired observation data are in segments, \( X = \{ x_i, t=1, \ldots, T \} \), and only two classes are present. The two speaker models are denoted by \( S_1, S_2 \). Assuming diagonal covariance matrix, the optimization in step 3) has a closed form solution for diagonal linear transformation of Gaussian mixture means. The adapted mean vector is given by:
\[ \hat{\mu} = A \mu + b \]  
(7)
where \( [A, b] \) is the extended transformation matrix. Assume that \( M_1 \) and \( M_2 \) are the total number of Gaussians in classes \( S_1 \) and \( S_2 \), and the segment \( X = \{ x_i, t=1, \ldots, T \} \) is classified as class \( S_i \). When the priors of two classes are equal, it can be shown that maximizing the margin in step 3) is equivalent to maximizing the following auxiliary function
\[ \phi(\theta, \hat{\theta}) = \sum_{i=1}^{M_i} p(m_i \mid x_i) \log \pi_{m_i} N(x_i \mid \hat{\mu}_{m_i}) \]  
(8)
\[ -\lambda \sum_{i=1}^{M_i} p(m_i \mid x_i) \log \pi_{m_i} N(x_i \mid \hat{\mu}_{m_i}) \]
where \( \hat{\theta} = \{ \hat{\theta}_1, \hat{\theta}_2 \} \) is the transformed parameter set to be estimated. In the case of diagonal covariance matrix, the diagonal transformation can be found by solving the following system of equations:
\[ \begin{align*}
\hat{\mu}^{\pi_{m_i}} & = \mu_{m_i} + \frac{1}{\sigma_{m_i}^2} \sum_{i=1}^{M_i} n_{m_i} \mu_{m_i} (\hat{\mu}_{m_i} - b_{m_i}^{\pi_{m_i}}) / \sigma_{m_i}^2 \\
\hat{\sigma}^{\pi_{m_i}} & = \sigma_{m_i} \sum_{i=1}^{M_i} n_{m_i} \sigma_{m_i}^2 / \sigma_{m_i}^2
\end{align*} \]  
(9)
\[ \begin{align*}
\hat{\mu}^{\pi_{m_i}} & = \mu_{m_i} + \frac{1}{\sigma_{m_i}^2} \sum_{i=1}^{M_i} n_{m_i} \mu_{m_i} (\hat{\mu}_{m_i} - a_{m_i}^{\pi_{m_i}}) / \sigma_{m_i}^2 \\
\hat{\sigma}^{\pi_{m_i}} & = \sigma_{m_i} \sum_{i=1}^{M_i} n_{m_i} \sigma_{m_i}^2 / \sigma_{m_i}^2
\end{align*} \]  
(10)
\[ \begin{align*}
\hat{\mu}^{\pi_{m_i}} & = \mu_{m_i} + \frac{1}{\sigma_{m_i}^2} \sum_{i=1}^{M_i} n_{m_i} \mu_{m_i} (\hat{\mu}_{m_i} - a_{m_i}^{\pi_{m_i}}) / \sigma_{m_i}^2 \\
\hat{\sigma}^{\pi_{m_i}} & = \sigma_{m_i} \sum_{i=1}^{M_i} n_{m_i} \sigma_{m_i}^2 / \sigma_{m_i}^2
\end{align*} \]  
(11)
\[ \begin{align*}
\hat{\mu}^{\pi_{m_i}} & = \mu_{m_i} + \frac{1}{\sigma_{m_i}^2} \sum_{i=1}^{M_i} n_{m_i} \mu_{m_i} (\hat{\mu}_{m_i} - a_{m_i}^{\pi_{m_i}}) / \sigma_{m_i}^2 \\
\hat{\sigma}^{\pi_{m_i}} & = \sigma_{m_i} \sum_{i=1}^{M_i} n_{m_i} \sigma_{m_i}^2 / \sigma_{m_i}^2
\end{align*} \]  
(12)
Note that when \( \lambda = 0 \), this algorithm is degenerated to conventional MLLR.

### 3. Online Model Adaptation

In online applications, it is desirable to use sequentially updated models in sufficient statistics computation and in data classification. For MAP, a quasi-Bayes learning algorithm was proposed where the sufficient statistics are accumulated into the prior parameters [8]. For online learning of MLLR, incremental EM algorithm is often used [9]. The incremental EM-like algorithm for LMLR as described in Section 2 differs from the incremental EM in that the incremental LMLR needs to store the accumulated statistics for each discriminative class, and hence requires more memory when the model space is large.

In online EM, a forgetting factor \( \rho(t) \) is often used to reduce the effect of inappropriate old statistics caused by the earlier inaccurate estimator [10]. The forgetting factor \( \rho(t) \) is defined as:
\[ \rho(t) = 1 - \frac{1 - c}{c(t + d - c)} \]  
(13)
where \( c \) and \( d \) are constants with \( 0 < c < 1 \) and \( d \geq 1 \). As \( t \to \infty \), \( \rho(t) \) approaches 1. MAP and linear regression methods have different properties. Linear regression transforms the model location when only applied to the mean; MAP adjusts the location gradually and also shrinks the shape. Therefore, they can be combined together to improve system performance. The combination can be performed at model level or classifier level. At the classifier level, outputs of individual classifiers are combined under some fusion strategies for hope of improving accuracy [11]. At the model level, combination of the model parameters aims at delivering more accurate models, which consequently improves classification accuracy.
The current study focuses on combination at the model level. The combined model is in the form of
\[ \tilde{\theta}_{\text{sum}} = \tau \tilde{\theta}_{\text{LMLR}} + (1 - \tau) \tilde{\theta}_{\text{MAP}} \]
where \( \tilde{\theta}_{\text{LMLR}} \) denotes the LMLR transformed model, \( \theta < 1 \) is a tuning factor, which is obtained empirically. The online adaptation algorithm is described in Algorithm 2.

**Algorithm 2.**

1. Initialize GMM parameter \( \theta^0 \), set \( k = 1 \);
2. Compute statistics for incoming segment \( X_t \):
   \[ S^{(n)}_{\text{LMLR}}(\theta^{i-1}) \] and
   \[ S^{(n)}_{\text{MAP}}(\theta^{i-1}) \] for \( x_t \in X \);
3. Estimate parameters using EM
   a. For LMLR, use
   \[ S^{(i)}_{\text{LMLR}} = \sum_{t} \left( \prod_{s} \rho(s) \right) S_{\text{LMLR}}(\theta^{(i-1)}) \]
   and
   \[ \text{update prior parameters, where} \ \rho(s) \ \text{is in the form of (13)}; \]
   b. For MAP, initialize \( S^{(i)}_{\text{MAP}} = S^{(n)}_{\text{MAP}}(\theta^{i-1}) \) and
   \[ \text{iteratively estimate model parameters and} \]
   update statistics;
4. Compute \( \tilde{\theta}^{(i)}_{\text{LMLR}} \) and \( \tilde{\theta}^{(i)}_{\text{MAP}} \) and update prior parameters as in incremental MAP;
5. Set \( \tilde{\theta}^{(i)} = \tau \tilde{\theta}^{(i)}_{\text{LMLR}} + (1 - \tau) \tilde{\theta}^{(i)}_{\text{MAP}} \) and export \( \tilde{\theta}^{(i)} \) to classifier to classify next segment, set \( k = k + 1 \).
6. Go to step 2) for new segment.

### 4. Doctor-Patient Classification

#### 4.1. Speech-pause Segmentation

Normally, telemedicine offices are quiet and signal-to-noise ratio (SNR) is high in speech recordings. In such a condition, pauses that occur at most speech turns, referred to as turn pauses, are characterized by low energy regions as well as non-negligible durations. For example, measured in a typical university telemedicine office, SNR is approximately 23 dB and the pause durations are mostly longer than 100 ms. Based on this characteristic of telemedicine speech, an energy and duration based segmentation algorithm is designed to separate speech segments by pauses. A short-time cepstral analysis is first performed. The 0th-order cepstral coefficients, or speech segments by pauses. A short-time cepstral analysis is on this characteristic of telemedicine speech, an energy and

#### 4.2. Cepstral Analysis and Statistical Segment Labeling

A speech feature vector \( x_t \) consists of \( P \) MFCCs, including \( c_{0,0} \), and their first-order temporal regressions over five analysis frames. Cepstral mean is computed for each segment and cepstral mean removal is applied. The speech feature \( x_t \) is assumed to have class-conditional distributions of Gaussian mixture densities:

\[ p(x_t; \theta_d) = \sum_{m=1}^{M} \alpha_{m,M} N(x_t; \mu_{d,m}, \Sigma_{d,m}) \]

for doctor and

\[ p(x_t; \theta_p) = \sum_{m=1}^{M} \alpha_{m,P} N(x_t; \mu_{p,m}, \Sigma_{p,m}) \]

for patient.

The parameters \( \theta_d \) and \( \theta_p \) of the two densities are initially estimated from labeled training data by maximum likelihood estimation (MLE) and they are adapted online using the proposed incremental adaptation algorithms.

For statistical segment labeling, the discriminant function is defined as the log likelihood ratio

\[ D(x; \Theta) = \log p(x \mid \theta_d) - \log p(x \mid \theta_p) \]

The reason of applying the energy thresholds \( 10^{-a} \) is that low energy frames may not carry reliable spectral information for segment classification.

Denote the label of speech segment \( X \) by \( C(X) \). The decision rule of the statistical classifier is simply the likelihood ratio test

\[ \text{If } D(X; \Theta) > 0 \text{, then } C(X) = \text{Doctor}; \]
\[ \text{else } C(X) = \text{Patient}. \]

#### 5. Experiments

##### 5.1. Evaluation Measures

The performance of the speech stream separation system was evaluated at the frame and the turn levels. The speech recordings were hand-labeled into speech turns, and each turn was labeled as doctor or patient. These manually generated labels were used as references in calculating error rates.

At the frame level, errors of miss and false alarm are counted, where a missed frame (MF) accounts for the error of mislabeling a doctor’s speech frame as patient’s, and a false-alarmed frame (FF) accounts for the error of mislabeling a patient’s speech frame as doctor’s. The frame error rate (FER) is defined as

\[ \text{FER} = \frac{\text{number of MFs} + \text{number of FFs}}{\text{total reference speech frames}} \times 100\% \]

At the turn level, a missed turn (MT) is defined as a doctor’s speech turn containing less than 75% frames labeled as doctor’s, and a false-alarmed turn (FT) is defined as a patient’s speech turn containing less than 75% frames labeled as patient’s. The turn error rate (SER) is then defined as

\[ \text{SER} = \frac{\text{number of MTs} + \text{number of FTs}}{\text{total reference speech turns}} \times 100\% \]

##### 5.2. Experimental Data

Two sets of telemedicine conversation data were collected from the University of Missouri’s telemedicine network. These data consisted of one female doctor’s conversations with two female patients where the patients’ speech were transmitted over the telemedicine network. Details of the two recorded are summarized in Table 1. In addition, 10-minute dictation speech was collected from the doctor in a computer
5.3. Experimental Results

The speech-pause segmentation energy threshold $\sigma_p$ was determined as 12.5 dB for both Test-1 and Test-2. MFCC analysis was performed with window size of 20 ms and 10 ms overlap between windows. Speech feature consisted of 16 MFCCs and their first-order derivatives. Both doctor and patient’s GMMs were chosen to have 16 Gaussians. The doctor’s GMM was trained by her offline dictation speech, and patient’s GMMs were trained using telemedicine speech of speakers excluding themselves.

In online discriminative adaptation, it is expected that the adjustment of models made at each step remains moderate of speakers excluding themselves. The performance of incremental LMLR was compared with incremental MLLR and MAP. The test results are shown in Table 2.

Table 2. Comparison of MLLR, MAP and LMLR

<table>
<thead>
<tr>
<th>SER/FER</th>
<th>No Adapt</th>
<th>MLLR</th>
<th>MAP</th>
<th>LMLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>11.8/7.03</td>
<td>2.73</td>
<td>4.55/2.97</td>
<td>2.73/3.03</td>
</tr>
<tr>
<td>Test 2</td>
<td>7.80/5.11</td>
<td>3.55/2.84</td>
<td>5.67/3.49</td>
<td>3.55/2.73</td>
</tr>
</tbody>
</table>

As we can see, LMLR produced improved results over MLLR, where approximately 6% error reduction on FER was achieved on average. It is also shown that the error patterns of MAP and linear regression are different. For example, in test 1, the SER of MAP was the largest while the FER remained the smallest. Therefore, there are potential benefits in combining both methods. Incremental LMLR and MAP were therefore combined, and the test results at different combination weight $\tau$ levels are listed in tables 3 and 4.

Table 3. SER of LMLR+MAP vs. $\tau$

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>2.73</td>
<td>2.73</td>
<td>2.73</td>
<td>2.73</td>
<td>2.73</td>
<td>4.55</td>
</tr>
<tr>
<td>Test 2</td>
<td>3.55</td>
<td>3.13</td>
<td>2.13</td>
<td>2.13</td>
<td>2.84</td>
<td>5.67</td>
</tr>
</tbody>
</table>

Table 4. FER of LMLR+MAP vs. $\tau$

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>3.03</td>
<td>2.75</td>
<td>2.63</td>
<td>2.29</td>
<td>2.26</td>
<td>2.97</td>
</tr>
<tr>
<td>Test 2</td>
<td>2.73</td>
<td>2.56</td>
<td>2.47</td>
<td>2.73</td>
<td>2.67</td>
<td>3.49</td>
</tr>
</tbody>
</table>

The results indicate that the combination of LMLR and MAP improved the performance, and the good range of combination weight $\tau$ was around 0.4-0.6, depending on the error pattern of LMLR and MAP in different test sets. When judged by human listeners, the segment and frame errors as indicated in tables 3 and 4 were basically not perceptible in the separated speech streams.

6. Conclusion

An unsupervised discriminative linear regression method, called largest margin linear regression (LMLR), is proposed to improve the performance of speech stream separation for telemedicine application. Online incremental adaptation based on combined MAP and LMLR methods are employed to update doctor and patient’s GMM models for accurate classification of speech segments. Experimental results on real telemedicine data show that LMLR is superior to MLLR and combining LMLR and MAP during online model adaptation is highly effective for the proposed telemedicine application.

7. Acknowledgements

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 DC03430-01A2 and the National Science Foundation under the grant NSF EIA 9911095.

8. References

[1] National Center for Health Statistics (NCHS), U.S. Department of Health and Human Services, Data from the National Health Interview Survey, Series 10, No. 188, Table 1, B, C, 1994.