IMPROVED JACOBIAN ADAPTATION FOR FAST ACOUSTIC MODEL ADAPTATION IN NOISY SPEECH RECOGNITION*

Ruhi Sarikaya and John H. L. Hansen
Robust Speech Processing Laboratory, The Center for Spoken Language Research
University of Colorado at Boulder, Boulder, CO 80303
http://cslr.colorado.edu  {sarikaya, jhhlh}@cslr.colorado.edu

ABSTRACT

This paper describes two algorithms to improve a previously proposed Jacobian adaptation (JA) technique for fast acoustic speech recognizer model adaptation in environmental noise. The first technique introduces a new bias term, that is a function of the reference noise estimate to account for the mismatch between the reference noise estimate and noise component of the noisy speech spectrum. This functional mismatch bias is quite general, and here we choose to represent it as a linear function of the reference noise estimate. The second algorithm uses a more accurate relationship between the log and linear spectral domain versions of the HMM parameters. The combination of these new techniques achieves an increase of between 3.3-10.9% in recognition accuracy in adapting from automobile highway noise (HWY1) to such low-frequency noise sources as IBM PC cooling fan noise (PS2), large city street noise (LCI) and HWY2 (a different highway noise with different automobile at different signal-to-noise ratio (SNR)) over the original Jacobian adaptation in context independent phone recognition on TIMIT database.

1. INTRODUCTION

Speech recognition system performance suffers from mismatch between training and test conditions. One factor that contributes to mismatch is additive environmental noise. A common goal of all techniques for robust speech recognition is to seek a reduction in this mismatch. Techniques for robust speech recognition can be grouped into three broad classes. The first one is feature based techniques, where speech features are modified in such a way that the effect of the noise is removed [3, 2]. The second class comprises of techniques which formulate inherently robust features to noise [1]. The third class of techniques are based on model compensation or adaptation [10, 7, 8]. Combining these techniques is also promising for robust speech recognition.

One example of model compensation techniques is parallel model combination (PMC) [7]. Although it is an effective adaptation technique, intense computational requirements limits its use in practical applications. Recently, a new technique has been proposed which nearly doubles the speed of PMC [6]. However, for large vocabulary continuous speech recognition (LVCSR) systems with over 50000 distributions, even the fast PMC is sufficient to adapt the models to rapidly changing environments. One alternative approach to fast model adaptation is a Jacobian Approach (JA) [8, 9]. The basis of the technique is related to vector Taylor series method [11] in the sense that the first-order coefficient of the series is equivalent to a Jacobian matrix. The Jacobian approach adapts the acoustic models trained in the reference noise condition to the target noise condition when the noise changes. The underlying idea for the Jacobian approach is to express the change in the noisy speech model given that the reference noise changes towards the target noise. This relationship is given in terms of a Jacobian matrix. In order to achieve effective recognition performance with JA, the target and reference noises should be close and the change in noisy speech statistics should be in the linear range of the Jacobian approximation. However, two of the assumptions in the derivation of the Jacobian adaptation over-simplifies the relationships between the parameters. The first one is between the noise component of the noisy speech and the reference noise, and the second is between the log and linear spectral domain versions of the transformed HMM parameters.

In this paper, we formulate two more sophisticated techniques, one for each over-simplifications, to improve recognition performance. The first algorithm introduces a new term in the noise component of the noisy speech spectrum to account for the mismatch between the actual reference noise spectrum and noise component of the noisy speech spectrum estimated from the HMM parameters. This term also adds extra flexibility to the system and generalizes the baseline JA technique. The second algorithm proposes a more accurate relationship between the log and linear power spectral domain versions of the noisy speech and noise models. A set of new Jacobian matrices are derived using these two algorithms.

The remainder of the paper is organized as follows. In Sec. 2 we review baseline Jacobian adaptation. Sec. 3 presents a generalized Jacobian adaptation followed by the formulation of a more accurate relationship between log and linear power spectral domains. In Sec. 5 we present a set of context independent phone recognition experiments on TIMIT database using the new as well as baseline JA techniques. Concluding remarks are presented in Sec. 6.

---

*This work was supported in part by a grant from the DARPA.
2. BASELINE JACOBIAN ADAPTATION

In this section we review the Jacobian approach given in [8, 9]. In this framework, acoustic models are assumed to be nonlinear functions of the speech and additive noise. If we have a set of acoustic models compensated or trained in reference noise, a change in noise will propagate to the model domain. Let us assume that we have two noise conditions, reference noise \( N_r \), and target noise \( N_t \). Furthermore, we assume that the difference between the statistics of \( N_r \) and \( N_t \) is small such that the change in noisy speech model statistics stays within the linear range of the Jacobian adaptation. The relation between acoustic models in the target noise condition and the reference condition is linked via a Jacobian matrix. The main assumption underlying the basic Jacobian method is that the change in noise condition should be small, in which case the method is very effective. Otherwise, the method does not work as well.

The underlying idea for Jacobian adaptation is a basic calculus theorem. If a vector, \( Y = (y_1, y_2, \ldots, y_n)^T \) is an analytic function of a vector, \( X = (x_1, x_2, \ldots, x_n)^T \), a small change, \( \Delta X \) in \( X \) causes a small change, \( \Delta Y \), in \( Y \). The relation between them is as follows:

\[
\Delta Y = \frac{\partial Y}{\partial X} \Delta X
\]

where \( \Delta Y = (\Delta y_1, \Delta y_2, \ldots, \Delta y_n)^T \), \( \Delta X = (\Delta x_1, \Delta x_2, \ldots, \Delta x_n)^T \), and the matrix \( \frac{\partial Y}{\partial X} \) is called the Jacobian matrix between \( Y \) and \( X \).

The relation between noisy speech cepstrum coefficients, clean speech and noise cepstra can be expressed as in Eqn. 2. Note that subscript \( r \) denotes that the noise source is the reference noise. The relationship between noisy speech and noise cepstra is nonlinear. Suppose that the vectors are long enough such that the Fourier matrix approximates the Fourier integral. In this case, following Eqn. 1, the relationship between small changes in the noisy speech and noise cepstra can be given as in Eqn. 3.

\[
C_{s+N_r} = F^* \log(\exp(FC_{s}) + \exp(FC_{N_r}))
\]

\[
\Delta C_{s+N_r} = \frac{\partial C_{s+N_r}}{\partial C_{N_r}} \Delta C_{N_r}
\]

Suppose that \( S, N_r, \) and \( S+N_r \) denote the speech, noise and noisy speech spectra respectively, in the linear spectral domain. Furthermore, let \( C_S, C_{N_r}, \) and \( C_{S+N_r} \) denote speech, noise and noisy speech cepstra, respectively. These parameters are related via the following set of equations,

\[
\begin{align*}
\log S &= FC_{s} \\
\log N_r &= FC_{N_r} \\
\log(S + N_r) &= FC_{s+N_r}
\end{align*}
\]

where \( F \) denotes the inverse discrete cosine transform (DCT) matrix. The Jacobian matrix between the noisy speech and noise cepstra can be written using the above relationships.

\[
\begin{align*}
\frac{\partial C_{s+N_r}}{\partial C_{N_r}} &= \begin{bmatrix}
\frac{\partial \log(S+N_r)}{\partial \log(S+N_r)} & \frac{\partial \log(S+N_r)}{\partial \log(S+N_r)} & \frac{\partial \log(S+N_r)}{\partial \log(S+N_r)} & \frac{\partial \log(S+N_r)}{\partial \log(S+N_r)}
\end{bmatrix} \\
F^* &= \begin{bmatrix}
\frac{\partial \log(N_r)}{\partial \log(N_r)} & \frac{\partial \log(N_r)}{\partial \log(N_r)} & \frac{\partial \log(N_r)}{\partial \log(N_r)} & \frac{\partial \log(N_r)}{\partial \log(N_r)}
\end{bmatrix}
\end{align*}
\]

where \( F^* \) denotes the discrete cosine transform matrix and \( I \) denotes the identity matrix. Therefore, when \( C_{N_r} \) changes into \( C_{N_t} \), the noisy speech spectrum, \( C_{S+N_r} \), changes into \( C_{S+N_t} \), where the subscript \( t \) stands for the target noise.

\[
\begin{align*}
C_{S+N_t} &= C_{S+N_r} + \frac{\partial C_{S+N_r}}{\partial C_{N_r}}(C_{N_t} - C_{N_r}) \\
&= C_{S+N_r} + J^t_r(C_{N_t} - C_{N_r})
\end{align*}
\]

Here, \( J^t_r \) denotes the Jacobian matrix for the reference noise.

3. GENERALIZED JACOBIAN ADAPTATION

One key assumption in the derivation of the Jacobian matrix is that the noise component of the noisy speech spectrum is equal to the linear power spectrum of the reference noise estimate. It is this assumption that connects the noisy speech cepstrum to reference noise cepstrum. However, in general, these terms are different because of the nonstationarity of the noise and maximum likelihood (ML) training. The noise component in the ML estimate of the noisy speech is different than the ML estimate of the noise alone. Therefore, in order to compensate for the difference between the two, we introduce an additive term which is a function of the noise estimate such that the two are equalized (this is not done mixture by mixture, but rather set globally). A new Jacobian matrix is derived taking this into consideration. Let us denote the noise component of the noisy speech spectrum as \( N^*_r \) rather than \( N_r \). Therefore, the \( \frac{\partial}{\partial N^*_r} \) term in Eqn. 5 is not equal to \( I \), and the baseline Jacobian matrix, \( J^* \) is:

\[
J^*_r = F^* \frac{N_r}{S + N^*_r} F
\]

In order to compensate for the difference between \( N^*_r \) and \( N_r \), we introduce an additive term \( f(N_r) \) which is a function of the noise estimate such that \( N^*_r + f(N_r) \approx N_r \). The derivation of the Jacobian matrix is given in Eqn. 8.

\[
\begin{align*}
J_r &= \frac{\partial C_{(S+N^*_r) + f(N_r)}}{\partial \log((S+N^*_r) + f(N_r))} \\
& \quad \frac{\partial \log((S+N^*_r) + f(N_r))}{\partial \log((S+N^*_r) + f(N_r))} \\
&= F^* \frac{N_r}{(S + N^*_r) + f(N_r)} F
\end{align*}
\]
Now, the next step is to estimate \( f(.) \). At this point we note the similarity between the parametric Wiener filtering approach [4] and our problem. In image restoration applications, using a parametric Wiener filter is shown to preserve high SNR frequency components [5]. Although, the denominator of Eqn. 8 has the same form as the generalized Wiener filter, the numerator is different and we are not applying a Wiener filter. However, the above result for image restoration suggests that in a noisy speech spectrum using a scaled version of the noise may emphasize high SNR frequency components.

Therefore, we simplify \( f(N_r) = \alpha N_r \), where \( \alpha \) is a constant and must be estimated. Next, the Jacobian matrix will be in the form given in Eqn. 9, with \( \alpha \) determined experimentally. In Fig. 1, we present recognition accuracy versus \( \alpha \), where accuracy is maximized within 0.4 \( \leq \alpha \leq 0.6 \). Comparable ranges of \( \alpha \) maximized recognition accuracy in adapting to other noise sources. The effective range of \( \alpha \) where recognition is high narrower for PS2 noise compared to HWY2 and LCI. Note that \( \alpha = 0 \) corresponds to the baseline Jacobian approach. As \( \alpha \to \infty \), \( [J_{ij}]_{ij} \to 0 \), and the initial models remain the same. This form of the parametric relationship provides extra flexibility for improved performance.

\[
\frac{\partial C}{\partial C_{N_r}} = F^* N_r (S + N_r) + \alpha N_r F
\]  

4. TRANSFORMATION OF DISTRIBUTIONS FROM LOG-SPECTRAL TO LINEAR SPECTRAL DOMAIN

The second algorithm proposed in this study concerns a more accurate relationship between the log and linear power spectral domain versions of the HMM parameters. In the baseline, as well as for generalized Jacobian adaptation, the relation between log-spectral and linear spectral domains for model statistics is assumed to be a simple logarithm operation. Although this is true for each observation point, it is an approximation for the transformation of model statistics into linear domain. We use a more accurate relationship previously proposed in the context of parallel model combination [7].

The \( i^{th} \) mean vector component in the log domain is given in terms of the mean in a linear spectral domain as:

\[
\mu_i = \log \mu_i - \frac{1}{2} \log \left( \frac{\Sigma_{ii}}{\mu_i^2} + 1 \right)
\]

where \( \Sigma_{ii} \) denotes the \( i^{th} \) diagonal element of the covariance matrix in the linear spectral domain.

Therefore, we can modify equations for the baseline Jacobian adaptation by incorporating this transformation into the derivation. For the sake of clarity, let us denote the log domain by \( \theta \). In all the equations, terms of the form \( \frac{\partial \log(X)}{\partial X} \) should be replaced by \( \frac{\partial \log(X)}{\partial \theta} \). In the following derivation, we set \( X = S + N \):

\[
\frac{\partial \theta}{\theta} \frac{\partial S}{S + N_r} = \frac{\theta}{\theta} \frac{\partial (S + N_r)}{S + N_r} \left( \log(S + N_r) - \frac{1}{2} \log \left( \frac{\Sigma_{DD} + N_r}{(S + N_r)^2} + 1 \right) \right)
\]

\[
= \frac{1}{S + N_r} \frac{2 \Sigma_{DD} + N_r + (S + N_r)^2}{\Sigma_{DD} + N_r + (S + N_r)^2}
\]

where \( \Sigma_{DD} \) is the covariance vector for noisy speech in linear spectral domain, which has the diagonal terms of the covariance matrix as its components. The same relationship holds for \( \frac{\partial \log(N_r)}{\partial \theta} \) as well.

5. EVALUATIONS

The task in the experimental evaluation is to perform context independent phone recognition using an 8 KHz band-limited version of the TIMIT database. We employ four types of noise sources: HWY1, HWY2, PS2 and LCI. HWY1 is recorded inside a Ford Taurus driving at 65 mph on highway and used as our reference noise. HWY2 is collected within a Chevy Blazer on asphalt road while driving 55 mph. PS2 is IBM PC cooling fan noise. LCI is the large city street traffic noise. HWY1 noise is selected as our reference noise and the training data is corrupted at 5 dB SNR. The adaptation is performed from HWY1 towards HWY2, PS2, and LCI. The SNRs for HWY2, PS2 and LCI are 20dB, 10dB and 20dB respectively. The noise observations are obtained from the first 8-10 frames (i.e., 80-100 msec) of a sample noisy speech test file.

The phone recognition system used in this study is based on continuous density HMMs. Here, 45 phones are modeled by a 3-state, left-to-right HMM topology. For each HMM state, a range of 4 to 16 mixture densities are used depending on available training data to characterize observation probability densities. The speech waveform is parameterized every 10 msec by a vector consisting of 12 static Mel-frequency cepstral coefficients (MFCC), 12 delta MFCC, energy (e.g., c0) and delta energy (e.g., delta c0). The data used for training and test are the standard training and the test sets of the TIMIT database.

A series of simulations were conducted to investigate the viability of the baseline and improved Jacobian adaptation (JA) algorithms for recognition. The idea is to adapt
Table 1: Recognition results for clean, matched and mismatched experiments for the noise sources considered in the simulations.

<table>
<thead>
<tr>
<th>Reference Noise: H1Y1 (10dB) Target Noise: PS2 (10dB)</th>
<th>Acc. (%)</th>
<th>Corr. (%)</th>
<th>Del. (%)</th>
<th>Ins. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline JA</td>
<td>39.6</td>
<td>41.5</td>
<td>17.7</td>
<td>7.3</td>
</tr>
<tr>
<td>ALG. 1 (α = 1.0)</td>
<td>37.5</td>
<td>44.8</td>
<td>14.3</td>
<td>7.3</td>
</tr>
<tr>
<td>ALG. 1 + ALG. 2</td>
<td>38.6</td>
<td>46.6</td>
<td>14.1</td>
<td>7.3</td>
</tr>
<tr>
<td>Reference Noise: H1Y1 (5dB) Target Noise: PS2 (10dB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline JA</td>
<td>40.4</td>
<td>47.5</td>
<td>15.4</td>
<td>7.2</td>
</tr>
<tr>
<td>ALG. 1 (α = 1.0)</td>
<td>44.6</td>
<td>51.8</td>
<td>13.2</td>
<td>7.0</td>
</tr>
<tr>
<td>ALG. 1 + ALG. 2</td>
<td>44.4</td>
<td>51.7</td>
<td>13.2</td>
<td>7.3</td>
</tr>
<tr>
<td>Reference Noise: H1Y1 (5dB) Target Noise: LC1 (20dB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline JA</td>
<td>38.9</td>
<td>46.7</td>
<td>15.0</td>
<td>7.8</td>
</tr>
<tr>
<td>ALG. 1 (α = 1.0)</td>
<td>44.6</td>
<td>60.7</td>
<td>14.0</td>
<td>6.1</td>
</tr>
<tr>
<td>ALG. 1 + ALG. 2</td>
<td>45.1</td>
<td>60.7</td>
<td>14.7</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Table 2: Jacobian adaptation of models trained with H1Y1 noise at 5dB SNR to PS2 (10dB), HW2 (20 dB), and LCI (20 dB) noise conditions. Only static mean is adapted.

Table 3: Jacobian adaptation of models trained with H1Y1 noise at 5dB SNR to PS2 (10dB), HW2 (20 dB), and LCI (20 dB) noise conditions. Only static mean is adapted.

1. REFERENCES


6. CONCLUSIONS

In this study we formulated two algorithms to improve Jacobian adaptation for HMM recognition in noise. These algorithms take advantage of the simplifying assumptions of the baseline technique and propose more accurate relationships between the parameters. The first of these techniques generalizes the baseline algorithm to a parametric form. The second technique defines a more accurate relationship between the log and linear power spectral domain versions of the HMM statistics. A set of new Jacobian matrices are derived using these two algorithms. Context independent phone recognition experiments on the TIMIT database showed about 3.8-10.9% improvement over the baseline technique in adapting from automobile highway (HWY) to various other low-frequency noise sources at varying SNRs.