A COMPARATIVE STUDY OF APPROXIMATIONS FOR PARALLEL MODEL COMBINATION OF STATIC AND DYNAMIC PARAMETERS

Yifan Gong

Speech Technologies Laboratory, DSP Solutions R&D Center, Texas Instruments
P.O. Box 660199, MS-8649, Dallas, TX 75266-0199, USA
Yifan.Gong@ti.com

ABSTRACT
Reducing mismatch between HMMs trained with clean speech and speech signals corrupted with background noise can be approached by speech distribution adaptation using parallel model combination (PMC). Accurate PMC has no closed-form expression, therefore simplification assumptions must be made in implementation.
Under three assumptions, i.e. log-normal, log-add and log-max, adaptation formula for log-spectral parameters are presented, both for static and dynamic parameters.
Experimental evaluation uses TI-DIGITS speech database corrupted with car noise at 0dB signal-to-noise ratio. The recognition performance of the above three types of simplification is established. It is shown that, the adaptation of both static and dynamic parameters gives as much as 30% lower WER compared to adapting only static parameters.
The findings and results presented in the paper provide a basis for trading-offs between recognition accuracy and computation requirement.

1. INTRODUCTION
We are given a set of speech models (HMM) trained with speech collected in a quiet environment and our task is to recognize speech utterances recorded in a noisy background. In such case a mismatch exists between the environments of models and the utterances. The mismatch may degrade substantially recognition performance [1].
This problem is of importance in applications where it is too expensive to collect training speech in the noisy environment, or the changing nature of the noisy background makes it impossible to have a collection covering all situations. Hands-free speech recognition in automobile is a typical case.
Parallel model combination [2] (PMC) can be used to to reduce the mismatch. PMC uses the HMM distribution of clean speech models and the noise distribution to give a maximum likelihood estimate of the corrupted-speech models.
Two advantages of PMC can be mentioned. Firstly, no speech data is required for compensation. Secondly, all the models are individually compensated.
As accurate PMC has no closed-form expression, simplification assumptions must be made in implementation. We discuss three assumptions for the adaptation of log-spectral parameters, which lead to different degrees of simplification.
Although our experiments are performed using MFCC parameters, the results can be directly applied to other feature parameters linearly transformed from log-spectral parameters.
PMC adaptation of dynamic parameters can be approached from two different directions.
The first one establishes a mismatch function for (difference-based) dynamic parameters [3]. It can be shown that the adapted dynamic parameters at time $\tau$ are a function of static parameters at time $\tau - w$, an undesired requirement for practical applications. Besides, the results doesn’t apply to dynamic parameters obtained by linear-regression. A solution to this problem which sums up several difference-based compensated dynamic parameters has been proposed [4]. However, only little improvement due to dynamic coefficients were reported.
The second uses continuous time derivative of static parameters as dynamic parameters [5]. This is an approximation to the discrete nature of dynamic parameters, but the resulting expression is extremely simple and easy to implement. We will pursue this direction in this paper.

2. APPROXIMATION OF PMC

2.1. Introduction
PMC deals with Gaussian distributions. In PMC, an independent noise model is estimated from noise samples collected in the new environment. Distribution by distribution, clean speech model and the noise model are then combined using a mismatch function, to obtain a corrupted speech model.
matched to the new environment. The mismatch function assumes that speech and noise are independent and additive in the time domain. The mismatch function for computing the mean of the corrupted model in the log DFT domain has the form:

$$
\hat{\mu}_{i,j}^{\text{log}} = \text{E} \{ \log(g \exp(s_{i,j}^{\text{log}} + h_{i,j}^{\text{log}}) + \exp(n_{i,j}^{\text{log}})) \} \tag{1}
$$

where $s_{i,j}^{\text{log}}$ and $n_{i,j}^{\text{log}}$ represent speech and noise observations in the log DFT domain and their statistics are obtained from appropriate speech and noise state pair. $h_{i,j}^{\text{log}}$ is a convolutive (in time domain) noise representing channel, transducer and some speaker characteristics, which will be omitted in this study. $g$ is a gain matching factor accounting for level difference between the clean speech and the noisy speech. Since Eq-1 does not have a closed form, several approximations may be used, which allows trading-off between accuracy and hardware requirement: log-normal approximation, log-add approximation, and log-max approximation. In the following sections, we will present PMC formula for each of the three cases, with the notation:

- $\hat{x}$ denotes estimate (adapted value) of parameters $x$, $\hat{x}$ denotes parameters $x$ of noise.
- $\text{lin}$ for linear domain parameters, $\text{log}$ for log spectral domain.

### 2.2. Log-normal approximation

#### 2.2.1. Static parameter

Log-normal approximation is based on the assumption that the sum of two log-normally distributed random variables is itself log-normally distributed. In the linear domain, the mean of the compensated model is computed as

$$
\hat{\mu}_{i,j}^{\text{lin}} = g \mu_{i,j}^{\text{lin}} + \hat{\mu}_{i,j}^{\text{lin}} \tag{2}
$$

$$
\hat{\Sigma}_{i,j}^{\text{lin}} = g^2 \Sigma_{i,j}^{\text{lin}} + \hat{\Sigma}_{i,j}^{\text{lin}} \tag{3}
$$

where $i, j$ are indices for the feature vector dimension, and $g$ accounts for the gain of speech produced in noise with respect to clean speech and, for speech and noise:

$$
\mu_{i,j}^{\text{lin}} = \exp(\mu_{i,j}^{\text{log}} + \frac{1}{2} \Sigma_{i,j}^{\text{log}}) \tag{4}
$$

$$
\Sigma_{i,j}^{\text{lin}} = \mu_{i,j}^{\text{lin}} \frac{\Sigma_{i,j}^{\text{log}}}{\mu_{i,j}^{\text{lin}}} \tag{5}
$$

The adapted mean and variance in log domain can be obtained by inverting the above equations:

$$
\mu_{i,j}^{\text{log}} = \log(\mu_{i,j}^{\text{lin}}) - \frac{1}{2} \log \left( \frac{\Sigma_{i,j}^{\text{lin}}}{\mu_{i,j}^{\text{lin}}} + 1 \right) \tag{6}
$$

$$
\Sigma_{i,j}^{\text{log}} = \log \left( \frac{\Sigma_{i,j}^{\text{lin}}}{\mu_{i,j}^{\text{lin}}} + 1 \right) \tag{7}
$$

This assumption allows to adapt covariance matrix. However, it requires the conversion of covariance matrix into linear DFT domain, which is computationally expensive.

#### 2.2.2. Dynamic parameter

To derive the adaptation equation for dynamic parameters, we further assume that the noise is stationary, i.e., in average:

$$
\frac{\partial \hat{\mu}_{i,j}^{\text{lin}}}{\partial t} = 0. \tag{8}
$$

Following the idea presented in section-1, the adapted dynamic log-spectral vector is:

$$
\Delta \mu_{i,j}^{\text{log}} = \frac{\partial \hat{\mu}_{i,j}^{\text{log}}}{\partial t} = g \frac{\beta_i}{\beta_i + 1} \frac{\alpha_i + 2}{\alpha_i + 1} \Delta \mu_{i,j}^{\text{log}} \tag{9}
$$

where

$$
\alpha_i = \frac{(\hat{\mu}_{i,j}^{\text{lin}})^2}{\Sigma_{i,j}^{\text{lin}}} \quad \beta_i = \frac{\mu_{i,j}^{\text{lin}}}{\hat{\mu}_{i,j}^{\text{lin}}} \tag{10}
$$

is the signal-to-noise ratio (in linear scale), and, finally,

$$
\Delta \mu_{i,j}^{\text{log}} \leq \frac{\partial \mu_{i,j}^{\text{log}}}{\partial t} \tag{11}
$$

is the dynamic parameter of the clean model.

### 2.3. Log-add approximation

#### 2.3.1. Static parameter

Log-add approximation is based on the assumption that the effect of variance of both speech and noise on the estimate can be ignored:

$$
\Sigma_{i,j} = 0. \tag{12}
$$

Taking the logarithm of Eq-2, we have:

$$
\mu_{i,j}^{\text{log}} = \log \left( g \exp(\mu_{i,j}^{\text{log}}) + \exp(\hat{\mu}_{i,j}^{\text{log}}) \right) \tag{13}
$$

Only mean vectors are adapted. As variance conversion is not performed, the scheme is more computational efficient.

#### 2.3.2. Dynamic parameter

Applying Eq-12 to Eq-9, we have:

$$
\Delta \mu_{i,j}^{\text{log}} = g \frac{\beta_i}{\beta_i + 1} \Delta \mu_{i,j}^{\text{log}} \tag{14}
$$

Notice that $\beta_i$ is the SNR in linear scale.
2.4. Log-max approximation

2.4.1. Static parameter

We point out that:

\[
\log (g \exp(a) + \exp(b)) \approx \log g + a \quad \text{if} \quad a \gg b \quad (15)
\]

\[
\log (g \exp(a) + \exp(b)) \approx \log b \quad \text{if} \quad b \gg a \quad (16)
\]

Under log-max approximation, Eq-13 can be approximated by:

\[
\hat{\mu}_i^{\log} = \max \left( \log g + \hat{\mu}_i^{\log}, \hat{\mu}_i^{\log} \right) \quad (17)
\]

This transformation is performed totally in the log domain and hence is fast, though less accurate. It can be shown that the maximum error compared to log-add is \(\log 2\).

2.4.2. Dynamic parameter

Taking the time derivative of Eq-17, we obtain:

\[
\Delta \hat{\mu}_i^{\log} = \Delta \hat{\mu}_i^{\log} \quad \text{if} \quad \log g + \hat{\mu}_i^{\log} > \hat{\mu}_i^{\log} \quad (18)
\]

\[
= \Delta \hat{\mu}_i^{\log} \quad \text{otherwise.} \quad (19)
\]

If we use the assumption Eq-8, then the result is even simpler:

\[
\Delta \hat{\mu}_i^{\log} = \Delta \hat{\mu}_i^{\log} \quad \text{if} \quad \log g + \hat{\mu}_i^{\log} > \hat{\mu}_i^{\log} \quad (20)
\]

\[
= 0 \quad \text{otherwise.} \quad (21)
\]

2.5. Implementation

To satisfy real-time adaptation requirement, we developed an on-line version of model combination scheme, referred to as on-line model combination (OMC). During speech pause, OMC procedure adapts a fraction of HMM distributions with a newly estimated noise statistics. Two extreme cases can be possible: only one Gaussian distribution is adapted at each sample frame, or whole set of Gaussians is adapted. OMC can use either of the above-mentioned assumptions, based on available computational resource. Noise estimation is based on a modified MAP estimation of noise mean and variance [6].

3. PERFORMANCE EVALUATION

3.1. Database

The TI-digits database, down-sampled to 8 kHz, is used for all the experiments. The digit sequences have 1-7 digits.

- The training set consists of 4229 digit sequences (13896 words) from the male speakers and 4385 digit sequences (14400 words) from the female speakers.

- The test set consists of 113 speakers, 57 of which are female speakers and 56 are male speakers. The test set consists of 3747 digit sequences (10225 words) from male speakers and 3815 digit sequences (10412 words) from the female speakers.

The noise used in the experiments is recorded inside a passanger car driven on a highway. The noise was scaled and added to the test and adaptation (section-3.3) data to simulate 0dB signal-to-noise ratio (SNR) conditions.

3.2. Feature vectors

The observation vectors consist of mel-frequency cepstral coefficients (MFCC) along with their regression-based first-order time derivative, derived at a frame rate of 20ms. Three types of parameters are tested in the experiments:

- **DFT-MFCC-13** The power spectrum is calculated by DFT, 13 MFCC coefficients.

- **LPC-MFCC-13** The power spectrum is calculated through LPC analysis, 13 MFCC coefficients.

- **DFT-MFCC-10** The power spectrum is calculated by DFT, 10 MFCC coefficients.

Dynamic parameters are calculated using linear-regression.

3.3. Experiments

We use Gaussian mixture HMM recognizer, with a maximum of 8 mixture per state. The speaker-independent (SI) word error rate (WER) for clean speech is 0.52%.

We report results of noisy speech recognition by PMC, with speaker-adapted models. The speaker-adapted models are obtained by MLLR [7] from the SI model set. The MLLR uses ten utterances from each test speaker, that are reserved for MLLR adaptation and not used for testing.

The noise recordings are strictly stationary. However, we have chosen to use a single state HMM to model the noise. The noise for model combination is estimated from noise-only frames immediately before each speech utterance. Static parameters are adapted according a model combination assumption described in section 2. However, dynamic parameters are adapted exclusively using log-add approximation in Eq-18 of section-2.3.

Both static and dynamic parameters are used in all tests, with two adaptation tests

- **S**: Only static parameters are adapted.

- **S+D**: Both static and dynamic parameters are adapted.

The results are shown in Table-1.
<table>
<thead>
<tr>
<th>Type</th>
<th>LPC-MFCC-13</th>
<th></th>
<th></th>
<th>DFT-MFCC-10</th>
<th></th>
<th></th>
<th>DFT-MFCC-13</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>RD%</th>
<th>RD%</th>
<th>RD%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>S+D</td>
<td>RD%</td>
<td>S</td>
<td>S+D</td>
<td>RD%</td>
<td>S</td>
<td>S+D</td>
<td>RD%</td>
<td>S</td>
<td>S+D</td>
<td>RD%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-max WER</td>
<td>1.25</td>
<td>0.99</td>
<td>20.8</td>
<td>1.59</td>
<td>1.06</td>
<td>33.3</td>
<td>1.28</td>
<td>0.91</td>
<td>28.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-max SER</td>
<td>3.13</td>
<td>2.47</td>
<td>21.1</td>
<td>3.97</td>
<td>2.74</td>
<td>31.0</td>
<td>3.27</td>
<td>2.29</td>
<td>30.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-add WER</td>
<td>1.10</td>
<td>0.87</td>
<td>20.9</td>
<td>1.44</td>
<td>0.96</td>
<td>33.3</td>
<td>1.14</td>
<td>0.81</td>
<td>28.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-add SER</td>
<td>2.79</td>
<td>2.17</td>
<td>22.2</td>
<td>3.65</td>
<td>2.47</td>
<td>32.3</td>
<td>2.90</td>
<td>2.06</td>
<td>29.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-nrm WER</td>
<td>1.07</td>
<td>0.93</td>
<td>13.1</td>
<td>1.42</td>
<td>0.94</td>
<td>33.8</td>
<td>1.11</td>
<td>0.84</td>
<td>24.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-nrm SER</td>
<td>2.75</td>
<td>2.31</td>
<td>16.0</td>
<td>3.61</td>
<td>2.42</td>
<td>33.0</td>
<td>2.79</td>
<td>2.09</td>
<td>25.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. WER as function of model combination assumption type and of parameter type. No variance adaptation, 0dB SNR, 20637 words (7562 strings) tested. SER: string error rate, WER: word error rate, RD: WER reduction from S to S+D.

4. SUMMARY OF OBSERVATIONS

- Log-add approximation, with dynamic parameters or not, gives comparable results than log-normal at a substantially lower computational cost.
- With static (only) parameter adaptation, LPC-MFCC gives slightly lower WER than DFT-MFCC.
- With static and dynamic parameter adaptation, DFT-MFCC gives about 10% lower WER than LPC-MFCC.
- DFT-MFCC-10 benefits the most dynamic parameter adaptation, with an error reduction of about 30%.
- DFT-MFCC-13 gives 10-20% lower WER than DFT-MFCC-10.
- With log-add approximation, the lowest WER obtained is 0.81% (SER 2.06%).

5. REFERENCES


