SCALING AND BIAS CODES FOR MODELING SPEAKER-ADAPTIVE DNN-BASED SPEECH SYNTHESIS SYSTEMS

Hieu-Thi Luong, Junichi Yamagishi (NII, Japan)

Most neural network-based speaker adaptation method can be classified as layer-based or input-based.

Systematically looking into the principle reveals the common elements shared between these approaches.

By first combining and then factoring these elements into scaling and bias codes, we can design more sophisticated speaker-adaptive models.

The concept is also useful for other tasks.

Multi-speaker acoustic model utilizing scaling and bias codes to model speaker transformations instead of using only conventional speaker code which is essentially a bias constant at every frame.

Related works

Feedforward Layer

\[ h_t = f(W_t h_{t-1} + c_t) \]

Layer-based/Finetuning

\[ h_t = f(W_t h_{t-1} + b_t) \]

E.g.: speaker dependent layer, linear network, low-rank plus diagonal, ...

Learning Hidden Unit Contribution

\[ h_t = \text{Diag} A_t^{(k)} \odot f(W_t h_{t-1} + c_t) \]

\[ h_{t+1} = f(W_{t+1} A_t^{(k)} h_t + c_{t+1}) \]

\[ h_{t+1} = f(W_{t+1} h_t + c_{t+1}) \]

Input codes

\[ h_t = f(W_t h_{t-1} + c_t + W_s^{(k)}) \]

\[ h_t = f(W_t h_{t-1} + c_t) \]

E.g.: one-hot vector, i-vector, d-vector, discriminant condition code, ...

Multi-level strategies with scaling and bias codes used separately at different layers and as residual transformation.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Size</th>
<th>Bias</th>
<th>Bottleneck</th>
</tr>
</thead>
<tbody>
<tr>
<td>bias code</td>
<td>-</td>
<td>64</td>
<td>-</td>
</tr>
<tr>
<td>scaling code</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>affine</td>
<td>32</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>level</td>
<td>32</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>bottle</td>
<td>64</td>
<td>32</td>
<td>512</td>
</tr>
</tbody>
</table>

Injection layer of speaker transformation either at a nonlinear intermediate layer or a linear layer near output.

Objective evaluation on English data VCTK corpus with 72 speakers used as background while evaluation is calculated on 8 speakers, 150 utterances in total. The first strategy is a multi-speaker task and the rests are adaptation tasks in which only speaker parameters are updated.

Subjective evaluation on Japanese data An in-house corpus with 235 speakers was used as the background. Another set of 20 speakers was used as the target with 200 utterances in totals up for evaluation. 189 native Japanese participated in the test to judge quality and similarity of the samples.

Conclusions

Most neural network-based speaker adaptation method can be classified as layer-based or input-based.

Systematically looking into the principle reveals the common elements shared between these approaches.

By first combining and then factoring these elements into scaling and bias codes, we can design more sophisticated speaker-adaptive models.

The concept is also useful for other tasks.

Most neural network-based speaker adaptation method can be classified as layer-based or input-based.

Systematically looking into the principle reveals the common elements shared between these approaches.

By first combining and then factoring these elements into scaling and bias codes, we can design more sophisticated speaker-adaptive models.

The concept is also useful for other tasks.

Existing adaptation techniques for neural networks have a similar mathematical principle.

The conventional input code is restricted by its constant bias nature. It should be expanded to scaling to create a new tool to model better speaker transformation.

Modeling speakers as a deep, non-linear transformation might not always be the best choices.

See website: www.hieuthi.com