Vocal Tract Length Perturbation for Speech Recognition with DNN-HMMs

- Navdeep Jaitly
- Geoffrey Hinton
Outline

- Background on Mel Filterbanks
- Vocal Tract Length Normalization
- Vocal Tract Length Perturbation
- Results
- Avenues for Exploration
Mel log Filterbanks

- Low Resolution pre-processing of spectrograms
Each frame of a spectrogram is processed by multiple filters, each of which look at a frequency subbands.
Some comments

- Filterbanks are just a linear layer of a neural networks – with a very specific, fixed architecture
  - fixed local filters, whose location, and window size depends on their center frequency
  - fixed weights (typically triangular)
Mel-Filterbanks are Fixed Layers

![Graph showing Mel scale vs. frequency](image-url)
Applying log to the output of the filters on raw spectrograms is very similar to max-pooling, followed by log because intensities in a raw spectrogram vary over many orders of magnitude, and the log is dominated by the maximum intensity frequency.
Vocal Tract Length Normalization

- Fixed Pre-processing of spectrograms to remove some degree of speaker variation
  - Parameterized by a warp factor which changes how and where the filters are applied, smoothly.

- Warping can be applied straight to the construction of the filterbanks by changing where the centers of the filters are located
Projection Matrices for different warp factors

Warp factors – 0.8, 1, 1.2

Mel scale
Some VTLN comments

- Requires some amount of training data per speaker to fit the warp factors.
- The normalized data "presumably" is more consistent so a better model can be built focussing on the "true underlying structure".
- Great for GMMs because it means we can get by with fewer gaussians.

*The data become more speaker independent*
Vocal Tract Length Perturbation

- Instead of building a preprocessing model that makes filterbanks speaker independent, make the model invariant to warp factors
  - Inject the variations into the data

- Strategy well applied on vision tasks to augment databases
  - Transform the data in reasonable ways and add to databases
  - Transformations must preserve classes
Algorithm - Training

```plaintext
procedure PERTURBED_FEATURES(lst_spec)
    lst_f ← []
    for each spec ∈ lst_spec
        α ← RANDOM_NUMBER_IN_RANGE(0.9, 1.1)
        do
            fb ← FILTERBANKS(α)
            APPEND(lst_f, LOG(fb * spec))
        end do
    end for
    return (lst_f)

main
    while stopping criterion not reached
        lst_spec ← LOAD_RAW_SPECTROGRAMS()
        do
            lst_f ← PERTURBED_FEATURES(lst_spec)
            TRAINMODEL(lst_f)
        end do
    end while
```

*Use random warp for each utterance in each epoch of training*
Combine posterior probability predictions from multiple warp factors and decode with HMM
**Results – Simple Decoding**

- Trained on TIMIT, warp factors generated with mean 1, stdev 0.1, truncated at 0.9, 1.1
- Simple decoding with warp factor = 1.0

<table>
<thead>
<tr>
<th># of layers</th>
<th>Without VTLP</th>
<th>With VTLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>21.9</td>
<td>21.5</td>
</tr>
<tr>
<td>4</td>
<td>21.6</td>
<td>20.9</td>
</tr>
<tr>
<td>5</td>
<td>21.4</td>
<td>21.3</td>
</tr>
<tr>
<td>6</td>
<td>21.0</td>
<td>20.9</td>
</tr>
<tr>
<td>7</td>
<td>21.6</td>
<td>20.9</td>
</tr>
</tbody>
</table>
### Results - Averaging

<table>
<thead>
<tr>
<th># of layers</th>
<th>Without averaging</th>
<th>With averaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>21.5</td>
<td>21.1</td>
</tr>
<tr>
<td>4</td>
<td>20.9</td>
<td>20.6</td>
</tr>
<tr>
<td>5</td>
<td>21.3</td>
<td>21.2</td>
</tr>
<tr>
<td>6</td>
<td>20.9</td>
<td>20.2</td>
</tr>
<tr>
<td>7</td>
<td>20.9</td>
<td>20.9</td>
</tr>
</tbody>
</table>

- VTLP trained model, without and with averaging at test time (over 5 warp factors 0.95-1.05)
## Results – Averaging with non-VTLP models

A model with no warps, without and with averaging at test time (over 5 warp factors 0.95-1.05)

<table>
<thead>
<tr>
<th># of layers</th>
<th>Without Averaging</th>
<th>With Averaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>21.9</td>
<td>22.0</td>
</tr>
<tr>
<td>4</td>
<td>21.6</td>
<td>21.7</td>
</tr>
<tr>
<td>5</td>
<td>21.4</td>
<td>21.8</td>
</tr>
<tr>
<td>6</td>
<td>21.0</td>
<td>21.3</td>
</tr>
<tr>
<td>7</td>
<td>21.6</td>
<td>21.6</td>
</tr>
</tbody>
</table>
Most Improving phones

ah
dx
eh
aa
d
Future Work

- Explore other variations around the idea of distorting filterbanks
  - Does warping really need to be linear?
- Explore ideas on how to combine predictions from multiple warp factors, and possibly use that in the training
- Connections to sampling in convolutions
- Large Vocabulary Tasks on larger databases