Multi-stream Recognition of Noisy Speech with Performance Monitoring

Ehsan Variani, Feipeng Li, Hynek Hermansky
†CLSP, Johns Hopkins University, Baltimore, MD 21218, USA

Abstract
The speech signal is decomposed into seven band-limited streams, and then fused to form 127 combinations (processing streams). A performance monitor is designed to predict the reliability of individual processing streams. The top \( N \) streams that are least affected by noise are averaged to yield a more reliable estimation.

1 Multi-stream speech recognition
The proposed multi-stream phoneme recognition system (refer to Fig. 1) takes a three-stage processing scheme: stream formation, stream selection, followed by a Viterbi decoder.

1.1 Stream Formation
• Speech decomposed into seven band-limited streams, each covers about three critical bands
• Signal encoded by FDLP feature, which characterizes the Hilbert envelope for over 200 ms
• An independent three-layer MLP is trained to classify the band-limited signal
• All combinations of seven band-limited streams are fused to form 127 processing streams

1.2 Stream Selection
• Compare the statistics of training and testing data for each stream
• Rank 127 processing streams based on the prediction of performance monitor
• Select the \( N \) best processing streams for further processing

1.3 Integration of Selected Streams
• Average the output of \( N \) best streams to reduce the variance of posterior probability

2 Performance Monitor
2.1 Divergence between Training and Testing

![Figure 2: Comparing test statistics and training statistics.](image)

\[
M(\Delta t) = \sum_{i=1}^{N} \frac{D_{\text{sym}}(P_i, P_{i+\Delta t})}{N - \Delta t}
\]

where \( D_{\text{sym}} \) is the symmetric KL divergence,

\[
D_{\text{sym}}(p, q) = \sum_{i=1}^{N} \log \left( \frac{p_i}{q_i} \right) + \sum_{i=1}^{N} \log \left( \frac{q_i}{p_i} \right)
\]

2.2 Mean Temporal Distance (MTD)

The Mean Temporal Distance \( M(\Delta t) \) is defined as the average symmetric Kullback-Leibler (KL) divergence between two vectors of phoneme posterior probabilities \( P_i \) separated by \( \Delta t \)

\[
M(\Delta t) = \frac{1}{N \cdot \Delta t} \sum_{i=1}^{N} D_{\text{sym}}(P_i, P_{i+\Delta t})
\]

2.3 Exp. 1: Rejection of Narrow-band Noise

• Speech corrupted by 1kHz pure tone noise at -20 dB SNR, causes /iy/ confusion (Fig. 3a)

![Figure 3: M(\Delta t) vs. SNR for clean speech, noisy speech, and other sounds](image)

3 Experiments
3.1 Exp. 1: Rejection of Narrow-band Noise

• Speech corrupted by various types of noise with SNR ranges from 0, 5, 10, to 15 dB

![Figure 4: Priorigram of noise speech w/o stream selection by PM](image)

3.2 Exp. 2: Recognition of Noisy Speech

• Speech corrupted by various types of noise with SNR ranges from 0, 5, 10, to 15 dB

![Figure 5: Relative PER vs. different values of \( N \)](image)

<table>
<thead>
<tr>
<th>Conditions</th>
<th>full-band</th>
<th>baseline</th>
<th>multi-stream-PM</th>
<th>multi-stream-hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>23.76</td>
<td>31.21</td>
<td>29.89</td>
<td>25.70</td>
</tr>
<tr>
<td>babble</td>
<td>57.10</td>
<td>52.80</td>
<td>49.68</td>
<td>42.85</td>
</tr>
<tr>
<td>subway</td>
<td>46.62</td>
<td>45.15</td>
<td>40.79</td>
<td>34.11</td>
</tr>
<tr>
<td>factory</td>
<td>68.1</td>
<td>69.87</td>
<td>67.10</td>
<td>59.91</td>
</tr>
<tr>
<td>restaurant</td>
<td>63.14</td>
<td>65.03</td>
<td>61.61</td>
<td>55.18</td>
</tr>
<tr>
<td>street</td>
<td>67.26</td>
<td>68.47</td>
<td>65.27</td>
<td>58.08</td>
</tr>
<tr>
<td>zjchall</td>
<td>70.67</td>
<td>71.16</td>
<td>68.67</td>
<td>61.85</td>
</tr>
<tr>
<td>car</td>
<td>54.32</td>
<td>48.76</td>
<td>40.24</td>
<td>34.30</td>
</tr>
</tbody>
</table>

Table 1: PER (%) of the proposed multi-stream with performance monitor