Introduction

• Supervised global SNR Estimation

• Assumptions:
  - Independence of speech and noise signals
  - Both speech and noise are zero-mean
  - Additive Noise

• Features that capture speech in a signal
  - Create Estimators based on features
  - I-vectors to capture channel information
  - Neural Network for nonlinear regression

Speech SNR

\[
\text{SNR} = 20 \cdot \log_{10} \left( \frac{\sum_{i=1}^{N} s^2(i)}{\sum_{i=1}^{N} n^2(i)} \right)
\]

\[
= 10 \cdot \log_{10} \frac{P(S)}{P(N)}
\]

\[
= 10 \cdot \log_{10} \frac{P(X) - P(N)}{P(N)}
\]

Features

• Energy
• Long-Term Signal Variability (LTSV)
• Pitch
• Voicing Probability

Experimental Setup

• TIMIT Speech Database
• DEMAND Noise Database (18 noises)
• Used 9 different SNR levels (-5dB-15dB)
• 300000 noisy utterances for training (2000*9*17)
• Leave One Out Approach
• 900 noisy utterances for testing
• Compared against 2 other methods

Conclusions and Future Work

• Accurate SNR Estimation
• Independent of noise type
• Outperforms other methods
• Depends on the availability of noise pool
• Explore features that capture both noise characteristics and SNR information

Results

<table>
<thead>
<tr>
<th></th>
<th>WADA</th>
<th>DNN Selection</th>
<th>Ch. Ad. DNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>KITCHEN</td>
<td>4.663</td>
<td>2.976</td>
<td>2.835</td>
</tr>
<tr>
<td>LIV. ROOM</td>
<td>3.641</td>
<td>2.413</td>
<td>1.358</td>
</tr>
<tr>
<td>METRO</td>
<td>7.126</td>
<td>4.761</td>
<td>2.902</td>
</tr>
<tr>
<td>PARK</td>
<td>5.644</td>
<td>3.356</td>
<td>2.116</td>
</tr>
<tr>
<td>STATION</td>
<td>3.121</td>
<td>1.732</td>
<td>1.141</td>
</tr>
<tr>
<td>TRAFFIC</td>
<td>4.567</td>
<td>3.599</td>
<td>1.936</td>
</tr>
<tr>
<td>RESTAURANT</td>
<td>3.345</td>
<td>2.451</td>
<td>1.918</td>
</tr>
<tr>
<td>CAFE</td>
<td>3.766</td>
<td>2.691</td>
<td>1.106</td>
</tr>
</tbody>
</table>