Enhancement of Very Noisy Speech Signals

Dan Ellis & Zhuo Chen
Team Swordfish / ICSI / Columbia

dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Speech Enhancement
2. Flat-Pitch Enhancement
3. Results & Future
1. Speech Enhancement

- Noisy speech is a challenge:
  - Surprise channel in surprise language

- How to distinguish speech and interference?
  - Energy peaks are speech (spectral subtraction)
  - Energy troughs are noise (Wiener, log-mmse)
  - Speech has a known form (Factorial HMM)
  - Voiced speech is periodic (Pitch-based)
Noisy Channel Detection

• MIC channels are in a different formant
• Even after resampling, noisy channels are very distinct:
KWS on Noisy Signals

• WER is very poor
• only nonzero thanks to common word ஆஹ் ("ah")

<table>
<thead>
<tr>
<th></th>
<th>Corr</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP1_204_13189__inLine</td>
<td>5.1</td>
<td>5.1</td>
<td>89.8</td>
<td>1.1</td>
<td>96.0</td>
</tr>
<tr>
<td>OP1_204_61440__inLine</td>
<td>7.9</td>
<td>4.8</td>
<td>87.4</td>
<td>0.0</td>
<td>92.1</td>
</tr>
<tr>
<td>OP1_204_73990__inLine</td>
<td>5.4</td>
<td>2.6</td>
<td>92.0</td>
<td>1.2</td>
<td>95.9</td>
</tr>
<tr>
<td>OP1_204_78161__inLine</td>
<td>15.1</td>
<td>54.4</td>
<td>30.5</td>
<td>6.0</td>
<td>90.8</td>
</tr>
<tr>
<td>OP1_204_90937__inLine</td>
<td>3.0</td>
<td>3.0</td>
<td>93.9</td>
<td>0.0</td>
<td>97.0</td>
</tr>
<tr>
<td>OP1_204_91808__inLine</td>
<td>2.4</td>
<td>10.0</td>
<td>87.5</td>
<td>1.2</td>
<td>98.8</td>
</tr>
</tbody>
</table>
### Spectral-Based Enhancement

- **Classic enhancement boosts loudest parts**
- evaluated over 6 MIC utterances of OPI_204_DEV

<table>
<thead>
<tr>
<th></th>
<th>Corr</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>6.8</td>
<td>10.6</td>
<td>82.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Wiener</td>
<td>8.8</td>
<td>17.5</td>
<td>73.7</td>
<td>1.5</td>
</tr>
<tr>
<td>log-MMSE</td>
<td>9.9</td>
<td>31.1</td>
<td>59.0</td>
<td>1.4</td>
</tr>
<tr>
<td>E.hist norrm</td>
<td>8.6</td>
<td>23.1</td>
<td>68.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>
RPCA Enhancement

• Decompose spectrogram into sparse + low-rank
• Sparse activation $H$ of dictionary $W$

$$\min_{H, L, S} \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \lambda_S \|S\|_1$$

$$+ \mathcal{I}_+(H)$$

s.t. $Y = WH + L + S$

• ASR benefits:

<table>
<thead>
<tr>
<th></th>
<th>Corr</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>6.8</td>
<td>10.6</td>
<td>82.6</td>
<td>0.7</td>
</tr>
<tr>
<td>log MMSE</td>
<td>9.9</td>
<td>31.1</td>
<td>59.0</td>
<td>1.4</td>
</tr>
<tr>
<td>L+WH</td>
<td>15.5</td>
<td>46.1</td>
<td>38.4</td>
<td>1.6</td>
</tr>
</tbody>
</table>
2. Pitch-Based Enhancement

- Voiced Speech has near-periodic waveform
  - Energy concentrated in harmonics
- Given pitch, keep only those harmonics?
  - time-varying filter
  - sinusoidal model
- Problems
  - pitch errors
  - filtering artefacts
  - unvoiced speech, graceful degradation

Denbigh & Zhao ’92

after Wang ’95
Pitch Estimation in Noise

• Conventional pitch trackers are based on periodic structure
  - e.g. finding peaks in autocorrelation
  - not robust to noise

• Classifier-based approach
  - don’t predefine nature of pitch
  - let a classifier learn from examples
Classification-based Pitch Tracker

- Subband Autocorrelation Classification (SAcC) Pitch Tracker:
  - Trained on noisy speech with true pitch targets

Cochlea filterbank

- Subband autocorrelation features
- PCA to reduce dimensions
SAcC Results

- **Excellent in-domain results**
  - at low SNRs
  - errors dominated by V/UV

- **Generalization is good**
  - between different RATS channels
Flat-Pitch Processing

- **Time-varying filtering is tricky**
  - if pitch variation and filter impulse response are on a similar time-scale

- **Solution:** Flatten the pitch
  - use local pitch estimate to resample
  - process constant-pitch
  - resampling is (near) invertible

![Diagram of flat-pitch processing](image)
Flat-Pitch Processing

- How to enhance flat pitch?
  - Wiener filtering
  - Comb filtering
  - "Phase Vocoder" emphasis
# Flat-Pitch Results

- Over 6 MIC utterances of OPI_204_MIC

<table>
<thead>
<tr>
<th></th>
<th>Corr</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>6.8</td>
<td>10.6</td>
<td>82.6</td>
<td>0.7</td>
</tr>
<tr>
<td>log MMSE</td>
<td>9.9</td>
<td>31.1</td>
<td>59.0</td>
<td>1.4</td>
</tr>
<tr>
<td>L+WH</td>
<td>15.5</td>
<td>46.1</td>
<td>38.4</td>
<td>1.6</td>
</tr>
<tr>
<td>flat_pitch-comb</td>
<td>10.0</td>
<td>25.2</td>
<td>64.9</td>
<td>0.7</td>
</tr>
<tr>
<td>MMSE+f_p-comb</td>
<td>13.6</td>
<td>42.5</td>
<td>43.9</td>
<td>2.8</td>
</tr>
<tr>
<td>f_p-comb+L+WH</td>
<td>15.0</td>
<td>52.4</td>
<td>32.6</td>
<td>3.1</td>
</tr>
</tbody>
</table>
Summary

• **Noisy Speech**
  • single distant mic in real-world environments

• **Enhancement**
  • boosting spectrogram energy that appears to be speech
  • low-rank + sparse dictionary exploits knowledge

• **Flat-Pitch enhancement**
  • trained noise-robust pitch classifier
  • dynamic resampling to flatten pitch for enhancement