TOWARDS MICROPHONE-INDEPENDENT SPEECH RECOGNITION

Alejandro Acero
Richard M. Stern

Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
BASELINE PERFORMANCE AND GOALS

Baseline performance

<table>
<thead>
<tr>
<th>TRAIN TEST</th>
<th>CLSTK</th>
<th>CLSTK</th>
<th>CRPZM</th>
<th>CRPZM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>85.3%</td>
<td>18.6%</td>
<td>36.9%</td>
<td>76.5%</td>
</tr>
</tbody>
</table>

Immediate challenges:
- Improve performance for cross conditions (robustness)
- Improve absolute CRPZM/CRPZM performance

Ultimate goal: *Microphone-independent* system
- Works well with a standard microphone and acceptably with the rest.
- Does not need data about the new microphone/environment
- Works in an uncontrolled environment.
- No system does this at the present time
MULTI-STYLE TRAINING

<table>
<thead>
<tr>
<th>TRAIN</th>
<th>CLSTK</th>
<th>CRPZM</th>
<th>MULTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test CLSTK</td>
<td>85.3%</td>
<td>36.9%</td>
<td>78.3%</td>
</tr>
<tr>
<td>Test CRPZM</td>
<td>18.6%</td>
<td>76.5%</td>
<td>69.7%</td>
</tr>
</tbody>
</table>

- Used in speaker independence, provides greater robustness (for "cross" conditions), but limits performance
- Better performance expected if we had a model for the degradation
A MODEL OF THE ENVIRONMENT

• Degraded speech is formed by passing "clean" (reference) speech through a filter and adding independent noise

• **Goal:** Find the parameters that undo these transformations

![Diagram](https://via.placeholder.com/150)
INDEPENDENT COMPENSATION FOR NOISE AND FILTERING

Spectral Equalization
  • EQUAL. (Stockham)

Noise suppression techniques
  • PSUB - Boll’s Power spectral subtraction
  • MSUB - Magnitude Spectral Subtraction
  • MMSE1 - Use a transformation curve that minimizes squared error between CLSTK and PZM

[ALEX - YOU PROBABLY COULD BE EVEN A LITTLE MORE VERBOSE HERE. DO EITHER MSUB OR MMSE1 RELATE TO BEROUTI? PORTER AND BOLL? IF SO, REFERENCE THEM, TOO.]
INDEPENDENT COMPENSATION FOR NOISE AND FILTERING

Observations

- Spectral Subtraction and Spectral Equalization interact non-linearly so a simple cascade of these algorithms does not work.

- By treating different frequencies independently we obtain frames that are not speech-like.
PERFORMANCE OF COMPENSATION SCHEMES

Training on Close-talking Microphone:

Training on Crown PZM Microphone:
SDCN ALGORITHM

SNR-Dependent Cepstral Normalization

\( w \) is chosen to minimize the mean-squared average difference between CLSTK and CRPZM cepstra for each SNR

Interpretation of \( w \):

- Equalization at high SNR
- Noise subtraction at low SNR

<table>
<thead>
<tr>
<th>TRAIN TEST</th>
<th>CLSTK</th>
<th>CLSTK PZM</th>
<th>CRPZM CLSTK</th>
<th>CRPZM CRPZM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>85.3%</td>
<td>18.6%</td>
<td>36.9%</td>
<td>76.5%</td>
</tr>
<tr>
<td>MMSEN</td>
<td>85.3%</td>
<td>66.4%</td>
<td>75.5%</td>
<td>72.3%</td>
</tr>
<tr>
<td>SDCN</td>
<td>85.3%</td>
<td>67.2%</td>
<td>76.4%</td>
<td>75.5%</td>
</tr>
</tbody>
</table>
Advantages

- Joint Compensation for noise and spectral tilt

- Very easy to implement, only $c_0$ and $c_1$ need to be normalized.

Disadvantages

- For every new microphone/environment, a new stereo database is needed to estimate the corresponding $w$ vectors, hence

- Not microphone independent
CDCN ALGORITHM

Codeword-dependent Cepstral Normalization

Estimation process:

1. ML estimate of \( q \) and \( n \). Find the parameters of the transformation that maximize the probability or alternatively minimize the overall VQ distortion. Use of the EM algorithm for convergence.

2. MMSE estimate of every cepstrum vector given \( q \) and \( n \)
PERFORMANCE OF COMPENSATION SCHEMES

Training on Close-talking Microphone:

- Word Accuracy

Training on Crown PZM Microphone:
BASELINE SPECTRA
CDCN SPECTRA
SPECTRAL TILT COMPARISON
CROSS MICROPHONE RECOGNITION

Performance using different microphones. In each case SPHINX had been trained with the CLSTK microphone

<table>
<thead>
<tr>
<th>Microphone Model</th>
<th>BASE</th>
<th>CDCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLSTK CRPCC160</td>
<td>82.4%</td>
<td>70.2%</td>
</tr>
<tr>
<td>BASE</td>
<td>81.0%</td>
<td>78.5%</td>
</tr>
<tr>
<td>CDCN</td>
<td>84.8%</td>
<td>41.8%</td>
</tr>
<tr>
<td>CDCN</td>
<td>83.3%</td>
<td>73.9%</td>
</tr>
<tr>
<td>CLSTK SENN518</td>
<td>87.2%</td>
<td>84.5%</td>
</tr>
<tr>
<td>BASE</td>
<td>82.2%</td>
<td>83.3%</td>
</tr>
<tr>
<td>CDCN</td>
<td>83.7%</td>
<td>71.4%</td>
</tr>
<tr>
<td>CDCN</td>
<td>81.5%</td>
<td>80.7%</td>
</tr>
<tr>
<td>CLSTK SENNME80</td>
<td>55.9%</td>
<td>56.3%</td>
</tr>
<tr>
<td>BASE</td>
<td>81.7%</td>
<td>72.2%</td>
</tr>
</tbody>
</table>

Carnegie Mellon School of Computer Science Speech Group
SUMMARY

- Desk-top microphones like the Crown PZM6fs increase the recognition error rate by allowing weak phonetic events to become confused with silences.

- **Microphone-independent** systems can be built by estimating the parameters of the transformation: noise and spectral tilt.

- A framework for speech normalization in the *cepstral domain* has been introduced.