CHiME Challenge:
Approaches to Robustness using Beamforming and Uncertainty-of-Observation Techniques

Dorothea Kolossa ¹, Ramón Fernandez Astudillo ², Alberto Abad ², Steffen Zeiler ¹, Rahim Saeidi ³, Pejman Mowlaee ¹, João Paulo da Silva Neto ², Rainer Martin ¹

¹ Institute of Communication Acoustics (IKA) Ruhr-Universität Bochum
² Spoken Language Laboratory, INESC-ID, Lisbon
³ School of Computing, University of Eastern Finland
Overview

- Uncertainty-Based Approach to Robust ASR
- Uncertainty Estimation by Beamforming & Propagation
- Recognition under Uncertain Observations
- Further Improvements
  - Training: Full-covariance Mixture Splitting
  - Integration: Rover
- Results and Conclusions
Introduction: Uncertainty-Based Approach to ASR Robustness

- Speech enhancement in time-frequency-domain is often very effective.
- However, speech enhancement itself can neither
  - remove all distortions and sources of mismatch completely
  - nor can it avoid introducing artifacts itself

Simple example: Time-Frequency Masking

Mixture
Introduction: Uncertainty-Based Approach to ASR Robustness

How can decoder handle such artificially distorted signals?

One possible compromise:

Problem: Recognition performs significantly better in other domains, such that missing feature approach may perform worse than feature reconstruction [1].

Introduction: Uncertainty-Based Approach to ASR Robustness

Solution used here:
Transform uncertain features to desired domain of recognition

\[ m(n) \xrightarrow{\text{STFT}} Y_{kl} \xrightarrow{\text{Speech Processing}} \tilde{X}_{kl} \xrightarrow{\text{Uncertainty Propagation}} \tilde{X}_{kl} \xrightarrow{M_{kl}} \text{Recognition Domain} \]
Introduction: Uncertainty-Based Approach to ASR Robustness

Solution used here:
Transform uncertain features to desired domain of recognition

\[ m(n) \rightarrow \text{STFT} \rightarrow Y_{kl} \rightarrow \text{Speech Processing} \rightarrow p(X_{kl} | Y_{kl}) \rightarrow \text{Uncertainty Propagation} \rightarrow \text{Missing Data HMM Speech Recognition} \]
Introduction: Uncertainty-Based Approach to ASR Robustness

Solution used here:
Transform uncertain features to desired domain of recognition

\[ m(n) \quad \rightarrow \quad \text{STFT} \quad \rightarrow \quad Y_{kl} \quad \rightarrow \quad \text{Speech Processing} \quad \rightarrow \quad p(X_{kl} | Y_{kl}) \quad \rightarrow \quad \text{Uncertainty Propagation} \quad \rightarrow \quad p(x_{kl}^c | Y_{kl}) \quad \rightarrow \quad \text{Uncertainty-based HMM Speech Recognition} \quad \rightarrow \quad \text{Recognition Domain} \]
Uncertainty Estimation & Propagation

- Posterior estimation here is performed by using one of four beamformers:
  - Delay and Sum (DS)
  - Generalized Sidelobe Canceller (GSC) [2]
  - Multichannel Wiener Filter (WPF)
  - Integrated Wiener Filtering with Adaptive Beamformer (IWAB) [3]


Posterior of clean speech, $p(X_{kl} \mid Y_{kl})$, is then propagated into domain of ASR

Feature Extraction
- STSA-based MFCCs
- CMS per utterance
- possibly LDA
Uncertainty Estimation & Propagation

- Uncertainty model: Complex Gaussian distribution
Uncertainty Estimation & Propagation

- Two uncertainty estimators:
  
a) Channel Asymmetry Uncertainty Estimation
  - Beamformer output input to Wiener filter
  - Noise variance estimated as squared channel difference
  - Posterior directly obtainable for Wiener filter [4]:

\[
\lambda_D = \text{DFT}\{(m_L(n) - m_R(n))^2\}
\]

\[
p(X_{kl}|Y_{kl}) = \mathcal{N}\left(\frac{\lambda_{X_{kl}}}{\lambda_{D_{kl}} + \lambda_{X_{kl}}} Y_{kl}; \frac{\lambda_{X_{kl}} \lambda_{D_{kl}}}{\lambda_{D_{kl}} + \lambda_{X_{kl}}}\right)
\]

Two uncertainty estimators:

b) Equivalent Wiener variance

- Beamformer output directly passed to feature extraction

\[ p(X_{kl}|Y_{kl}) = \mathcal{N}(Y_{kl}, \lambda_{kl}) \]

- Variance estimated using ratio of beamformer input and output, interpreted as Wiener gain

Uncertainty Propagation

- Uncertainty propagation from [5] was used
  - Propagation through absolute value yields MMSE-STSA
  - Independent log normal distributions after filterbank assumed

- Posterior of clean speech in cepstrum domain assumed Gaussian
- CMS and LDA transformations simple

Recognition under Uncertain Observations

- Standard observation likelihood for state $q$ mixture $m$:
  \[ p(x | \mu_{q,m}, \Sigma_{q,m}) = N(x; \mu_{q,m}, \Sigma_{q,m}) \]

- Uncertainty Decoding:
  \[ p(\mu_x | \mu_{q,m}, \Sigma_{q,m}, \Sigma_x) = N(\mu_x; \mu_{q,m}, \Sigma_{q,m} + \Sigma_x) \]


- Modified Imputation:
  \[ p(\mu_x | \mu_{q,m}, \Sigma_{q,m}, \Sigma_x) = \mathcal{N}(\hat{x}; \mu_{q,m}, \Sigma_{q,m}) \]
  
  \[ \text{with } \hat{x} = (\Sigma_{q,m} + \Sigma_x)^{-1}(\Sigma_{q,m}\mu_x + \Sigma_x\mu_{q,m}) \]


- Both uncertainty-of-observation techniques collapse to standard observation likelihood for $\Sigma_x = 0$. 

Further Improvements

- Training: Informed Mixture Splitting
  - Baum-Welch Training is only optimal locally -> good initialization and good split directions matter.
  - Therefore, considering covariance structure in mixture splitting is advantageous:

\[ \text{split along maximum variance axis} \]
Further Improvements

- Training: Informed Mixture Splitting
  - Baum-Welch Training is only optimal locally -> good initialization and good split directions matter.
  - Therefore, considering covariance structure in mixture splitting is advantageous:

  ![Diagram](split along first eigenvector of covariance matrix)
Further Improvements

- Integration: Recognizer output voting error reduction (ROVER)
  - Recognition outputs at word level are combined by dynamic programming on generated lattice, taking into account
    - the frequency of word labels and
    - the posterior word probabilities
  - We use ROVER on 3 jointly best systems selected on development set.

Results and Conclusions

- **Evaluation:**
  - Two scenarios are considered, clean training and multicondition (‘mixed’) training.
  - In mixed training, all training data was used at all SNR levels, artificially adding randomly selected noise from noise-only recordings.
  - Results are determined on the development set first.
  - After selecting the best performing system on development data, final results are obtained as *keyword accuracies* on the *isolated sentences* of the *test set*.
Results and Conclusions

- JASPER Results after clean training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean:</td>
<td>30.33</td>
<td>35.42</td>
<td>49.50</td>
<td>62.92</td>
<td>75.00</td>
<td>82.42</td>
</tr>
<tr>
<td>Official Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JASPER*</td>
<td>40.83</td>
<td>49.25</td>
<td>60.33</td>
<td>70.67</td>
<td>79.67</td>
<td>84.92</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* JASPER uses full covariance training with MCE iteration control. Token passing is equivalent to HTK.
## Results and Conclusions

- **JASPER Results after clean training**

<table>
<thead>
<tr>
<th>Test</th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean: Official Baseline</td>
<td>30.33</td>
<td>35.42</td>
<td>49.50</td>
<td>62.92</td>
<td>75.00</td>
<td>82.42</td>
</tr>
<tr>
<td>JASPER Baseline</td>
<td>40.83</td>
<td>49.25</td>
<td>60.33</td>
<td>70.67</td>
<td>79.67</td>
<td>84.92</td>
</tr>
<tr>
<td>JASPER + BF* + UP</td>
<td>54.50</td>
<td>61.33</td>
<td>72.92</td>
<td>82.17</td>
<td>87.42</td>
<td>90.83</td>
</tr>
</tbody>
</table>

* Best strategy here: 
Delay and sum beamformer + noise estimation + modified imputation
Results and Conclusions

- HTK Results after clean training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean:</td>
<td>30.33</td>
<td>35.42</td>
<td>49.50</td>
<td>62.92</td>
<td>75.00</td>
<td>82.42</td>
</tr>
<tr>
<td>Official Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTK + BF* + UP</td>
<td>42.33</td>
<td>51.92</td>
<td>61.50</td>
<td>73.58</td>
<td>80.92</td>
<td>88.75</td>
</tr>
</tbody>
</table>

* Best strategy here: Wiener post filter + uncertainty estimation
Results and Conclusions

- Results after clean training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean: Official Baseline</td>
<td>30.33</td>
<td>35.42</td>
<td>49.50</td>
<td>62.92</td>
<td>75.00</td>
<td>82.42</td>
</tr>
<tr>
<td>HTK + BF + UP</td>
<td>42.33</td>
<td>51.92</td>
<td>61.50</td>
<td>73.58</td>
<td>80.92</td>
<td>88.75</td>
</tr>
<tr>
<td>HTK + BF* + UP + MLLR</td>
<td>54.83</td>
<td>65.17</td>
<td>74.25</td>
<td>82.67</td>
<td>87.25</td>
<td>91.33</td>
</tr>
</tbody>
</table>

* Best strategy here:
Delay and sum beamformer + noise estimation
## Results and Conclusions

### Overall Results after clean training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean: Official Baseline</td>
<td>30.33</td>
<td>35.42</td>
<td>49.50</td>
<td>62.92</td>
<td>75.00</td>
<td>82.42</td>
</tr>
<tr>
<td>JASPER Baseline</td>
<td>40.83</td>
<td>49.25</td>
<td>60.33</td>
<td>70.67</td>
<td>79.67</td>
<td>84.92</td>
</tr>
<tr>
<td>JASPER + BF + UP</td>
<td>54.50</td>
<td>61.33</td>
<td>72.92</td>
<td>82.17</td>
<td>87.42</td>
<td>90.83</td>
</tr>
<tr>
<td>HTK + BF + UP</td>
<td>42.33</td>
<td>51.92</td>
<td>61.50</td>
<td>73.58</td>
<td>80.92</td>
<td>88.75</td>
</tr>
<tr>
<td>HTK + BF + UP + MLLR</td>
<td>54.83</td>
<td>65.17</td>
<td>74.25</td>
<td>82.67</td>
<td>87.25</td>
<td>91.33</td>
</tr>
<tr>
<td>ROVER (JASPER + HTK)*</td>
<td>57.58</td>
<td>64.42</td>
<td>76.75</td>
<td>86.17</td>
<td>88.58</td>
<td>92.75</td>
</tr>
</tbody>
</table>

* (JASPER +DS + MI) & (HTK+GSC+NE) & (JASPER+WPF+MI)
### Results and Conclusions

#### JASPER Results after multicondition training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multicondition:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTK Baseline</td>
<td>63.00</td>
<td>72.67</td>
<td>79.50</td>
<td>85.25</td>
<td>89.75</td>
<td>93.58</td>
</tr>
<tr>
<td>JASPER Baseline</td>
<td>64.33</td>
<td>73.08</td>
<td>81.75</td>
<td>85.67</td>
<td>89.50</td>
<td>91.17</td>
</tr>
</tbody>
</table>
Results and Conclusions

- JASPER Results after multicondition training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicondition: HTK Baseline</td>
<td>63.00</td>
<td>72.67</td>
<td>79.50</td>
<td>85.25</td>
<td>89.75</td>
<td>93.58</td>
</tr>
<tr>
<td>JASPER Baseline</td>
<td>64.33</td>
<td>73.08</td>
<td>81.75</td>
<td>85.67</td>
<td>89.50</td>
<td>91.17</td>
</tr>
<tr>
<td>JASPER + BF* + UP</td>
<td>73.92</td>
<td>79.08</td>
<td>86.25</td>
<td>89.83</td>
<td>91.08</td>
<td>93.00</td>
</tr>
</tbody>
</table>

* best JASPER setup here: Delay and sum beamformer + noise estimation + modified imputation + LDA to 37d
## Results and Conclusions

### JASPER Results after multicondition training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicondition:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTK Baseline</td>
<td>63.00</td>
<td>72.67</td>
<td>79.50</td>
<td>85.25</td>
<td>89.75</td>
<td>93.58</td>
</tr>
<tr>
<td>JASPER Baseline</td>
<td>64.33</td>
<td>73.08</td>
<td>81.75</td>
<td>85.67</td>
<td>89.50</td>
<td>91.17</td>
</tr>
<tr>
<td>JASPER + BF* + UP</td>
<td>73.92</td>
<td>79.08</td>
<td>86.25</td>
<td>89.83</td>
<td>91.08</td>
<td>93.00</td>
</tr>
<tr>
<td>as above, but 39d</td>
<td>+0.58%</td>
<td>-0.25%</td>
<td>-2.16%</td>
<td>-1.41%</td>
<td>-2.0%</td>
<td>-0.5%</td>
</tr>
</tbody>
</table>

* best JASPER setup here: Delay and sum beamformer + noise estimation + modified imputation + LDA to 37d
## Results and Conclusions

- **HTK Results after multicondition training**

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multicondition:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTK Baseline</td>
<td>63.00</td>
<td>72.67</td>
<td>79.50</td>
<td>85.25</td>
<td>89.75</td>
<td>93.58</td>
</tr>
<tr>
<td><em><em>HTK + BF</em> + UP</em>*</td>
<td>67.92</td>
<td>77.75</td>
<td>84.17</td>
<td>89.00</td>
<td>91.00</td>
<td>92.75</td>
</tr>
</tbody>
</table>

* best HTK setup here: Delay and sum beamformer + noise estimation
Results and Conclusions

- HTK Results after multicondition training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicondition:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTK Baseline</td>
<td>63.00</td>
<td>72.67</td>
<td>79.50</td>
<td>85.25</td>
<td>89.75</td>
<td>93.58</td>
</tr>
<tr>
<td>HTK + BF + UP</td>
<td>67.92</td>
<td>77.75</td>
<td>84.17</td>
<td>89.00</td>
<td>91.00</td>
<td>92.75</td>
</tr>
<tr>
<td>HTK + BF* + UP + MLLR</td>
<td>68.25</td>
<td>79.75</td>
<td>84.67</td>
<td>89.58</td>
<td>91.25</td>
<td>92.92</td>
</tr>
</tbody>
</table>

* best HTK setup here: Delay and sum beamformer + noise estimation
Results and Conclusions

- Overall Results after multicondition training

<table>
<thead>
<tr>
<th></th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicondition:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTK Baseline</td>
<td>63.00</td>
<td>72.67</td>
<td>79.50</td>
<td>85.25</td>
<td>89.75</td>
<td>93.58</td>
</tr>
<tr>
<td>JASPER Baseline</td>
<td>64.33</td>
<td>73.08</td>
<td>81.75</td>
<td>85.67</td>
<td>89.50</td>
<td>91.17</td>
</tr>
<tr>
<td>JASPER + BF + UP</td>
<td>73.92</td>
<td>79.08</td>
<td>86.25</td>
<td>89.83</td>
<td>91.08</td>
<td>93.00</td>
</tr>
<tr>
<td>HTK + BF + UP</td>
<td>67.92</td>
<td>77.75</td>
<td>84.17</td>
<td>89.00</td>
<td>91.00</td>
<td>92.75</td>
</tr>
<tr>
<td>HTK + BF + UP + MLLR</td>
<td>68.25</td>
<td>79.75</td>
<td>84.67</td>
<td>89.58</td>
<td>91.25</td>
<td>92.92</td>
</tr>
<tr>
<td>ROVER (JASPER + HTK )*</td>
<td>74.58</td>
<td>80.58</td>
<td>87.92</td>
<td>90.83</td>
<td>92.75</td>
<td>94.17</td>
</tr>
</tbody>
</table>

* (JASPER +DS + MI + LDA ) & (JASPER+WPF, no observation uncertainties) & (HTK+DS+NE)
Results and Conclusions

Conclusions

- Beamforming provides an opportunity to estimate not only the clean signal but also its standard error.
- This error - the observation uncertainty - can be propagated to the MFCC domain or an other suitable domain for improving ASR by uncertainty-of-observation techniques.
- Best results were attained for uncertainty propagation with modified imputation.
- Training is critical, and despite strange philosophical implications, observation uncertainties improve the behaviour after multicondition training as well.
- Strategy is simple & easily generalizes to LVCSR.
Thank you!
Further Improvements

- Training: MCE-Guided Training
  - Iteration and splitting control is done by minimum classification error (MCE) criterion on held-out dataset.
  - Algorithm for mixture splitting:
    - initialize split distance $d$
    - while $m < \text{numMixtures}$
      - split all mixtures by distance $d$ along 1st eigenvector
      - carry out re-estimations until accuracy improves no more
      - if $\text{acc}_m \geq \text{acc}_{m-1}$
        - $m = m+1$
      - else
        - go back to previous model
        - $d = d/f$