Robust ASR on Mobile Devices with Small Array

Iván López Espejo, A. M. Peinado, and A. M. Gomez
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain
{iloes,amp,amgg}@ugr.es

The University of Sheffield
6-3-2015, Sheffield
Main research lines of the group:

- **Robust speech recognition on mobile environments.**
  - Robust ASR on mobile devices with small microphone array.
- Robust transmission of speech and video.
- Ultrasonic non-destructive testing.
- Signal processing in proteomics.
Work done so far...

An overview of the work done so far

---

Work done by me et al. (in chronological order):


Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Introduction and motivation

New ASR upswing

The use of ASR applications has notably increased due to the latest portable electronic devices:

- Great amount of apps (search-by-voice, IPA, dictation, etc.).

Noise-robust ASR in smartphones

- It is crucial to tackle with noisy environments.
- We can take benefit from the novel dual-mic feature.
In a close-talk position:

- Speech power at the primary mic tends to be greater than at the secondary one.
- **Far field noise**: Noise power received at both mics is similar.
- **Our goal**: Estimating the clean speech power spectrum at the primary channel by using the information at both channels.
Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Dual-channel signal model

- We consider additive noise:
  \[ y_i(m) = x_i(m) + n_i(m), \text{ where } i = 1, 2 \]
  indicates the mic (channel).

- Assuming that speech and noise are independent:
  \[
  |Y_1(k, t)|^2 = |X_1(k, t)|^2 + |N_1(k, t)|^2 \\
  |Y_2(k, t)|^2 = |X_2(k, t)|^2 + |N_2(k, t)|^2
  \]
Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Minimum mean square noise (MMSN) feature enhancer

• Minimum mean square noise (MMSN) feature enhancer is defined as
  \[ |\hat{X}_1(k, t)|^2 = w_k^T \left( \begin{array}{c} |Y_1(k, t)|^2 \\ |Y_2(k, t)|^2 \end{array} \right). \]

• The speech power in the second channel is related with the speech power in the first one through a time-invariant factor \( A_{21}(k) \):
  \[ |Y_2(k, t)|^2 = A_{21}(k)|X_1(k, t)|^2 + |N_2(k, t)|^2. \]

• Weights are computed by using the well-known MVDR (minimum variance distortionless response) approach:
  \[ w_k = \frac{\Phi_{N,k}^{-1}(1,A_{21}(k))^T}{(1,A_{21}(k))\Phi_{N,k}^{-1}(1,A_{21}(k))^T}. \]
Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Dual-channel spectral subtraction (DCSS)

• We can also relate noise power spectra at both channels:
  \[ |Y_1(k, t)|^2 = |X_1(k, t)|^2 + G_{12}(k)|N_2(k, t)|^2. \]

• Dual-channel spectral subtraction (DCSS) estimator:
  \[ |\hat{X}_1(k, t)|^2 = \frac{|Y_1(k, t)|^2 - G_{12}(k)|Y_2(k, t)|^2}{1 - G_{12}(k)A_{21}(k)}. \]

• \( G_{12}(k) \) is estimated by minimizing
  \[ \mathbb{E} [ (|N_1(k, t)|^2 - G_{12}(k)|N_2(k, t)|^2)^2 ] : \]
  \[ \hat{G}_{12}(k) = \frac{\hat{\phi}_{N,k}(1,2)}{\hat{\phi}_{N,k}(2,2)}. \]
Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

The AURORA2-2C-CT database

- **Test A:** Bus, babble, car and pedestrian street
- **Test B:** Cafe, street, bus and train stations
- **SNRs:** \{-5,0,5,10,15,20\} dB and clean
Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

GMM-HMM
(trained with clean speech)

- **PLD**: speech enhancer for smartphones with dual-microphone.
- MMSN and DCSS have a similar performance when $A_{21}(k) \to 0$:
  \[
  \begin{align*}
  &w_1(k)^{\{MMSN,DCSS\}} \to 1 \\
  &w_2(k)^{\{MMSN,DCSS\}} \to -\frac{\phi_{N,k(1,2)}}{\phi_{N,k(2,2)}}
  \end{align*}
  \]

\[
|\hat{X}_1(k,t)|^2 = w_1(k)|Y_1(k,t)|^2 + w_2(k)|Y_2(k,t)|^2
\]
Another possible approach to noise-robust ASR: spectral reconstruction

A BINARY MASK IS NEEDED
A DNN Approach for Mask Estimation on Dual-Mic Smartphones

DNN-based mask estimation system

Features:
\[ \mathbf{y} = \begin{pmatrix} y(t - L) \\ \vdots \\ y(t + L) \end{pmatrix}, \]

where
\[ y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} \]

- Input dim.: \( d = 2 \cdot M \cdot (2L + 1) \times 1 \)

Target:
- Oracle binary mask vector for \( y_1(t) \)
- Output dim.: \( M \times 1 \)
- 7 dB SNR threshold
A DNN Approach for Mask Estimation on Dual-Mic Smartphones

Experiments and results

<table>
<thead>
<tr>
<th></th>
<th>WAcc (%)</th>
<th>Wrong mask bins (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test A</td>
<td>Test B</td>
</tr>
<tr>
<td>Baseline</td>
<td>67.96</td>
<td>59.78</td>
</tr>
<tr>
<td>AFE</td>
<td>82.71</td>
<td>76.37</td>
</tr>
<tr>
<td>Oracle+TGI</td>
<td>96.67</td>
<td>94.41</td>
</tr>
<tr>
<td>SPLICE</td>
<td>82.03</td>
<td>72.72</td>
</tr>
<tr>
<td>T-SNR+TGI</td>
<td>81.21</td>
<td>72.87</td>
</tr>
<tr>
<td>DNN+TGI</td>
<td><strong>88.10</strong></td>
<td><strong>78.07</strong></td>
</tr>
</tbody>
</table>

GMM-HMM (trained with clean speech)
Soft-Mask Weighting for Robust ASR on Smartphones with a Dual-Mic

Description of the approach

- We follow a Wiener filter approach:
  \[ |\hat{X}_1(k, t)|^2 = \hat{H}_1^2(k, t)|Y_1(k, t)|^2. \]

- \( \hat{H}_1^2(k, t) = \left( \frac{\hat{\xi}_1(k, t)}{\hat{\xi}_1(k, t) + 1} \right)^2 \) may be seen as a spectral weighting soft-mask (b).

- We exploit the PLD by assuming that
  \( \varphi_{X_1}(k, t) \gg \varphi_{X_2}(k, t) \) and \( \varphi_{N_2}(k, t) \approx \varphi_{N_1}(k, t) \):
  \[ \hat{\xi}_1(k, t) = \max \left( \frac{\varphi_{Y_1}(k, t)}{\varphi_{Y_2}(k, t)} - 1, 0 \right). \]

- We apply a post-processing to improve the soft-mask:
  1. Slight contrast by using a sigmoid function (c).
  2. Median and Gaussian filtering to improve the spectro-temporal coherence (d).
Robust ASR on Mobile Devices with Small Array

Iván López Espejo

Where I come from

Work done so far...

EUSIPCO’14

LNCS’14

InterSPEECH ’15

About to submit...

Conclusions

Soft-Mask Weighting for Robust ASR on Smartphones with a Dual-Mic

Description of the approach

• We follow a Wiener filter approach:
  \[ |\hat{X}_1(k, t)|^2 = \hat{H}_1^2(k, t)|Y_1(k, t)|^2. \]

• \[ \hat{H}_1^2(k, t) = \left( \frac{\hat{\xi}_1(k, t)}{(\hat{\xi}_1(k, t) + 1)} \right)^2 \] may be seen as a spectral weighting soft-mask.

• We exploit the PLD by assuming that \[ \varphi_{X_1}(k, t) \gg \varphi_{X_2}(k, t) \] and \[ \varphi_{N_2}(k, t) \approx \varphi_{N_1}(k, t): \]
  \[ \hat{\xi}_1(k, t) = \max \left( \frac{\varphi_{Y_1}(k, t)}{\varphi_{Y_2}(k, t)} - 1, 0 \right). \]

• We apply a post-processing to improve the soft-mask:
  1. Slight contrast by using a sigmoid function.
  2. Median and Gaussian filtering to improve the spectro-temporal coherence.

From top to bottom:

1. Clean
2. Noisy
3. Soft-mask
4. Enhanced
Soft-Mask Weighting for Robust ASR on Smartphones with a Dual-Mic

Results

<table>
<thead>
<tr>
<th>Tech./SNR (dB)</th>
<th>-5</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>Clean</th>
<th>Av. (-5 to 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>18.15</td>
<td>31.85</td>
<td>56.11</td>
<td>82.78</td>
<td>94.72</td>
<td>97.76</td>
<td>99.13</td>
<td>63.56</td>
</tr>
<tr>
<td>SMW</td>
<td>26.23</td>
<td>51.76</td>
<td>77.03</td>
<td>89.49</td>
<td>94.19</td>
<td>96.09</td>
<td>98.40</td>
<td>72.47</td>
</tr>
<tr>
<td>MMSN</td>
<td>24.16</td>
<td>46.31</td>
<td>74.78</td>
<td>90.66</td>
<td>96.14</td>
<td>97.87</td>
<td>98.90</td>
<td>71.65</td>
</tr>
<tr>
<td>DCSS</td>
<td>24.37</td>
<td>46.69</td>
<td>75.06</td>
<td>90.65</td>
<td>96.03</td>
<td>97.64</td>
<td>99.13</td>
<td>71.74</td>
</tr>
<tr>
<td>New</td>
<td>28.36</td>
<td>52.59</td>
<td>79.65</td>
<td>92.39</td>
<td>96.68</td>
<td>98.04</td>
<td>99.11</td>
<td>74.62</td>
</tr>
<tr>
<td>VTS-1</td>
<td>44.25</td>
<td>72.75</td>
<td>89.69</td>
<td>95.44</td>
<td>97.71</td>
<td>98.49</td>
<td>99.09</td>
<td>83.06</td>
</tr>
<tr>
<td>SMW+VTS-1</td>
<td>29.37</td>
<td>56.52</td>
<td>79.75</td>
<td>90.19</td>
<td>94.08</td>
<td>95.72</td>
<td>98.13</td>
<td>74.27</td>
</tr>
<tr>
<td>MMSN+VTS-1</td>
<td>56.15</td>
<td>81.05</td>
<td>92.41</td>
<td>96.60</td>
<td>98.15</td>
<td>98.65</td>
<td>98.99</td>
<td>87.17</td>
</tr>
<tr>
<td>DCSS+VTS-1</td>
<td>56.22</td>
<td>81.04</td>
<td>92.40</td>
<td>96.57</td>
<td>98.21</td>
<td>98.63</td>
<td>99.09</td>
<td>87.18</td>
</tr>
<tr>
<td>New+VTS-1</td>
<td>61.14</td>
<td>81.43</td>
<td>92.05</td>
<td>95.89</td>
<td>97.82</td>
<td>98.48</td>
<td>99.05</td>
<td>87.80</td>
</tr>
</tbody>
</table>

Table: WAcc results in terms of percentage. Results are averaged across all types of noise in test sets A and B.
Power Spectrum Enhancement for Noise-Robust ASR with Small Mic Arrays

Motivation

Facts...
- Several types of devices can be used.
- Devices can have more than two mics arranged in different ways.
- Devices can be used in different and variable positions.

Therefore...
- We generalize our previous work to $C$ mics and a variable position.
  - MMSN $\rightarrow$ P-MVDR (power MVDR)
  - DCSS $\rightarrow$ MSS (multichannel spectral subtraction)
Power Spectrum Enhancement for
Noise-Robust ASR with Small Mic Arrays

Speech gain vector and P-MVDR and MSS equations

**P-MVDR:**
\[
|\hat{X}_1(k, t)|^2 = \left( \frac{\Phi^{-1}_{k,t} A_{k,t}}{A_{k,t}^T \Phi^{-1}_{k,t} A_{k,t}} \right)^T Y_{k,t}
\]

**MSS:**
\[
|\hat{X}_1(k, t)|^2 = \frac{Y_{k,t}^T \Gamma_{k,t} A_{k,t} - Y_{k,t}^T A_{k,t} \cdot ||G_{k,t}||^2}{(A_{k,t}^T G_{k,t})^2 - ||A_{k,t}||^2 \cdot ||G_{k,t}||^2}
\]
\[
\Gamma_{k,t} = G_{k,t} \cdot G_{k,t}^T
\]
\[
G_{k,t} = (1, G_{21}(k, t), ..., G_{C1}(k, t))^T
\]

- Use position or acoustics may be variable.
- We developed an MMSE-based estimator to estimate
  \[
  A_{k,t} = (1, A_{21}(k, t), ..., A_{C1}(k, t))^T
  \]
on a frame-by-frame basis.
Power Spectrum Enhancement for Noise-Robust ASR with Small Mic Arrays

About our experiments and results

- We created the AURORA2-2C-FT database emulating a smartphone with a dual-mic but in far-talk conditions.
- We created validation test datasets with real noisy data for both close-talk and far-talk conditions.
- Our recognition results showed the success of our developments in all cases.
Conclusions and future work

- Multichannel information can be exploited to improve ASR performance.
- There is little work on robust ASR with small mic arrays → We should be able to achieve further improvements.
- We are very interested in obtaining new and good results on the CHiME-3 database.
Thanks for your attention and any questions?

Contact:
Iván López Espejo
Department of Signal Theory, Telematics and Communications
University of Granada
E-mail: iloes@ugr.es