Low-Resource Keyword Spotting for Hearing Assistive Devices

Iván López-Espejo

Signal and Information Processing (SIP)

ivl@es.aau.dk

Thursday 5th March, 2020
Overview

1. Project Overview

2. Works

3. Robustness Against External Speakers

4. Improved Robustness Against External Speakers
Project Overview

Motivation

- Manual operation of hearing assistive devices (HADs) is cumbersome in a number of situations.

- To assist in addressing this issue, voice interfaces are envisioned as a means for handling and operating HADs in a practical manner.
Objectives

- Research and development of keyword spotting (KWS) systems for HADs:
  1. Personalization.
  2. Robustness against noise.
  3. Low memory and low computational complexity.

To accomplish these objectives, we explore...

1. ...the combined use of multi-microphone signals from HADs along with signal processing and the latest deep learning techniques.
2. ...the utilization of user-specific aspects, e.g., voice characteristics or head-related acoustics of the specific user.

We expect to contribute to enhance the life quality of hearing-impaired people.


Robustness Against External Speakers

Introduction and Motivation

- **KWS systems for HADs must be robust against external speakers**, that is, the user must be the only one allowed to trigger actions on her/his HAD.

- We proposed HAD user (speaker)-dependent KWS drawing from a state-of-the-art small-footprint KWS system based on deep residual learning and dilated convolutions (res15) [1].

**Two tasks:** KWS and own-voice/external speaker detection.

The sigmoid layer outputs a probability $P \left( S_u \mid \tilde{V}, \theta \right)$ that the input $\tilde{V}$ corresponds to an utterance said by the HAD user $S_u$.

KWS prediction $P \left( \mathcal{W}_c \mid \tilde{V}, \theta \right)$ from $\tilde{V}$ is considered if $P \left( S_u \mid \tilde{V}, \theta \right) > P_{THR}$ ($P_{THR} = 0.5$).
We created a **two-microphone hearing aid speech database** from the Google Speech Commands Dataset (GSCD).

HAD user own-voice signals were generated by filtering 75% of the GSCD through a single own-voice transfer function (OVTF).

External speaker signals were created by filtering the remaining 25% of the GSCD through head-related transfer functions (HRTFs).

Apart from the **unknown word** class, **10 keywords** were considered: “yes”, “no”, “up”, “down”, “left”, “right”, “on”, “off”, “stop” and “go”.
Robustness Against External Speakers

Results

- **Accuracy results (%)** with 95% confidence intervals.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Training data</th>
<th>Input type</th>
<th>Own-voice subset</th>
<th>External speaker subset</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>res15 (KWS only)</td>
<td>Front and rear mics</td>
<td>Overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Front</td>
<td>Multi-task</td>
<td>Own and external voice</td>
<td>97.49 (\pm) 1.02</td>
<td>80.38 (\pm) 5.23</td>
<td>93.02 (\pm) 0.76</td>
</tr>
<tr>
<td>Rear</td>
<td>Multi-task</td>
<td>Own and external voice</td>
<td>97.28 (\pm) 1.08</td>
<td>79.03 (\pm) 5.06</td>
<td>92.51 (\pm) 0.68</td>
</tr>
<tr>
<td>Dual</td>
<td>Multi-task</td>
<td>Own and external voice</td>
<td>99.60 (\pm) 0.22</td>
<td>96.22 (\pm) 1.61</td>
<td>98.72 (\pm) 0.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keyword spotting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Own-voice subset</td>
</tr>
<tr>
<td>94.21 (\pm) 0.39</td>
</tr>
<tr>
<td>94.28 (\pm) 0.37</td>
</tr>
<tr>
<td>94.48 (\pm) 0.25</td>
</tr>
<tr>
<td>94.59 (\pm) 0.32</td>
</tr>
</tbody>
</table>

- **The OVTF and HRTFs are more similar (in terms of MFCC Euclidean distance)** at angles where we see a relative drop in performance.

DET curves for own-voice/external speaker detection.

External speaker detection accuracy as a function of the angle of external speakers.

Iván López-Espejo (SIP)
While the created two-microphone hearing aid speech database comprises speech signals uttered by many different speakers, impulse responses for its generation were only measured on a single actual person.

Impulse responses are user-dependent, as these characterize physical features, e.g., head size and shape.

We created a new speech corpus with impulse responses measured on multiple persons wearing a hearing aid: **multi-user database**.

**Problem!** Performance loss in terms of KWS accuracy: from 94.86% ± 0.39 to 80.45% ± 0.55.
Towards reducing the performance loss:

- The relative position of the users’ mouth w.r.t. the hearing aid microphones is virtually time-invariant and different from that of an external speaker:
  - Spectral magnitude features for KWS.
  - Phase difference information (GCC-PHAT-based coefficients) for own-voice/external speaker detection.

- Use of the perceptually-motivated constant-Q transform: at lower (higher) frequencies the frequency (time) resolution is higher.
Improved Robustness Against External Speakers

Results

Accuracy results (%) with 95% confidence intervals.

<table>
<thead>
<tr>
<th>Multi-user database</th>
<th>Own-voice/External speaker detection</th>
<th>Keyword spotting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Own-voice subset</td>
<td>External speaker subset</td>
</tr>
<tr>
<td>Baseline</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MFCC-80×1</td>
<td>92.64 ± 1.39</td>
<td>55.36 ± 4.43</td>
</tr>
<tr>
<td>MFCC-40×2</td>
<td>97.03 ± 1.81</td>
<td>87.18 ± 2.06</td>
</tr>
<tr>
<td>STFT-S</td>
<td>98.60 ± 0.95</td>
<td>95.03 ± 1.10</td>
</tr>
<tr>
<td>CQT-S</td>
<td>98.44 ± 0.87</td>
<td>92.12 ± 2.39</td>
</tr>
<tr>
<td>STFT-S+GCC</td>
<td>98.61 ± 1.30</td>
<td>96.40 ± 1.21</td>
</tr>
<tr>
<td>CQT-S+GCC</td>
<td>99.49 ± 0.47</td>
<td>98.67 ± 0.36</td>
</tr>
</tbody>
</table>

DET curves for own-voice/external speaker detection.

External speaker detection accuracy as a function of the angle of external speakers.
Thanks for your attention!