Performance evaluation of voice assistant devices

ETS! Workshop on Multimedia Quality in Virtual, Augmented, or other Realities.

S. Isabelle, Knowles Electronics

May 10, 2017
Voice-controlled assistant devices are rapidly growing in popularity
- Amazon Echo: 8.2 million devices from late 2014 to early 2017 (CIRP)
- Google Home in US November 2016, UK April 2017

Driven in large part by recent success of cloud-based voice recognition and natural language processing preceding internet search and home automation.

Local acoustics/signal processing play significant role in user experience.

Develop consistent and repeatable evaluation methods to support new market.
Focus on Voice Wake

- Current devices offer ‘speak to wake’ keyword/keyphrase detection
- Locally processed, prior to buffering longer speech strings for e.g., internet search
- Key performance metrics:
  - **True Positive Rate**: detecting keyword/phrase when actually present
    - Measured for range of acoustic ambient conditions, speech levels, distances to device
    - Including when device is already generating sound (e.g., playing music): “Barge In”
  - **False Detections**: incorrectly declaring detection when keyword/phrase is not present
Devices and Talkers

**Devices:**
- Three devices: two commercially available, one pre-commercial
  - Intended for table-top usage
  - can be addressed from any direction

**Speech corpus:**
- 30 talkers, each speaking two trials of the keyphrase for each of the three devices: total of 60 test utterances per device
  - Recordings made in anechoic environment with low ambient noise of 16.4 dB SPL(A)
  - Instrumentation microphone within 20 cm of talker’s mouth
  - Keyphrase is 3 to 4 syllables in duration
  - Not all talkers were native speakers of American English

- Speech reproduced using artificial mouth of a Head and Torso Simulator (ITU-T P.58)
  - Normalization of reproduced level at Mouth Reference Position (MRP)
  - Level tested over a range from 91 down to 55 dB SPL at MRP
Device Set up

- Test chamber: 6.7 x 4.6 x 2.9 m

- Acoustically treated to mimic a quiet domestic living room:
  - RT$_{60}$ = 420 msec
  - C$_{80}$ = 12.7 dB
  - C$_d$ = 1.35 m
  - L = 25 dB SPL(A)

- Distance HATS to DUT = 1, 3, 5 m

- Loudspeakers to generate background noise (babble)

- Echo signal (music) played through DUT
Results: TPR versus distance, with/without noise

- Babble is 57 dB SPL (A)
- SNR of +6 dB at the DUT with HATS at 1 m.
- TPR based on 60 trials
- Error bars show confidence intervals based on binomial distribution
Results: TPR versus speech level

TPR based on 60 trials

The information contained in this document is confidential and/or proprietary to Knowles Corporation and/or its affiliates. Please do not share this document or the information contained herein with anyone outside of Knowles Corporation or its affiliates, without first obtaining permission from an authorized representative of Knowles Corporation or its affiliates, as applicable, © 2016, Knowles Electronics, LLC, Itasca, IL USA. All Rights Reserved. Knowles and the logo are trademarks of Knowles Electronics, LLC.
Results: TPR with echo, 1 m

- TPR based on 60 trials
- DUT playing music
- Speech to Echo Ratio (SER) adjusted by changing volume control on DUT, to match SER across DUTs.
- Highest SER obtained at low setting of volume control.
- Highest SER about +3 dB
- Low SER is about -20 dB
Results: TPR with echo, 5 m

- TPR based on 60 trials
- DUT playing music
- Speech to Echo Ratio (SER) adjusted by changing volume control on DUT, to match SER across DUTs.
- Highest SER obtained at low setting of volume control.
False Detections

- DUTs placed in front of loudspeaker playing a broadcast talk-radio program
- No keyphrases intentionally presented.
- Table shows number of false wakeups in 20 hours of testing:

<table>
<thead>
<tr>
<th>Device</th>
<th>False Wakeups</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
</tr>
</tbody>
</table>
Results show keyphrase detection in quiet and noise, at a range of distances, and a range of levels of speech.

Several factors can be identified for further investigation
• Impact of room and relative geometry of sources and devices within the room
• Impact of background noise reproduction
• Other performance metrics, such as task completion

This class of device is expanding in the marketplace, and correspondingly, the need for standardized methods of evaluation is increasing.