Keyword Spotting for Hearing Assistive Devices Robust to External Speakers

Iván López-Espejo[1], Zheng-Hua Tan[1] and Jesper Jensen[1,2]

[1]Department of Electronic Systems, Aalborg University, Denmark
[2]Oticon A/S, Denmark
{ivl,zt,jje}@es.aau.dk, jesj@oticon.com

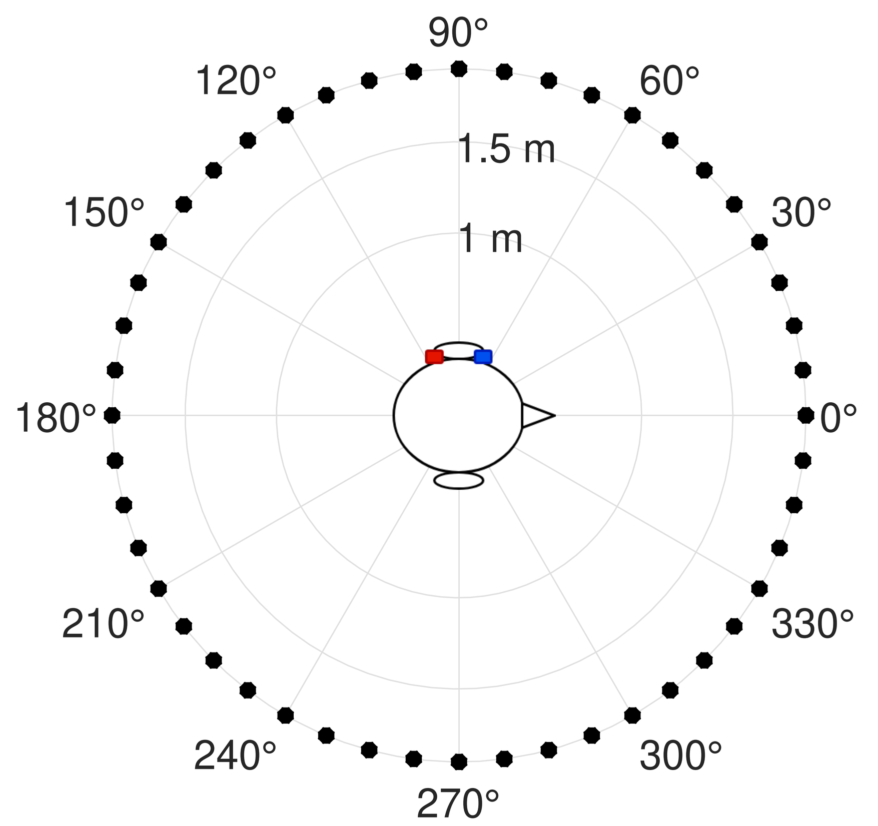
AALBORG UNIVERSITY DENMARK

CASPR Centre for Acoustic Signal Processing Research

Introduction and Motivation

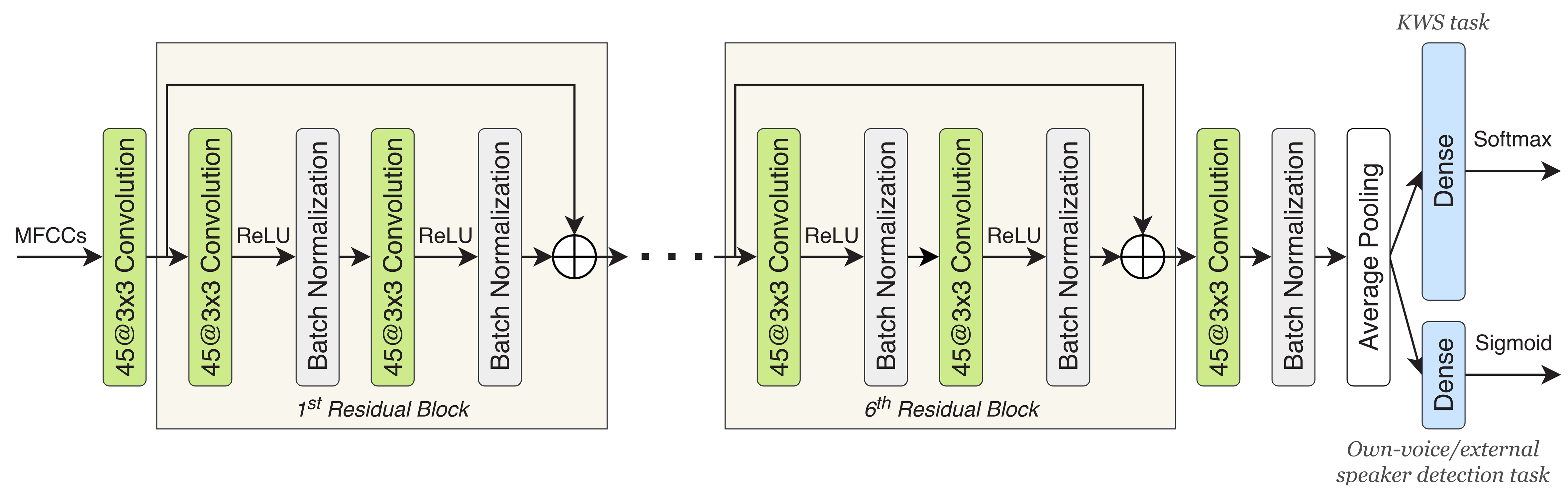
- Keyword spotting (**KWS**) may allow a hearing impaired person to initiate certain actions on her/his hearing assistive device (**HAD**), e.g., increasing the volume.
- KWS systems for HADs must...
  - ...have low memory and computational complexity (small footprint).
  - ...be robust against external speakers, that is, the user must be the only one allowed to trigger actions on her/his HAD.
- In general, current KWS systems are speaker-independent.
- **We propose 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].

Experimental Framework



- We create a **two-microphone hearing aid speech database** from the Google Speech Commands Dataset (**GSCD**) [2].
- HAD user own-voice signals are generated by filtering 75% of the GSCD through a single own-voice transfer function (**OVTF**).
- External speaker signals are created by filtering the remaining 25% of the GSCD through head-related transfer functions (**HRTFs**).
- Speakers do not overlap across the training (80%), validation (10%) and test (10%) sets.
- Apart from the *unknown word* class, **10 keywords** are considered: "yes", "no", "up", "down", "left", "right", "on", "off", "stop" and "go".

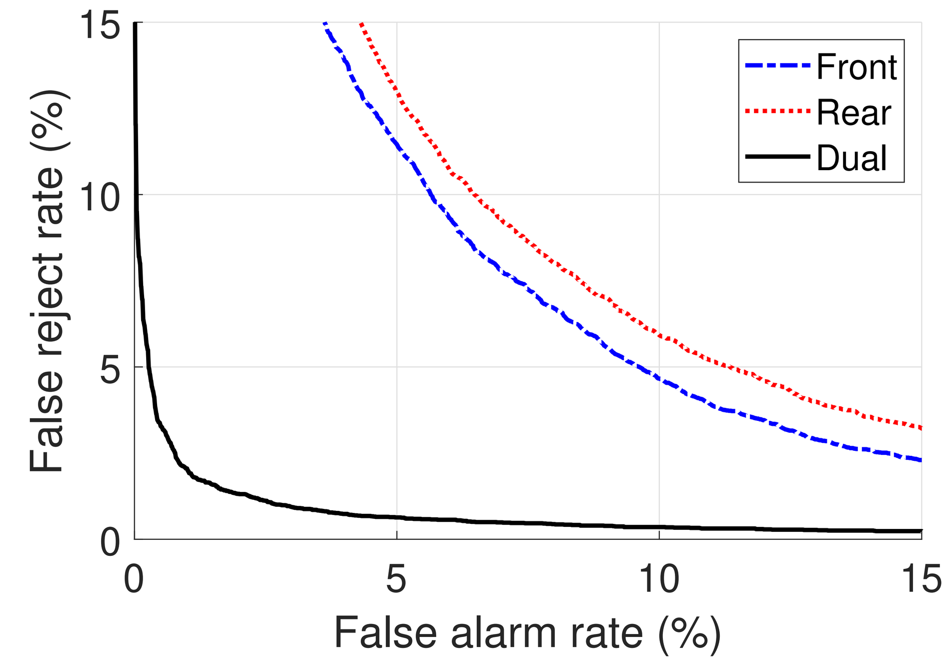
Multi-task Learning for KWS and Own-Voice/External Speaker Detection

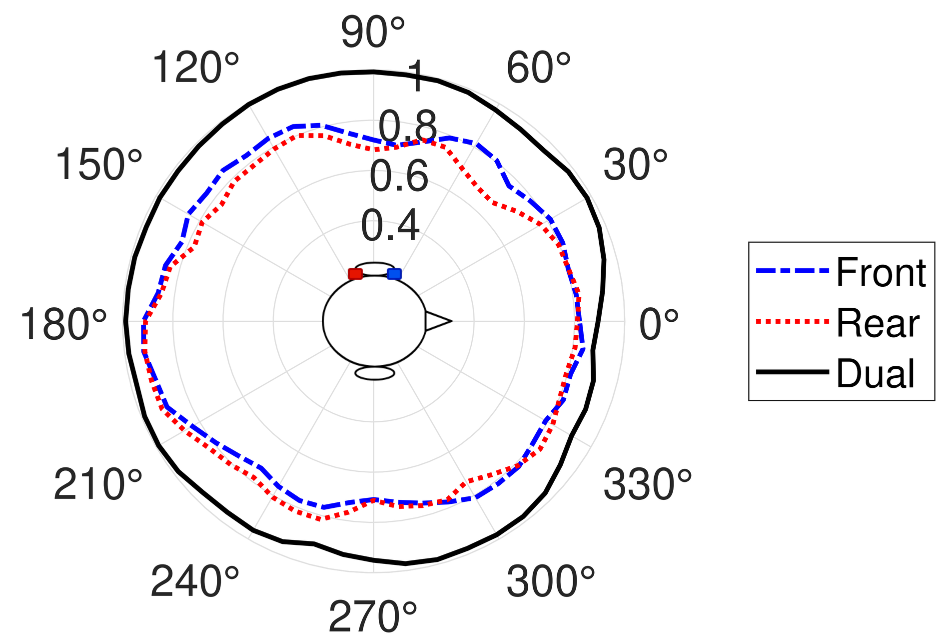


- **Two tasks:** KWS and own-voice/external speaker detection.
- We experimented with dynamic task prioritization for loss weight selection.
- The sigmoid layer outputs a probability that the input MFCCs correspond to an utterance said by the HAD user.
- KWS prediction from MFCCs is considered if and only if the sigmoid layer output is above 0.5.

Results

	Architecture	Training data	Input type	Own-voice/External speaker detection: *Own-voice subset*	*External speaker subset*	*Overall*	Keyword spotting: *Own-voice subset*	*Overall*
Baseline	`res15` (KWS only)	Own voice	Front and rear mics	—	—	—	94.21 ± 0.39	71.87 ± 0.30
Front	Multi-task	Own and external voice	Front mic	97.49 ± 1.02	80.38 ± 5.23	93.02 ± 0.76	94.28 ± 0.37	89.48 ± 0.74
Rear	Multi-task	Own and external voice	Rear mic	97.28 ± 1.08	79.03 ± 5.06	92.51 ± 0.68	94.48 ± 0.25	89.29 ± 0.55
Dual	Multi-task	Own and external voice	Front and rear mics	99.60 ± 0.22	96.22 ± 1.61	98.72 ± 0.29	94.59 ± 0.32	94.86 ± 0.39





- **Accuracy results (%)** with 95% confidence intervals.
- *Left plot*: Detection error trade-off curves for own-voice/external speaker detection.
- *Right plot*: External speaker detection accuracy as a function of the position (angle) of external speakers.

Observations:

- The OVTF and HRTFs are more similar (in terms of MFCC Euclidean distance) at angles where we see a relative drop in performance (right plot).
- Own-voice/external speaker detection is crucial for good KWS performance (check overall *vs.* own-voice subset KWS accuracy).

Conclusions

- Proposed approach outperforms state-of-the-art small-footprint KWS systems with a negligible increase in the number of parameters of the model.
- **Code available at** `http://www.ugr.es/~iloes/codes/MultitaskKWS.zip`

References

[1] Raphael Tang and Jimmy Lin. Deep residual learning for small-footprint keyword spotting. In *Proceedings of ICASSP 2018, April 15-20, Calgary, Canada*, pages 5484–5488, 2018.

[2] Pete Warden. Speech Commands: A dataset for limited-vocabulary speech recognition. *arXiv:1804.03209v1*, 2018.