Low-Resource Keyword Spotting for Hearing Assistive Devices

Iván López-Espejo

Signal and Information Processing (SIP)

ivl@es.aau.dk

Thursday 5^{th} March, 2020

Overview

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.

- 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.

Works

Iván López-Espejo, Zheng-Hua Tan and Jesper Jensen: "Keyword Spotting for Hearing Assistive Devices Robust to External Speakers", in *Proc. of INTERSPEECH 2019*, Graz (Austria), 2019.

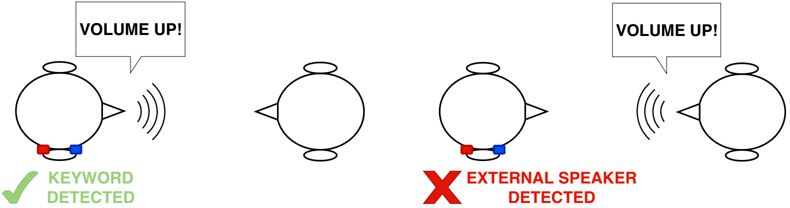
Iván López-Espejo, Zheng-Hua Tan and Jesper Jensen: "Improved External Speaker-Robust Keyword Spotting for Hearing Assistive Devices", submitted to *IEEE Transactions on Audio, Speech and Language Processing*.

Iván López-Espejo, Zheng-Hua Tan and Jesper Jensen: "Exploring Filterbank Learning for Keyword Spotting", submitted to *EUSIPCO 2020*, Amsterdam (The Netherlands), 2020.

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.

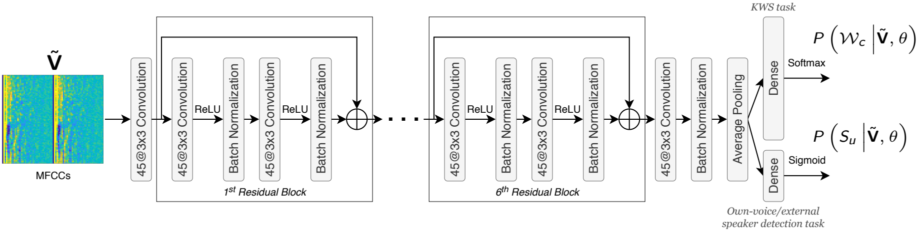


- **We proposed HAD user (speaker)-dependent KWS** drawing from a state-of-the-art small-footprint KWS system based on deep residual learning and dilated convolutions (`res15`) [1].

[1] Raphael Tang and Jimmy Lin: "Deep residual learning for small-footprint keyword spotting", in *Proc. of ICASSP 2018*, Calgary (Canada), 2018.

Robustness Against External Speakers

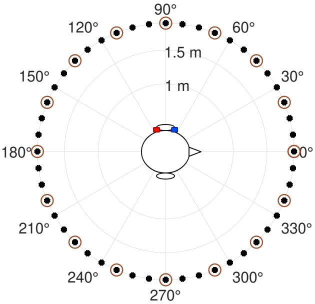
Multi-task Learning

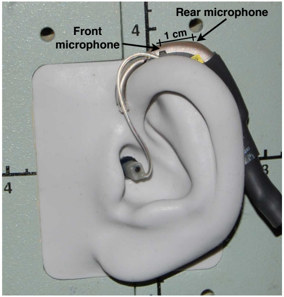


- **Two tasks:** KWS and own-voice/external speaker detection.
- The sigmoid layer outputs a probability $P\left(S_u \middle| \tilde{\mathbf{V}}, \theta\right)$ that the input $\tilde{\mathbf{V}}$ corresponds to an utterance said by the HAD user S_u.
- KWS prediction $P\left(\mathcal{W}_c \middle| \tilde{\mathbf{V}}, \theta\right)$ from $\tilde{\mathbf{V}}$ is considered if $P\left(S_u \middle| \tilde{\mathbf{V}}, \theta\right) > P_{THR}$ ($P_{THR} = 0.5$).

Robustness Against External Speakers

Experimental Framework

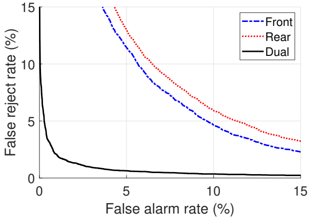




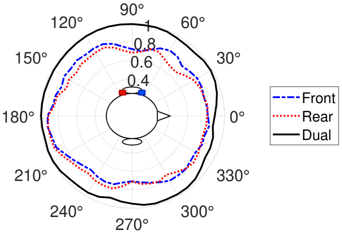
- We created a **two-microphone hearing aid speech database** from the Google Speech Commands Dataset (**GSCD**).
- HAD user own-voice signals were generated by filtering 75% of the GSCD through a single own-voice transfer function (**OVTF**).
- External speaker signals were created by filtering the 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".

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

	Architecture	Training data	Input type	Own-voice/External speaker detection			Keyword spotting	
				Own-voice subset	*External speaker subset*	*Overall*	*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



DET curves for own-voice/external speaker detection.



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

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

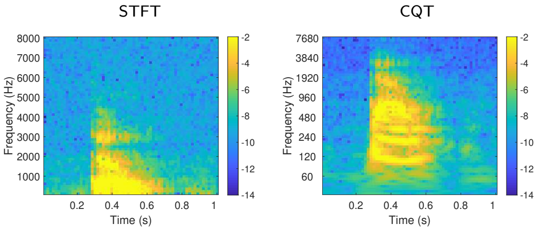
Improved Robustness Against External Speakers

Improved Experimental Framework

- 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% $\pm$ 0.39 to 80.45% $\pm$ 0.55.

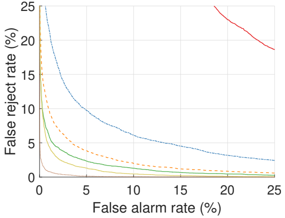
Towards reducing the performance loss:

- 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.

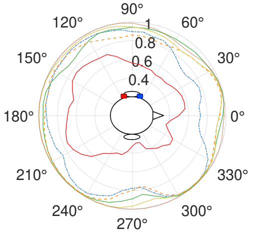


Accuracy results (%) with 95% confidence intervals.

		Own-voice/External speaker detection			Keyword spotting	
		Own-voice subset	*External speaker subset*	*Overall*	*Own-voice subset*	*Overall*
Multi-user database	Baseline	—	—	—	93.81 ± 0.27	73.88 ± 0.23
	MFCC-80×1	92.64 ± 1.39	55.36 ± 4.43	84.26 ± 0.45	93.27 ± 0.30	**80.45 ± 0.55**
	MFCC-40×2	97.03 ± 1.81	87.18 ± 2.06	94.81 ± 1.20	94.32 ± 0.21	90.78 ± 1.16
	STFT-S	98.60 ± 0.95	95.03 ± 1.10	97.80 ± 0.53	94.30 ± 0.34	93.59 ± 0.64
	CQT-S	98.44 ± 0.87	92.12 ± 2.39	97.02 ± 0.44	94.60 ± 0.31	93.19 ± 0.52
	STFT-S+GCC	98.61 ± 1.30	96.40 ± 1.21	98.11 ± 0.93	94.23 ± 0.57	93.77 ± 0.99
	CQT-S+GCC	99.49 ± 0.47	98.67 ± 0.36	99.31 ± 0.33	94.81 ± 0.26	**95.34 ± 0.32**



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!