Filterbank Learning for Noise-Robust Small-Footprint Keyword Spotting

Iván López-Espejo[1,2], Ram C. M. C. Shekar[2], Zheng-Hua Tan[1], Jesper Jensen[1,3] and John H. L. Hansen[2]

[1]Department of Electronic Systems, Aalborg University, Denmark
[2]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA
[3]Oticon A/S, Denmark

{ivl,zt,jje}@es.aau.dk, {ramcharan.chandrashekar,john.hansen}@utdallas.edu, jesj@oticon.com

Funded by the European Union

Introduction

- *In spite of recent attempts*, the replacement of solid *handcrafted* speech features by learnable features that are able to yield better **keyword spotting (KWS)** performance has not been achieved
- Much of the spectral information is redundant when it comes to the recognition of a set of few keywords
- We prove that filterbank learning outperforms handcrafted speech features for KWS *as long as the number of filterbank channels is drastically reduced* $\longrightarrow$ **KWS performance-energy consumption trade-off**

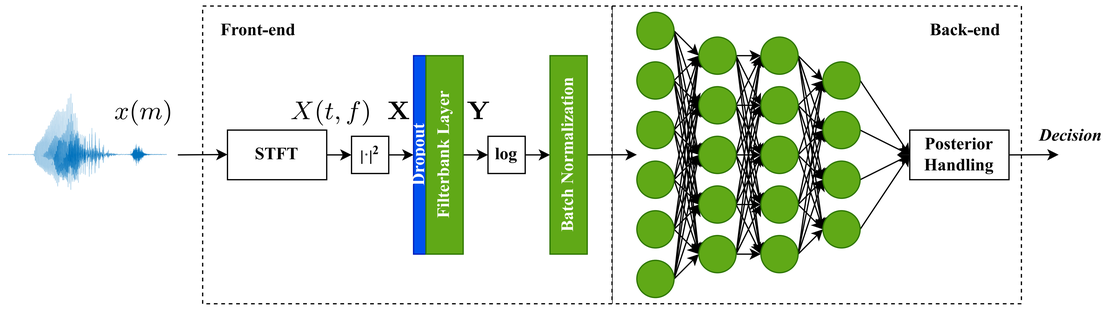
Experimental Setup

- We employ a *noisy* version of the **Google Speech Commands Dataset**

Noise type	SNR (dB) -10	-5	0	5	10	15	20	Clean
White noise			TRAINING AND VALIDATION SETS					
Babble								
Machine gun								
F-16 cockpit								
Vehicle interior								
Factory$_1$								
Bus								
Pedestrian street								
Factory$_2$								
Buccaneer jet cockpit			TEST SET					
Café								
Street junction								

- The number of linear frequency bins is F = 241, and the dropout rate, 0.4
- **W** is initialized by a Mel filterbank
- The back-end is trained to model **11** different **classes**:
  1. The **10 keywords** “yes”, “no”, “up”, “down”, “left”, “right”, “on”, “off”, “stop” and “go”
  2. The filler (i.e., non-keyword) class

Filterbank Learning Methodology



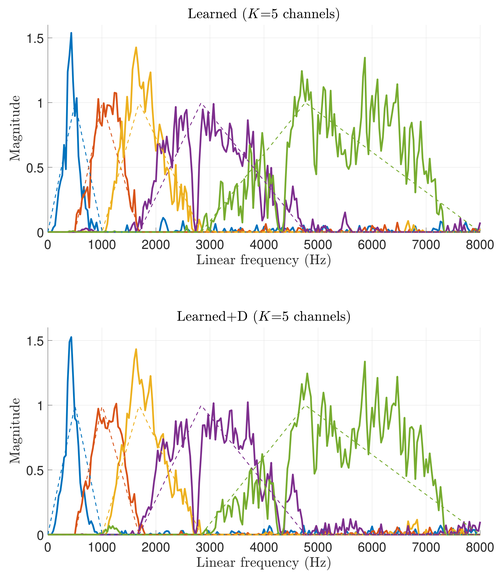
- The filterbank layer implements $\mathbf{Y} = \mathbf{X} \cdot \max(\mathbf{W}, 0)$, where the K-ch. filterbank matrix $\mathbf{W} \in \mathbb{R}^{F \times K}$ is optimized *jointly* along with the back-end
- The goal of **dropout** is to improve robustness and generalization of individual filterbank channels
- For acoustic modeling (*back-end*), we use a deep residual convolutional neural network integrating dilated convolutions

Results and Discussion

- Statistically significant **accuracy** (%) improvements w.r.t. log-Mel (Learned) are indicated in **boldface** (underline)
- The **number of multiplications** of the back-end (**#Mult.**) exhibits a *strong positive linear relationship* with the **energy consumption** of the KWS system

#Ch. K	#Mult.	Method	SNR (dB) - Seen Noises -10	-5	0	5	10	15	20	Clean	SNR (dB) - Unseen Noises -10	-5	0	5	10	15	20	Clean
40	895M	log-Mel	51.26	67.61	82.27	90.48	93.43	94.08	95.36	96.09	37.17	61.38	78.86	86.87	91.92	93.66	94.58	96.13
		Learned	49.61	66.93	82.46	90.58	93.96	94.23	95.41	95.58	35.41	58.84	77.94	86.82	92.24	93.47	94.70	95.60
		Learned+D	51.21	68.16	83.50	90.29	93.89	93.84	95.43	95.85	38.14	63.29	78.96	87.35	92.29	93.86	94.90	96.20
10	188M	log-Mel	45.41	63.09	79.30	87.85	91.93	92.90	94.59	95.70	32.00	55.67	74.34	83.94	90.23	92.29	93.91	95.07
		Learned	47.37	64.15	80.10	88.86	92.71	93.55	94.88	95.87	33.54	55.91	74.85	84.67	91.22	92.48	94.22	95.31
		Learned+D	46.72	63.57	80.39	88.36	92.13	93.29	94.71	95.58	32.50	57.39	75.14	84.57	90.67	**93.23**	94.39	95.72
8	141M	log-Mel	44.64	60.51	76.01	85.00	90.29	91.88	93.31	94.59	28.25	49.94	68.92	80.82	87.45	90.40	92.12	93.93
		Learned	45.70	62.22	77.97	86.06	90.75	91.45	92.97	94.15	29.65	**53.11**	**72.24**	**84.04**	**90.33**	**92.16**	**93.40**	94.24
		Learned+D	46.23	**63.67**	78.99	**87.58**	90.94	92.32	94.13	**95.70**	30.52	**55.02**	**73.76**	**84.62**	**90.45**	**92.94**	**94.00**	**94.82**
7	118M	log-Mel	42.90	60.72	76.28	84.49	89.18	90.31	92.51	94.23	27.88	48.20	68.51	80.53	87.06	90.40	92.72	94.82
		Learned	44.08	62.27	76.91	85.53	89.54	91.01	92.68	94.49	30.47	**51.80**	**72.12**	82.08	88.05	90.79	92.53	93.74
		Learned+D	45.53	61.35	78.00	**87.00**	**90.80**	**91.43**	93.26	94.88	31.32	**54.68**	**73.30**	**83.94**	**89.50**	**92.09**	92.99	94.32
5	71M	log-Mel	40.12	56.52	73.31	82.56	87.13	88.41	89.78	92.13	24.52	46.84	66.26	79.93	85.03	88.59	90.59	92.50
		Learned	**42.92**	**59.40**	75.34	**84.69**	88.26	89.57	**91.52**	**93.55**	**28.34**	**50.47**	**70.11**	80.82	**86.99**	89.41	91.22	92.74
		Learned+D	**44.66**	**61.76**	**76.93**	**84.57**	**88.77**	**90.07**	**91.79**	**93.91**	**28.39**	**51.29**	**70.04**	**82.54**	**87.67**	**91.27**	**92.26**	**93.81**

- **40-ch. log-Mel** $\rightarrow$ **8-ch. Learned+D:** 3.5% accuracy loss **&** 6.3$\times$ energy consumption reduction
- **8-ch. log-Mel** $\rightarrow$ **5-ch. Learned+D:** Same accuracy **&** 2$\times$ energy consumption reduction



- *Filterbank learning adapts to noise spectral characteristics* to offer a higher degree of robustness to noise