Microsoft

Spatial audio feature discovery with convolutional neural networks

Etienne Thuillier, Hannes Gamper, Ivan J. Tashev
Microsoft Research Redmond

Need

Better understanding of elevation cues

… for efficient & immersive spatial audio rendering

Problem

Listening test studies

… *do not scale to large sample sizes*

… *are exposed to human error*

… *provide highly compressed data*

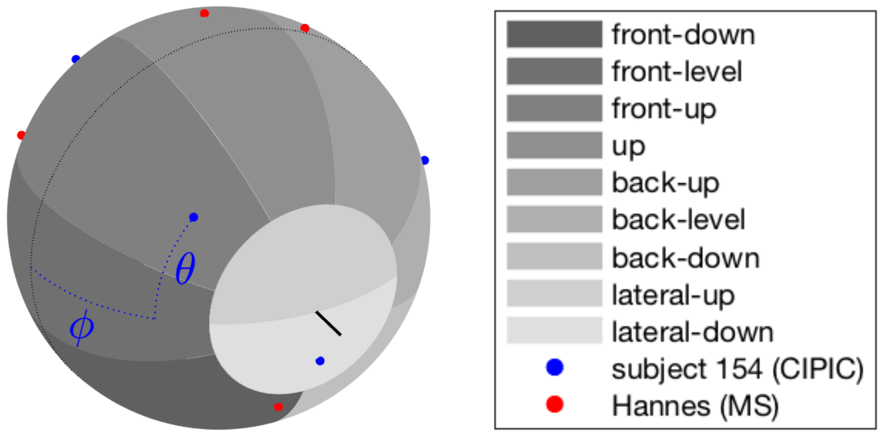
Proposed approach (2 steps)

1) Train a CNN to classify spatialized sounds

Dataset	Year	Subjects	Pairs/subj.
Aachen	2016	46	2304
Microsoft	2015	252	400
RIEC	2014	105	865
ARI	2010	135	1150
CIPIC	2001	45	1250

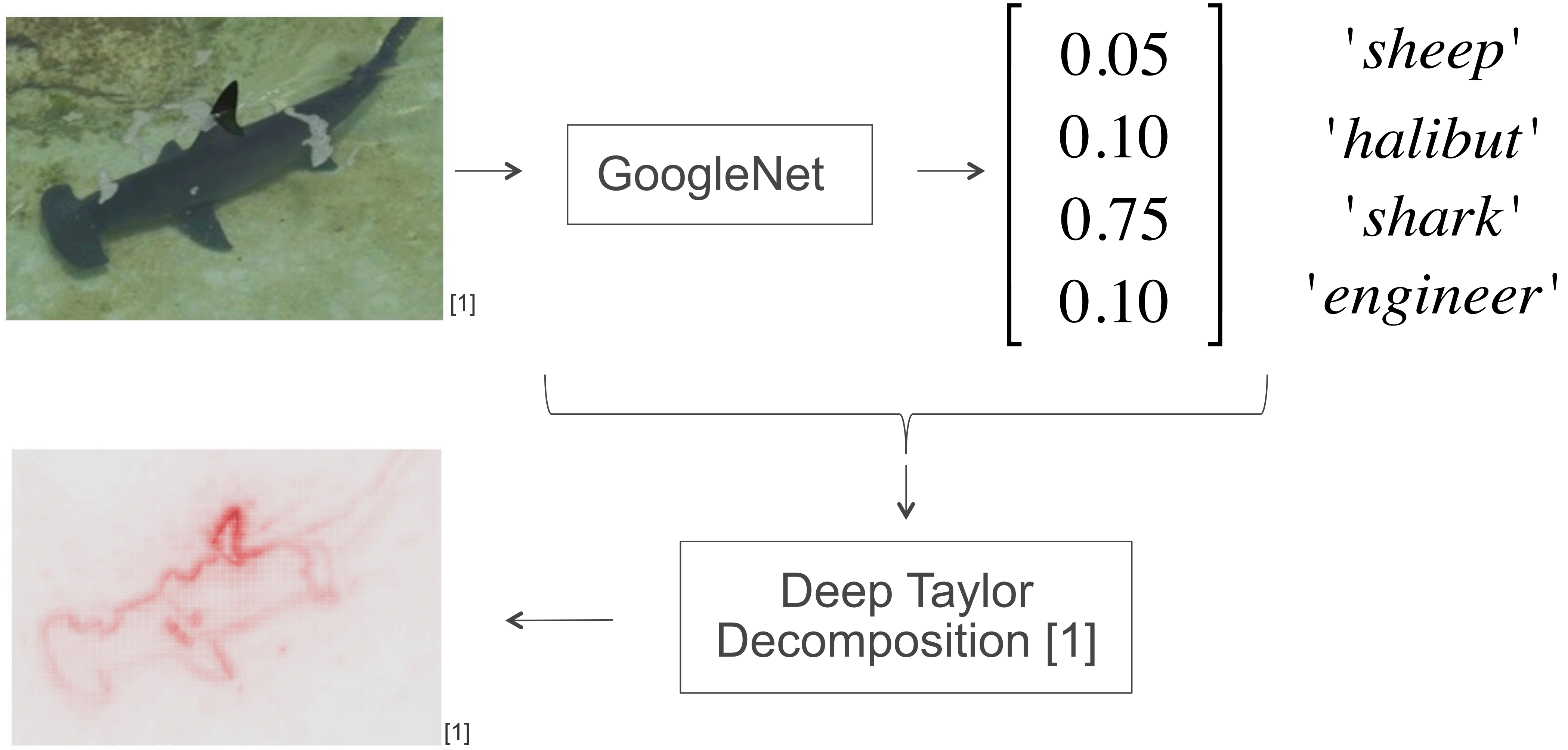
sample HRIR pair

spatialise 50 ms
white noise burst



~500k HRIR pairs from 583 subjects!

2) Apply an explanation technique



[1]

GoogleNet

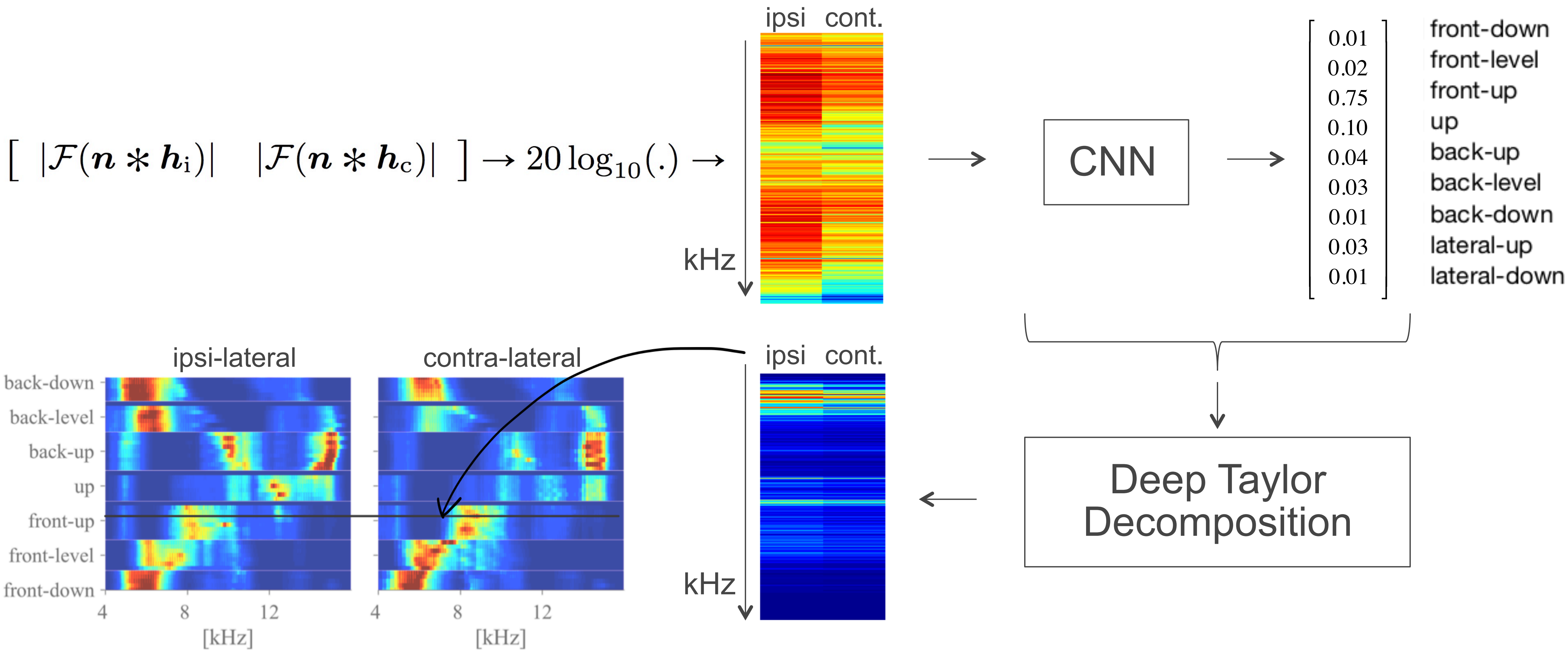
$$\begin{array}{cc} \begin{bmatrix} 0.05 \\ 0.10 \\ 0.75 \\ 0.10 \end{bmatrix} & \begin{array}{c} '\mathit{sheep}' \\ '\mathit{halibut}' \\ '\mathit{shark}' \\ '\mathit{engineer}' \end{array} \end{array}$$

Deep Taylor Decomposition [1]

[1]

[1] G. Montavon et al., “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

2) Apply an explanation technique



$$\left[\ |\mathcal{F}(\boldsymbol{n} * \boldsymbol{h}_{\mathrm{i}})| \quad |\mathcal{F}(\boldsymbol{n} * \boldsymbol{h}_{\mathrm{c}})|\ \right] \rightarrow 20\log_{10}(.) \rightarrow$$

CNN

0.01	front-down
0.02	front-level
0.75	front-up
0.10	up
0.04	back-up
0.03	back-level
0.01	back-down
0.03	lateral-up
0.01	lateral-down

Deep Taylor Decomposition

Models

Models

		input	**conv. 1**	conv. 2	conv. 3	**conv. 4**	filters/layer	stride	dense	output layer
WB	$>$ **300 Hz**	1005×2	**25×2**	11×1	11×1	**10×1**	4	2	216×9	soft-max
HP	$>$ **4 kHz**	1005×2	**25×1**	11×1	11×1	**10×2**	4	2	216×9	soft-max

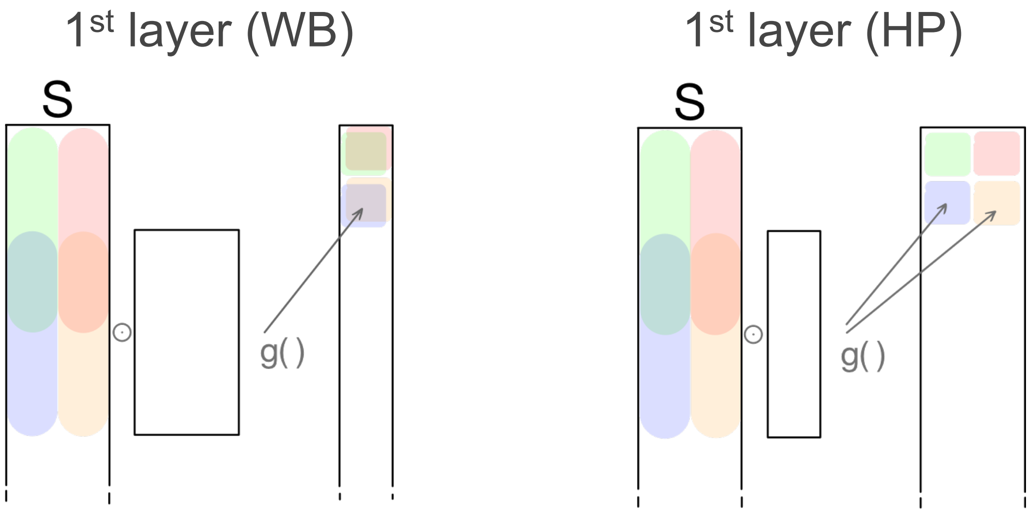
< 3k trained parameters

> 500 training points parameter

~ 500 subjects in train set

~ 30% classification error (test)

	CE [%]	RMSE [deg]	MAE [deg]	r
random	91.3	74.5	59.5	0.65
[15]	-	25.2	-	0.85
[27]	-	-	22.3	0.82
[28]	-	-	≈25	-
WB	45.1	43.2	16.5	0.90


1st layer (WB)

1st layer (HP)

⊙ element-wise multiply

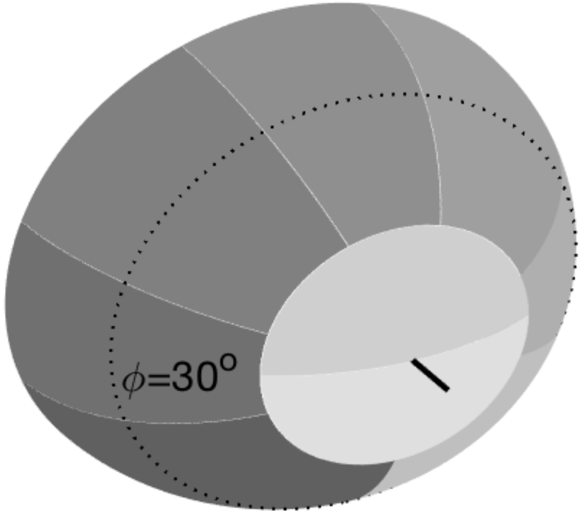
g( ) sum and rectify

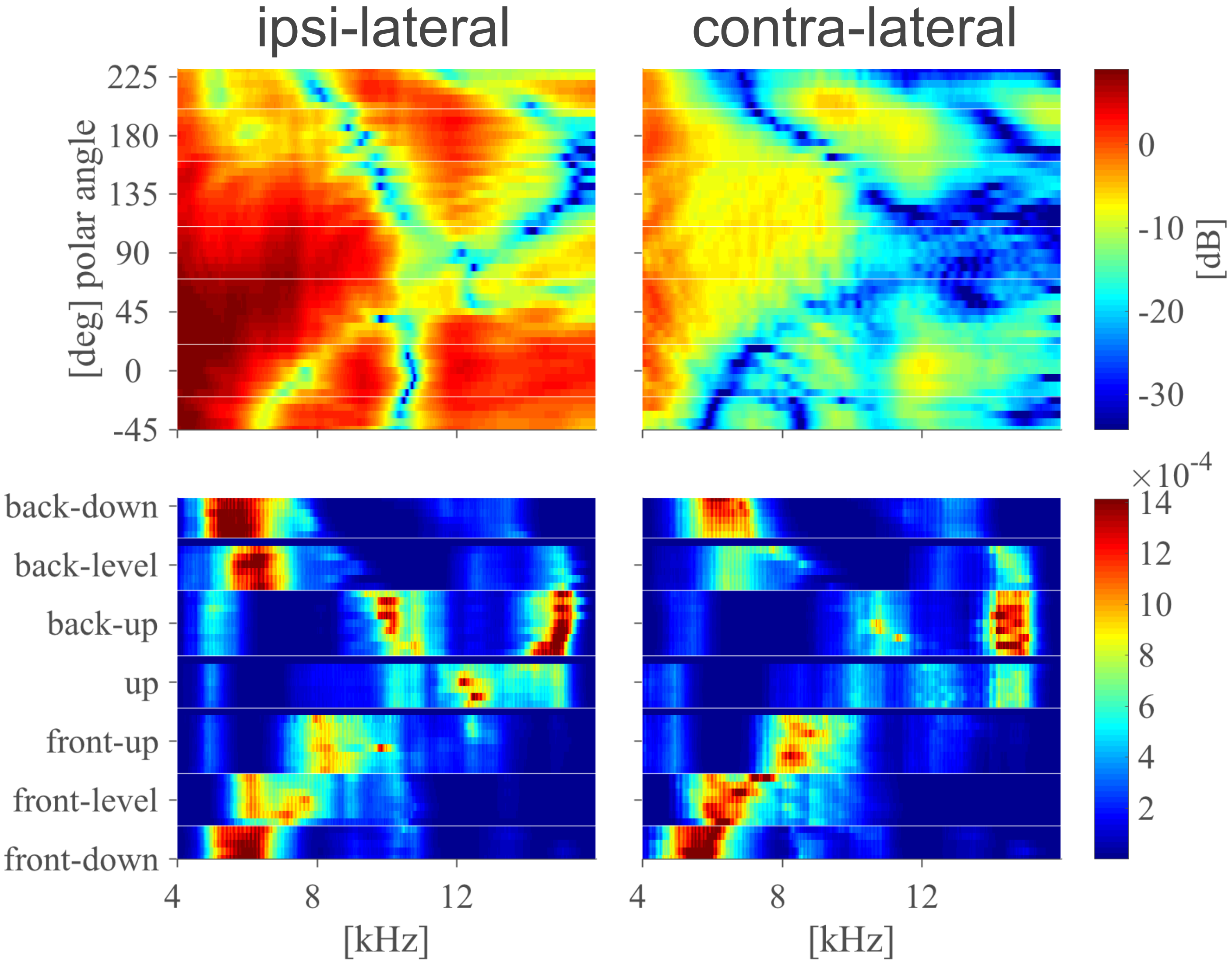
Results

HP model

Cone of confusion at 30^{o} lat.

CIPIC's subject 154

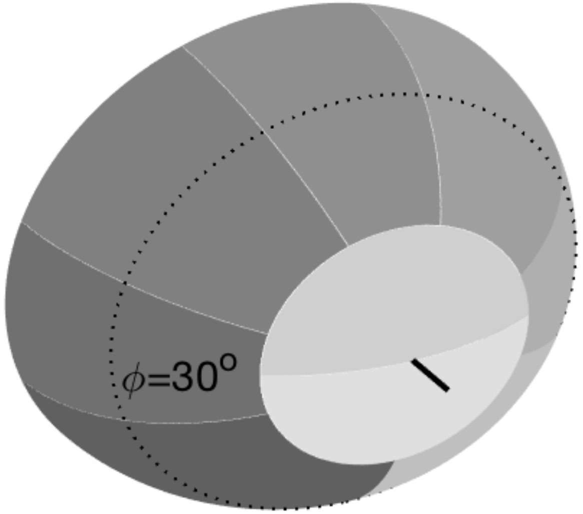


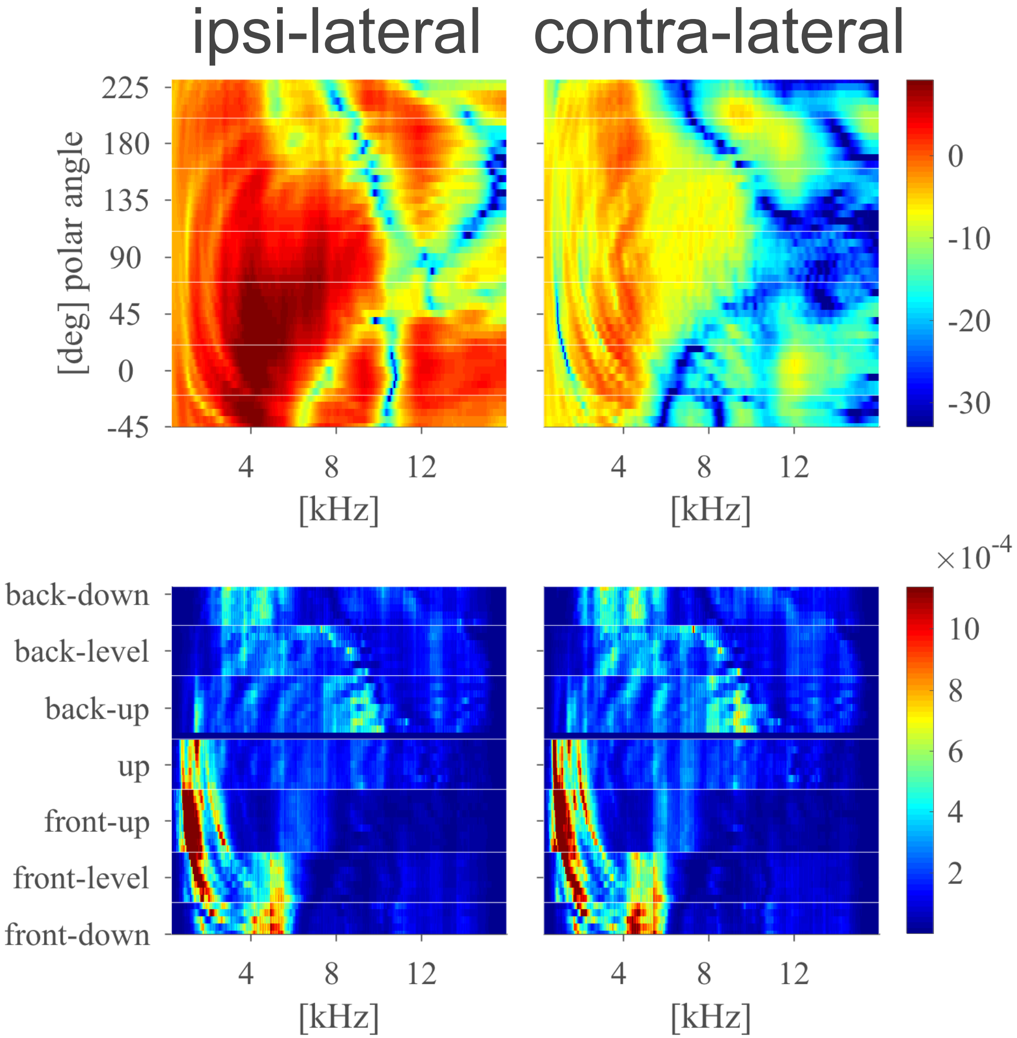


WB model

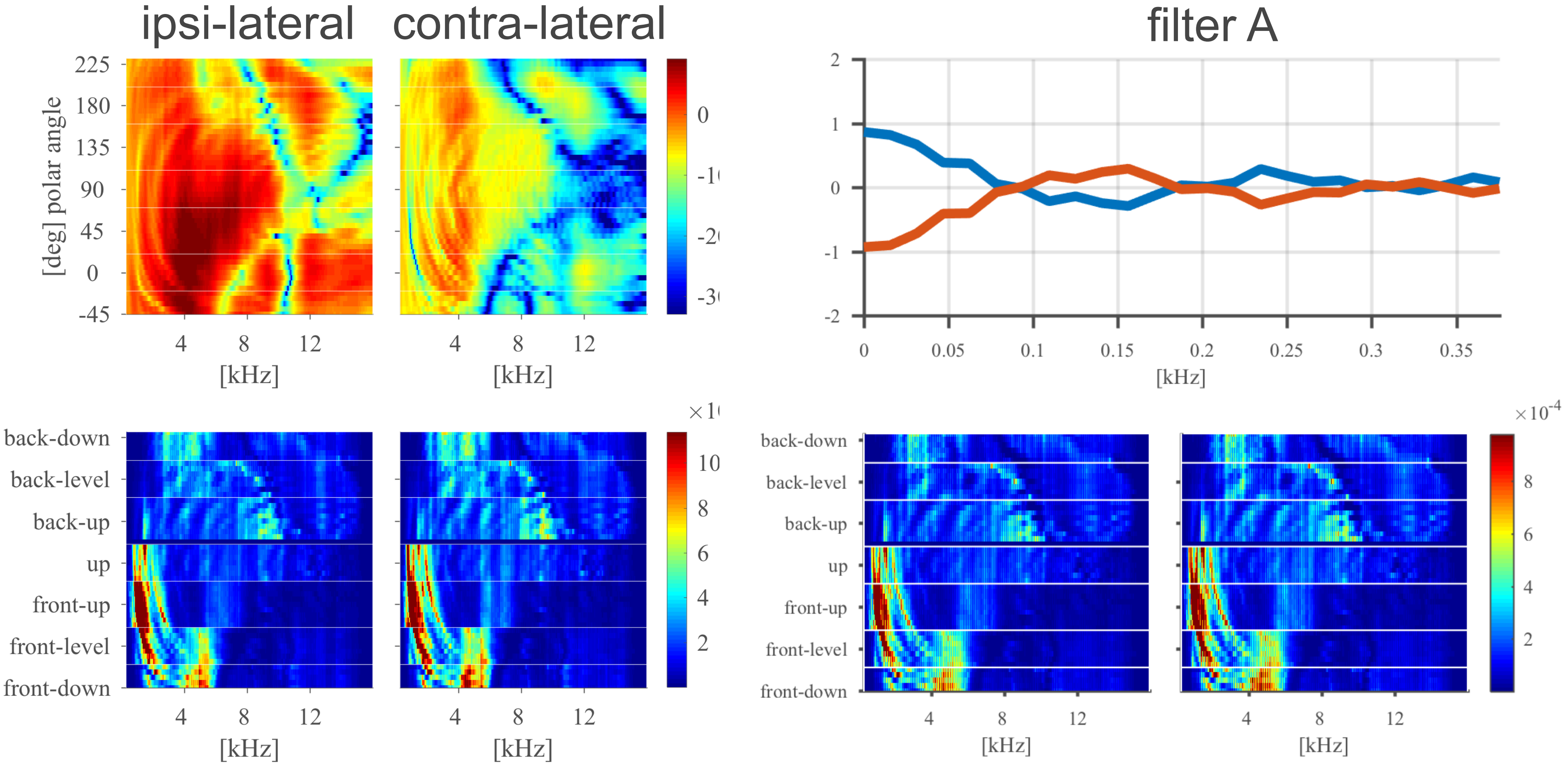
Cone of confusion at 30^o lat.

CIPIC's subject 154





WB model

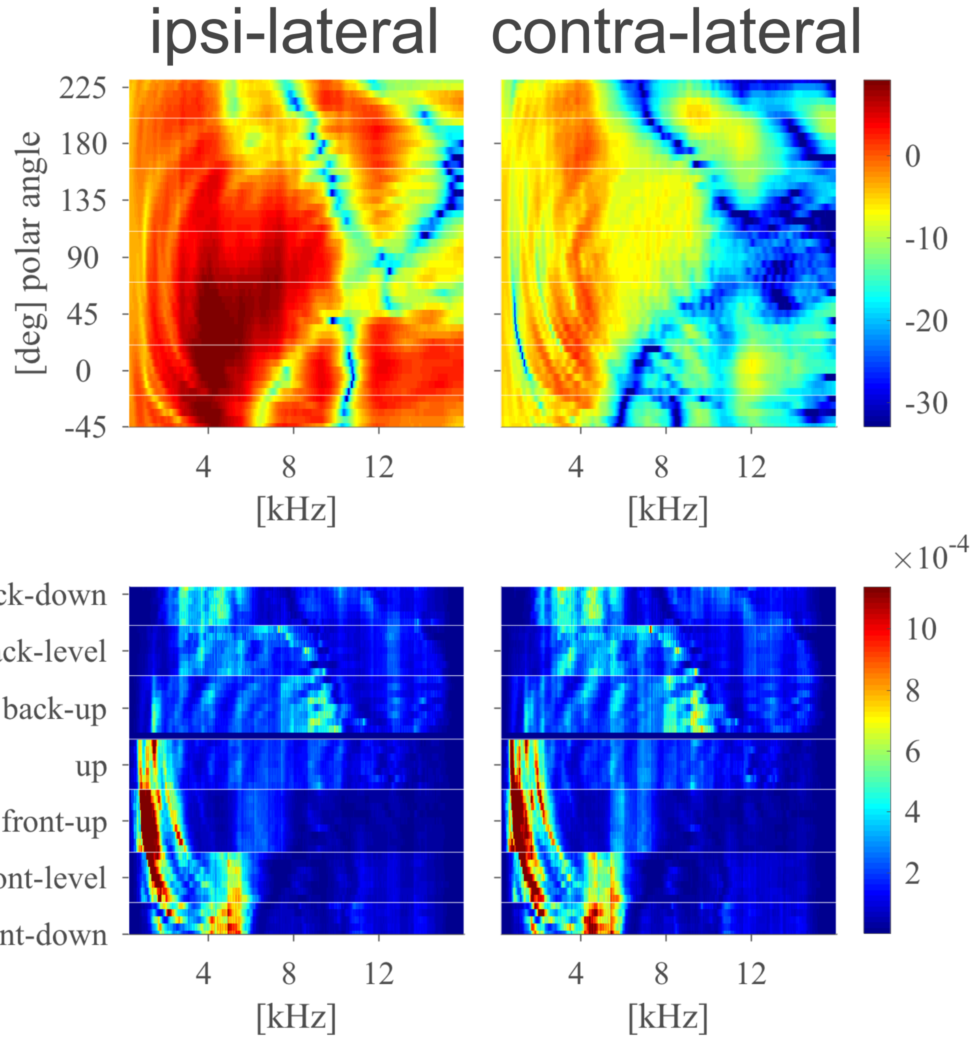


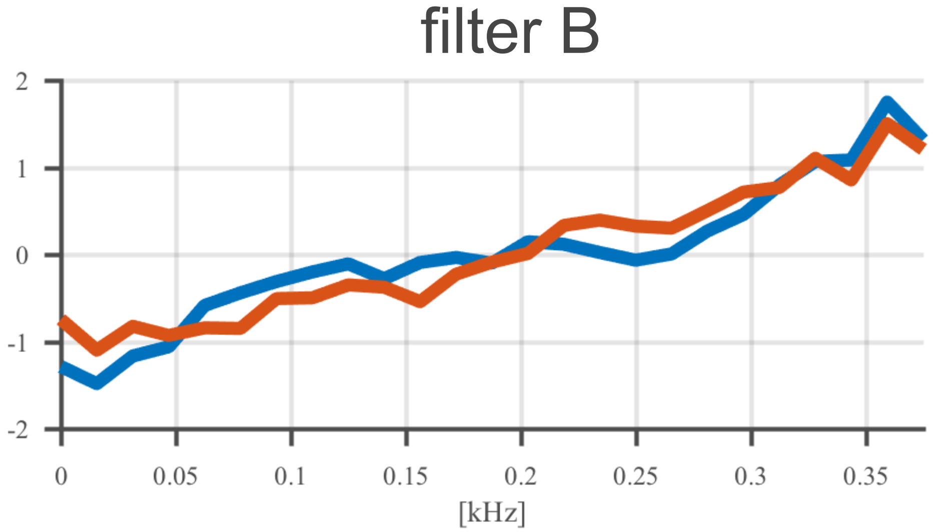
ipsi-lateral

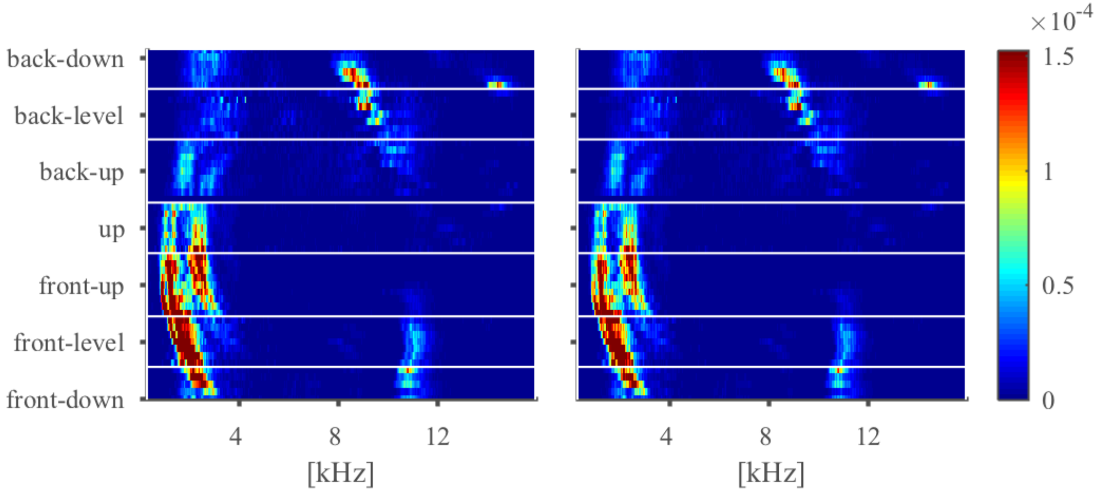
contra-lateral

filter A

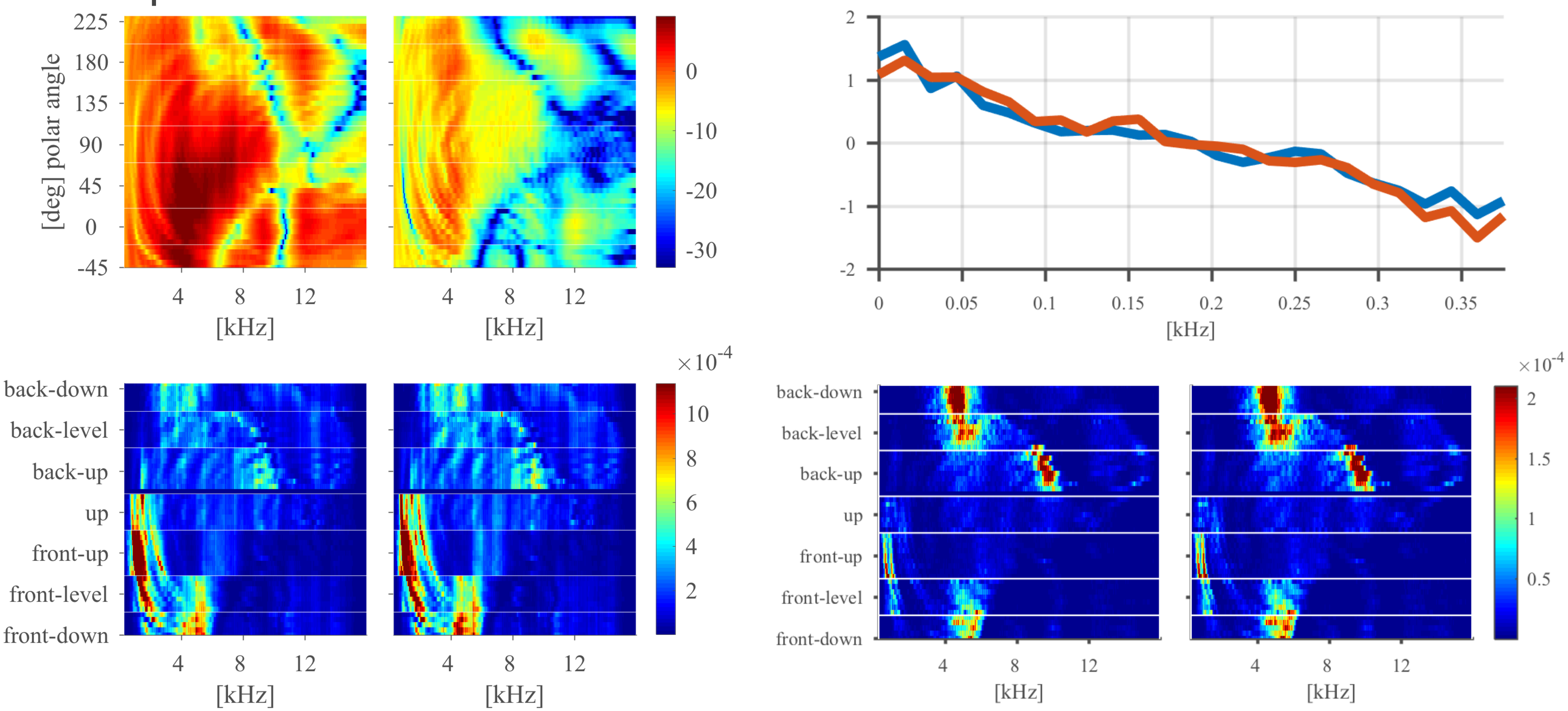
WB model


ipsi-lateral
contra-lateral
225
180
135
90
45
0
-45
[deg] polar angle
4
8
12
[kHz]
0
-10
-20
-30
back-down
back-level
back-up
up
front-up
front-level
front-down
$\times 10^{-4}$
10
8
6
4
2


filter B
2
1
0
-1
-2
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
[kHz]


back-down
back-level
back-up
up
front-up
front-level
front-down
4
8
12
[kHz]
$\times 10^{-4}$
1.5
1
0.5
0

WB model



ipsi-lateral contra-lateral

filter C

Conclusion

Summary and conclusion

- (Spatial) audio feature discovery using neural network explanation techniques

- Rudimentary models seem to learn cues similar to ones reported in listening test experiments

- Deep Taylor Decomposition seems useful for discovering/visualizing audio features

- Results to be confirmed using a perceptually-grounded front-end

- Lots of future work…!

Thank you!