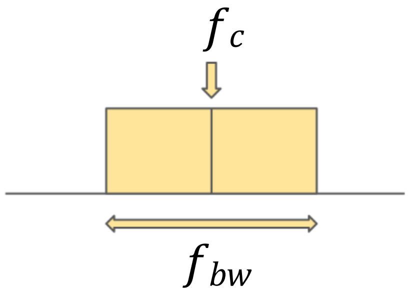
Low-Level Physiological Implications of End-to-End Learning of Speech Recognition

Louise Coppieters

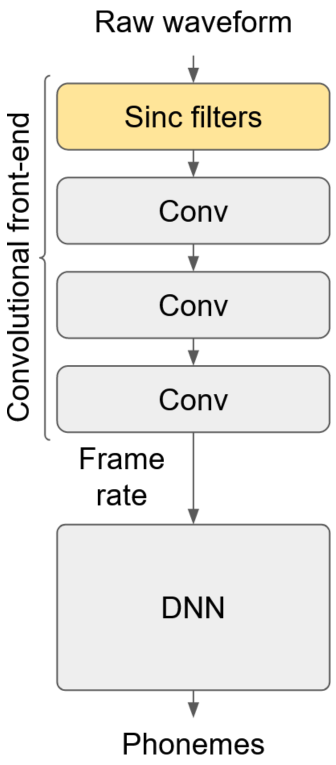
SincNet Model

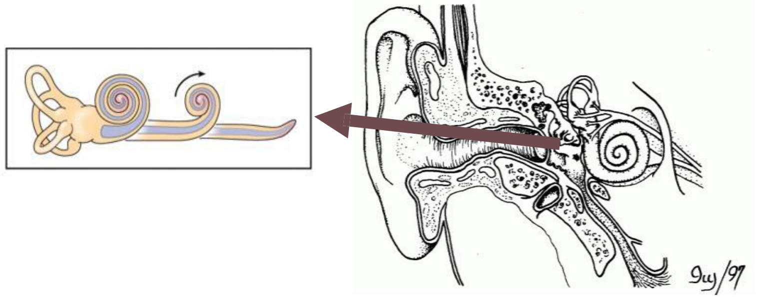


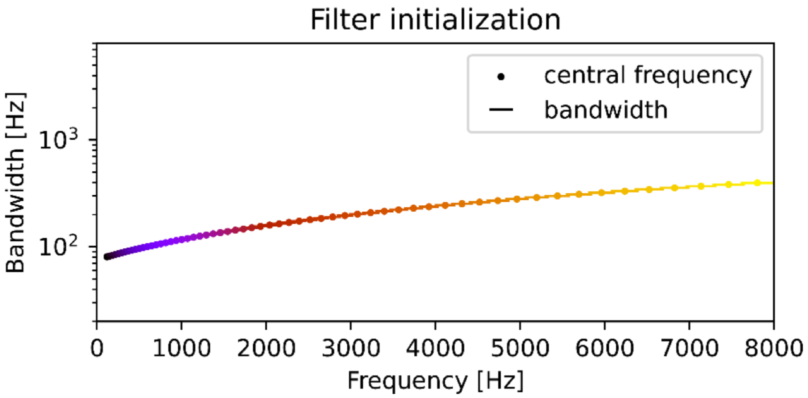
- Input: raw waveform
- 4-layer CNN: The first layer is made of filters defined by two trainable parameters:

$$h[n] = sinc\ (2\pi f2n) - sinc(2\pi f1n)$$

- 5-layer DNN

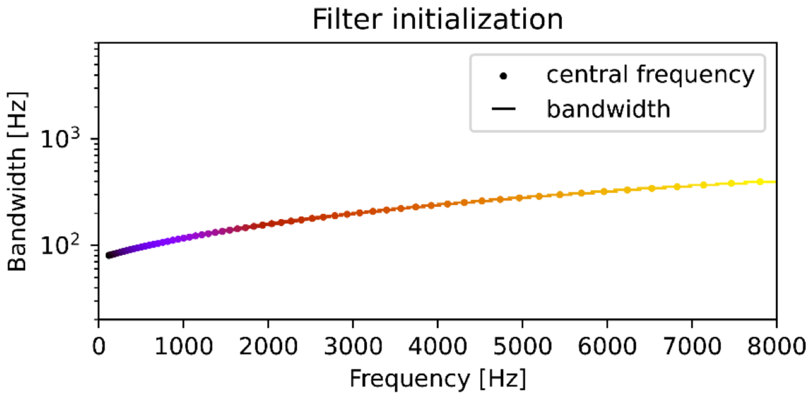


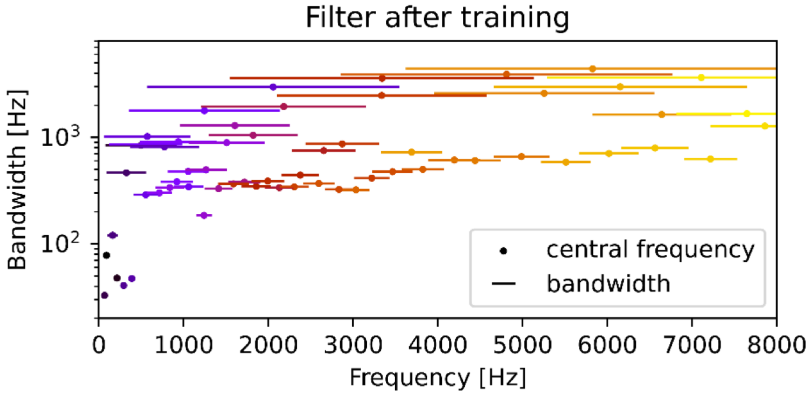




Coppieters de Gibson, L., & Garner, P. N. (2022). Low-Level Physiological Implications of End-to-End Learning of Speech Recognition.

Model characteristic results



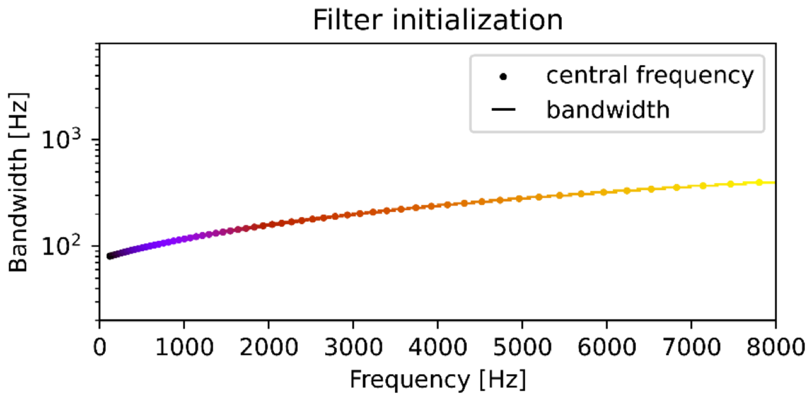


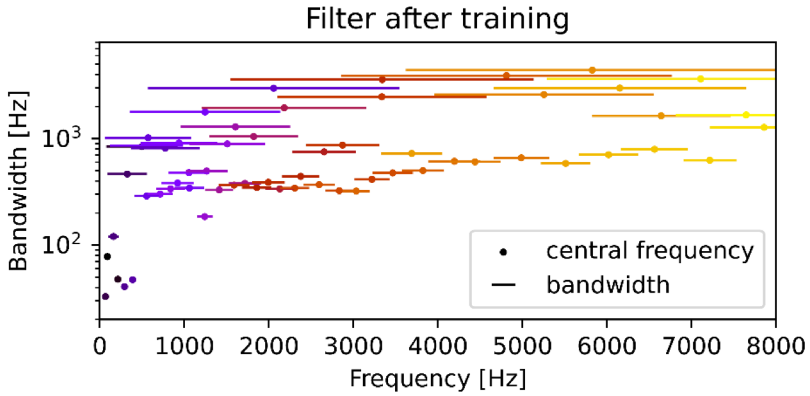
Sinc-Layer Num. filters	CNN-layers	Narrow band filters	PER [%]
128	60 – 60 – 60	39	17.1
100	60 – 60 – 60	45	17.1
80	60 – 60 – 60	38	17.2
60	60 – 60 – 60	32	17.4
40	60 – 60 – 60	27	17.5
30	60 – 60 – 60	24	17.5

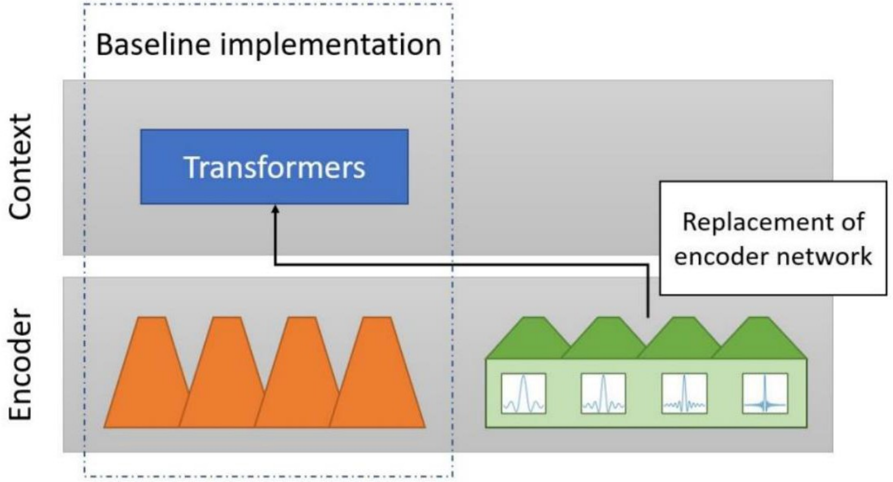
Initialized to	Compared to			$\cdot 10^{-3}$
Scale – filters	Mel	Bark	ERB	Greenwood
Mel – 128	2.3	4.7	7.0	8.6
Mel – 60	1.8	4.4	7.0	8.8
Mel – 40	2.2	3.9	6.5	8.4
Mel – 30	2.0	4.3	7.1	9.1
Bark – 30	2.5	3.7	6.2	8.2
ERB – 30	3.0	2.9	5.5	7.6
Greenwood - 30	3.7	6.8	9.5	11.6

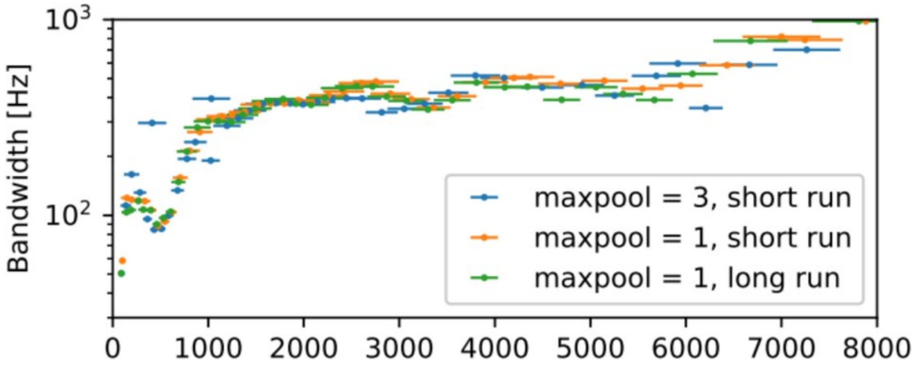
Coppieters de Gibson, L., & Garner, P. N. (2022). Low-Level Physiological Implications of End-to-End Learning of Speech Recognition.

SincNet integrated and trained within wav2vec2


Filter initialization


Filter after training





Coppieters de Gibson, L., & Garner, P. N. (2022). Low-Level Physiological Implications of End-to-End Learning of Speech Recognition.