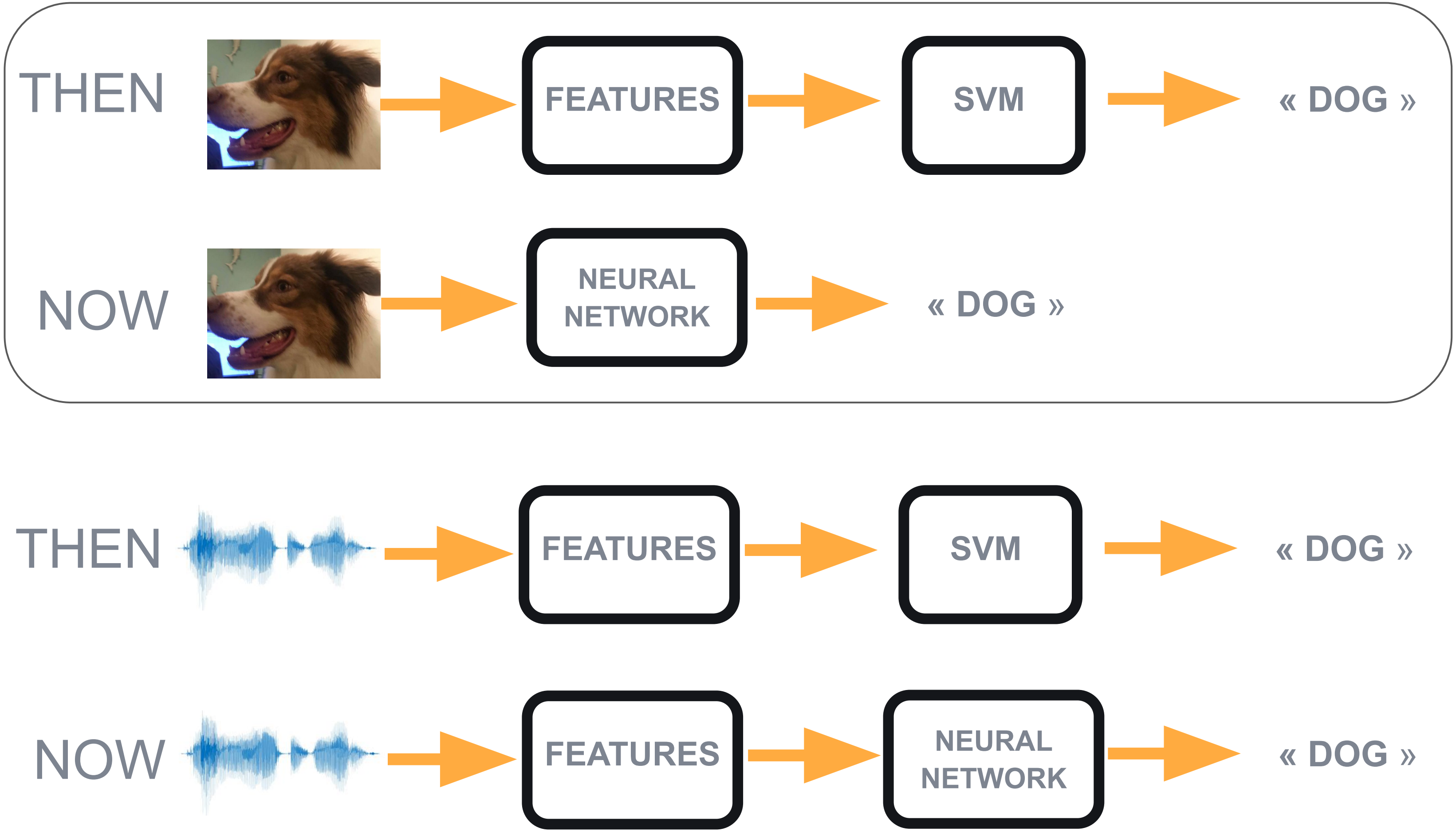
Google AI
Brain Team

LEAF: A Learnable Frontend for Audio Classification

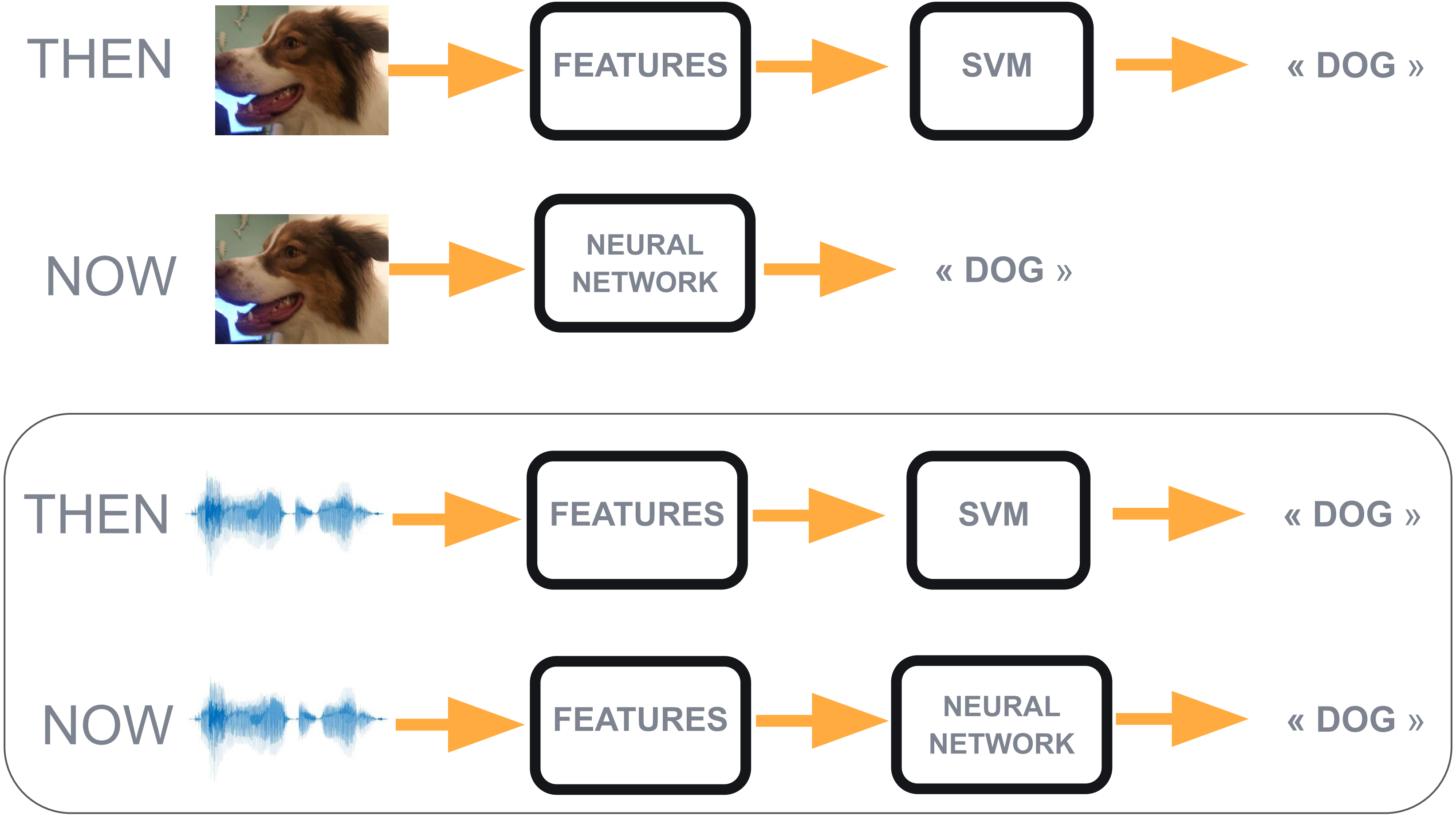
Neil Zeghidour, Olivier Teboul, Félix de Chaumont-Quitry, Marco Tagliasacchi
Google Research

Computer vision vs audio classification



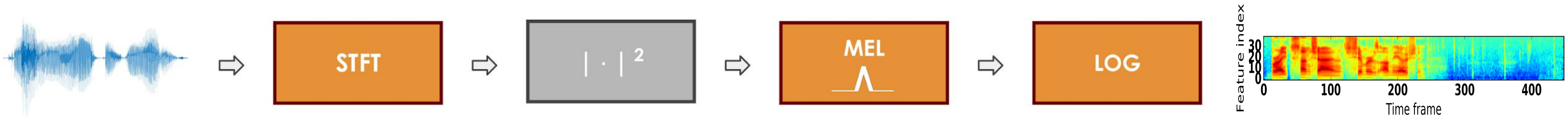
Computer vision vs audio classification

Google AI
Brain Team



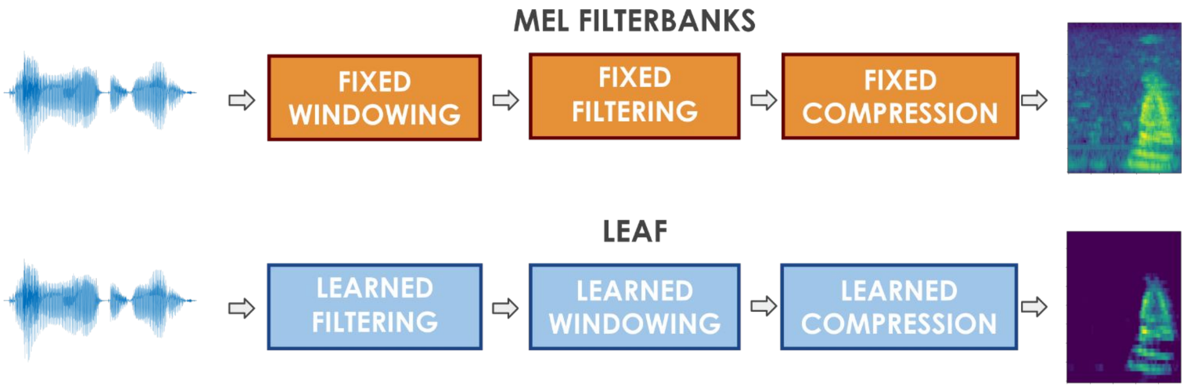
Standard audio features: mel-filterbanks

- Typical features are mel-filterbanks, that replicate human perception:
  - Compute a spectrogram
  - Pass it through mel(odic) filters (log sensitivity to pitch)
  - Pass it through a logarithmic compression (log sensitivity to loudness)



- Limitations:
  - Many banks of filters, compressions have been proposed
  - Not clear when matching human perception is good
- Solution:
  - Test several combinations with trial and error
  - Let the neural network learn all these operations

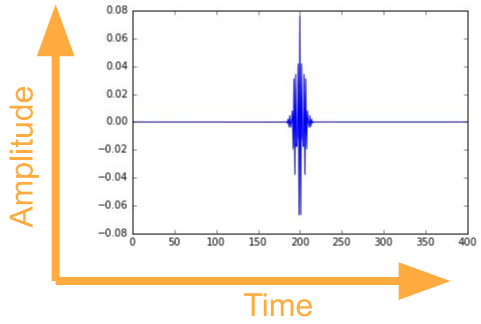
LEAF: A LEarnable Audio Frontend



Learning filtering: Gabor 1D-convolution

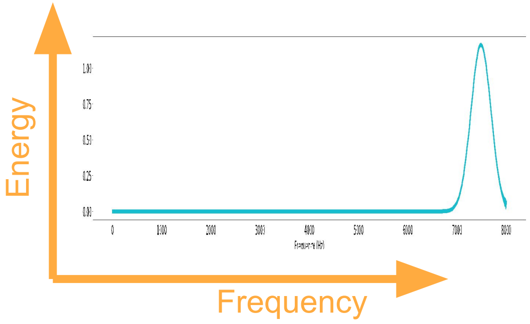
- A Gabor filter has the following expression in time domain:

$$\varphi_n(t) = e^{i2\pi\eta_n t}\frac{1}{\sqrt{2\pi}\sigma_n}e^{-\frac{t^2}{2\sigma_n^2}}$$



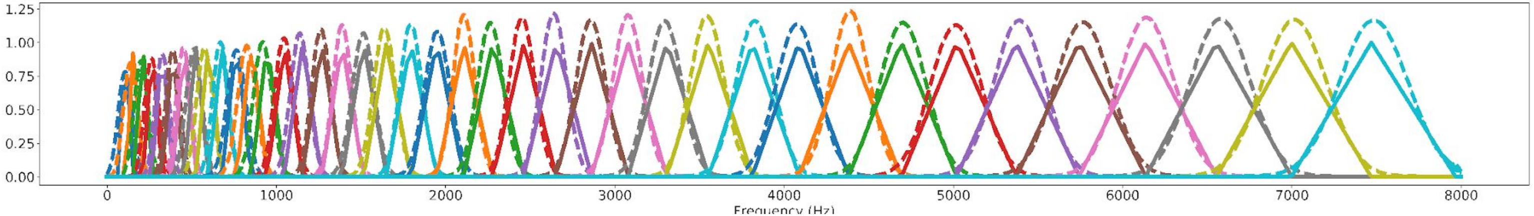
- And in frequency domain:

$$\hat{\varphi}_n(\xi) \propto \sqrt{\sigma_n}e^{-\frac{1}{2}\sigma_n^2(\xi-\eta_n)^2}$$

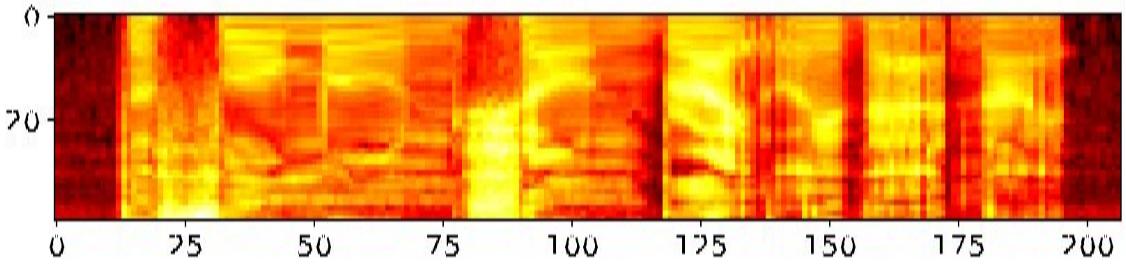


- We learn its center and bandwidth

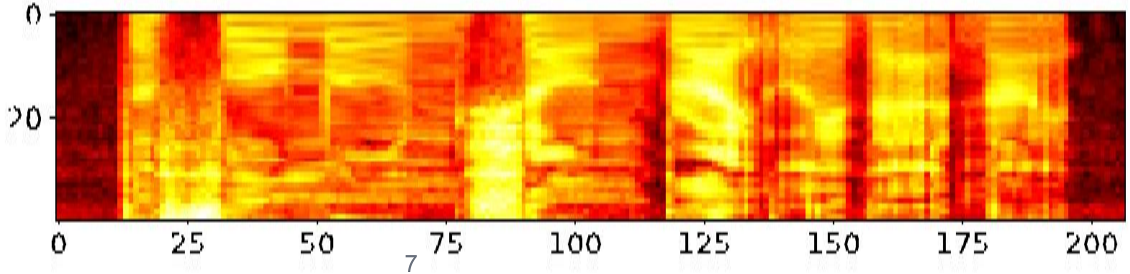
We can approximate the mel scale and then learn a new scale



Mel-filterbanks

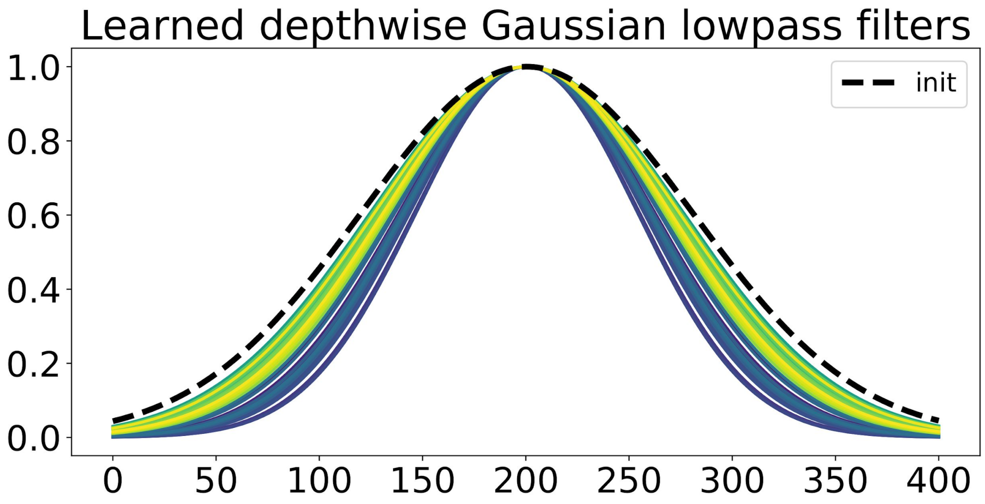


Gabor 1D-convolution
(initialization)



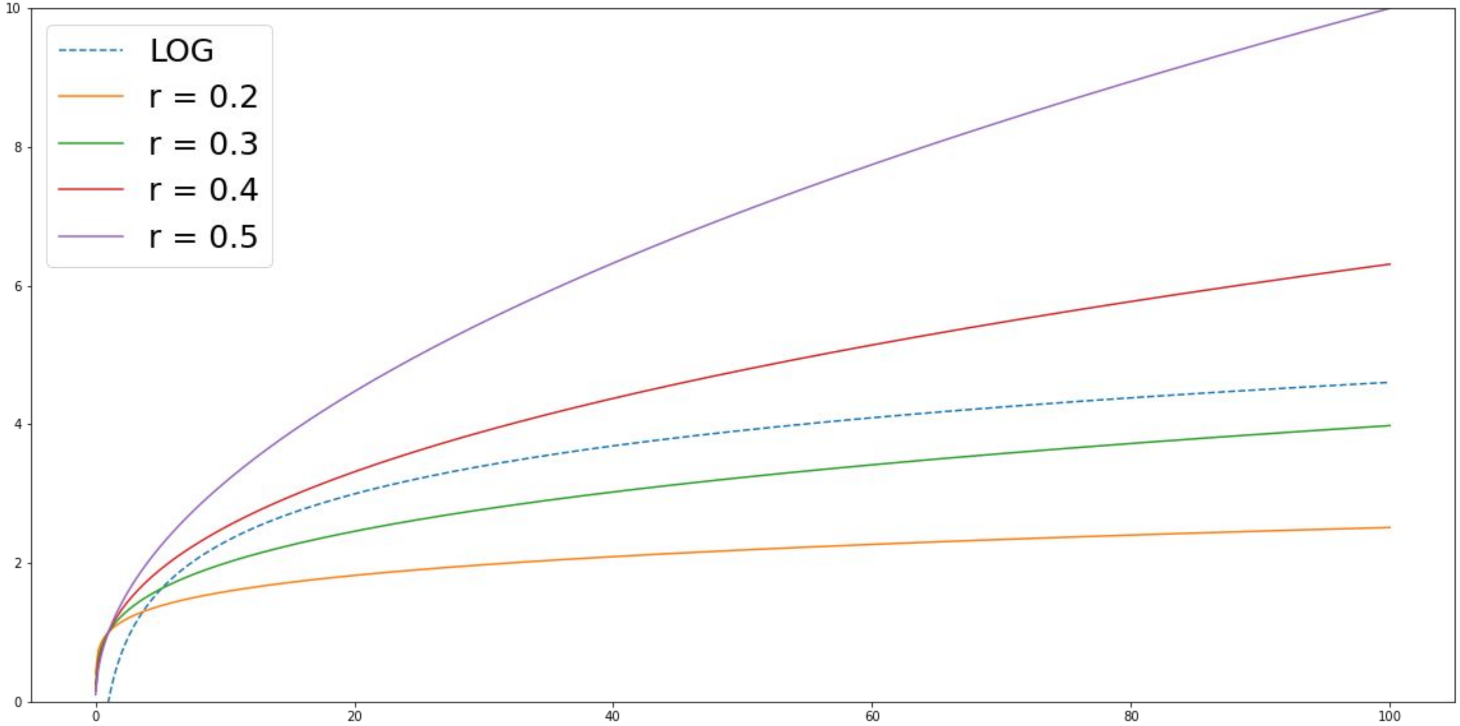
Windowing: Channelwise Gaussian Pooling

- We learn the width of the windowing function
- The wider the window, the more we remove high frequencies



Learning the compression

- Instead of a logarithm we can learn the "r" in x^(1/r)
- We can learn to compress more or less than a logarithm



Classification performance

- Train a model to recognize many kinds of sounds at once
- Metric: accuracy (% of time we predict the right category)
- LEAF outperforms on average
- But mel filterbanks are a very strong baseline

Table 3: Test accuracy (%) for multi-task classification.

Task	Mel	TD-fbanks	SincNet	LEAF
Acoustic scenes	$\mathbf{99.1} \pm 0.5$	98.3 ± 0.6	91.0 ± 1.4	98.9 ± 0.5
Birdsong detection	81.3 ± 0.9	$\mathbf{82.3} \pm 0.9$	78.8 ± 0.9	81.9 ± 0.9
Emotion recognition	24.1 ± 2.1	24.4 ± 2.1	26.2 ± 2.1	$\mathbf{31.9} \pm 2.3$
Speaker Id. (LBS)	$\mathbf{100.0} \pm 0.0$	$\mathbf{100.0} \pm 0.0$	$\mathbf{100.0} \pm 0.0$	$\mathbf{100.0} \pm 0.0$
Music (instrument)	$\mathbf{70.7} \pm 0.6$	66.3 ± 0.6	67.4 ± 0.6	70.2 ± 0.6
Music (pitch)	88.5 ± 0.4	86.4 ± 0.4	81.2 ± 0.5	$\mathbf{88.6} \pm 0.4$
Speech commands	$\mathbf{93.6} \pm 0.3$	89.5 ± 0.4	91.4 ± 0.4	$\mathbf{93.6} \pm 0.3$
Language Id.	64.9 ± 0.5	58.9 ± 0.5	60.8 ± 0.5	$\mathbf{69.6} \pm 0.5$
Average	77.8 ± 0.7	75.8 ± 0.7	74.6 ± 0.8	$\mathbf{79.3} \pm 0.7$

Large scale audio classification

- AudioSet = 1M audio sequences from 527 classes
- We report AUC and d-prime averaged over classes (multi-label classification)

Table 4: Test AUC and d-prime on Audioset, with the number of learnable parameters per frontend.

		EfficientNetB0		CNN14 (ours)		CNN14 (Kong et al., 2019)	
Frontend	#Params	AUC	d-prime	AUC	d-prime	AUC	d-prime
Mel	0	0.966	2.58	0.973	2.72	0.973	2.73
Mel-PCEN	256	0.966	2.58	0.973	2.72	-	-
Wavegram	300 k	0.950	2.34	0.962	2.51	0.968	2.61
TD-fbanks	51 k	0.962	2.50	0.972	2.70	-	-
SincNet	256	0.959	2.47	0.970	2.66	-	-
SincNet+	448	0.966	2.58	0.973	2.72	-	-
LEAF	448	0.969	2.63	**0.974**	**2.74**	-	-

Meet us at our poster!

- Poster session 1
- May 3rd, 2021 1am-3am (PDT)
- We released our code on github:
  https://github.com/google-research/leaf-audio/tree/master/leaf_audio