tinyML® Research Symposium

Enabling Ultra-low Power Machine Learning at the Edge

March 27, 2023

www.tinyML.org

How tiny can analog filterbank features be made for ultra-low-power on-device keyword spotting?

Subhajit Ray, Xinghua Sun,
Nolan Tremelling, Maria Gordiyenko,
Peter Kinget

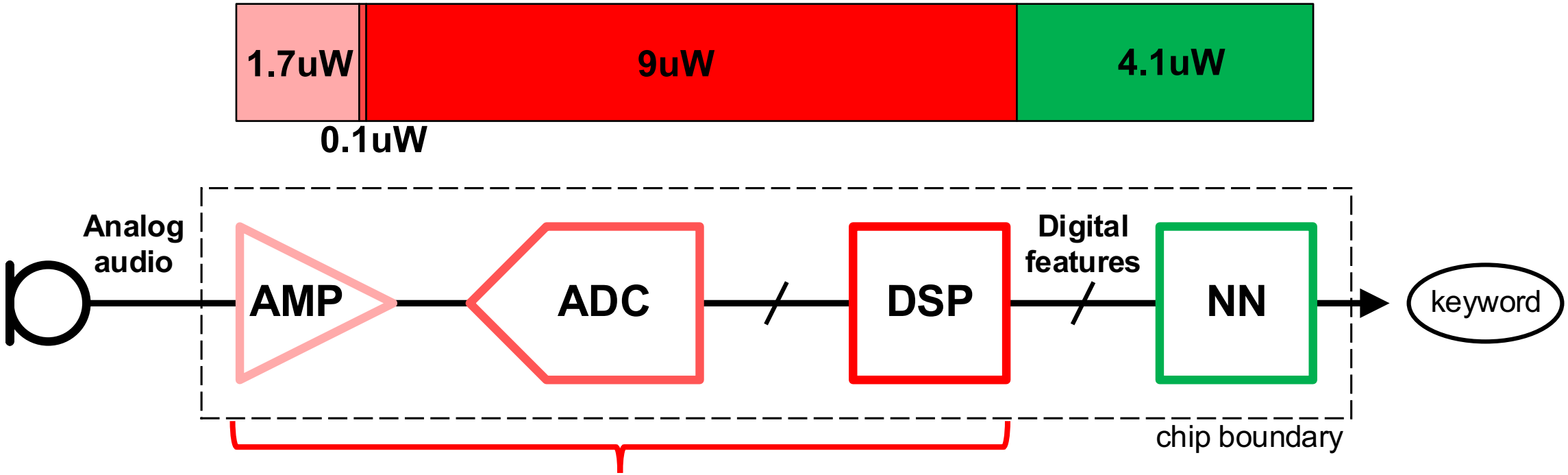
Columbia University, New York, United States

3/27/23

An example of a state-of-the-art fully-integrated keyword-spotting chip

[JSSC'20_GiraldoVerhelst]: 91% accuracy, 10 words → 16.1uW…

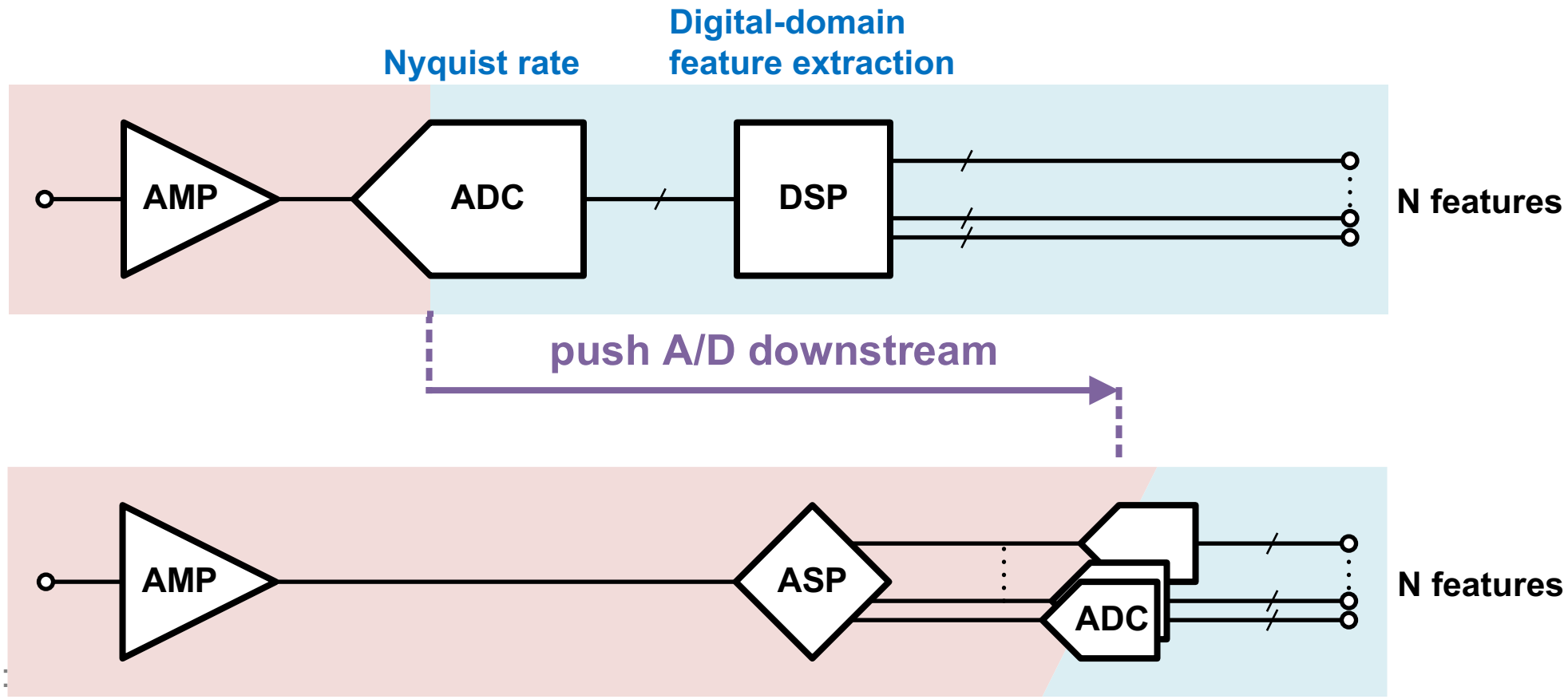
- …but a decade-long lifetime on a coin cell battery* would require <1uW



Frontend feature extractor is the bottleneck, relative to the backend classifier

* SR927: 60mAh, 1.55V, 9.5mm x 2.7mm
** Bar chart excludes the 1.2uW of the "sound detector" block (including its leakage)

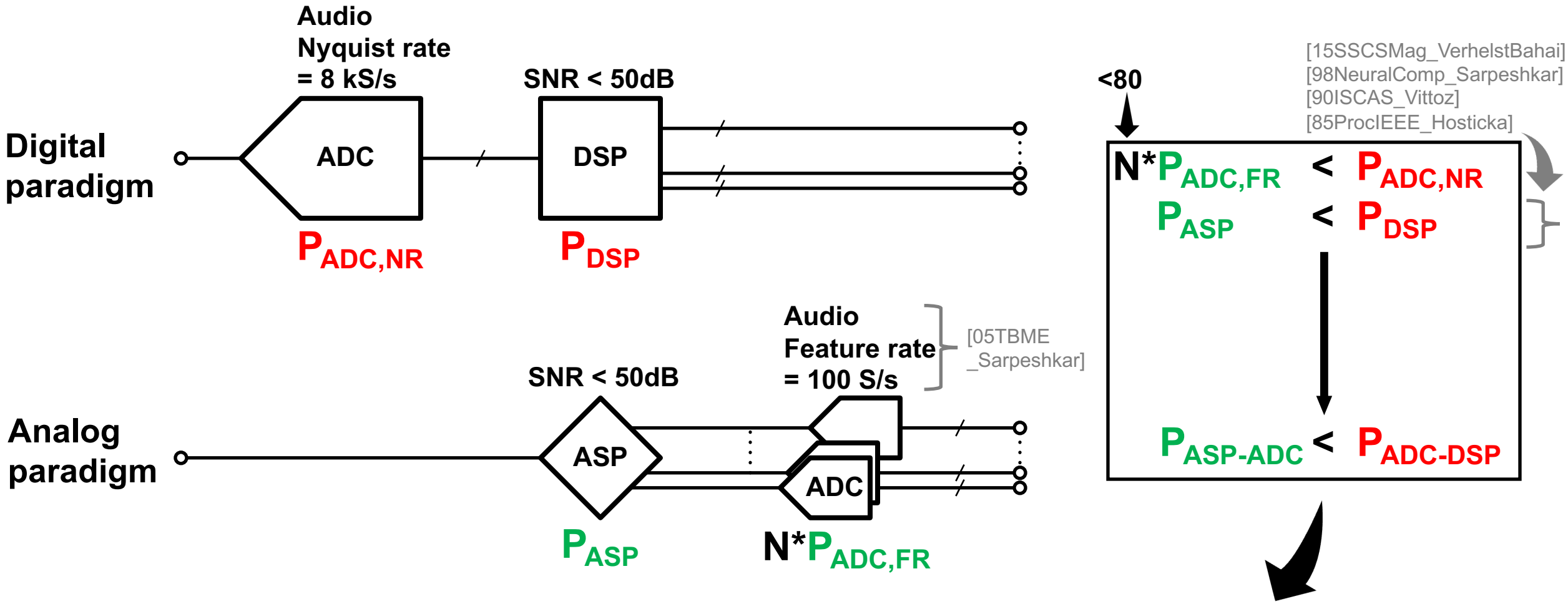
Frontend feature extractor: digital vs analog paradigms



Early examples for audio:
[83JSSC_BuiMichel]
[05TBME_Sarpeshkar]

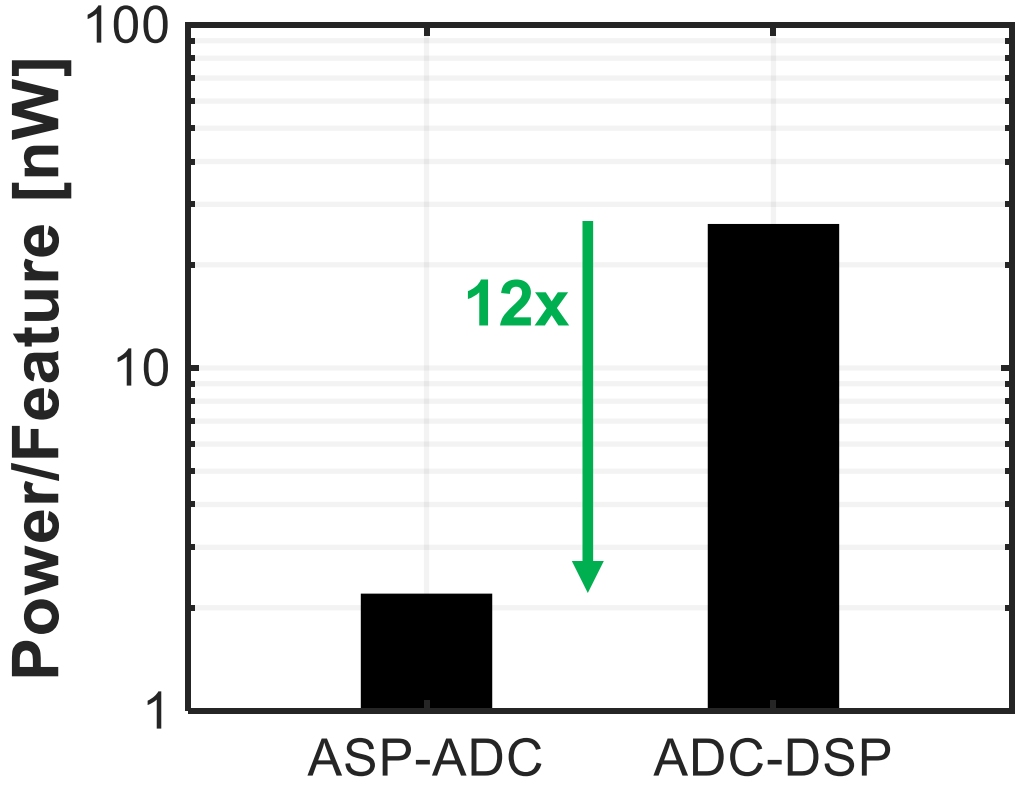
[15SSCSMag_VerhelstBahai]

Digital vs analog paradigms: *in principle*



Analog *should be* more power-efficient than digital

Digital vs analog paradigms: *in practice, for audio feature extractors*



[JSSC'21_YangSeok]

ADC: [VLSI'18_BadamiVerhelst]
DSP: [ESSCIRC'21_ZhuLu]

Analog *is indeed* more power-efficient than digital

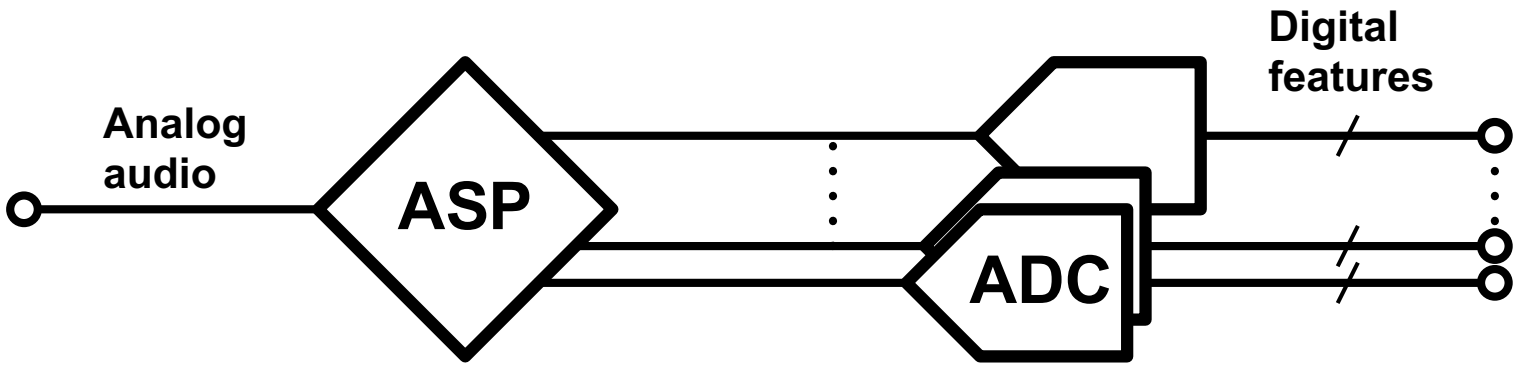
Analog audio feature extraction can be more power-efficient than *digital*, but…

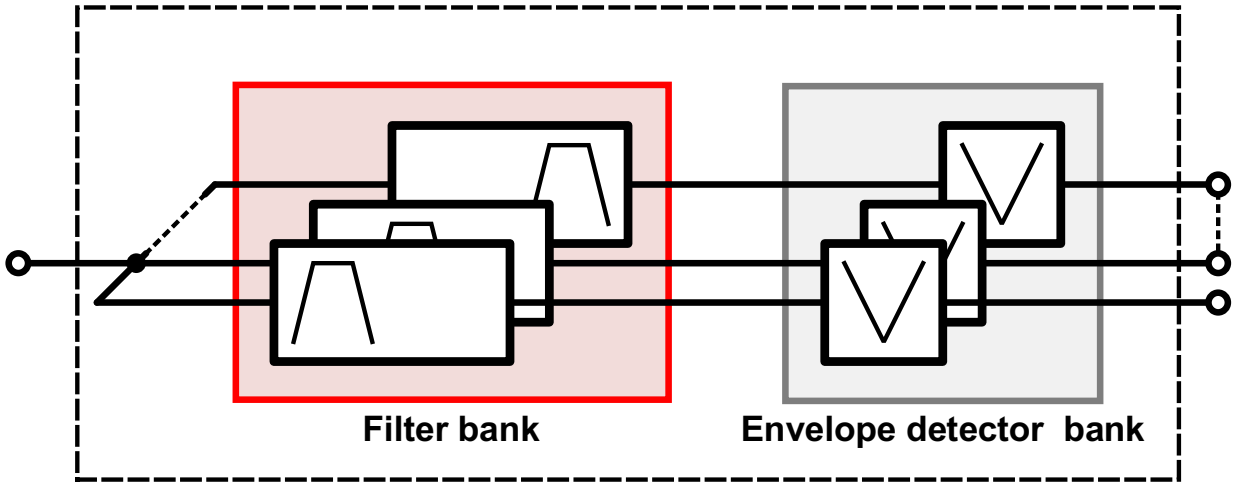
…it does suffer, significantly, from variability;
however, jury is still out as to whether this is important.

General architecture of analog audio feature extractor

Filterbank parameters

- **$N_{filters}$**
- f_{min}
- **f_{max}**
- Dist. of f_c's
- **Q**
- Dist. of Q's
- Filter order

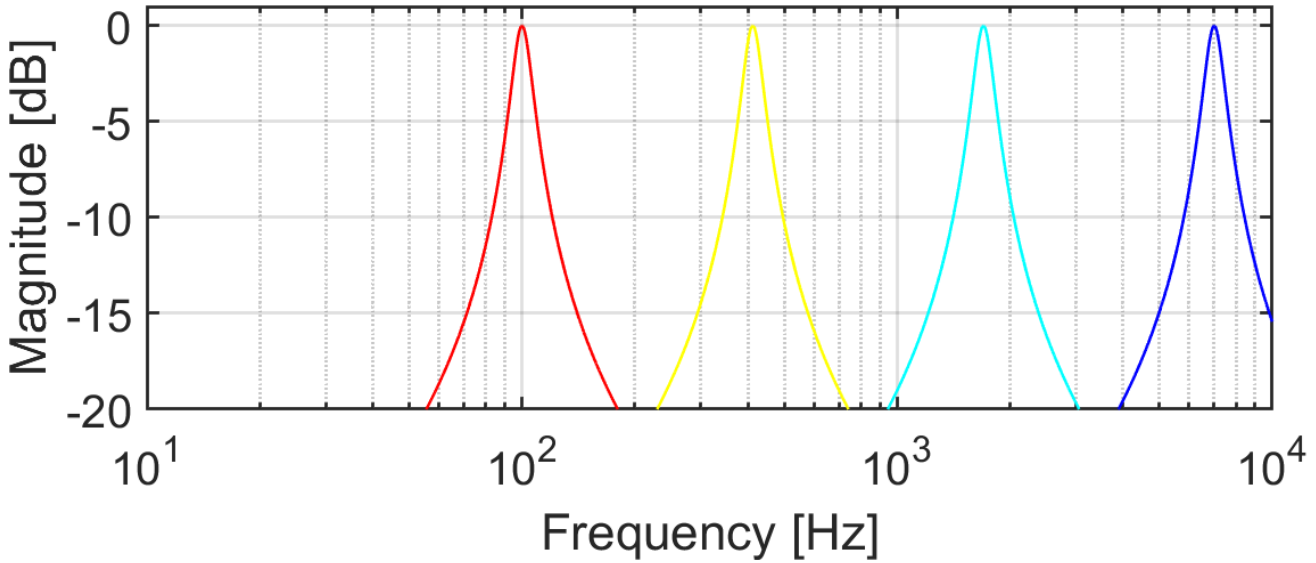




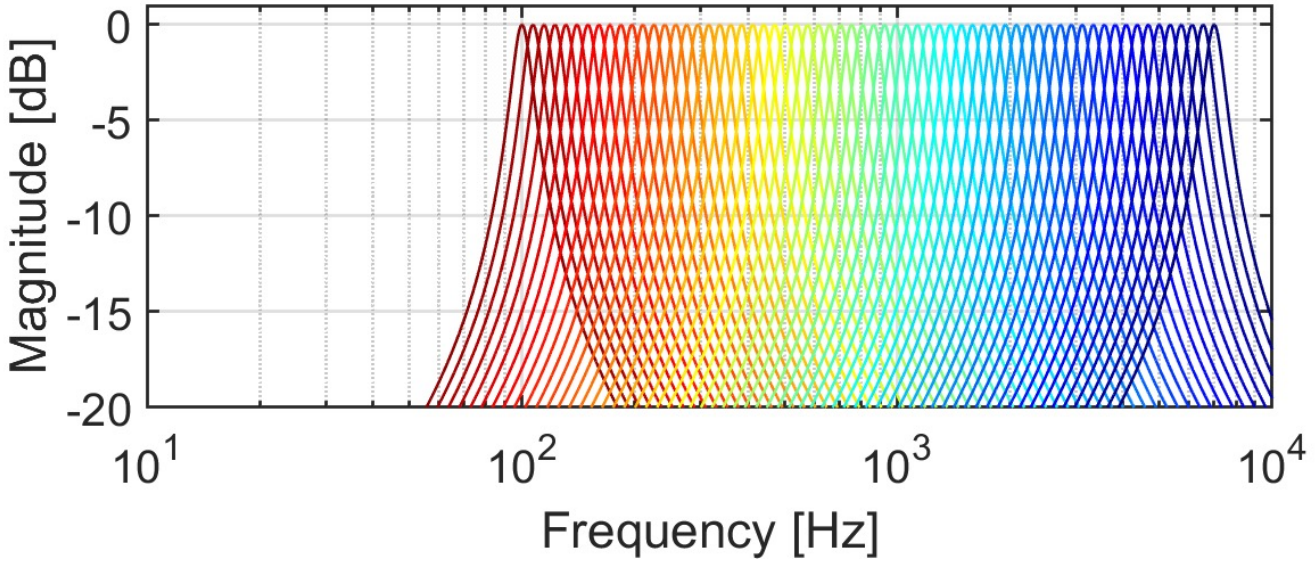
The filterbank is a critical block

Architectural parameters of filterbank: $N_{filters}$

Small: 4



Large: 64



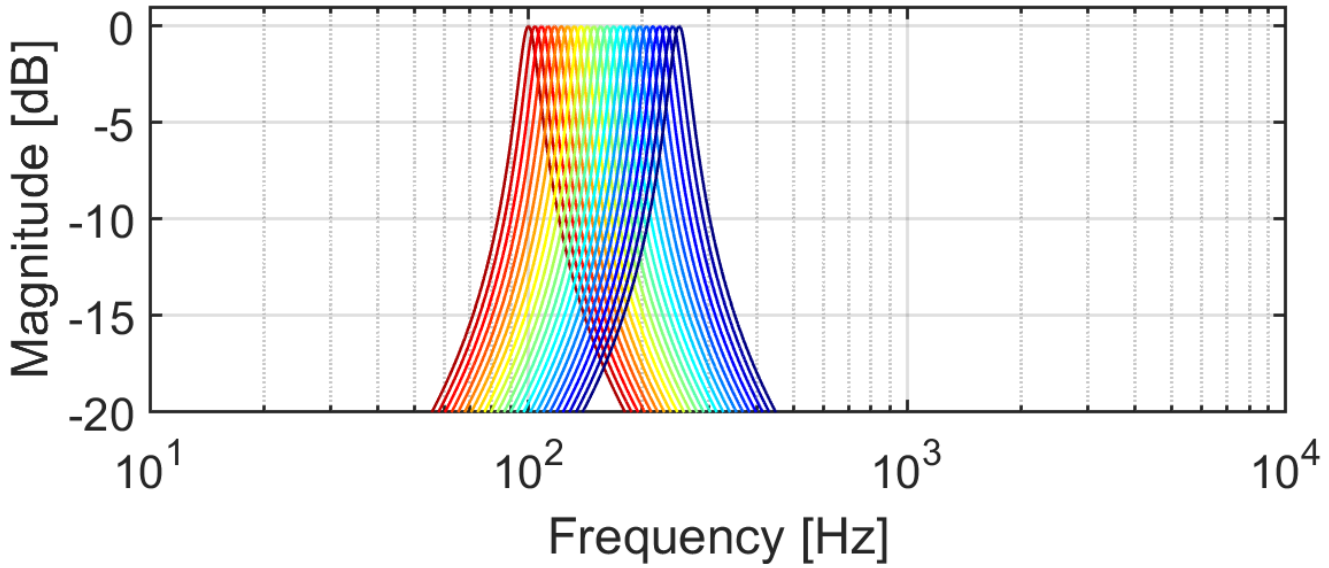
$P_{filterbank} \propto N_{filters}$

to first order

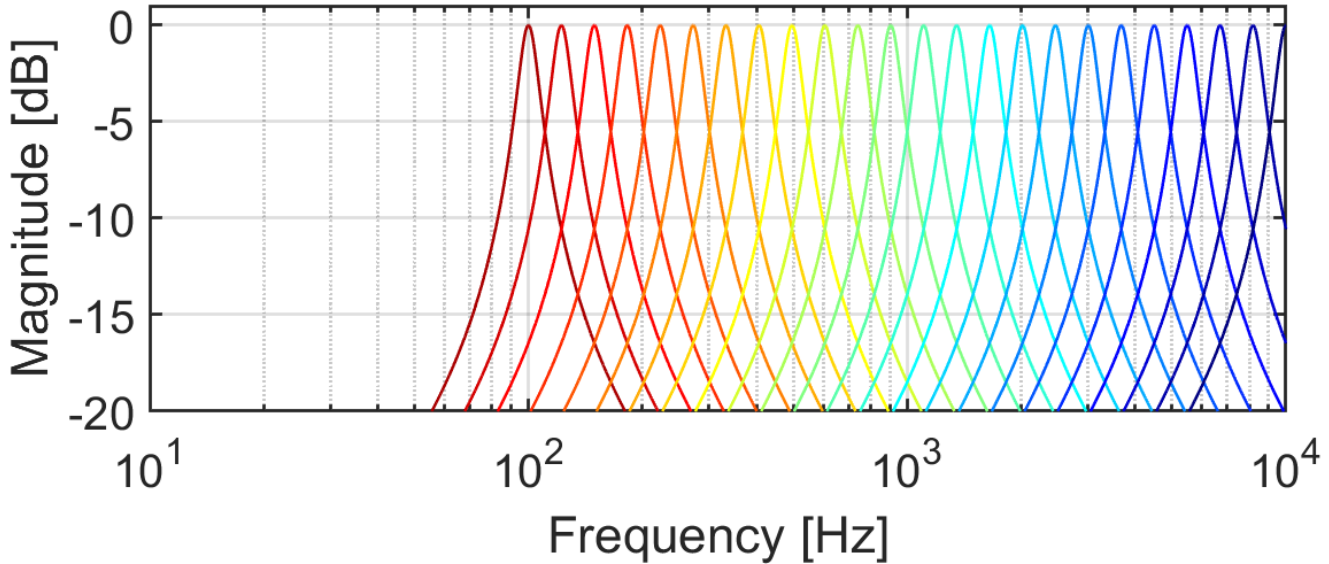
These are mathematical responses

Architectural parameters of filterbank: f_{max}

Small: 250Hz



Large: 10,000Hz



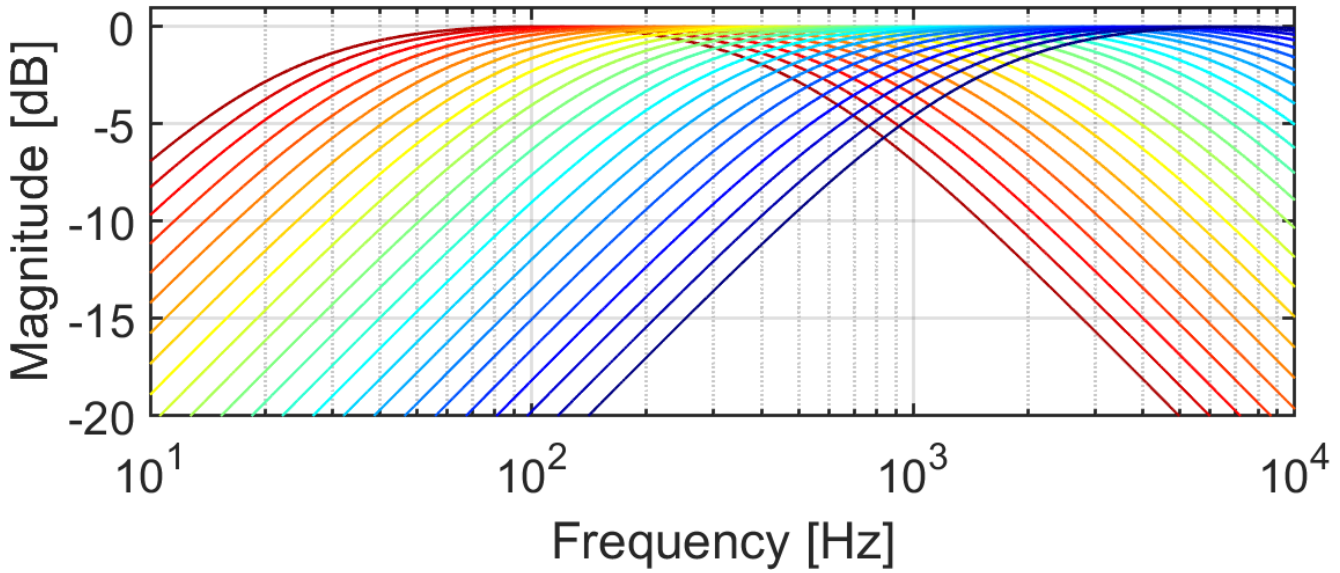
$P_{filterbank} \propto f_{max}$

for gm-C filters in weak inversion, which audio filters are in

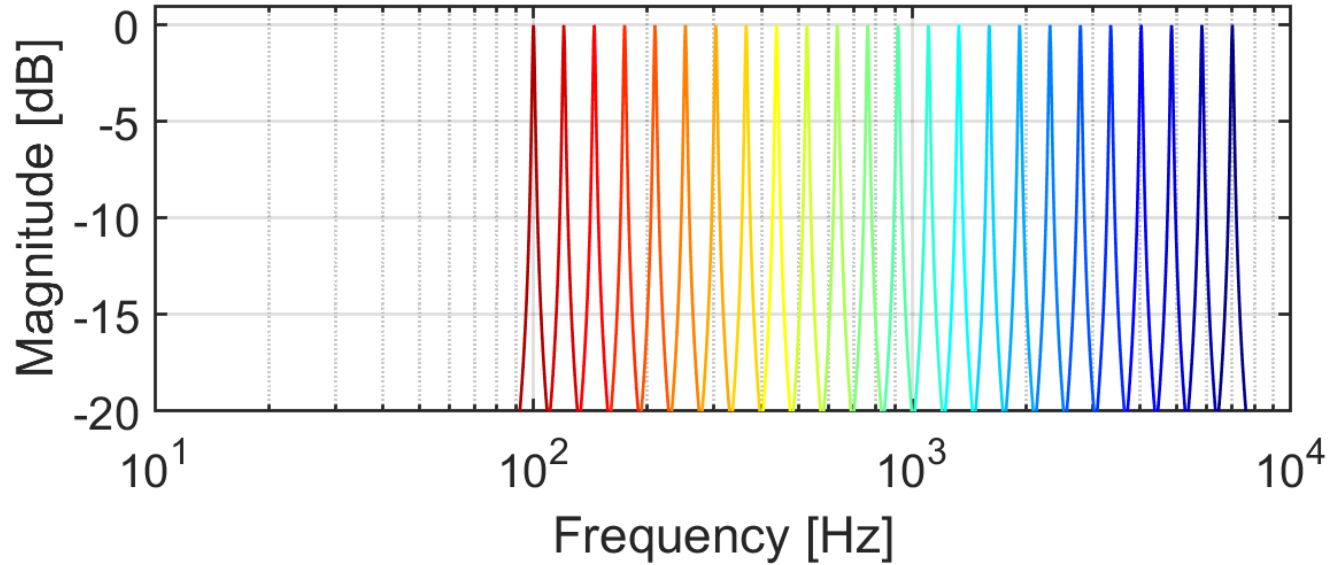
These are mathematical responses

Architectural parameters of filterbank: Q_{filter}

Small: 0.2



Large: 60

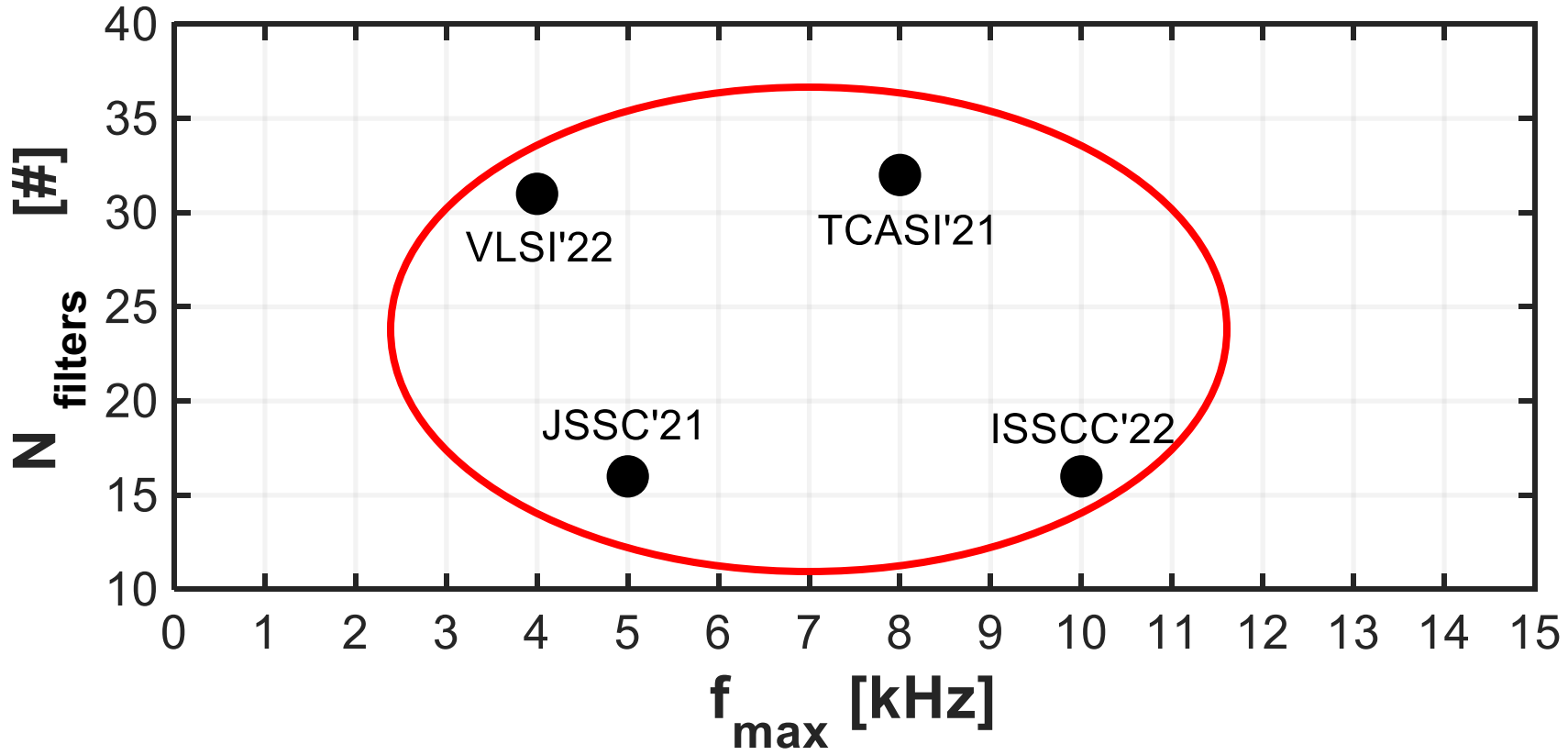


$P_{filterbank} \propto Q_{filter}$

for gm-C filters
[18Tsividis_SSCSMAG]

These are mathematical responses

State-of-the-art analog audio feature extractor chips in terms of architectural parameters

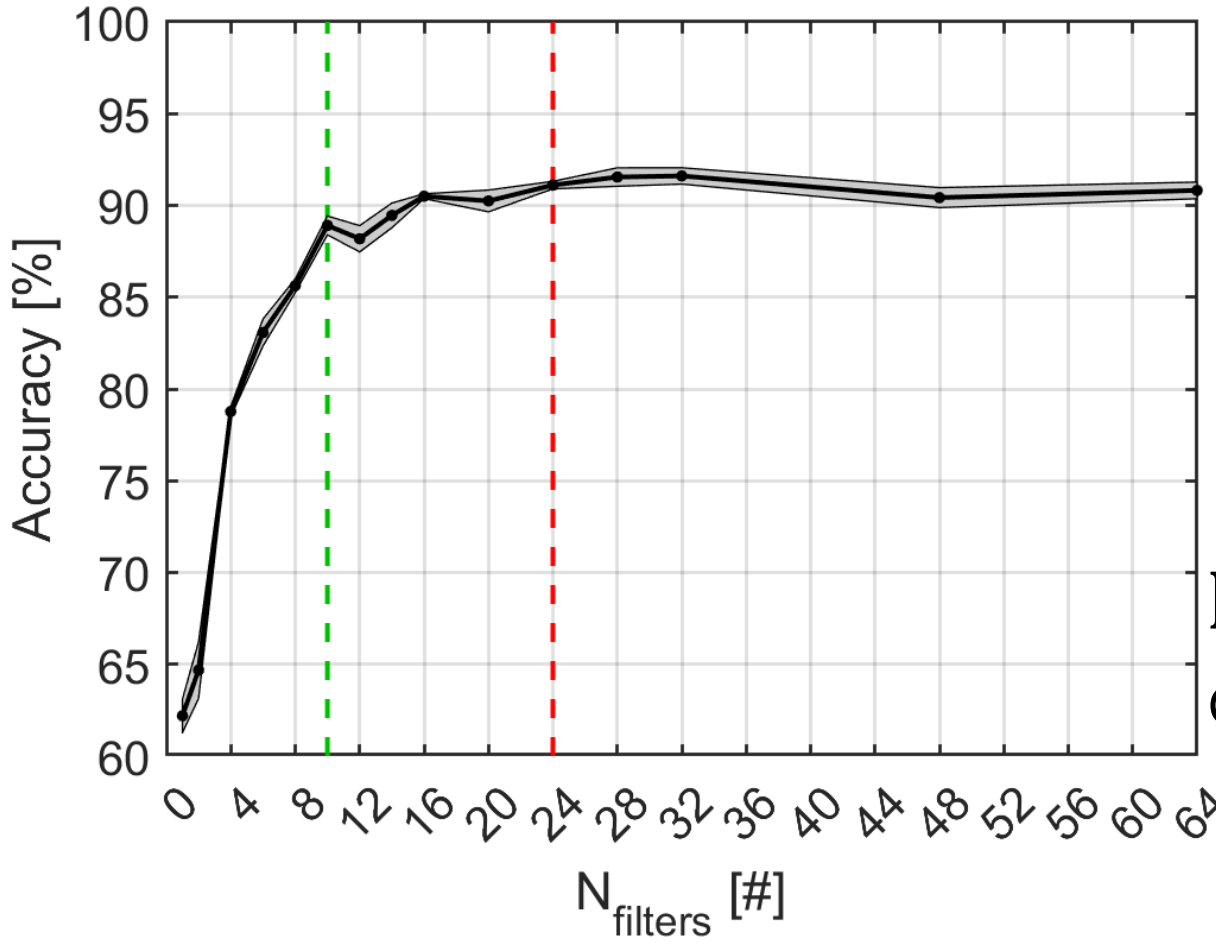


Large spread, although end task is the same—KWS

$N_{filters}$ sweep

This is simulation, not measurement:

- MATLAB frontend feature extractor [custom]
- MATLAB backend neural network [18ICAASP_TangLin]
- Google Speech Commands Dataset, 10 keywords



Red: "typical" choice

Green: proposed "tiny" choice

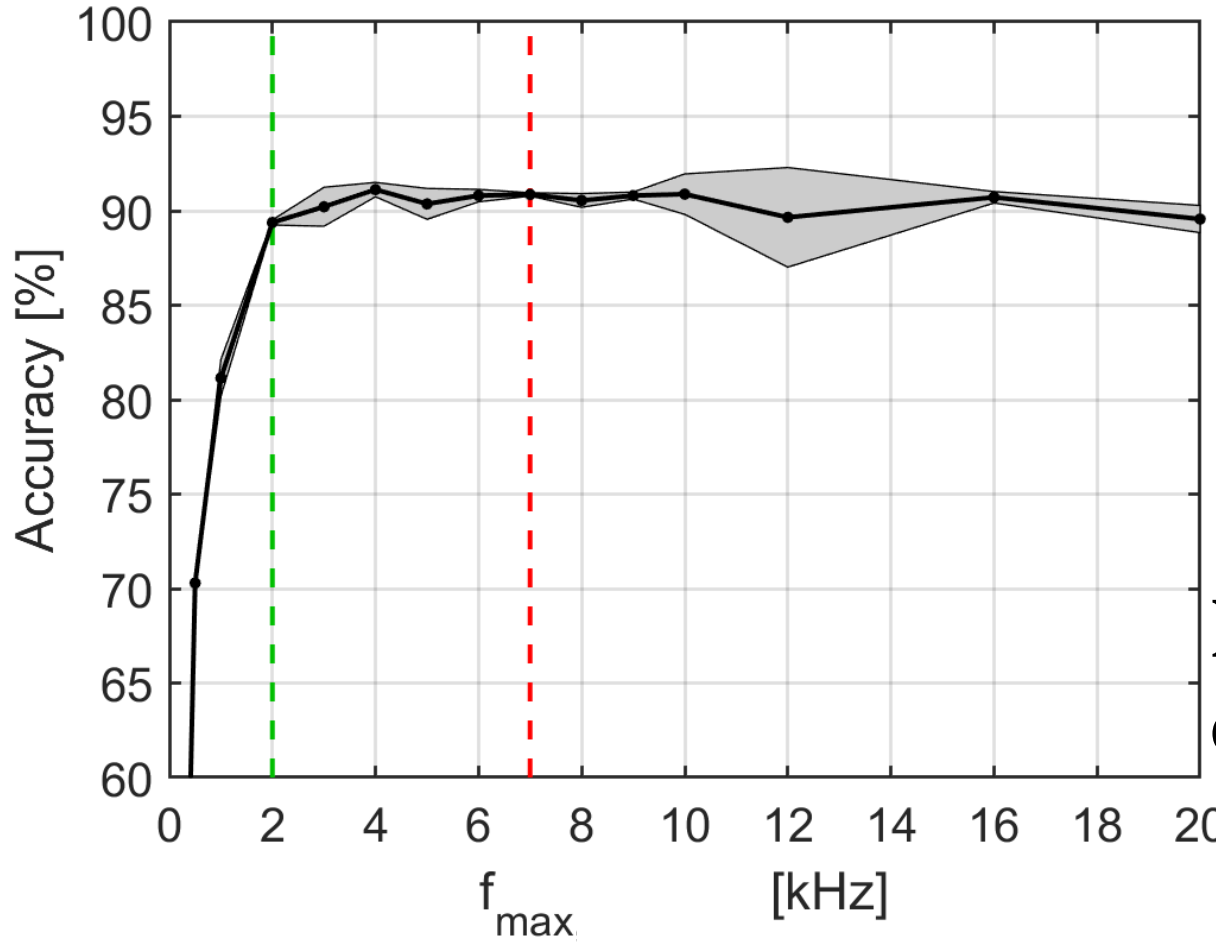
$P_{filterbank} \propto N_{filters}$

Can reduce $N_{filters}$ by 2.4x from 24 to 10

f_{max} sweep

This is simulation, not measurement:

- MATLAB frontend feature extractor [custom]
- MATLAB backend neural network [18ICAASP_TangLin]
- Google Speech Commands Dataset, 10 keywords



Red: "typical" choice

Green: proposed "tiny" choice

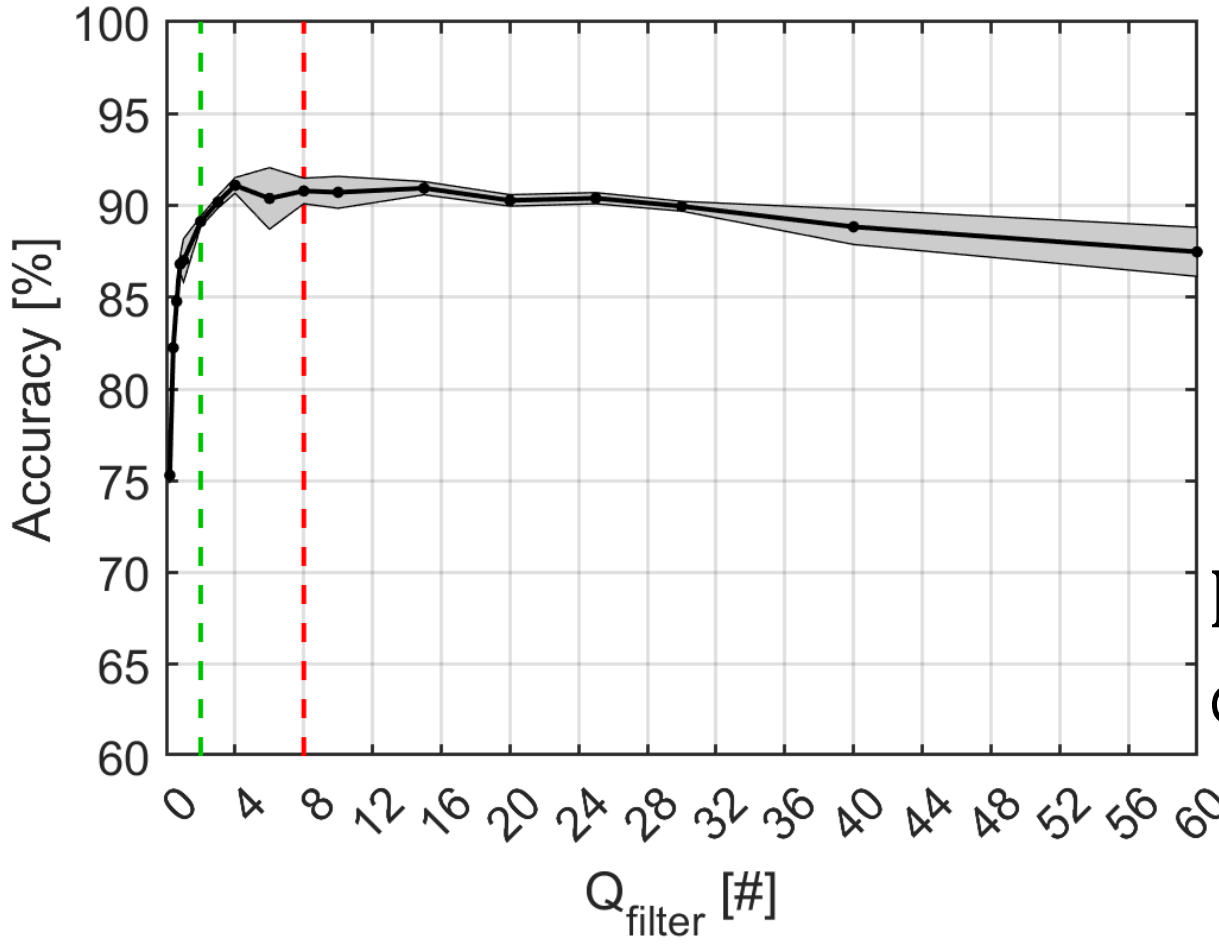
$$P_{filterbank} \propto f_{max}$$

Can reduce f_{max} by 3.5x from 7kHz to 2kHz

Q_{filter} sweep

This is simulation, not measurement:

- MATLAB frontend feature extractor [custom]
- MATLAB backend neural network [18ICAASP_TangLin]
- Google Speech Commands Dataset, 10 keywords



Red: "typical" choice

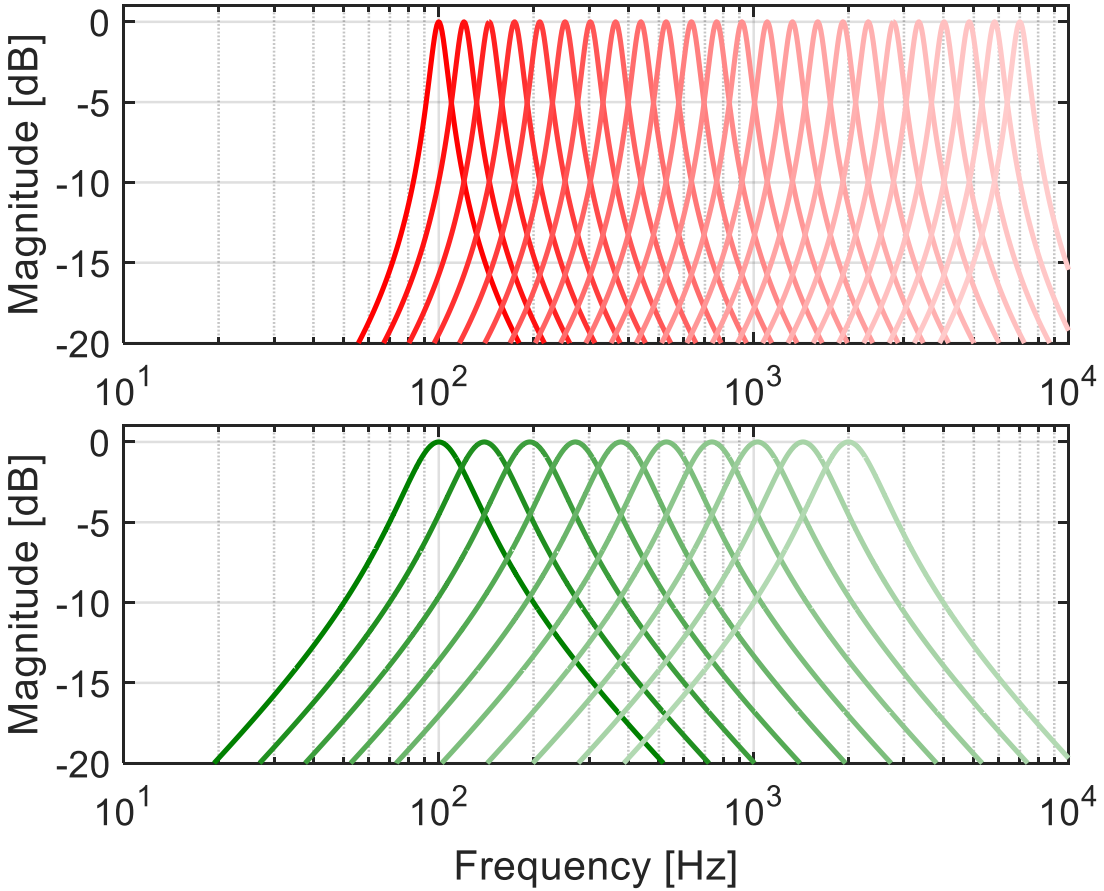
Green: proposed "tiny" choice

$$P_{filterbank} \propto Q_{filter}$$

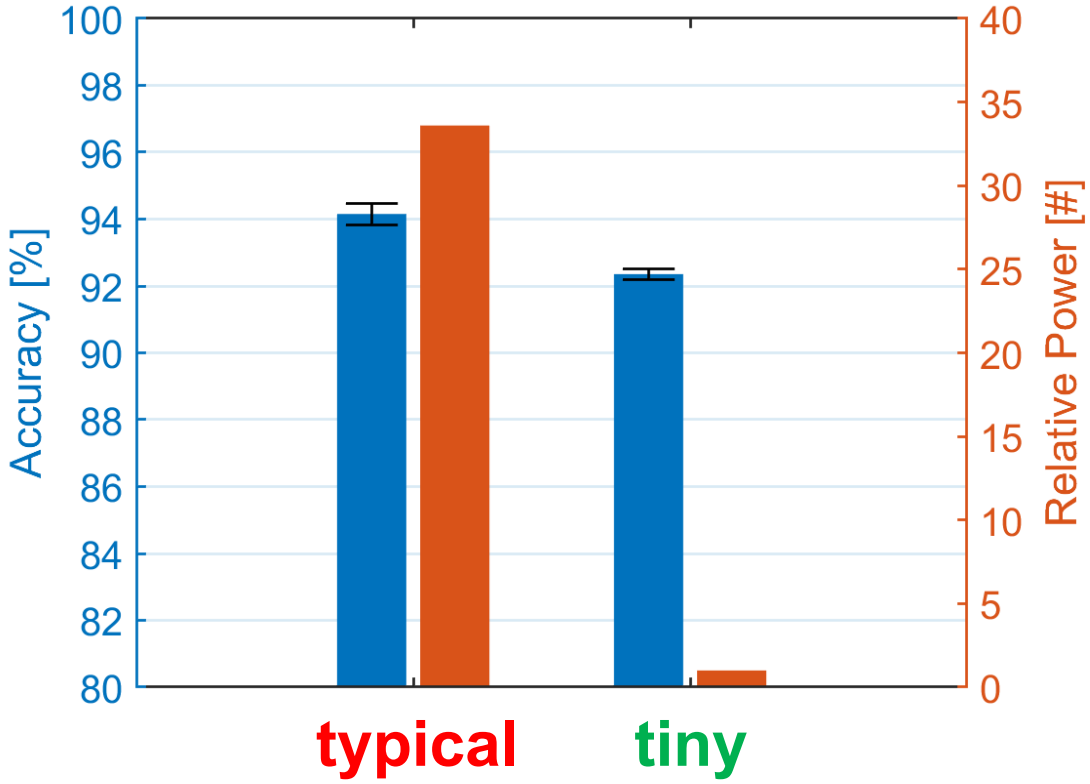
Can reduce Q_{filter} by 4x from 8 to 2

Comparison between typical and tiny filterbanks

Frequency responses



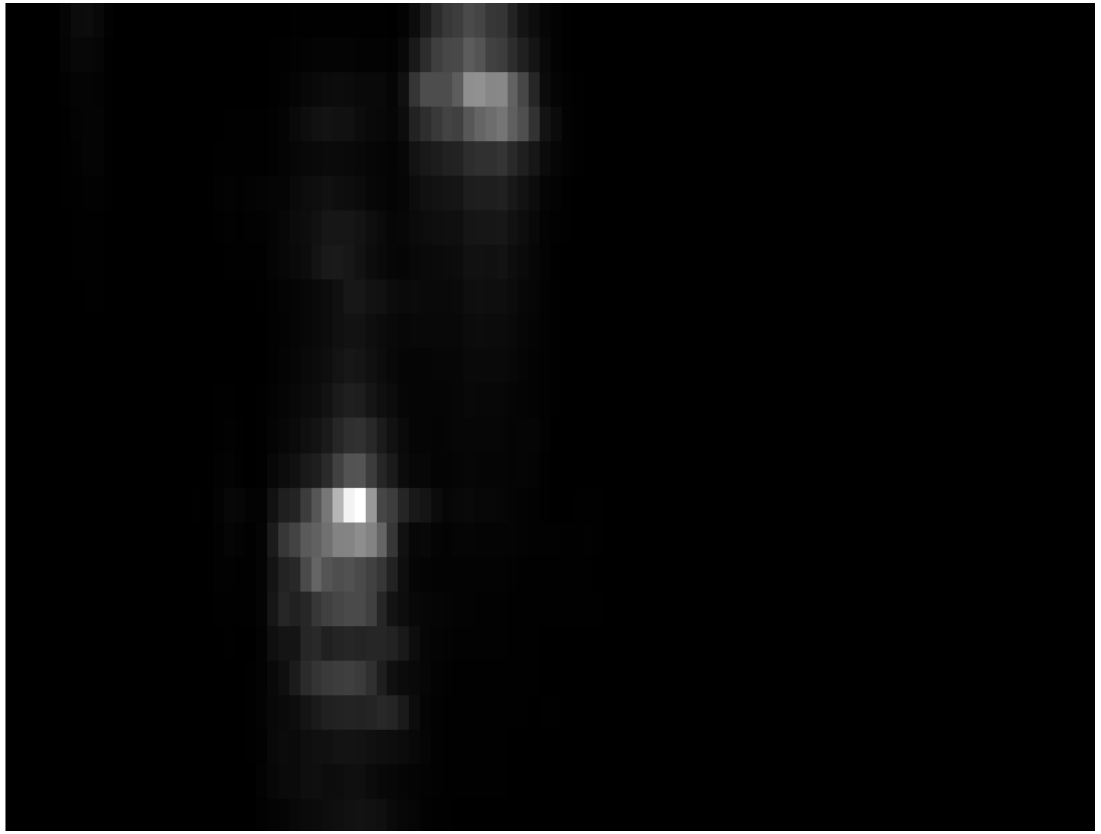
10-KWS accuracy and relative power



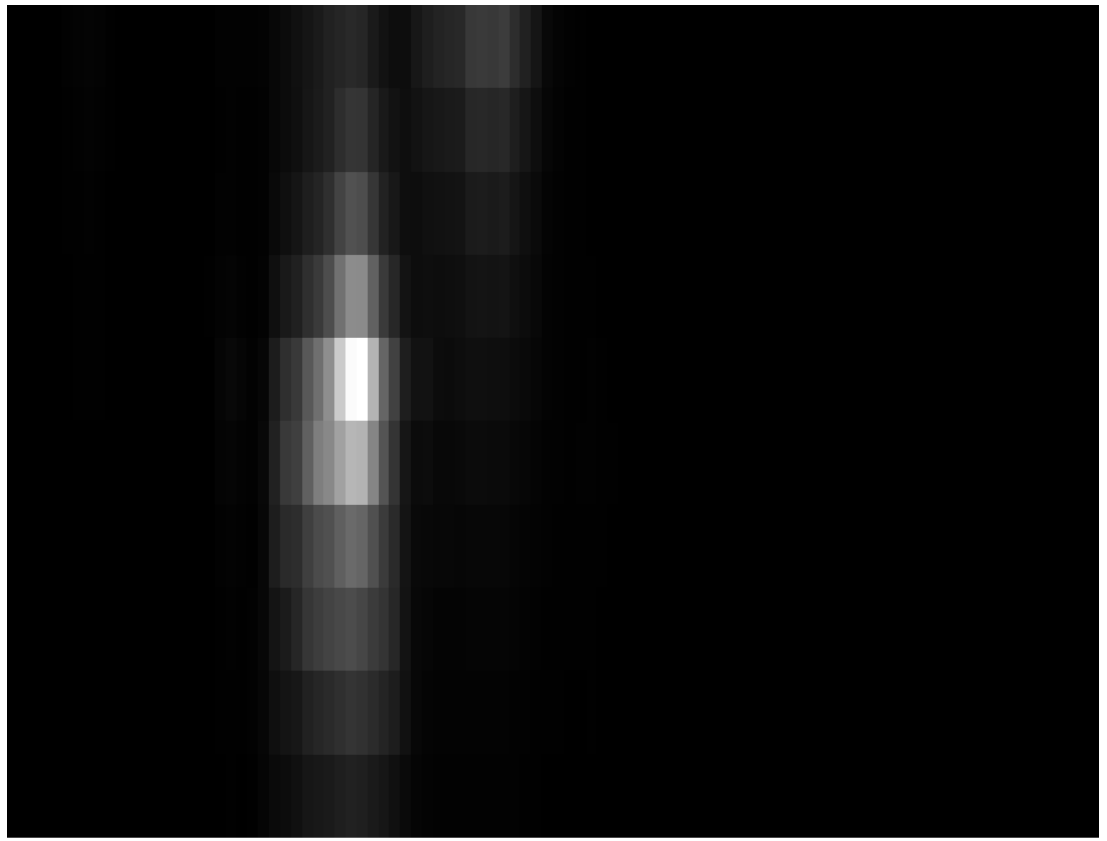
33.6x power savings for 1.8% acc penalty

Intuition?
…observe typical vs tiny spectrograms

typical



tiny



Scaling down the filterbank preserves the essential “features”

Summary

- Analog audio feature extraction is a power-efficient paradigm that is re-gaining interest in the integrated circuit design community.

- But there is currently little consensus on how to set its architectural parameters, namely those of the filterbank.

- And we show that there is a lot of opportunity to scale down the filterbank save power while maintaining KWS accuracy.
  - In particular, 33.6x power savings in filterbank for 1.8% downstream acc penalty.

- **This research is just a first, and small step…**

Future work

- Investigate effect of analog non-idealities on KWS accuracy
  - How low can the filter SNR be pushed?
  - How much filter nonlinearity be tolerated?

- Zoom out to system-level
  - In addition to filterbank, consider microphone, amplifier, envelope detectors, feature-rate ADCs, and neural network

- Repeat using a contemporary, state-of-the-art neural network

Acknowledgments

- Xinghua Sun is thanked for initial simulations.
- Nolan Tremelling and Maria Gordiyenko are thanked for more simulations.
- Ray Xu is thanked for simulation support.
- Rebecca Zhang is thanked for neural network implementation support.

- Prof. Peter Kinget is thanked for advising.

- This research was supported in part by NSF 1704899 and Analog Devices.

Questions?

Copyright Notice

www.tinyml.org