tinyML® Foundation

Enabling Ultra-low Power Machine Learning at the Edge

tinyML Summit April 22 - 24, 2024

www.tinyML.org

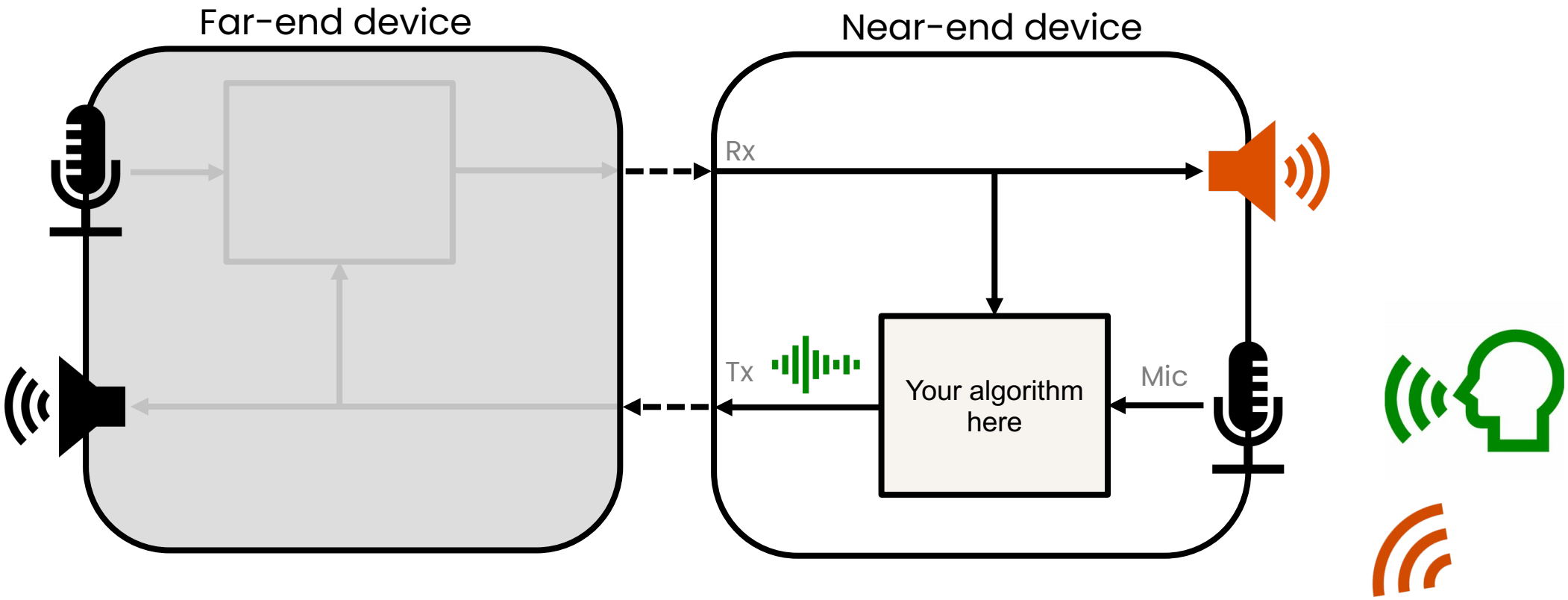
Embedded Joint Acoustic Echo Cancellation and Noise Suppression

Francesco Castelli

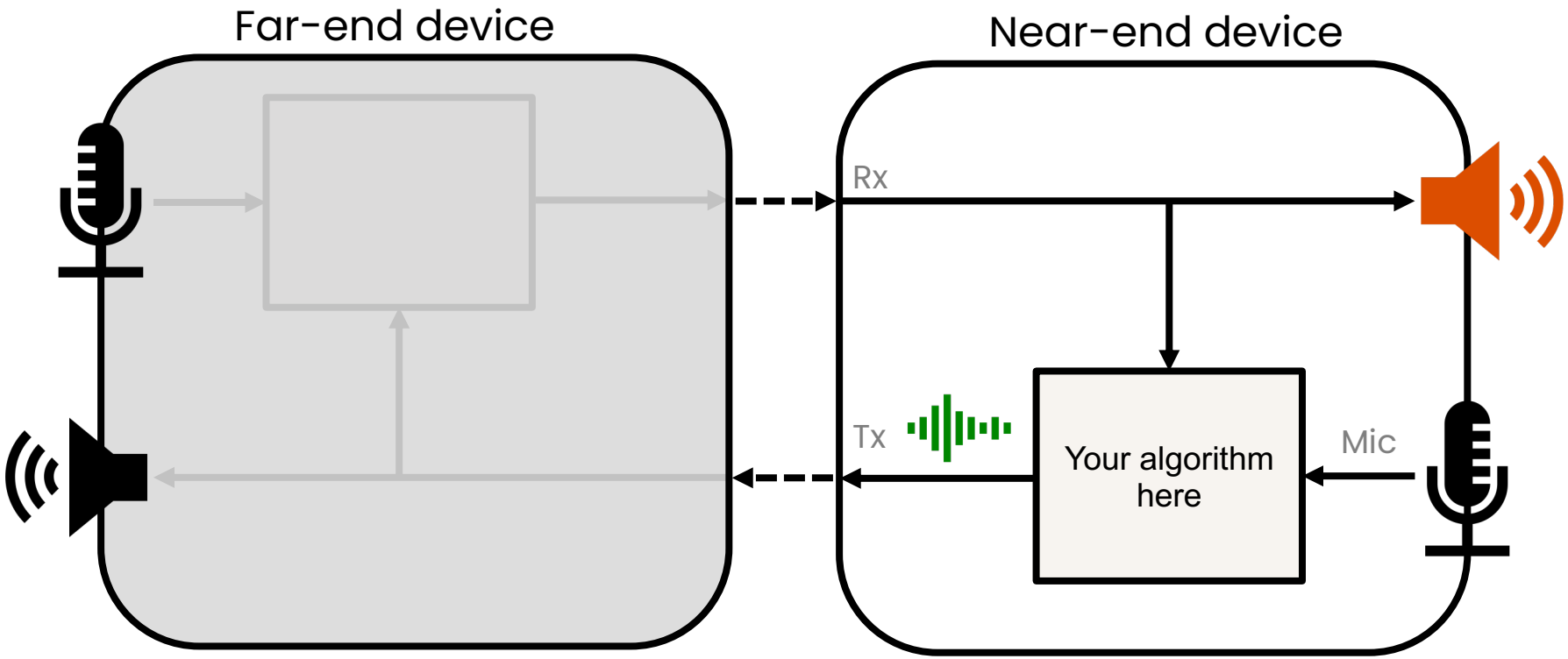
DSP Engineer, Voice & Audio Team, NXP

Speech enhancement – Two problems in one

Strict latency requirements



Speech enhancement – Acoustic Echo Cancellation

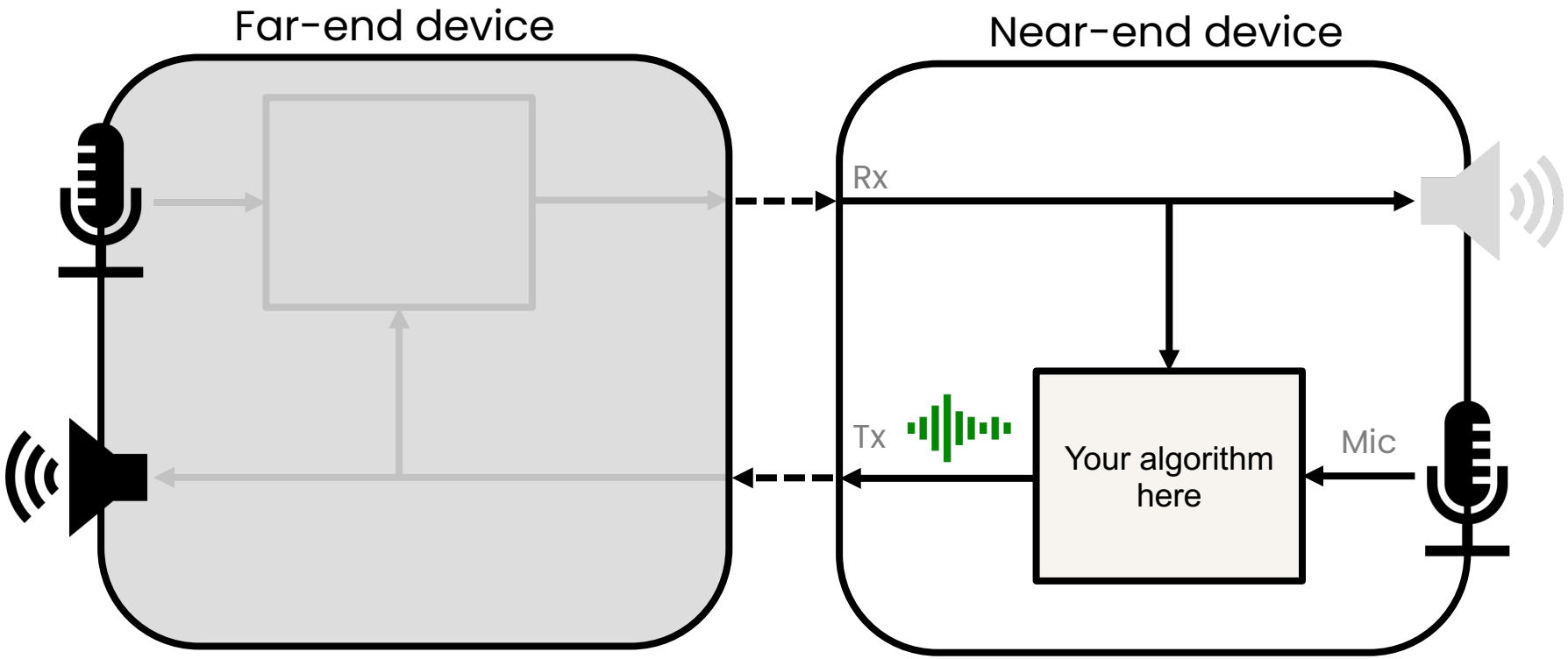


Loudspeaker non-linearity

Rx and microphone signal delay

Reverberation

Speech enhancement – Noise Suppression

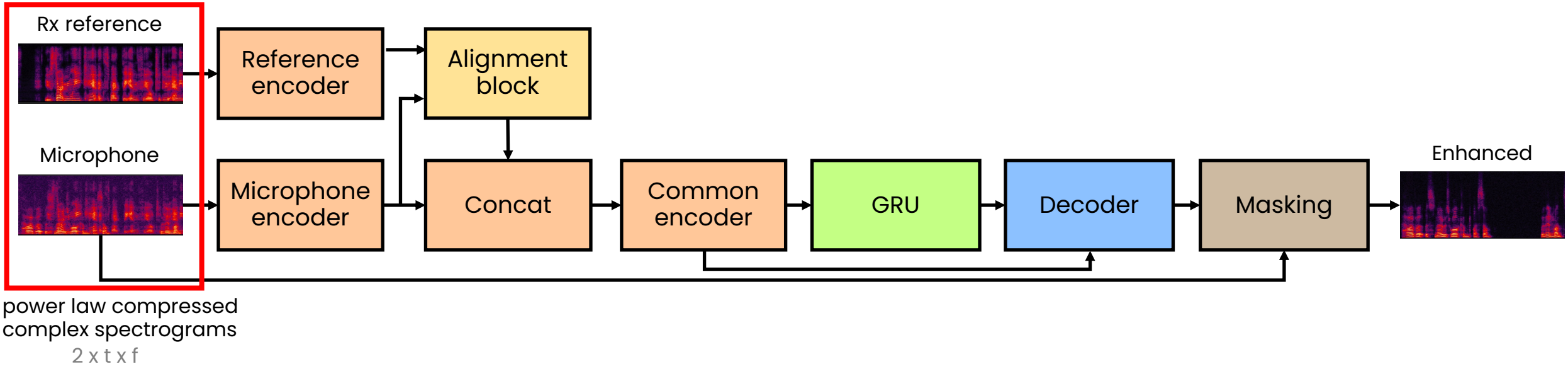


Low SNRs

Impulsive noise

Non-stationary noise

State Of The Art - DeepVQE

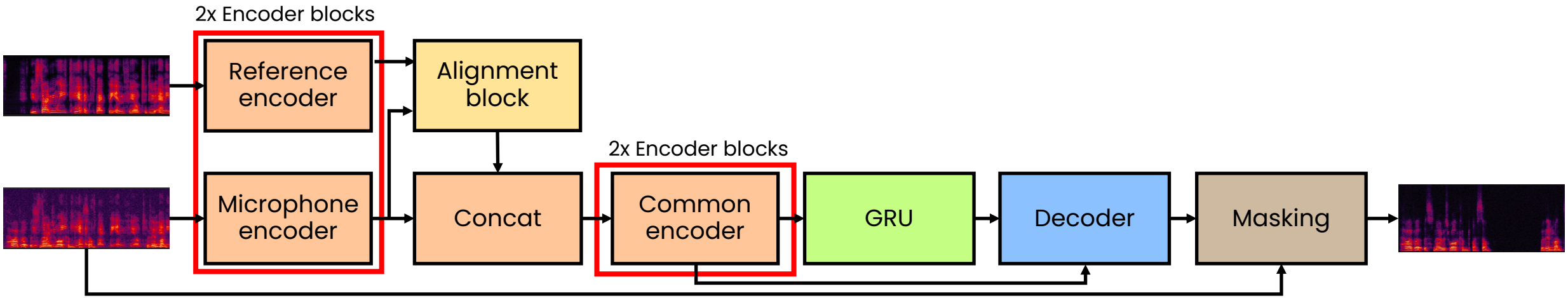


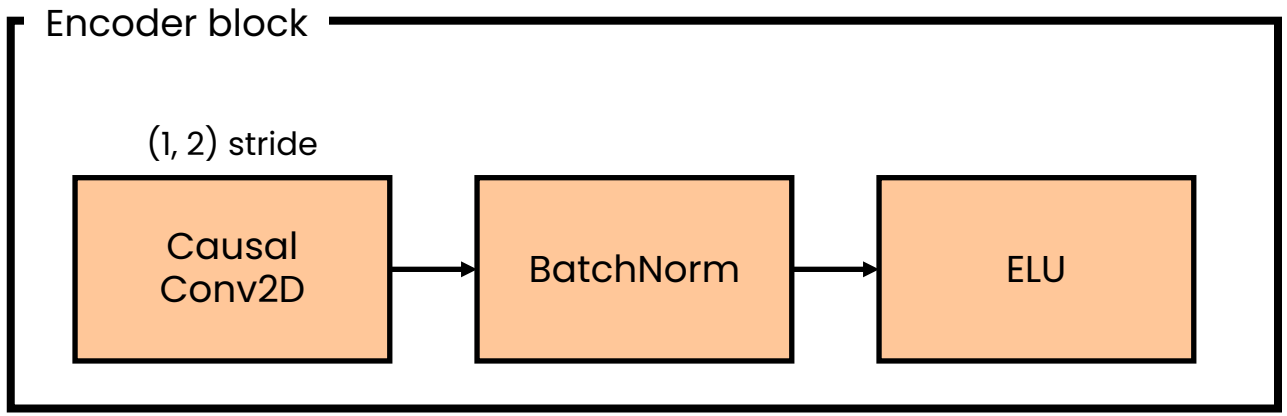
	Sample Rate	Win size	Hop size
paper	24 kHz	20 ms	10 ms
ours	16 kHz	32 ms	16/8 ms

Algorithmic latency

Indenbom, E., Ristea, N. C., Saabas, A., Parnamaa, T., Guzvin, J., & Cutler, R. (2023). DeepVQE: Real Time Deep Voice Quality Enhancement for Joint Acoustic Echo Cancellation, Noise Suppression and Dereverberation

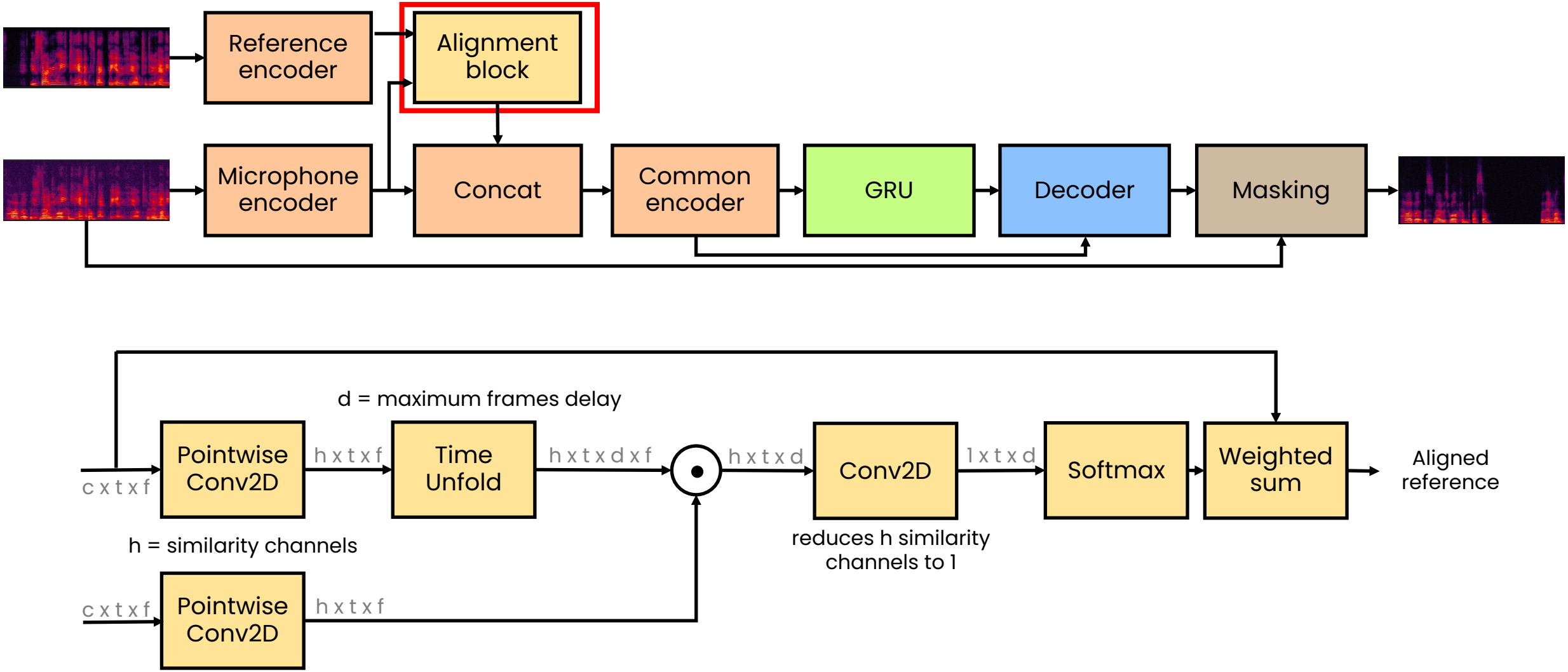
DeepVQE - Encoders





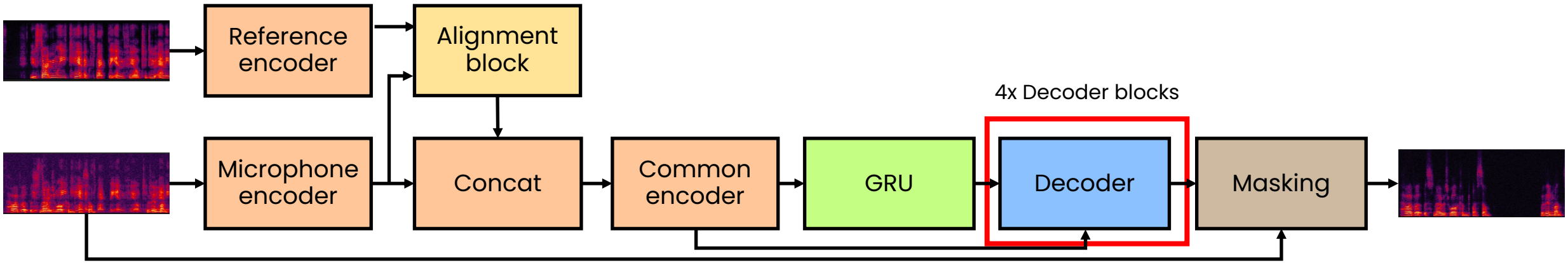
Indenbom, E., Ristea, N. C., Saabas, A., Parnamaa, T., Guzvin, J., & Cutler, R. (2023). DeepVQE: Real Time Deep Voice Quality Enhancement for Joint Acoustic Echo Cancellation, Noise Suppression and Dereverberation

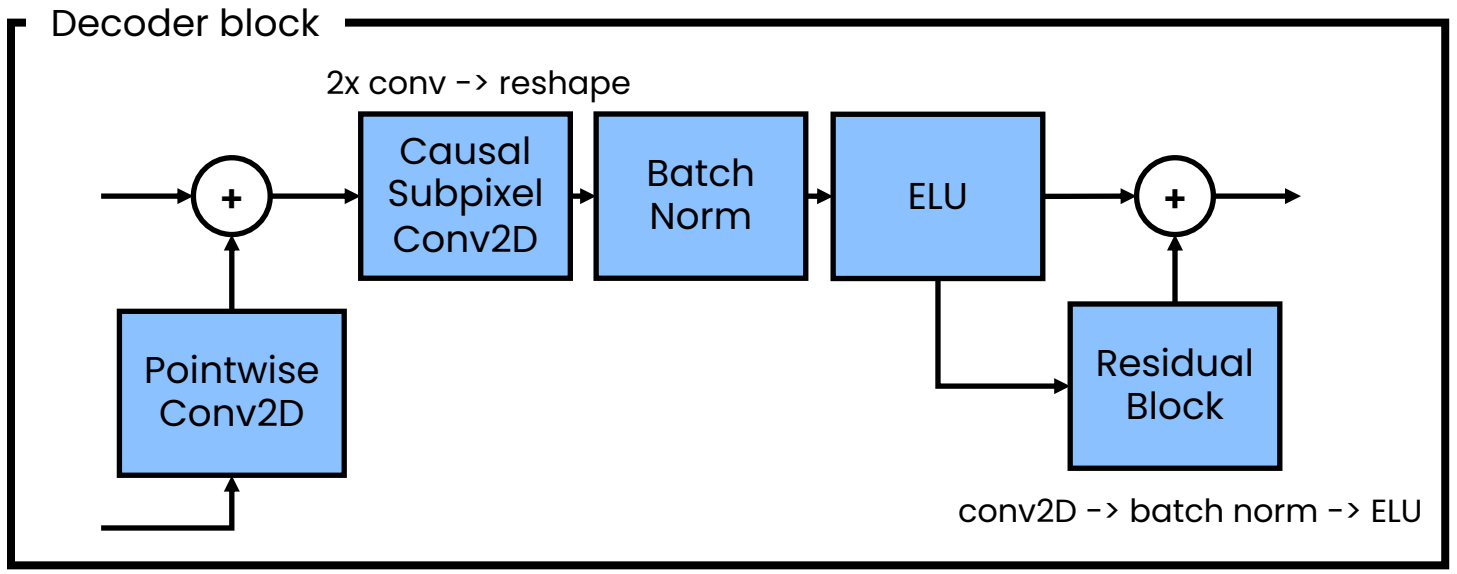
DeepVQE – Alignment Block



Indenbom, E., Ristea, N. C., Saabas, A., Parnamaa, T., Guzvin, J., & Cutler, R. (2023). DeepVQE: Real Time Deep Voice Quality Enhancement for Joint Acoustic Echo Cancellation, Noise Suppression and Dereverberation

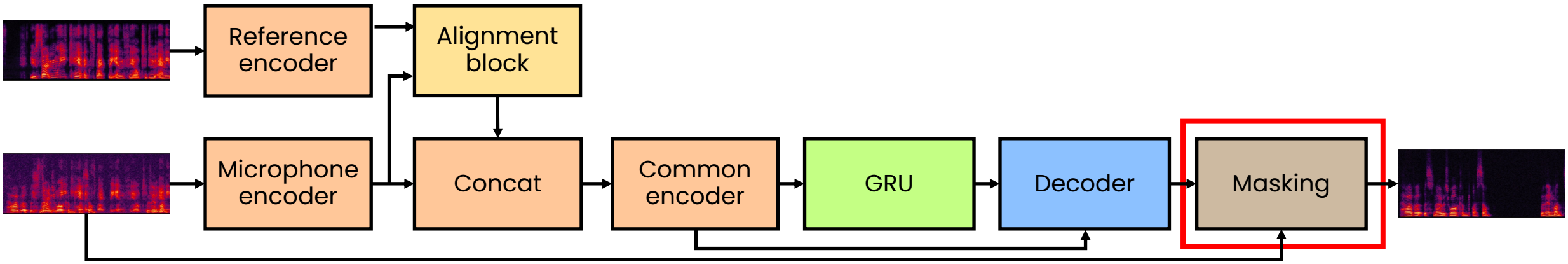
DeepVQE – Decoder

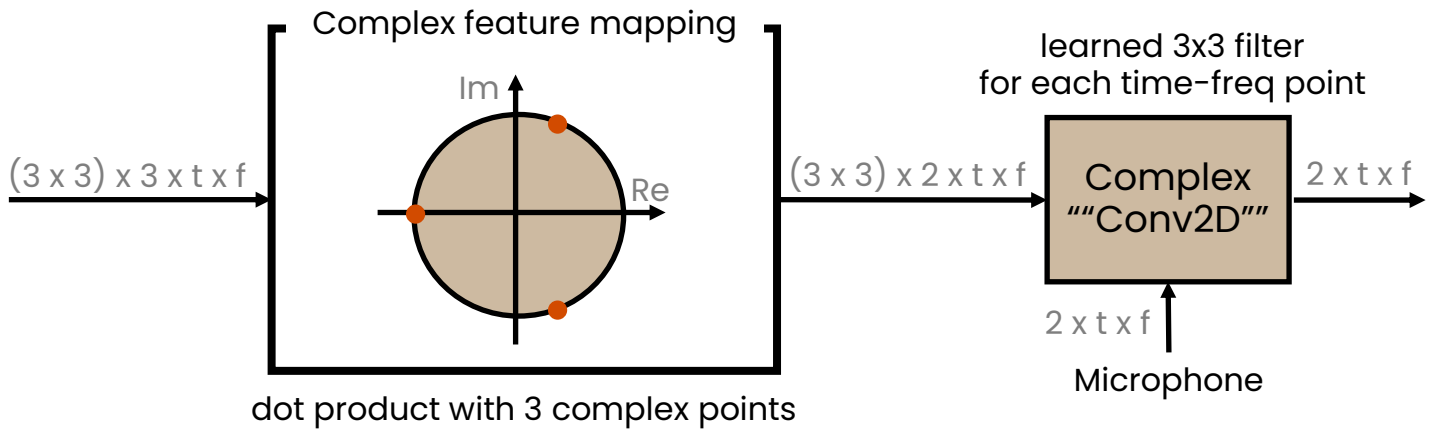




Indenbom, E., Ristea, N. C., Saabas, A., Parnamaa, T., Guzvin, J., & Cutler, R. (2023). DeepVQE: Real Time Deep Voice Quality Enhancement for Joint Acoustic Echo Cancellation, Noise Suppression and Dereverberation

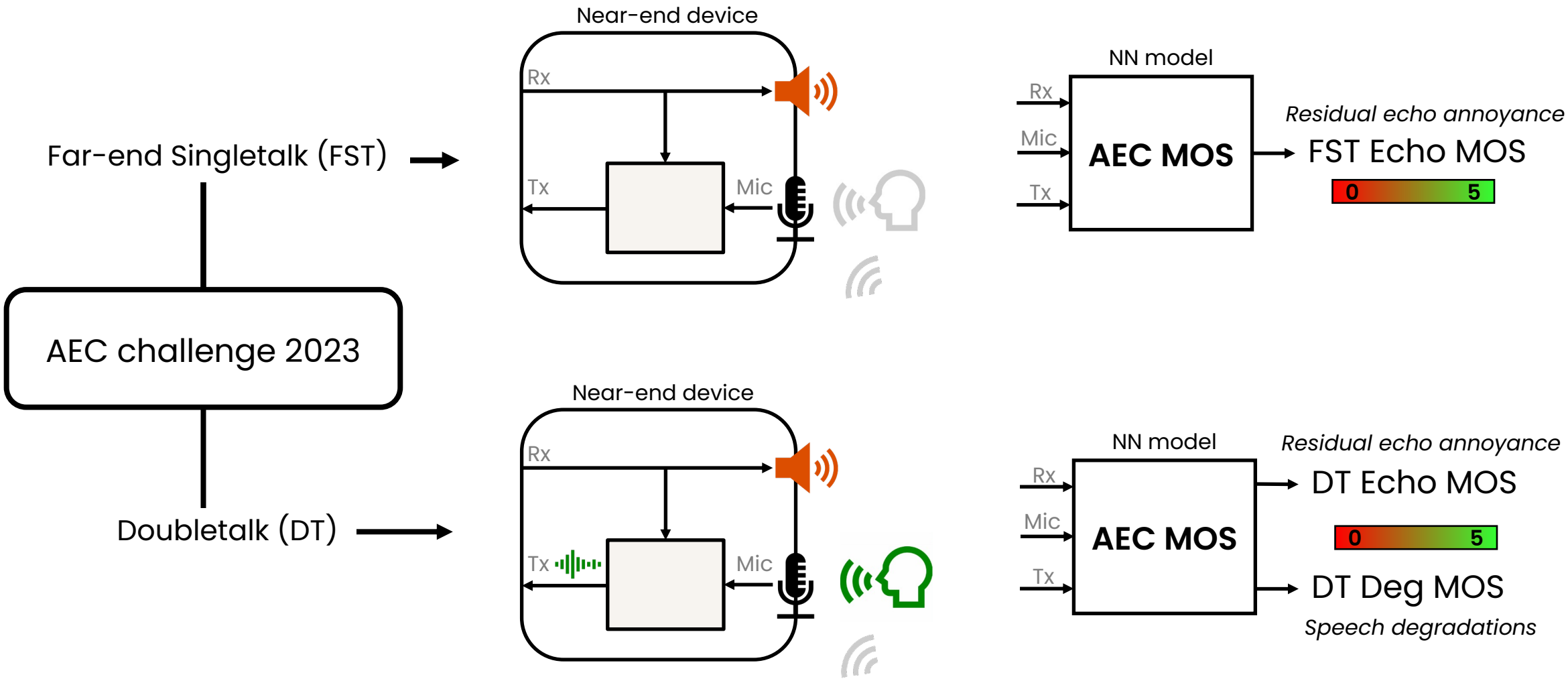
DeepVQE – Masking





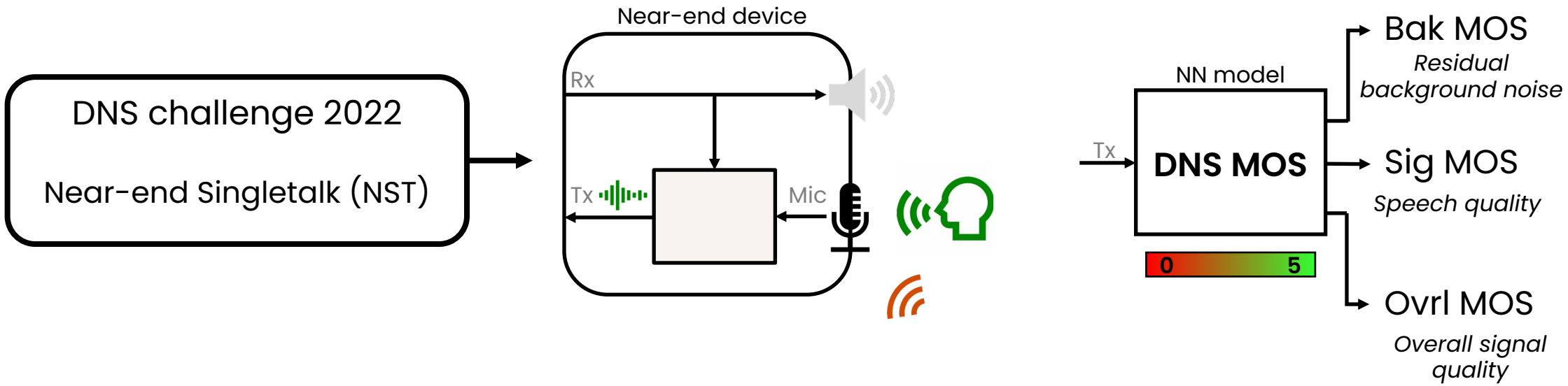
Indenbom, E., Ristea, N. C., Saabas, A., Parnamaa, T., Guzvin, J., & Cutler, R. (2023). DeepVQE: Real Time Deep Voice Quality Enhancement for Joint Acoustic Echo Cancellation, Noise Suppression and Dereverberation

AEC evaluation – Scenarios and metrics



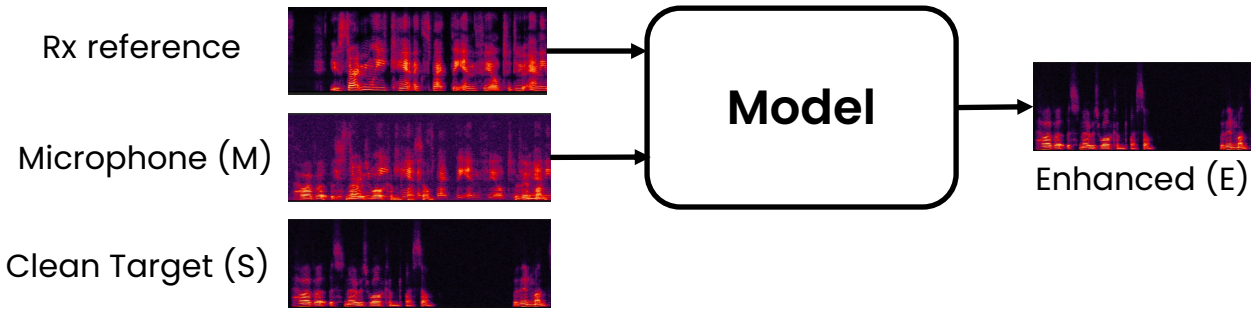
Purin, M., Sootla, S., Sponza, M., Saabas, A., & Cutler, R. (2022, May). AECMOS: A speech quality assessment metric for echo impairment. In ICASSP 2022-2022

NS evaluation – Scenario and metrics



Reddy, C. K., Gopal, V., & Cutler, R. (2022, May). DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In ICASSP 2022-2022

Training - Loss function



Complex L2 loss

$$L_{CSDR}(A,\hat{A}) = \frac{\sum_k \left| |A|^p e^{j\theta A} - |\hat{A}|^p e^{j\theta \hat{A}} \right|^2}{\sum_k \left| |A|^p e^{j\theta A} \right|^2}$$

Magnitude only L2 loss

$$L_{MSDR}(A,\hat{A}) = \frac{\sum_k \left| |A|^p - |\hat{A}|^p \right|^2}{\sum_k \left| |A|^p \right|^2}$$

$$L_{SDR}(A,\hat{A}) = \alpha L_{CSDR}(A,\hat{A}) + (1-\alpha) L_{MSDR}(A,\hat{A})$$

Speech loss — Echo and noise loss

$$L = \sum_n \beta L_{SDR}(S,E) + (1-\beta) L_{SDR}(M-S, M-E)$$

$$\beta = \sum_k |S|^2 / \sum_k |M|^2$$

First results – Evaluation

			Echo Cancellation performances			Noise Suppression performances		
			AEC MOS			DNS MOS		
Model	**Params (k)**	**MACs (M)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**
Unprocessed	-	-	2.19	2.09	4.05	3.49	2.11	2.31
			Echo only suppression	Doubletalk		Speech distortions	Noise suppression	Overall quality

First results - DeepVQE

			AEC MOS			DNS MOS		
Model	**Params (k)**	**MACs (M)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**
Unprocessed	-	-	2.19	2.09	4.05	3.49	2.11	2.31
DeepVQE-s (paper)	590	9.64*	4.61	4.62	4.02	3.60	4.10	3.30
DeepVQE-s (ours)	610	10.28	4.67	4.61	4.07	3.54	4.08	3.28

Slightly better echo cancellation

Slightly worst noise suppression

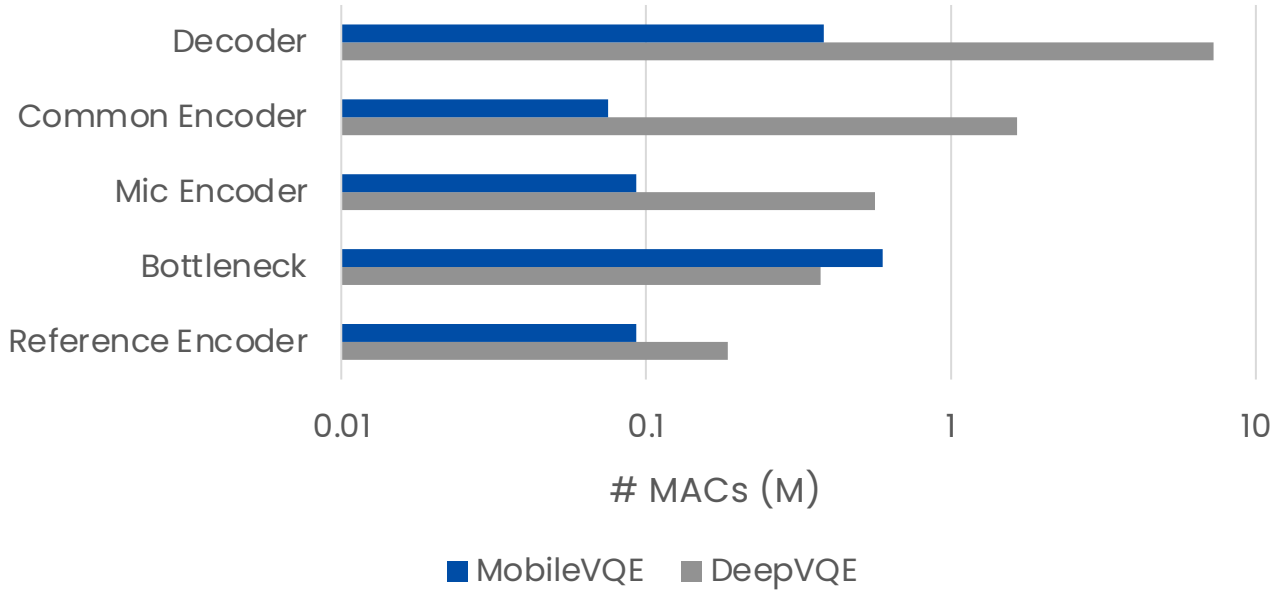
- Smaller batch sizes: Batch Norm -> Layer Norm
- Frame delay d=1s

* Estimated

First optimizations - MobileVQE

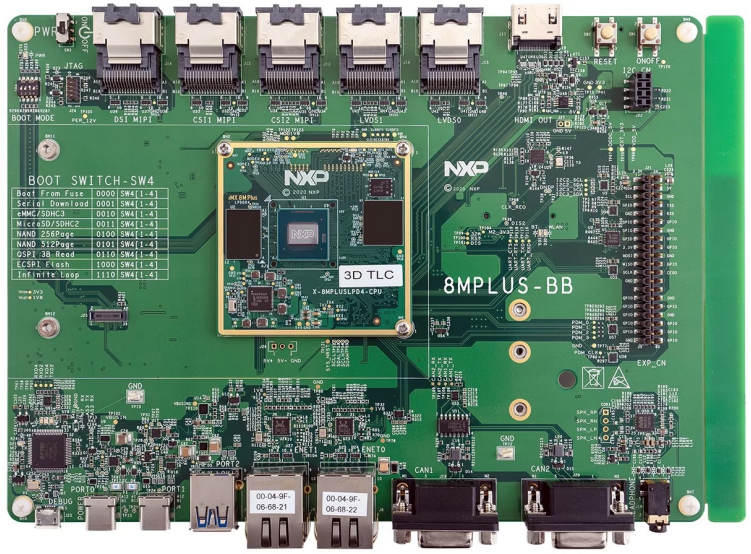
			AEC MOS			DNS MOS		
Model	**Params (k)**	**MACs (M)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**
Unprocessed	-	-	2.19	2.09	4.05	3.49	2.11	2.31
DeepVQE-s (paper)	590	9.64*	4.61	4.62	4.02	3.60	4.10	3.30
DeepVQE-s (ours)	610	10.28	4.67	4.61	4.07	3.54	4.08	3.28
MobileVQE	635	1.34	4.68	4.49	3.95	3.39	3.95	3.11

- Conv2d -> Depthwise Separable Conv2D
- No decoder residual blocks
- Frame delay d=0.5s



* Estimated

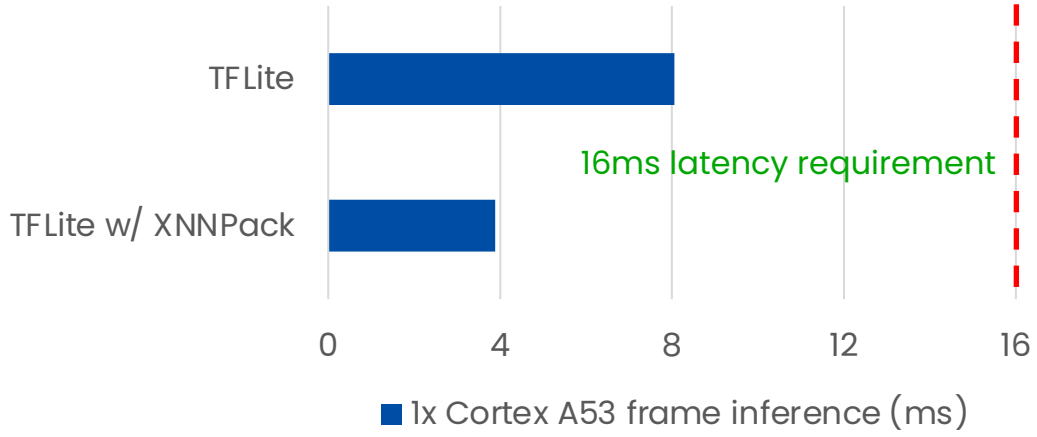
Integration - MobileVQE



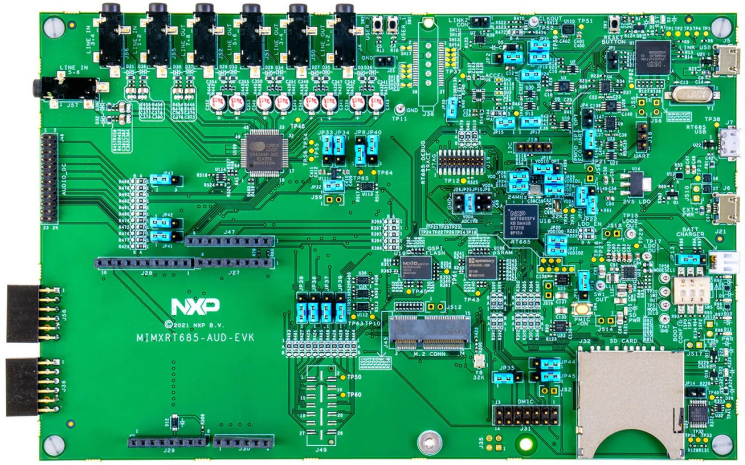
i.MX 8M Plus EVK

i.MX 8M Plus: High end NXP MPU
- 4x Arm® Cortex® A53 (1.8 GHz)
- NPU (2.3 TOPS)

- ➢ 1 core of Arm® Cortex® A53
- ➢ FP32 model, 16ms hop size
- ➢ TFLite runtime with XNNPack
- ➢ GStreamer audio-visual pipeline integration



The real target



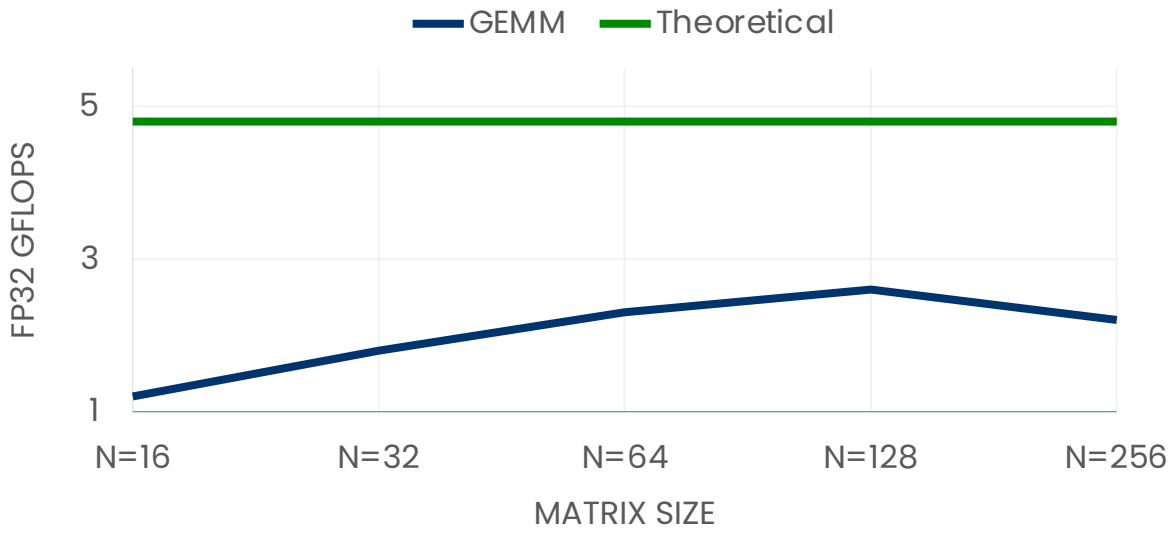
IMXRT600-AUD-EVK

i.MX RT600: dual-core NXP MCU

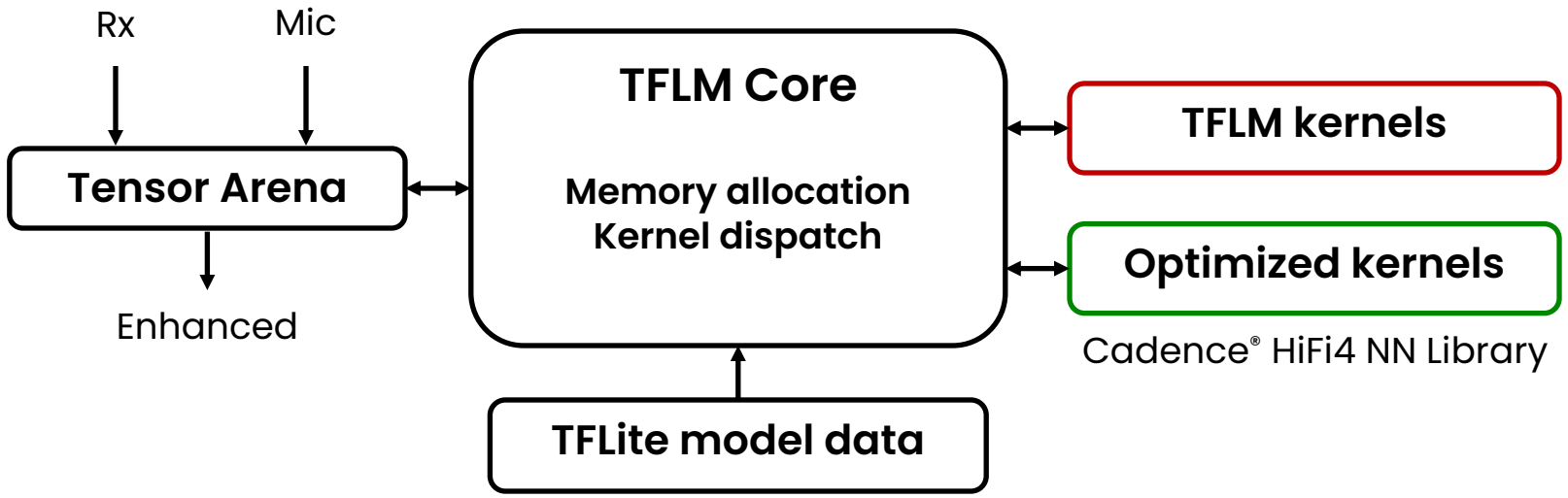
- Arm® Cortex® M33 (300 MHz)
- Cadence® Tensilica® HiFi 4 DSP (600 MHz)
- 4.5 MB shared on-chip SRAM

Cadence® Tensilica® HiFi 4 DSP:

- Two 2-way SIMD VFPU : 4 FP32 MACs/cycle
- Fixed-Point: 8 32x16 or 16x16 MACs/cycle
- C/C++ intrinsics
- Cadence® HiFi4 NN library



i.MX RT600 – DSP NN runtime



Model optimizations – FP32 performances

Echo Cancellation performances

Noise Suppression performances

			AEC MOS			DNS MOS			HiFi4 DSP
Params (K)	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**

FP32 Tensor arena

FP32 DSP frame inference @ 600MHz

Model optimizations – Smaller model

				AEC MOS			DNS MOS			HiFi4 DSP
Model	**Params (k)**	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**
DeepVQE-s (ours)	610	10.28	-	4.67	4.61	4.07	3.54	4.08	3.28	-
MobileVQE	635	1.34	-	4.68	4.49	3.95	3.39	3.95	3.11	-
Cut parameters	147	0.86	770	4.53	4.34	3.81	3.31	3.84	3.01	13.19

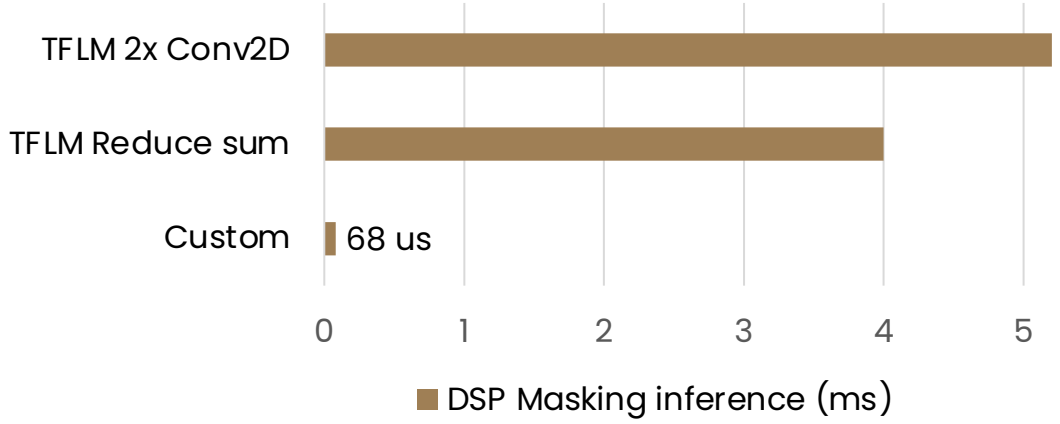
- Too large: 635k FP32 -> 2.54 MB
- Bottleneck: 598k / 635k -> 94%
- Frame delay d=500 ms

- 147k FP32 -> 588 KB
- Bottleneck: 102k / 147k -> 69%
- Frame delay d=250 ms

Model optimizations - Custom masking layer

				AEC MOS			DNS MOS			HiFi4 DSP
Model	**Params (k)**	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**
DeepVQE-s (ours)	610	10.28	-	4.67	4.61	4.07	3.54	4.08	3.28	-
MobileVQE	635	1.34	-	4.68	4.49	3.95	3.39	3.95	3.11	-
Cut parameters	147	0.86	770	4.53	4.34	3.81	3.31	3.84	3.01	13.19
Custom impls	147	0.86	690	4.53	4.34	3.81	3.31	3.84	3.01	7.19

- Masking MACs: 2.3k
- TFLM: Split, Concatenation, Transposition
- HiFi4 intrinsics: batched complex dot product



Model optimizations - ReLu

				AEC MOS			DNS MOS			HiFi4 DSP
Model	**Params (k)**	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**
DeepVQE-s (ours)	610	10.28	-	4.67	4.61	4.07	3.54	4.08	3.28	-
MobileVQE	635	1.34	-	4.68	4.49	3.95	3.39	3.95	3.11	-
Cut parameters	147	0.86	770	4.53	4.34	3.81	3.31	3.84	3.01	13.19
Custom impls	147	0.86	690	4.53 ↓	4.34 ↓	3.81	3.31 ↓	3.84 ↓	3.01	7.19 ↓↓
ELU -> ReLu	147	0.86	690	4.57	4.49	3.79	3.26	3.93	3.00	4.04

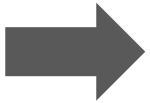
More “aggressive” model

- ELU: default TFLM kernel
- ReLu: HiFi4 FP32 optimized kernel

Model optimizations – Faster model

				AEC MOS			DNS MOS			HiFi4 DSP
Model	**Params (k)**	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**
DeepVQE-s (ours)	610	10.28	-	4.67	4.61	4.07	3.54	4.08	3.28	-
MobileVQE	635	1.34	-	4.68	4.49	3.95	3.39	3.95	3.11	-
Cut parameters	147	0.86	770	4.53	4.34	3.81	3.31	3.84	3.01	13.19
Custom impls	147	0.86	690	4.53	4.34	3.81	3.31	3.84	3.01	7.19
ELU -> ReLu	147	0.86 ↓	690 ↓↓	4.57	4.49 ↓	3.79	3.26 ↓	3.93 ↓	3.00	4.04
Cut MACs	139	0.54	455	4.56	4.45	3.87	3.28	3.82	2.98	2.99 ↓

- Skip Conv2Ds: 120k MACs
- Last decoder block: 117k MACs
- Reference/Mic encoders: 252k MACs



- Symmetrical model: no skip Conv2D
- Masking layer 27 -> 18: 80k MACs
- Reference/Mic encoders: 94k MACs

Model optimizations - TinyVQE

				AEC MOS			DNS MOS			HiFi4 DSP
Model	**Params (k)**	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**
DeepVQE-s (ours)	610	10.28	-	4.67	4.61	4.07	3.54	4.08	3.28	-
MobileVQE	635	1.34	-	4.68	4.49	3.95	3.39	3.95	3.11	-
Cut parameters	147	0.86	770	4.53	4.34	3.81	3.31	3.84	3.01	13.19
Custom impls	147	0.86	690	4.53	4.34	3.81	3.31	3.84	3.01	7.19
ELU -> ReLu	147	0.86	690	4.57	4.49	3.79	3.26	3.93	3.00	4.04
Cut MACs	139	0.54	455 ↓	4.56	4.45 ↓	3.87	3.28 ↓	3.82 ↓	2.98 ↓	2.99 ↓
TinyVQE	114	0.48	420	4.55	4.41	3.81	3.26	3.80	2.95	2.32

- Remove layer norm
- Longer training runs

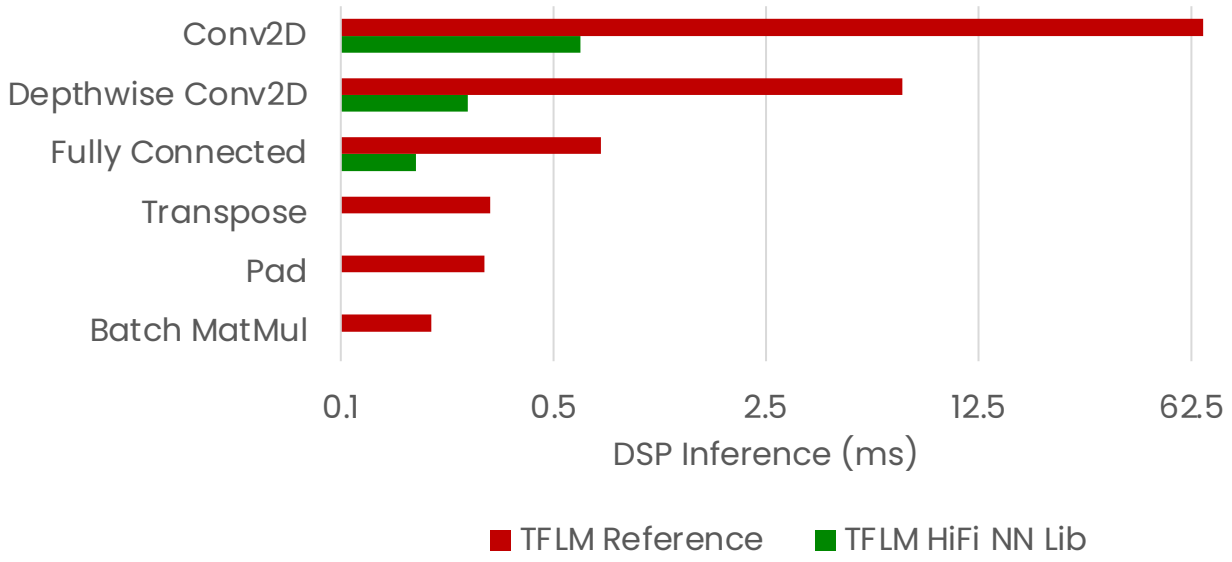
Model optimizations - Bonus

				AEC MOS			DNS MOS			HiFi4 DSP
Model	**Params (k)**	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**
DeepVQE-s (ours)	610	10.28	-	4.67	4.61	4.07	3.54	4.08	3.28	-
MobileVQE	635	1.34	-	4.68	4.49	3.95	3.39	3.95	3.11	-
Cut parameters	147	0.86	770	4.53	4.34	3.81	3.31	3.84	3.01	13.19
Custom impls	147	0.86	690	4.53	4.34	3.81	3.31	3.84	3.01	7.19
ELU -> ReLu	147	0.86	690	4.57	4.49	3.79	3.26	3.93	3.00	4.04
Cut MACs	139	0.54	455	4.56	4.45	3.87	3.28	3.82	2.98	2.99
TinyVQE	114	0.48	420	4.55	4.41	3.81	3.26	3.80	2.95	2.32
Bonus	92	0.45	418	4.54	4.24	3.63	3.27	3.79	2.92	2.26

- Not enough echo suppression

TinyVQE - Summary

				AEC MOS			DNS MOS			HiFi4 DSP
Model	**Params (k)**	**MACs (M)**	**Memory (KB)**	**FST Echo**	**DT Echo**	**DT Deg**	**Sig**	**Bak**	**Ovrl**	**Inference (ms)**
DeepVQE-s (ours)	610	10.28	-	4.67	4.61	4.07	3.54	4.08	3.28	-
TinyVQE	114	0.48	420	4.55	4.41	3.81	3.26	3.80	2.95	2.32



- Strong tiny model
- Suitable for wearables use cases
- Next step: 16x8 QAT
  - ≈ x4 smaller model
  - ≈ x2 DSP frame inference speed up

nxp.com

Copyright Notice

www.tinyml.org