A Model for Every User and Budget: Label-Free and Personalized Mixed-Precision Quantization

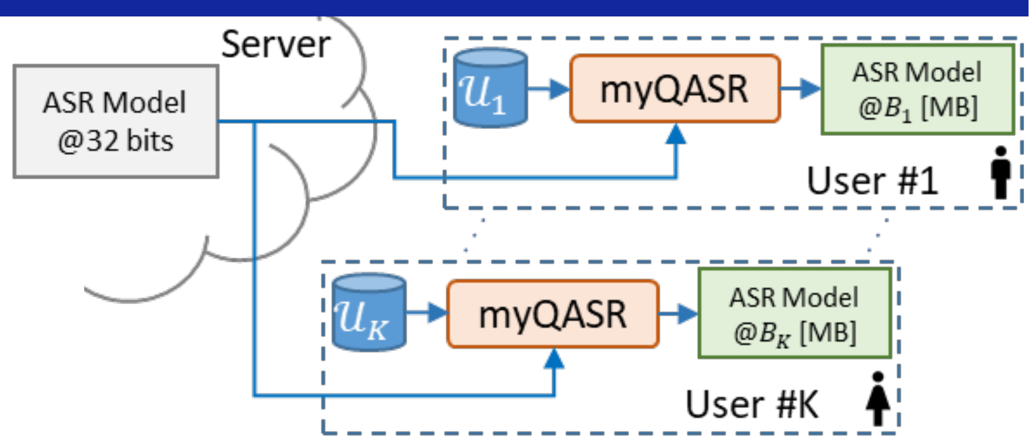
Samsung Research

INTERSPEECH 2023

Edward Fish[1,2], Umberto Michieli[1], Mete Ozay[1]
[1] Samsung Research UK - n.surname@samsung.com
[2] University of Surrey - edward.fish@surrey.ac.uk

Summary

Desiderata:
- ASR models need to fit on resource-limited devices
- ASR models on device should work better for the target users
- Target memory requirement specified in MB



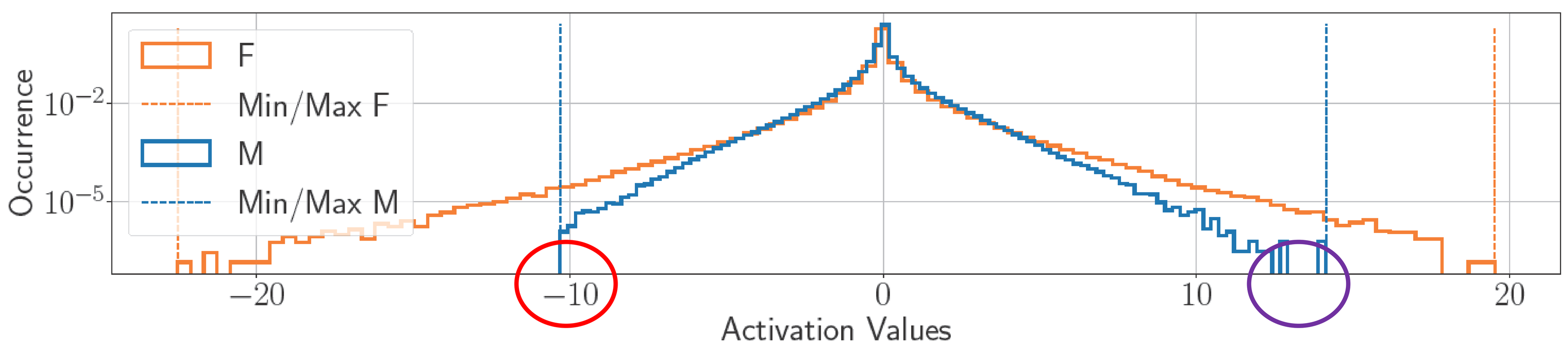
Our Solution (myQASR):
Mixed-precision post-training quantization method generating personalized compressed models for diverse users under any memory requirement.

Main Ideas:
- Layer-wise sensitivity detection
- Model calibration

} On a small unlabelled dataset from users

Motivation

Activation profile of different users is different
→ Models for different users require different compression
Example: First layer of wav2vec2, Male vs. Female



Method

3 MAIN STEPS:

1. Sensitivity Detection

Algorithm 1: Sensitivity detection of myQASR.

Data: B memory budget in MB, M model size in MB ($M > B$), $\mathcal{W}$ model parameters, and $\mathcal{U}$ unlabelled user samples.
Result: Array $\mathbf{b}$ of selected bit depths.
$\mathbf{b} \leftarrow \{32, \dots, 32\}$ *// initialize to FP.*
Compute median activations $\mathbf{a}$ over $\mathcal{U}$ ($a_l, \forall l \in [L]$);
$\hat{\mathbf{q}} \leftarrow$ argsort($\mathbf{a}$) *// get sorted list of layer indices.*
while $M > B$ **do**
  for l *in* $\hat{\mathbf{q}}$ **do**
    $b_l - = 1$ *// reduce l-th layer bit depth by one.*
    $M =$ `ComputeModelSize` ($\mathbf{b}, \mathcal{W}$)
    if $M <= B$ **then return** *bit depth array* $\mathbf{b}$;
def `ComputeModelSize` ($\mathbf{b}, \mathcal{W}$) **:**
  $\forall (b_l, W_l)$ in $(\mathbf{b}, \mathcal{W})$: qParams += $(b_l / 8) \times |W_l|$
return qParams / 1024^2 *// model size in MB.*

Forwards pass and save median of activations for each layer

Activation strength used as a proxy for sensitivity

2. Model Quantization
We quantize both **weights** and **activations**, via:

$$Q(\theta_l, b_l) = [\text{round}(\theta_l / S_l) - Z_l]_{b_l}$$

Z_l: zero-point correction
S_l: scaling factor

Weights have Gaussian distribution → Can use standard $S_l = 2^{b_l - 1}$
Activations do not follow Gaussian distribution→ Need for step 3

3. Activations' Calibration

A. *myQASR:* uses layer-wise min (X_l^m) and max (X_l^M)

$$S_l = (X_l^M - X_l^m)/(2^{b_l - 1}),$$
$$Z_l = -2^{b_l - 1} - \text{round}(X_l^m / S_l).$$

B. *myQASR-Hessian*: minimizes the distance between quantized and FP outputs of each layer scaled by its impact on the task loss

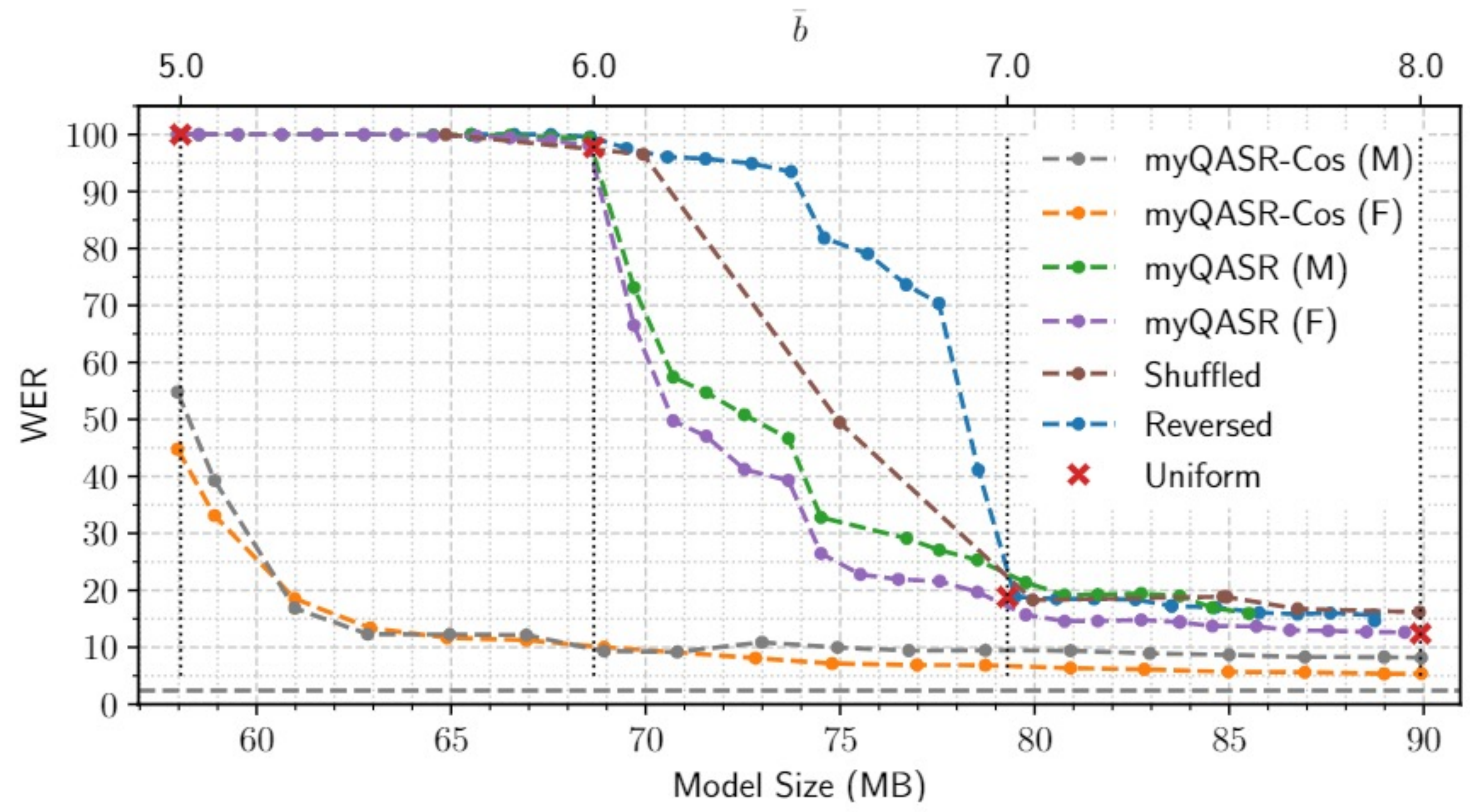
C. *myQASR-Cosine*: minimizes the cosine distance between quantized and FP outputs of each layer

Results

3 USE CASES:

1. Gender personalization

Original: multi-gender model → Quantized: optimized for specific gender



WER of W2V2-B on LS-F. Original model is 360MB.

2. Language personalization

Original: multi-language model → Quantized: optimized for specific language

Language Test \ Language Calibration	ca	de	en	fr	ja	ko	nl	pl	pt	ru	No Calib
ca	8.10	8.78	9.12	9.14	8.59	9.10	8.76	9.34	8.74	9.27	36.00
de	17.38	17.19	17.19	17.65	17.36	17.19	17.74	17.74	17.38	17.31	46.50
en	12.52	12.45	11.69	12.78	12.45	12.52	12.65	12.52	12.35	12.29	75.46
fr	11.85	11.61	11.93	11.02	11.19	11.96	11.93	11.13	11.11	12.53	40.95
ja	14.80	14.49	15.15	15.00	14.55	15.18	14.83	15.11	15.30	14.90	30.56
ko	19.28	19.46	21.53	19.73	19.73	19.12	19.37	21.08	20.81	19.64	25.38
nl	11.70	11.87	11.81	11.23	12.16	11.87	10.99	12.46	11.64	11.87	24.27
pl	12.79	12.61	13.47	13.40	12.54	12.93	12.97	12.61	12.82	12.61	32.67
pt	10.19	9.91	9.98	9.89	9.98	10.14	10.37	9.98	9.86	9.96	37.08
ru	9.62	9.55	9.62	10.09	9.42	9.38	9.96	10.12	10.49	9.72	20.28

WER on FLEUR with myQASR-Whisper-L.

3. Speaker personalization

Original: multi-speaker model → Quantized: optimized for specific speaker

Speaker ID Test \ Speaker ID Calibration	1	2	3	4	5	6	7	8	9	10	No Calib
1	90.9	90.9	90.9	81.8	90.9	81.8	90.9	90.9	90.9	81.8	90.9
2	78.6	100	100	85.7	78.6	85.7	100	78.6	78.6	71.4	92.9
3	91.7	91.7	100	91.7	91.7	91.7	83.3	91.7	91.7	83.3	91.7
4	52.6	54.4	64.9	75.4	45.6	50.9	52.6	49.1	50.9	45.6	57.9
5	83.3	83.3	91.7	83.3	91.7	75.0	91.7	83.3	91.7	83.3	83.3
6	93.3	100	100	93.3	86.7	100	100	86.7	93.3	80.0	93.3
7	75.0	75.0	87.5	68.8	75.0	62.5	93.8	68.8	62.5	75.0	81.3
8	50.0	80.0	80.0	60.0	40.0	60.0	80.0	100	60.0	60.0	60.0
9	73.3	66.7	80.0	73.3	73.3	73.3	60.0	73.3	80.0	73.3	60.0
10	75.0	66.7	91.7	75.0	58.3	75.0	75.0	75.0	66.7	91.7	75.0

ACC on GSC with myQASR-W2V2-L-C.

Conclusion

- New task: personalized post-training model quantization to bring large speech models on low-resource devices with performance targeted for the final end user.
- New method: myQASR, a versatile personalized quantization scheme to compress large speech models to any memory budget.
- myQASR features:
  - Uniformity constraint to evaluate layer sensitivity,
  - (optional) Hessian guidance to set quantization scaling parameters,
  - A few user-specific unlabelled samples to drive the quantization process,
  - PTQ: personalizing the model performance with no fine-tuning.