INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

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

*Edward Fish*[1,2,*], *Umberto Michieli*[1], *Mete Ozay*[1]

[1]Samsung Research UK
[2]University of Surrey

edward.fish@surrey.ac.uk, {u.michieli, m.ozay}@samsung.com

## Abstract

Recent advancement in Automatic Speech Recognition (ASR) has produced large AI models, which become impractical for deployment in mobile devices. Model quantization is effective to produce compressed general-purpose models, however such models may only be deployed to a restricted sub-domain of interest. We show that ASR models can be personalized during quantization while relying on just a small set of unlabelled samples from the target domain. To this end, we propose myQASR, a mixed-precision quantization method that generates tailored quantization schemes for diverse users under any memory requirement with no fine-tuning. myQASR automatically evaluates the quantization sensitivity of network layers by analysing the full-precision activation values. We are then able to generate a personalised mixed-precision quantization scheme for any pre-determined memory budget. Results for large-scale ASR models show how myQASR improves performance for specific genders, languages, and speakers.

**Index Terms**: ASR, Compression, Quantization, Transformers.

## 1. Introduction

Automatic Speech Recognition (ASR) models improve user experience when interacting with mobile devices while widening the accessibility of such technologies [1, 2]. Recent innovation in ASR has focused on large and multi-purpose (*e.g.*, multilingual) transformer-based architectures [3, 4]. However, there has been less consideration for how these models can be effectively compressed and deployed for a diverse range of users and devices, whilst preserving privacy (*i.e.*, with anonymized data).

Model quantization can be divided into three categories [5, 6]: i) Quantization Aware Training (QAT) where the quantized model is fine-tuned on labelled data to recover accuracy [7, 8, 9, 10, 11, 12, 13], ii) Post-Training Quantization (PTQ) where labelled data is used to calibrate quantization parameters without training [14, 15, 16], and iii) Data-Free PTQ (DF-PTQ) where data is unlabelled (*i.e.*, label-free) or not available at all [17, 18]. We focus on label-free PTQ. DF-PTQ has been tackled in the computer vision (CV) domain via layerwise knowledge distillation (KD) between original and quantized layers for compressing CNNs [19, 20, 21]. These methods have been extended to Transformers for both vision [22, 23, 24] and ASR [11, 25, 26, 27], adjusting for GELU and multi-headed attention. In a deployed scenario, computational and time expensive KD methods are not practical as a copy of the full precision (FP) network is required. When unlabelled data is available, a few methods [14, 28] proposed to minimize distance between both quantized and FP weights and activations, or to minimize a task loss with respect to the quantization parameters. While these methods are promising in CV [28], their effectiveness has not been explored in ASR. More recently, some data-free methods use statistics of input data to generate synthetic data used to fine-tune quantized models [29, 30]; however, these methods cannot account for out-of-domain data which occur in deployment setups and require further training. In [31], this problem is approached via a diverse sample generation scheme; however, the method still requires fine-tuning on synthetic data, is sensitive to hyper-parameters controlling the skew of sample distributions, and has not been applied to ASR tasks.

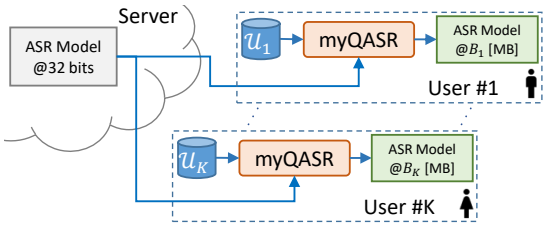


Figure 1: *Overview of myQASR. A large model is quantized according to users' audio data and their device storage budget.*

Another desirable property of personalized compressed ASR models, is the ability to set any target model size while preserving accuracy. Mixed precision (MP) quantization, where each layer is quantized to a different bit depth, allows the target model size to be controlled. Second-order information [32], NAS [33], and generative methods [34] are effective for low-bit MP settings in CV, but they either require multiple models to be stored in memory concurrently [33], or computationally expensive sensitivity detection [32, 34]. Current MP methods also require several hyper-parameters to search optimum bit-depth combinations or to set the min/max bit depths, which does not permit for fine-grained interpolation between model sizes since extreme values are set *a priori* and layer size is not considered.

To address these problems, we present myQASR, a system for personalized compression of ASR models, which performs fast layer-wise sensitivity detection to identify MP bit depths for a range of models (*e.g.*, large multi-purpose transformer models), user traits, and memory constraints. The scenario targeted by myQASR is depicted in Fig. 1, where a general-purpose FP ASR model is quantized for user $k \in [K]$ given a small dataset of unlabelled private user samples $\mathcal{U}_k$ and target storage budget in MB, $B_k$. Using $\mathcal{U}_k$, we find a good approximation of weight sensitivity to quantization by observing the median values of FP activations. Our approach is motivated by the activation change among different users (*e.g.*, male and female in Fig. 2): as such, models for different users require different quantization schemes to find a better compression-accuracy trade-off. We then experiment with some calibration methods to find the optimum scale and zero-point for the quantization

* Research completed during internship at Samsung Research UK.



10.21437/Interspeech.2023-61

function given the selected bit depths and statistics of data.

We report experimental validation on recent state-of-the-art architectures (*e.g.*, Wav2Vec2 [3] and Whisper [4]) with data segmented according to certain properties (*i.e.*, gender, language, and speaker identity) to demonstrate how our personalized model compression performs over a heterogeneous target.

To summarize, the main contributions of our work are: 1) We introduce myQASR: to our knowledge, the first method for personalized PTQ of ASR models. Our method requires only a few unlabelled user samples to adjust the quantization parameters without any fine-tuning. 2) myQASR breaks the common assumption of setting a minimum and a maximum value for the bit depth, and, instead, it relies on a uniformity constraint to guide the quantization process. 3) The uniformity constraint evaluates layer sensitivity in linear time complexity to identify candidate layers that can be further quantized to meet any predefined memory budget constraint to the nearest KB. 4) myQASR is the first ASR quantization approach that quantizes all parts of the network and supports integer bit shifting operations for matrix multiplication to ease on-device deployment.

## 2. Method

For simplicity, we drop the target user index $k$. Given a network pre-trained on a dataset composed of multiple data subsets and parametrized by $\mathcal{W} = \{W_l\}_{l=1}^{L}$ with $l \in [L]$ layers in the network, we aim to quantize the network to meet any storage budget $B$ minimizing the error rate for a specific target subset $\mathcal{U}$, of which only a few unlabeled samples are available, with $|\mathcal{U}| \leq 32$ in our experiments. myQASR can be employed in two stages: 1) layer-wise sensitivity detection - where we perform inference on $\mathcal{U}$ and collect statistics of the raw model, and 2) calibration, where we adjust network parameters based on $\mathcal{U}$.

**Sensitivity Detection.** To select MP bit-depths $\mathbf{b} \in \mathbb{Z}_{+}^{L}$, we compute outputs of each Conv and Linear layer[1] (*e.g.*, by inserting observers). We run inference over the unlabeled target dataset $\mathcal{U}$ storing the median values of output activations, $\mathbf{a} \in \mathbb{R}_{+}^{L}$, such that $a_l \triangleq \mathbf{a}[l]$ is the median of outputs obtained at the $l$-th layer $\mathbf{o}_l$. We empirically observed a positive correlation between the median of activations and quantization error of the layer with respect to the input samples. As shown in [35], at low bit-depths, uniform distribution of weights can reduce quantization error, thus the median measures distribution skew at FP which relates to quantization sensitivity. Then, we proceed to select the bit depths $\mathbf{b}$ according to the memory budget $B$, as described in Alg. 1, where $|W_l|$ counts parameters of $\mathbf{o}_l$.

myQASR performs inference only once with a small number of samples to select the MP scheme, making our method more memory and computationally efficient than distance- or loss- based metrics, such as [32, 36, 14, 22]. Once inference is performed, we take the absolute of median values from each layer (to measure how much they differ from 0 regardless of their sign) and search for the quantization scheme by reducing each layer by one bit until the budget is reached while retaining the highest degree of uniformity among bit depths between layers. This uniformity constraint during the sensitivity detection enables removal of any additional hyper-parameters that are required by other methods [37, 38, 23, 34] which set *a priori* range of bit widths for compression, *i.e.*, $\min(\mathbf{b})$ and $\max(\mathbf{b})$ denoting the minimum and maximum valued bits in $\mathbf{b}$.

**Quantization & Calibration.** Once we have obtained the optimum bit depth, we proceed with quantization. For each layer,

[1] Self-attention weights (*i.e.*, Queries, Keys, and Values) are considered separately, so they are quantized individually depending on budget.

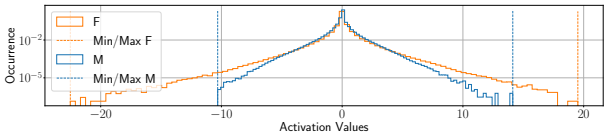


Figure 2: *Distribution of activations from the first convolution layer of Wav2Vec2 on female (F) and male (M) data.*

**Algorithm 1:** Sensitivity detection of myQASR.

**Data:** $B$ memory budget in MB, $M$ model size in MB ($M > B$), and $\mathcal{W}$ model parameters.
**Result:** Array $\mathbf{b}$ of selected bit depths.
$\mathbf{b} \leftarrow \{32, \ldots, 32\}$ *// initialize to FP.*
Compute median activations $\mathbf{a}$ over $\mathcal{U}$ $(a_l, \forall l \in [L])$;
$\hat{\mathbf{q}} \leftarrow \text{argsort}(\mathbf{a})$ *// 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.*
    $M = \texttt{ComputeModelSize}(\mathbf{b}, \mathcal{W})$
    **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.*

we quantize both weights, $W_l$, and inputs, $X_l$, using a quantization function $Q(\theta_l, b_l)$ where $\theta_l \in \{W_l, X_l\}$. The objective is to restrict the FP values of $\theta_l$ to finite integer values by scaling and rounding, as defined by

$$Q(\theta_l, b_l) = [\text{round}(\theta_l / S_l) - Z_l]_{b_l}, \tag{1}$$

where $\text{round}(\cdot)$ is the integer rounding operation, $Z_l$ corrects the zero point of the quantized output, $[\cdot]_{b_l}$ is the representation of $\cdot$ with $b_l$ bits, and $S_l$ is the scaling factor.

In standard uniform quantization [5], $S_l$ is defined by the maximum available values given by $b_l$, *i.e.*, $S_l = 2^{b_l - 1}$. Uniform quantization function $Q(\theta_l, b_l)$ is directly applicable for quantization of weights $W_l$, as they follow a Gaussian distribution. However, activations do not [35] (Fig. 2), so there may exist $S_l$ values which can minimize the quantization error more effectively. We employ three methods to find appropriate $S_l$ to scale activations, which we describe next.

1) The first method, called myQASR, inserts observers in the network to track the layer-wise minimum ($X_l^m$) and maximum ($X_l^M$) values of the input tensors at FP. The layer scale $S_l$ and zero point $Z_l$ for the input tensor are then obtained by

$$S_l = (X_l^M - X_l^m)/(2^{b_l - 1}), \tag{2}$$

$$Z_l = -2^{b_l - 1} - \text{round}(X_l^m / S_l). \tag{3}$$

We also experiment with other $S_l$, by minimizing the distance between quantized and FP output of layers as described in [39, 14]. In this setting, we define a range of possible values for $S_l$ defined by the minimum and maximum values given by $b_l$, and then take the distance between the FP output (*i.e.*, output vector $\mathbf{o}_l$) and quantized output ($\hat{\mathbf{o}}_l$) where $\hat{\mathbf{o}}_l = Q(W_l, b_l)^T Q(\mathbf{o}_{l-1}, b_l)$. In ablation studies, we evaluate a number of distance metrics for this calibration stage and find that the cosine distance is the most effective.

2) The second method (called myQASR-Hess) is driven by the recent observations [22] where quantization of GELU and softmax outputs benefits from asymetric non-uniform quantization schemes due to their non-Gaussian distribution. In [22], the authors utilise two quantization ranges per layer, $R_1^l$ and $R_2^l$ with scaling factors $S_{R_1^l}$ and $S_{R_2^l}$, where $R_1^l = [-2^{k-1} S_{R_1^l}, 0]$

and $R_2^l = [0, -2^{k-1}S_{R_2^l}]$ for post-GELU activations. Finding the optimum $S_{R_1^l}$ and $S_{R_2^l}$ can be performed via a linear search where the objective is to minimize the distance between quantized and FP output of each layer scaled by its impact on the task loss $\mathcal{L}(\hat{y}, y)$. Denoting $\mathbf{\Delta}_l = \hat{\mathbf{o}}_l - \mathbf{o}_l$, the Hessian-based calibration optimization for layer $l$ is defined by

$$\min_{S_{R_1^l}, S_{R_2^l}} \mathbb{E}_{\mathcal{U}} \left[ \mathbf{\Delta}_l^T \operatorname{diag} \left[ \left( \frac{\partial \mathcal{L}_{\mathcal{U}}}{\partial \mathbf{o}_i} \right)^2 \right]_{i=1}^{L} \mathbf{\Delta}_l \right]. \quad (4)$$

3) The third method (myQASR-Cosine) calibrates the model minimizing the cosine distance by

$$\min_{S_{R_1^l}, S_{R_2^l}} \mathbb{E}_{\mathcal{U}} \left[ \frac{\hat{\mathbf{o}}_l \cdot \mathbf{o}_l}{\|\hat{\mathbf{o}}_l\| \cdot \|\mathbf{o}_l\|} \right]. \quad (5)$$

As we will see, myQASR-Cosine provides the best results for personalized quantization at the cost of a remarked increase in calibration time and memory consumption which may not be available in certain deployment scenarios.

## 3. Experimental Analyses

**Datasets.** We employ 3 datasets, one for each personalization task. To demonstrate our method, we partition data into subsets. However, myQASR requires no *a priori* assumption on data split. We replicate a deployment scenario, where myQASR sees only a small amount of unlabelled target data from a user. *1) Gender-wise Personalization.* LibriSpeech (LS) [40] contains $\sim 1k$ hours of 16kHz English speech derived from audiobooks. We perform experiments on *test-clean*, creating Male (M) and Female (F) partitions, splitting audio data by speaker gender. *2) Language-wise Personalization.* FLEURS [41] contains 102 languages, each with $\sim$ 12 hours of speech. We select 10 among the top performing languages for the Whisper architecture. For each experiment, we randomly sample 32 full spoken sentences for calibration. *3) Speaker-wise Personalization.* Google Speech Commands (GSC) [42] is designed to evaluate effectiveness of keyword spotting models. The dataset contains one second long audio files from multiple speakers. We use the test partition v0.01, which contains 3081 spoken words. For calibration, we select 5 words from each speaker and test performance on all the available test data per speaker.

**Models.** We use Wav2Vec2 (W2V2) [3] and Whisper [4] to evaluate our method. For all models, we quantize all weights and activations. On LS, we use a pre-trained W2V2 base (B) model fine-tuned on 960 hours of clean English audio data from LS. For Keywords Spotting (KWS), we use a W2V2-Large-Conformer (W2V2-L-C), as described in [43], pre-trained on GSC. Whisper [4] is a multi-lingual sequence to sequence (encoder/decoder) Transformer model pre-trained on 680K hours of multi-lingual supervised audio data. We experiment on the Whisper-L, *i.e.*, the large variant. We use Word/Character Error Rate (WER/CER) for ASR tasks, and we use accuracy (ACC) for KWS. Average bit depth is denoted by $\bar{b}$.

**Main Results on Gender.** Fig. 3 shows WER of W2V2-B pre-trained on multi-gender data, quantized for a female user and tested on LS-F. The plot spans an increasing memory budget from 60MB (*i.e.*, $\bar{b} \approx 5$) to 90MB (*i.e.*, $\bar{b} \approx 8$). Original (*i.e.*, FP, non-quantized) model performance indicates the lower bound for WER (gray dashed line). Uniform quantization yields competitive results, however cannot meet fine-grained memory requirements. Our simplest myQASR calibrated on female data, myQASR (F), improves accuracy and can meet any desired target model size. We argue on the usefulness of our sensitivity method since: myQASR (F) shows significant benefits

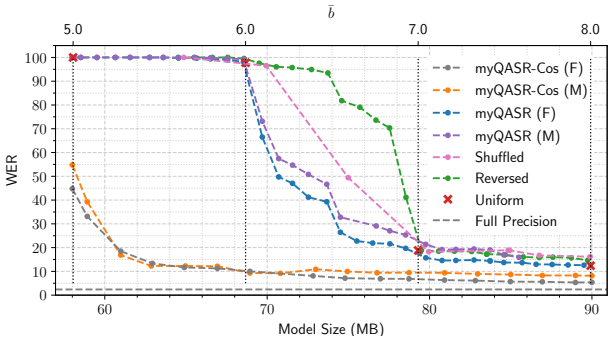


Figure 3: *WER of W2V2-B on LS-F. Original model is 360MB.*

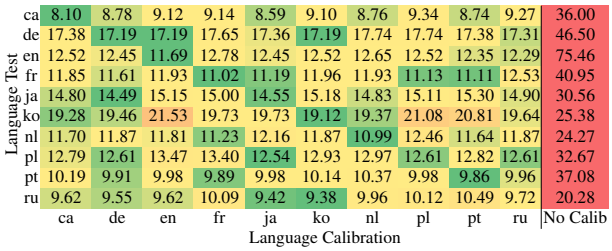


Figure 4: *WER on FLEURS with myQASR-Whisper-L.*

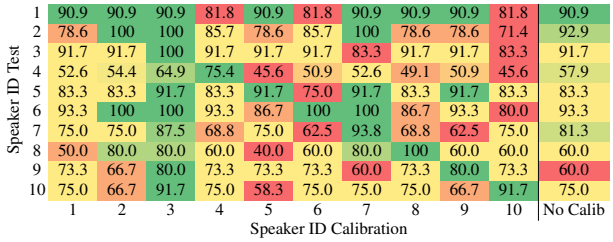


Figure 5: *KWS ACC on GSC with myQASR-W2V2-L-C.*

compared to myQASR (M), *i.e.*, quantizing the model according to male data; myQASR (F) outperforms its shuffled (*i.e.*, bit-depths shuffled) or reversed (*i.e.*, bit-depths reversed) versions by large margin. Cosine-based calibration brings large benefits, reducing the gap from the FP model. Nonetheless, calibration on female data still outperforms calibration on male data.

**Main Results on Language.** In Fig. 4, we show personalized compression in multi-lingual settings. We take the pre-trained multi-lingual Whisper-L model and calibrate bit-depths and activation ranges using just 32 samples of unlabelled data. Each language label represents a tune and test split, and we show that calibrating bit depths and activations for the same language leads to improved results. Although we obtain better results on the same language used for calibration (on-diagonal results), we remark that the resulting model still achieves competitive results on other languages (off-diagonal results); thus being able to predict also on such languages. In the worst case (*i.e.*, Russian), our method is outperformed by calibration on other languages. However, it shows a relative gain of $0.9\%$ compared to the average of other-language results. In the best case (*i.e.*, Catalan), our method outperforms the average of other-language by $10.9\%$ relative gain. On average, our same-language myQASR yields $66.2\%$ better results than standard uniform quantization with no calibration (12.5% vs. 36.9% WER), and $4.2\%$ better results than other-language quantization (13.0% WER).

**Main Results on Speaker.** In Fig. 5, we show ACC for our myQASR applied to a W2V2-L-C [43] compressed from 2.4GB to 375MB (*i.e.*, $\bar{b} = 5$bits). We partition GSC by speaker ID and evaluate on each ID with calibration data from different speakers. We show that, when sensitivity and calibration anal-

Table 1: *Ablation on min-max (mm) MP bit depths selection. Size: in MB and min-max values within brackets (min-max), my: myQASR.*

| | size | $\text{WER}_{\text{mm}}$ | $\text{WER}_{\text{my}}$ | size | $\text{WER}_{\text{mm}}$ | $\text{WER}_{\text{my}}$ | size | $\text{WER}_{\text{mm}}$ | $\text{WER}_{\text{my}}$ | size | $\text{WER}_{\text{mm}}$ | $\text{WER}_{\text{my}}$ | size | $\text{WER}_{\text{mm}}$ | $\text{WER}_{\text{my}}$ | size | $\text{WER}_{\text{mm}}$ | $\text{WER}_{\text{my}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **M** | 82.5 (5-7) | 6.6 | 4.7 | 87.7 (6-7) | 5.6 | 4.3 | 81.9 (4-8) | 6.9 | 6.6 | 87.5 (5-8) | 6.4 | 4.3 | 93.2 (6-8) | 5.4 | 4.1 | 98.4 (7-8) | 4.2 | 4.2 |
| **F** | 82.3 (5-7) | 7.4 | 5.3 | 97.7 (6-7) | 5.5 | 4.9 | 82.1(4-8) | 7.1 | 6.2 | 87.5 (5-8) | 7.0 | 4.9 | 93.0 (6-8) | 5.3 | 4.7 | 98.2 (7-8) | 4.6 | 4.6 |

ysis is performed on the same speaker, we achieve optimum performance. For example, in the best case (*i.e.*, speaker $\#7$), we achieve $100\%$ ACC when compression is personalized for that speaker, compared with $40\%$ ACC when personalized for another speaker from the same dataset, even though keywords are the same. On average, our same-speaker myQASR yields 17.5% higher results than standard uniform quantization with no calibration (92.3% vs. 78.6% ACC), and 19.6% higher results than other-speaker quantization (77.2% ACC).

# 4. Ablation Study

Ablation is performed on W2V2-B compressed via myQASR-Cosine to 75MB, i.e $\bar{b} = 6.5$, (unless otherwise stated) on gender data calibrated and tested on the same split.

**Bit Depth Selection & Uniformity Constraint.** Tab. 1 shows a comparison between our uniform constraint and the common min-max method [32, 34], where min-max values are chosen for depths according to a linear interpolation mapping highest (lowest) activation value to the min (max) bit depth. This approach leads to significantly lower results at fixed target compression ratios than ours. This shows the advantage of enforcing some uniformity among layers in the network. Using min-max bit depths also requires two more hyper-parameters that we avoid thanks to the sensitivity evaluation scheme, as discussed next.

**Sensitivity** is evaluated in Tab. 2. We grouped methods as reduction-based or distance-based. Reduction-based methods compute an aggregate measure of the distribution of activations obtained using the original model, namely: average, median (ours), max, max of the absolute and standard deviation. Distance-based methods compute a distance measure between layer-wise activations obtained using the quantized and original model, namely: L1, L2, Spectral norm, Frobenius norm, and KL divergence. As in the reduction setting, the values are sorted in order of increasing distance and then used to assign bit depths per layer. Reduction-based methods are more practical than distance-based ones as they do not require both the quantized and original model, and can achieve performance comparable to distance-based methods. Among reduction-based approaches, median provides the best results. We reason that the median provides a measure of distribution skew at FP which correlates with quantization sensitivity as described in [35].

**Calibration** is evaluated in Tab. 3. To evaluate it, we use a number of distance functions which compute the quantization error between FP and quantized activations. For each metric, apart from myQASR variations, we generate $t$ possible quantization scales per each layer and minimize the error defined by the distance metric ($t = 100$, as in [22]). Cosine and Hessian weighted myQASR perform the best but have a high computational search time. L1 achieves competitive results, however its WER is outperformed by myQASR-Hess with a similar calibration time. myQASR, which performs a simple min/max calibration as described in Sec. 2, provides a trade-off between accuracy and computational time, and does not require linear search or any additional hyper-parameters as the other methods.

We study the amount of unlabelled target data needed in Tab. 4 and verify that a few samples are sufficient for calibration, as reported in [44]. For robustness, we choose 32 samples, since variability of results is minimized (*i.e.*, low standard deviation). Similarly, we perform personalized compression via 5 samples for speakers and 32 for languages.

Table 2: *Ablation on the sensitivity scheme. We take only the measures used in compared methods for a fair comparison.*

| **Reduction** | **Male** | | **Female** | | **Distance** | **Male** | | **Female** | |
|---|---|---|---|---|---|---|---|---|---|
| | **WER** | **CER** | **WER** | **CER** | | **WER** | **CER** | **WER** | **CER** |
| Avg | 7.5 | 2.4 | 7.3 | 2.3 | L1 [22] | 7.3 | 2.3 | 7.2 | **2.2** |
| Median (ours) | **6.6** | **2.1** | **7.1** | **2.2** | L2 [22] | 7.2 | 2.2 | 7.2 | **2.2** |
| Max | 7.1 | 2.2 | 7.8 | 2.4 | SN [14] | 7.3 | 2.3 | 7.3 | **2.2** |
| Max Abs | 7.2 | 2.3 | 8.4 | 2.7 | Frob [36] | 7.3 | 2.3 | 7.2 | **2.2** |
| Std | 7.5 | 2.4 | 7.4 | 2.3 | KL [22] | 7.3 | 2.3 | 8.4 | 2.7 |

Table 3: *Ablation on the calibration scheme. Time (sec) measures calibration procedure only.*

| | **Male** | | | **Female** | | |
|---|---|---|---|---|---|---|
| | **WER** | **CER** | **Time** | **WER** | **CER** | **Time** |
| None | 87.5 | 67.2 | 0 | 66.2 | 83.0 | 0 |
| L1 [22] | 8.2 | 2.7 | 158 | 9.8 | 2.8 | 147 |
| L2 [22] | 63.7 | 3.4 | 155 | 57.5 | 29.4 | 156 |
| LinW L2 [22] | 89.5 | 62.9 | 154 | 92.2 | 70.2 | 155 |
| SqW L2 [22] | 94.7 | 81.0 | 154 | 97.1 | 82.9 | 155 |
| myQASR | 28.8 | 28.7 | **7** | 22.7 | 22.6 | **7** |
| myQASR-Hess | 7.9 | 2.5 | 161 | 7.9 | 2.4 | 154 |
| myQASR-Cosine | **6.6** | **2.1** | 172 | **7.1** | **2.2** | 171 |

Table 4: *Ablation on $|\mathcal{U}|$, results averaged over 5 seeds.*

| $\lvert\mathcal{U}\rvert$ | **4** | **8** | **16** | **32** | **64** | **128** |
|---|---|---|---|---|---|---|
| **WER (M)** | **6.1**$\pm$2.0 | 7.0$\pm$0.4 | 6.9$\pm$0.8 | 6.6$\pm$**0.1** | 8.4$\pm$0.8 | 8.5$\pm$1.4 |
| **WER (F)** | 7.1$\pm$1.6 | 7.0$\pm$1.0 | **6.6**$\pm$1.0 | 7.1$\pm$**0.2** | 6.9$\pm$0.8 | 6.8$\pm$0.5 |

Table 5: *Ablation on activation quantization.*

| | **Male** | | | | | | | **Female** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Act bits** | **4** | **6** | **8** | **10** | **12** | **16** | **32** | **4** | **6** | **8** | **10** | **12** | **16** | **32** |
| **WER** | 69.0 | 8.1 | **6.6** | 7.0 | 7.0 | 7.1 | 7.1 | 70.0 | 8.5 | **7.1** | 7.5 | 7.4 | 7.4 | 7.4 |
| **CER** | 34.7 | 2.7 | 2.2 | 2.2 | 2.2 | **2.1** | **2.1** | 3.6 | 2.7 | **1.9** | 2.3 | 2.3 | 2.3 | 2.3 |

**Activation** quantization is evaluated in Tab. 5 (also weights are quantized). As expected, we observe performance improvements when quantization decreases. We select 8-bit quantization following previous approaches [36] and verify that it represents a good trade-off between compression and accuracy. We note how female-specific models are more sensitive to quantization which is a reflection of an original bias of the FP model.

# 5. Conclusions

We introduced a new task of personalized model quantization to bring large ASR transformers on low-resource devices such as mobile and edge devices with performance targeted for the final end user. To address this, we propose myQASR, a versatile personalized quantization scheme to compress large ASR models to any memory budget. myQASR couples a uniformity constraint to evaluate layer sensitivity with optional Hessian guidance to set quantization scaling parameters. It requires only a few user-specific unlabelled samples to drive the quantization process personalizing the model performance with no fine-tuning. Our work provides a baseline for future research in personalized compression focusing on greater accessibility and performance for diverse users and devices.

# 6. References


[1] D. Yu and L. Deng, *Automatic speech recognition*. Springer, 2016, vol. 1.

[2] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *MTA*, vol. 80, no. 6, pp. 9411–9457, 2021.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv:2212.04356*, 2022.

[5] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," *arXiv:2103.13630*, 2021.

[6] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021.

[7] H. D. Nguyen, A. Alexandridis, and A. Mouchtaris, "Quantization aware training with absolute-cosine regularization for automatic speech recognition." in *Interspeech*, 2020, pp. 3366–3370.

[8] Y. Mishchenko, Y. Goren, M. Sun, C. Beauchene, S. Matsoukas, O. Rybakov, and S. N. P. Vitaladevuni, "Low-bit quantization and quantization-aware training for small-footprint keyword spotting," in *ICMLA*. IEEE, 2019, pp. 706–711.

[9] T. Allenet, D. Briand, O. Bichler, and O. Sentieys, "Disentangled loss for low-bit quantization-aware training," in *CVPR*, 2022, pp. 2788–2792.

[10] S. Ding, P. Meadowlark, Y. He, L. Lew, S. Agrawal, and O. Rybakov, "4-bit conformer with native quantization aware training for speech recognition," *arXiv:2203.15952*, 2022.

[11] A. Bie, B. Venkitesh, J. Monteiro, M. Haidar, M. Rezagholizadeh *et al.*, "A simplified fully quantized transformer for end-to-end speech recognition," *arXiv:1911.03604*, 2019.

[12] K. Zhen, H. D. Nguyen, R. Chinta, N. Susanj, A. Mouchtaris, T. Afzal, and A. Rastrow, "Sub-8-Bit Quantization Aware Training for 8-Bit Neural Network Accelerator with On-Device Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 3033–3037.

[13] A. Fasoli, C.-Y. Chen, M. Serrano, S. Venkataramani, G. Saon, X. Cui, B. Kingsbury, and K. Gopalakrishnan, "Accelerating Inference and Language Model Fusion of Recurrent Neural Network Transducers via End-to-End 4-bit Quantization," in *Proc. Interspeech 2022*, 2022, pp. 2038–2042.

[14] Z. Liu, Y. Wang, K. Han, W. Zhang, S. Ma, and W. Gao, "Post-training quantization for vision transformer," *NeurIPS*, vol. 34, pp. 28 092–28 103, 2021.

[15] M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, and T. Blankevoort, "Up or down? adaptive rounding for post-training quantization," in *ICML*, 2020, pp. 7197–7206.

[16] I. Hubara, Y. Nahshan, Y. Hanani, R. Banner, and D. Soudry, "Accurate post training quantization with small calibration sets," in *ICML*, 2021, pp. 4466–4475.

[17] M. Nagel, M. v. Baalen, T. Blankevoort, and M. Welling, "Data-free quantization through weight equalization and bias correction," in *ICCV*, 2019, pp. 1325–1334.

[18] K. Choi, D. Hong, N. Park, Y. Kim, and J. Lee, "Qimera: Data-free quantization with synthetic boundary supporting samples," *NeurIPS*, vol. 34, pp. 14 835–14 847, 2021.

[19] N. Aghli and E. Ribeiro, "Combining weight pruning and knowledge distillation for cnn compression," in *CVPR*, 2021, pp. 3191–3198.

[20] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *IJCV*, vol. 129, no. 6, pp. 1789–1819, 2021.

[21] J. Jin, C. Liang, T. Wu, L. Zou, and Z. Gan, "KDLSQ-BERT: A quantized bert combining knowledge distillation with learned step size quantization," *arXiv:2101.05938*, 2021.

[22] Z. Yuan, C. Xue, Y. Chen, Q. Wu, and G. Sun, "Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization," in *ECCV*. Springer, 2022, pp. 191–207.

[23] Z. Li and Q. Gu, "I-vit: integer-only quantization for efficient vision transformer inference," *arXiv:2207.01405*, 2022.

[24] Y. Lin, T. Zhang, P. Sun, Z. Li, and S. Zhou, "FQ-ViT: Post-training quantization for fully quantized vision transformer," in *IJCAI*, 2022, pp. 1173–1179.

[25] I. Zharikov, I. Krivorotov, V. Alexeev, A. Alexeev, and G. Odinokikh, "Low-bit quantization of transformer for audio speech recognition," in *Neuroinformatics*. Springer, 2022, pp. 107–120.

[26] N. Wang, C.-C. Liu, S. Venkataramani, S. Sen, C.-Y. Chen, K. El Maghraoui, V. Srinivasan, and L. Chang, "Deep compression of pre-trained transformer models," in *NeurIPS*, 2022.

[27] G. Odinokikh, "Low-bit quantization of transformer," in *Neuroinformatics*, vol. 1064. Springer Nature, 2022, p. 107.

[28] Y. Li, S. Xu, B. Zhang, X. Cao, P. Gao, and G. Guo, "Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer," *arXiv:2210.06707*, 2022.

[29] S. Kim, A. Gholami, Z. Yao, N. Lee, P. Wang, A. Nrusimha, B. Zhai, T. Gao, M. W. Mahoney, and K. Keutzer, "Integer-only zero-shot quantization for efficient speech recognition," in *ICASSP*. IEEE, 2022, pp. 4288–4292.

[30] Z. Yao, R. Y. Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers," *arXiv:2206.01861*, 2022.

[31] X. Zhang, H. Qin, Y. Ding, R. Gong, Q. Yan, R. Tao, Y. Li, F. Yu, and X. Liu, "Diversifying sample generation for accurate data-free quantization," in *CVPR*, 2021, pp. 15 658–15 667.

[32] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Hawq: Hessian aware quantization of neural networks with mixed-precision," in *ICCV*, 2019, pp. 293–302.

[33] C. Zhao, T. Hua, Y. Shen, Q. Lou, and H. Jin, "Automatic Mixed-Precision Quantization Search of BERT," in *IJCAI*, 2021.

[34] Y. Cai, Z. Yao, Z. Dong, A. Gholami, M. W. Mahoney, and K. Keutzer, "ZeroQ: A novel zero shot quantization framework," in *CVPR*, 2020, pp. 13 169–13 178.

[35] H. Yu, T. Wen, G. Cheng, J. Sun, Q. Han, and J. Shi, "Low-bit quantization needs good distribution," pp. 680–681, 2020.

[36] S. B. Eryilmaz and A. Dundar, "Understanding how orthogonality of parameters improves quantization of neural networks," *IEEE TNNLS*, pp. 1–10, 2022.

[37] W. Chen, P. Wang, and J. Cheng, "Towards mixed-precision quantization of neural networks via constrained optimization," in *ICCV*, 2021, pp. 5350–5359.

[38] Z. Dong, Z. Yao, D. Arfeen, A. Gholami, M. W. Mahoney, and K. Keutzer, "Hawq-v2: Hessian aware trace-weighted quantization of neural networks," *NeurIPS*, vol. 33, pp. 18 518–18 529, 2020.

[39] D. Wu, Q. Tang, Y. Zhao, M. Zhang, Y. Fu, and D. Zhang, "EasyQuant: Post-training quantization via scale optimization," *arXiv:2006.16669*, 2020.

[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[41] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," *arXiv:2205.12446*, 2022.

[42] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv:1804.03209*, 2018.

[43] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "FAIRSEQ S2T: Fast speech-to-text modeling with FAIRSEQ," *arXiv:2010.05171*, 2020.

[44] "NNI Documentation," https://nni.readthedocs.io/en/latest/_modules/nni/compression/pytorch/quantization/observer_quantizer.html, Accessed: 2023-01-25.