Lightweight Neural Networks for Speech Emotion Recognition using Layer-wise Adaptive Quantization

Tushar Shinde, Ritika Jain, Avinash Kumar Sharma
School of Engineering and Science, Indian Institute of Technology Madras Zanzibar, Tanzania
shinde@iitmz.ac.in

Abstract

Speech Emotion Recognition (SER) systems are essential in advancing human-machine interaction. While deep learning models have shown substantial success in SER by eliminating the need for handcrafted features, their high computational and memory requirements, alongside intensive hyper-parameter optimization, limit their deployment on resource-constrained edge devices. To address these challenges, we introduce an optimized and computationally efficient Multilayer Perceptron (MLP)-based classifier within a custom SER framework. We further propose a novel, layer-wise adaptive quantization scheme that compresses the model by adjusting bit-width precision according to layer importance. This layer importance is calculated based on statistical measures such as parameter proportion, entropy, and weight variance within each layer. Our approach achieves an optimal balance between model size reduction and performance retention, ensuring that the quantized model maintains accuracy within acceptable limits. Traditional fixed-precision methods, while computationally simple, are less effective at reducing model size without compromising performance. In contrast, our scheme provides a more interpretable and computationally efficient solution. We evaluate the proposed model on standard SER datasets using features such as Mel-Frequency Cepstral Coefficients (MFCC), Chroma, and Mel-spectrogram. Experimental results demonstrate that our adaptive quantization method achieves performance competitive with state-of-the-art models while significantly reducing model size, making it highly suitable for deployment on edge devices.

1 Introduction

Speech Emotion Recognition (SER) is a challenging task that aims to identify emotional states conveyed in human speech. It has a wide range of applications, including virtual assistants, social robots, lie detection, call center analytics, mental health monitoring, fitness tracking, and human-computer interaction (1). Recent advancements in SER have been driven by deep learning methods, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks (2), and transformers (3). These models have improved emotion classification accuracy due to their ability to leverage large annotated datasets for training and evaluation. Furthermore, the development of large language models (LLMs) (e.g., GPT-3, GPT-4, PALM, and Gemini) (4; 5; 6) and pre-trained speech models (e.g., wav2vec (7), HuBERT (8), wavLM (9), Whisper (10), and Conformer (11; 12)) has underscored the need for efficiency improvements in SER systems. However, the large size of these models, often containing billions of parameters, poses challenges related to computational complexity, deployment cost, and environmental impact (13). Optimizing model architectures and developing hardware-aware solutions are thus essential for real-time SER applications.

4th NeurIPS Efficient Natural Language and Speech Processing Workshop (ENLSP-IV 2024).

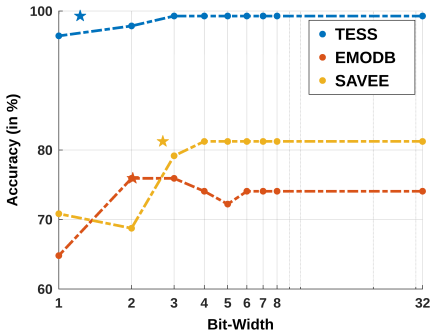


Figure 1: Comparison of model accuracy between various fixed-quantization methods and our layer-wise adaptive quantization approach (AQ) across three benchmark datasets: TESS, EMODB, and SAVEE. *(The performance of proposed LAQ approach is indicated by the star pattern).*

Feature selection plays a critical role in capturing the emotional information within speech, which can vary significantly in both intensity and expression. Among commonly used features, Mel-frequency cepstral coefficients (MFCCs) are prominent (14), along with Mel-spectrograms and Chroma features (3; 15). CNNs and LSTMs, often paired with attention mechanisms, are widely used in SER models (16; 17; 18; 19). For instance, Zhao et al. (17) employed an attention-based Bidirectional LSTM model to capture spatio-temporal emotional features, while other studies have adopted multimodal approaches integrating both audio and text data to enhance classification accuracy (20; 21). Despite their success, these models typically involve a large number of parameters, resulting in significant model size and computational demand (22).

To address these limitations, various compression techniques have been proposed, including model pruning (23; 24), low-rank factorization (25), knowledge distillation (26), and quantization (27; 28). Although much of the work on model compression has focused on image datasets such as MNIST (29) and CIFAR (30), some studies (31; 32; 33; 34; 35) have begun exploring these techniques in SER contexts. Our work extends these efforts by proposing an adaptive quantization scheme tailored to SER neural networks, aiming to reduce model size by lowering bit precision based on layer importance. Quantization is particularly valuable for model compression due to its efficient storage and memory utilization (28). Most current quantization methods use a uniform bit-width precision across layers, which can result in accuracy losses in complex tasks (36). Although mixed-precision quantization has been explored as a solution, it presents challenges in determining optimal bit-widths per layer, often leading to suboptimal trade-offs between compression and accuracy (37).

In this paper, we propose a lightweight Multilayer Perceptron (MLP) model with only three hidden layers to recognize various emotions in speech, including anger, fear, disgust, happiness, surprise, sadness, and neutrality. We introduce a framework for adaptive layer-wise quantization, as illustrated in Fig. 1, to demonstrate the effectiveness of our approach compared to fixed-quantization methods. As expected, model accuracy declines with higher quantization levels in fixed-precision methods; however, our approach retains accuracy while achieving significantly lower average bit-widths.

The main contributions of this work are as follows:

- We propose a lightweight MLP model with three fully connected layers, containing approximately $169K$ parameters, for speech emotion recognition. The model is evaluated on popular benchmark datasets, showing performance improvements over previous work.
- We introduce an efficient layer importance computation based on statistical measures such as parameter proportion, entropy, and variance.
- We develop an adaptive layer-wise quantization method that assigns different bit-widths to individual layers based on their importance, employing an iterative search algorithm to fine-tune bit precision based on an adaptive margin threshold for each layer.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 provides a detailed description of the proposed method. Section 4 outlines the experimental setup, and Section 5 presents the results. Finally, Section 6 concludes the paper.

2 Related Work

The advent of deep learning has enabled the development of models capable of automatically learning representations from raw audio data, significantly enhancing recognition accuracy in Speech Emotion Recognition (SER) (38; 39; 40). Prominent architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are well-suited for capturing the temporal dependencies intrinsic to speech signals. For example, Zhao et al. (17) employed an attention-based Bidirectional Long Short-Term Memory (LSTM) model that enhances SER by capturing essential spatio-temporal features. Similarly, Du et al. (18) demonstrated that attention mechanisms significantly improve SER performance, highlighting their effectiveness in processing temporal and contextual information in speech. Li et al. (20) demonstrated substantial performance gains using temporal alignment mean-max pooling mechanism to capture the subtle and fine-grained emotions implied in every utterance, providing a context-rich understanding of emotions.

Feature extraction and selection techniques also play a crucial role in the performance of SER (60; 61). A study by Krishnan et al. (61), utilized empirical mode decomposition and non-linear features to improve the SER effectiveness. Further, studies by Islam et al. (52) and Swain et al. (53) explored various handcrafted features for enhancing SER accuracy. Further, some studies have explored transfer learning for the task of emotion recognition in speech. Latif et al. (62) applied transfer learning to improve SER accuracy in cross-corpus and cross-language scenarios, while Wen et al. (63) and Ye et al. (65) developed frameworks that integrate transfer learning and gated multi-scale temporal convolution for cross-corpus SER tasks.

Though deep architectures have expanded SER capabilities, simpler models such as Multilayer Perceptron (MLP) remain relevant due to their flexible architecture and reasonable performance. The main issue with the traditional MLP is its huge number of parameters and model-size, thereby limiting the application in resource-constrained environments (43; 42). This challenge has spurred research into compression techniques aimed at developing lightweight yet accurate MLP classifiers for SER. Compression techniques such as pruning, quantization, and knowledge distillation have become central to optimizing SER models. Although model compression is well-explored in image classification, its application in SER is relatively less explored. Among them, quantization reduces model memory usage by decreasing weight precision, with uniform and dynamic quantization techniques showing promising results in reducing storage needs (46). However, quantization can introduce errors that degrade accuracy, especially in deeper networks. Despite the progress made with compression techniques, many SER models continue to face challenges related to model size and inference speed. Existing methods frequently apply uniform quantization across all layers, which can lead to suboptimal performance in complex tasks by disregarding the varying importance of individual layers (36). Additionally, most approaches rely on fixed-precision quantization, which limits the potential to balance compression and accuracy effectively (37).

To address these limitations, our work introduces a lightweight MLP architecture with an adaptive layer-wise quantization approach. This method selects different bit-widths for individual layers based on their importance, reducing model size while maintaining the performance. Our approach aims to make SER models both efficient and practical for real-world applications where computational resources are limited.

3 Methodology

3.1 Proposed Framework

Figure 2 depicts our novel framework for efficient speech emotion recognition, integrating audio feature extraction, a compact multi-layer perceptron (MLP) model, and an adaptive quantization approach to optimize performance and minimize model size. The framework begins with raw audio input, from which features like MFCC, Mel-spectrogram, and chroma are extracted. These features are then processed by a lightweight MLP architecture, designed to balance performance with efficiency. An adaptive quantization strategy, guided by layer importance, optimizes the precision of weights in each layer, reducing model size without sacrificing accuracy. This process consists of (i) computing layer importance, (ii) performing iterative bit precision optimization, and (iii) applying layer-wise quantization. The quantized model is thus compressed to retain high performance while being more memory-efficient. The framework's final stage involves classifying the speech sample into predefined

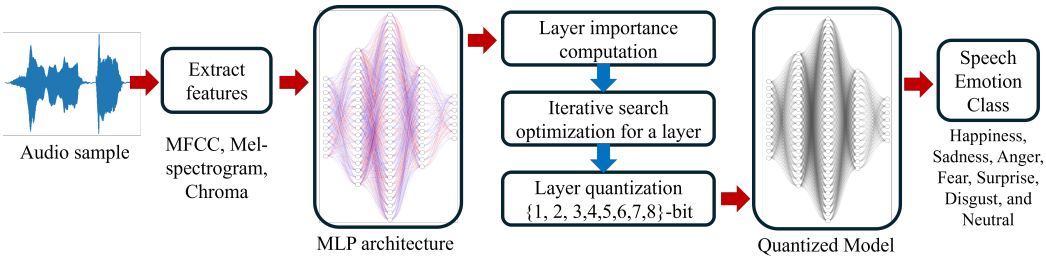


Figure 2: The framework of the proposed LAQ approach.

emotion categories. By combining effective feature extraction with an optimized MLP architecture and a layer-adaptive quantization (LAQ) approach, this framework supports resource-constrained deployments and real-world applications.

3.2 Layer Importance-Guided Adaptive Quantization (LAQ) Strategy

Our Layer Importance-Guided Adaptive Quantization (LAQ) strategy introduces a layer-wise quantization approach and an iterative optimization process aimed at minimizing accuracy loss while reducing model size. This strategy enables the deployment of deep neural networks (DNNs) on resource-limited devices by selecting bit precision per layer based on its relative importance, thus retaining performance close to original accuracy. The primary objective is to adaptively assign bit-widths across layers to achieve a balance between model size and accuracy. Motivated by the observation that layers contribute unequally to the model's final accuracy (48), we analyzed quantization sensitivity per layer by selectively quantizing each layer while maintaining 8-bit precision for others. This revealed that layers vary in sensitivity to quantization, which we leverage to assign different bit-widths based on computed importance. Given the challenge of optimal quantization order, we introduce a layer ranking mechanism grounded in layer importance.

3.2.1 Layer Importance Computation

Layer importance is computed by evaluating the following three metrics, each designed to balance bit precision with model accuracy:

Parameters Proportion: The parameter count per layer significantly influences model size. Layers with more parameters are prioritized for quantization to minimize model size, while layers with fewer parameters retain higher precision. We define the parameter proportion as:

$$N_P(l) = \frac{\text{Parameters in layer } l}{\text{Total parameters in the model}} \tag{1}$$

Normalized Entropy: The entropy of a layer indicates its information content, with higher entropy suggesting a need for greater bit precision. We compute the normalized entropy as:

$$N_E(l) = \frac{\text{Entropy of layer } l}{\text{Bit-precision of the model}} \tag{2}$$

Normalized Variance: The weight distribution within each layer affects its quantization sensitivity. Layers with higher variance may require advanced quantization to maintain performance. The normalized variance is defined as:

$$N_V(l) = \log\left(e - 1 + \frac{\text{Variance of layer } l}{\max_k \left(\text{Variance of layer } k\right)}\right) \tag{3}$$

Layer Importance Calculation: The importance of a layer l is quantified as a weighted sum of the above criteria, where:

$$Importance(l) = w_P \cdot N_P(l) + w_E \cdot (1 - N_E(l)) + w_V \cdot N_V(l) \tag{4}$$

Here, w_P, w_E, and w_V are weights for each criterion in the importance score, with $N_P(l)$ denoting the proportion of parameters, $N_V(l)$ reflecting weight variance, and $1 - N_E(l)$ adjusting importance inversely with entropy. These weights determine the relative impact of each criterion on layer importance.

3.2.2 Layer-wise Adaptive Quantization

Bit-width selection per layer is optimized through a search process that aims to minimize overall bit-width without sacrificing accuracy. Layers are ranked by importance, and quantization starts from the highest-ranked layers. A bit-width search begins from the minimum value and continues until performance degradation remains within a predefined threshold margin $T_{margin}(l)$, which adapts to layer importance as follows:

$$T_{margin}(l) = T_{margin} \times \frac{Importance(l)}{Importance(l_{prevQuantized})} \tag{5}$$

where $T_{margin}(l)$ is the threshold margin for layer l, $Importance(l)$ represents the importance of the current layer l, and $Importance(l_{prevQuantized})$ is the importance of the previously quantized layer. The importance ratio ensures controlled adaptation of the threshold margin, allowing efficient compression with minimal accuracy impact. This iterative adjustment optimally balances bit precision per layer.

4 Experimental Setup

4.1 Datasets

To evaluate the proposed Layer-wise adaptive quantization (LAQ) method, we conducted experiments on three benchmark Speech Emotion Recognition (SER) datasets: EMODB, SAVEE, and TESS. The EMODB dataset (49) consists of emotional expressions from 10 German speakers (5 male, 5 female), covering seven emotions: anger, boredom, disgust, fear, happiness, sadness, and neutral. The SAVEE dataset (50) includes recordings from 4 British male speakers expressing seven emotions: neutral, anger, disgust, fear, happiness, sadness, and surprise. Lastly, the TESS dataset (51) features recordings from two female actresses (aged 26 and 64) in seven emotions: anger, disgust, happiness, sadness, neutral, pleasant surprise, and fear.

4.2 Implementation Details

In this work, instead of directly using raw audio signals as input to the Multi-Layer Perceptron (MLP) model (52), we followed existing studies (53) to extract representative audio features. Specifically, we extracted Mel-frequency cepstral coefficients (MFCCs), Mel-spectrogram, and chroma features. The MLP model was implemented in PyTorch, with its hyperparameters detailed in Table 1. All experiments were conducted on the Kaggle platform, utilizing CPUs for model training. The dataset was split into training (80%), validation (10%), and test (10%) sets, with stratification applied to ensure balanced representation of each emotion across splits.

Hyperparameters	**Values**
Optimizer	Adam
Loss function	Cross-Entropy Loss
Batch size	32
Regularization	Early stopping (patience=5 epochs)
Learning rate	0.001
Dropout rate	0.1

Table 1: Training hyperparameters of the MLP model.

4.3 Evaluation Metrics

The primary evaluation metric for model performance is classification accuracy, enabling comparison with existing methods. Additionally, we report the average bit-width as a metric to assess model size, especially in quantized models, capturing the trade-off between model size and performance. The average bit-width provides a measure of bit-precision used across all layers of the model, indicating the extent of model compression. For a model with L layers, the average bit-width $\bar{b}$ is calculated as:

$$\bar{b} = \sum_{l=1}^{L} N_P(l) \cdot b(l) \tag{6}$$

where $N_P(l)$ is the parameter proportion for layer l, and $b(l)$ is the bit-width of parameters in layer l. Thus, $\bar{b}$ represents the weighted average bit-width across all layers in the model.

4.4 Model Architecture

In this study, we explored multiple lightweight MLP architectures with varying depths and widths to find an optimal balance between complexity and performance. The tested configurations are shown in the Table 2:

Number of Hidden Layers (HL)	Configuration
2	[256, 64]
3	[256, 512, 64]
5	[256, 512, 256, 128, 64]
7	[256, 512, 384, 256, 128, 96, 64]
9	[256, 512, 384, 256, 192, 128, 96, 72, 64]

Table 2: Configuration of different MLP architectures.

5 Results

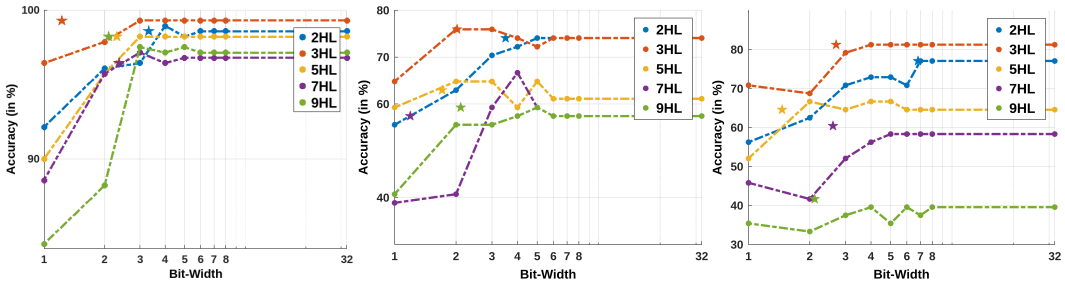
To validate the effectiveness of the proposed LAQ method, we conducted experiments using three speech emotion recognition (SER) datasets: TESS, EMODB, and SAVEE. This section presents the results of various configurations of the Multi-Layer Perceptron (MLP) architecture, specifically focusing on hidden layer (HL) variants ranging from 2HL to 9HL. Our first objective is to identify the optimal HL configuration for subsequent experiments. Following this, we detail the results of our adaptive layer-wise quantization method, evaluating the performance of quantized models against their full-precision counterparts, with an emphasis on accuracy and model size. Layers were ranked based on their importance, as computed using Equation (4), and each layer was quantized sequentially according to this ranking. We tested several bit precisions, including 8, 7, 6, 5, 4, 3, 2, and 1 bits, selecting the lowest bit precision that maintained model accuracy within a specified margin $T_{margin}(l)$ for each layer l.

5.1 Hidden Layer Configuration

The results in Fig. 3 illustrate the relationship between model accuracy and bit-width for various configurations of hidden layers (HL), specifically 2HL, 3HL, 5HL, 7HL, and 9HL, across the three datasets: TESS, EMODB, and SAVEE. The analysis reveals that increasing the number of hidden layers from 2 to 3 consistently enhances model performance. However, beyond the 3HL configuration, a noticeable decline in accuracy occurs as the depth increases to 5HL, 7HL, and 9HL. This trend indicates a diminishing return on accuracy with added complexity, leading us to select the 3HL configuration for further experiments. Additionally, our proposed method, indicated by star markers in the plot, demonstrates superior accuracy at lower bit-widths across all datasets.

5.2 Layer Importance Weights Configuration

We further evaluated the impact of different weight parameters used in the layer importance computations. Table 3 compares the average bit-width and accuracy across various layer importance weight



(a) TESS (b) EMODB (c) SAVEE

Figure 3: Comparison of model accuracy vs. bit-width for different variants of MLP. Performance of the proposed LAQ method is highlighted with star markers.

Layer Importance Weights			TESS		EMODB		SAVEE	
W_P	W_E	W_V	Bit-width	Accuracy (%)	Bit-width	Accuracy (%)	Bit-width	Accuracy (%)
0	0	0	2.412	99.29	2.375	74.07	1.933	81.25
1	0	0	2.347	99.29	2.375	74.07	1.933	81.25
0	1	0	2.412	99.29	2.891	75.93	1.933	81.25
0	0	1	1.771	99.29	2.025	75.93	1.936	81.25
0.5	0.5	0	2.347	99.29	2.375	74.07	1.933	81.25
0.5	0	0.5	1.226	99.29	2.025	75.93	2.696	81.25
0	0.5	0.5	1.771	99.29	2.025	75.93	1.936	81.25
0.33	0.33	0.33	1.226	99.29	2.025	75.93	2.696	81.25
0.11	0.01	0.88	**1.226**	99.29	**2.000**	75.93	**1.933**	81.25

Table 3: Comparison of average bit-width (b in bits) and accuracy (%) across various combinations of statistical parameters used in layer importance computation.

configurations (W_P, W_E, W_V) for the TESS, EMODB, and SAVEE datasets. For TESS, accuracy remains consistently high at 99.29% across configurations, with average bit-widths ranging from 1.226 to 2.412, indicating robustness to quantization variations. Conversely, EMODB and SAVEE exhibit greater sensitivity to layer importance weight configurations; for instance, EMODB achieves peak accuracy (75.93%) at 2.891 bits with $W_E = 1$, while SAVEE maintains an accuracy of 81.25% with a balanced weight of 0.33 across parameters. Although selecting optimal layer importance weights is challenging, our empirical analysis suggests that the configuration of $W_P = 0.11$, $W_E = 0.01$, and $W_V = 0.88$ yielded the best results, achieving the highest accuracies at the lowest bit-widths across all datasets. These findings indicate that dataset-specific optimal combinations for effective quantization warrant further investigation.

5.3 Comparison of Fixed-Quantization with the proposed LAQ approach

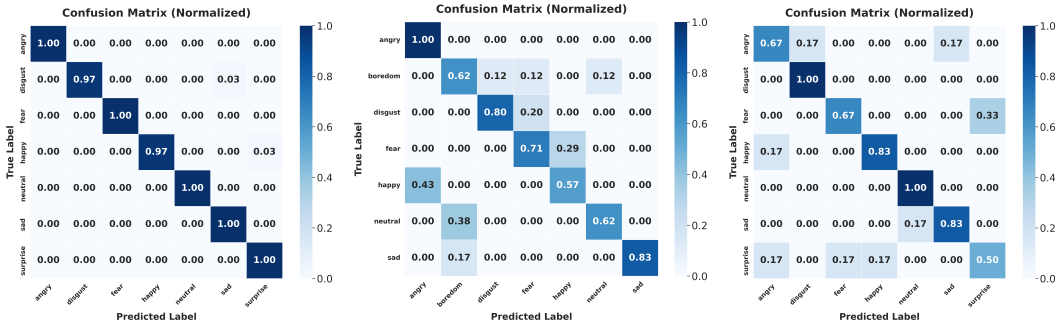
We present the results for various fixed-quantization methods alongside our proposed LAQ method. Table 4 compares accuracy and model size, represented as average bit-width ($\bar{b}$), for different quantization levels applied to the baseline model across the TESS, EMODB, and SAVEE datasets.

The baseline model, which operates at full 32-bit precision, achieves high accuracy rates of 99.29% on the TESS dataset, 74.07% on the EMODB dataset, and 81.25% on the SAVEE dataset. However, its large model size of 676 KB for each dataset underscores the necessity for more efficient quantization techniques. Fixed-bit quantization was applied, ranging from 8-bit down to 1-bit, leading to substantial reductions in model size while maintaining comparable accuracy levels at higher bit precisions. For instance, on the TESS dataset, the 8-bit quantization retained the baseline accuracy of 99.29% while reducing the model size to 169 KB. Accuracy decreased to 97.86% at 4-bit and 96.43% at 1-bit, with the model size reducing to 21 KB. A similar trend was observed on the EMODB dataset, where accuracy was maintained at 74.07% across 8-bit, 7-bit, and 6-bit quantization, with a slight decline to 72.22% at 5-bit and further decreases to 64.81% at 1-bit, along with a reduction in model size. On the SAVEE dataset, 8-bit quantization achieved an accuracy of 81.25%, with a gradual decrease to 68.75% at 2-bit and 70.83% at 1-bit, coupled with a reduction in model size.

In contrast, the proposed LAQ method, which assigns varying bit-widths to different layers based on their importance, achieves near-baseline accuracy while substantially reducing both the average bit-width and model size. For the TESS dataset, the proposed method attains an accuracy of 99.29%

	TESS		EMODB		SAVEE	
Model	Size (KB)	Accuracy (%)	Size (KB)	Accuracy	Size (KB)	Accuracy
32-bit Baseline	676	99.29	676	74.07	676	81.25
8-bit Fixed Quantization	169	99.29	169	74.07	169	81.25
7-bit Fixed Quantization	147	99.29	147	74.07	147	81.25
6-bit Fixed Quantization	126	99.29	126	74.07	126	81.25
5-bit Fixed Quantization	105	99.29	105	72.22	105	81.25
4-bit Fixed Quantization	84	99.29	84	74.07	84	81.25
3-bit Fixed Quantization	63	99.29	63	75.93	63	79.17
2-bit Fixed Quantization	42	97.86	42	75.93	42	68.75
1-bit Fixed Quantization	21	96.43	21	64.81	21	70.83
Proposed LAQ	25	**99.29**	43	**75.93**	40	**81.25**
Avg. Bit-Width ($\bar{b}$)	1.23		2.00		1.93	

Table 4: Comparison of model size (KB) and accuracy (%) across different quantization variants applied to the baseline model. The average bit-width per parameter ($\bar{b}$) is also reported (in bits).



(a) TESS (b) EMODB (c) SAVEE

Figure 4: Confusion matrices for test data across individual emotions for three benchmark datasets.

with an average bit-width of 1.23 and a model size of 25 KB. On the EMODB dataset, it achieves an accuracy of 75.93% with an average bit-width of 2.00 and a model size of 43 KB. Similarly, for the SAVEE dataset, the proposed method achieves an accuracy of 81.25% with an average bit-width of 1.93 and a model size of 40 KB. These results highlight the effectiveness of the proposed adaptive quantization approach in maintaining high accuracy while significantly reducing model size, establishing it as a more efficient alternative to fixed-bit quantization methods for neural networks.

The confusion matrix for the TESS dataset, as shown in Fig. 4a, reveals that the LAQ model achieves perfect classification (1.00) for the emotions of anger, fear, neutrality, sadness, and surprise. However, there is a minor misclassification, with 3% of happy samples incorrectly identified as surprise. This confusion likely stems from subtle similarities in the vocal expressions associated with these positive emotions. Likewise, the LAQ model demonstrates acceptable performance across various emotions for both the EMODB (Fig. 4b) and SAVEE (Fig. 4c) datasets.

5.4 Comparison of Proposed LAQ approach with existing studies

Finally, we compare existing approaches with the proposed LAQ method. Table 5 summarizes the performance of the proposed method alongside several studies on the TESS, EMODB, and SAVEE datasets. For the TESS dataset, the proposed LAQ method achieves the highest accuracy of 99.29% with only 169K parameters, surpassing other models such as CNN (52) (98.0%) and Vision Transformer (3) (98.0%), both of which have significantly larger parameter counts. Additionally, it outperforms methods like SVM (14) (96%) and Quaternion-valued CNN (67) (97%). For the EMODB dataset, the proposed LAQ model demonstrates strong performance with an accuracy of 75.93%, surpassing several methods, including ANN (56) (74.6%) and DBN (62) (72.4%), while maintaining a lower parameter count. For the SAVEE dataset, the proposed method achieves an accuracy of 81.25%, outperforming various models such as TIM-Net (55) (77.3%) and DCNN (60) (82.1%), which have higher parameter counts. Moreover, the average bit-width ($\bar{b}$) for each parameter is reduced to 1.22 for TESS, 2.00 for EMODB, and 1.93 for SAVEE using our layer-importance-based adaptive quantization approach. These results highlight the efficiency of the proposed method, achieving state-of-the-art accuracy with fewer parameters compared to more complex models.

Model	Year	TESS	Model	Year	EMODB	Model	Year	SAVEE
CNN (52)	2024	98.0% (4M)	Logistic Model Tree (54)	2020	80.0% (58M)	TIM-Net (55)	2023	77.3% (10M)
LSTM (52)	2024	77.0% (2M)	ANN (56)	2014	74.6% (1M)	TSP+INCA (57)	2021	83.4% (75M)
Transformer (58)	2023	98.2% (100M)	DNN (59)	2011	79.1% (7M)	DCNN (60)	2020	82.1% (62M)
EMD+LDA (61)	2021	93.3% (-)	DBN (62)	2018	72.4% (3M)	CPAC (63)	2022	83.7% (7M)
Vision Transformer (3)	2024	98% (4M)	MCNN (64)	2017	50% (1.3M)	GM-TCN (65)	2022	**83.9%**(-)
SVM (14)	2016	96% (-)	DCNN-DTPM (14)	2017	76.3% (5.2M)	PSOBBO+ELM (66)	2017	62.5% (2M)
Quaternion CNN (67)	2023	97% (5M)	RDBN (68)	2017	**82.3%** (3M)	RDBN (68)	2017	53.6% (3M)
Proposed LAQ	-	**99.29%**(169K)	**Proposed LAQ**	-	75.93% (169K)	**Proposed LAQ**	-	81.25% (169K)

Table 5: Comparison of the proposed LAQ with existing studies on three benchmark SER datasets: TESS, EMODB, and SAVEE. *(Classification accuracy (in %) and number of model parameters (in brackets) are provided for each dataset).*

Overall, the results underscore the potential of our adaptive layer-wise quantization approach for deploying deep neural networks on resource-constrained edge devices. By optimizing the bit-width adaptively, our method ensures that models remain both lightweight and efficient while retaining high accuracy, making it a valuable framework for practical applications where computational and memory resources are limited.

6 Conclusion

This study introduces a lightweight multilayer perceptron (MLP) neural network featuring a layer importance-guided adaptive quantization scheme. The network comprises three hidden layers and is specifically designed for the classification of seven emotions across three benchmark speech emotion recognition (SER) datasets. We extracted three audio features—Mel-frequency cepstral coefficients (MFCC), Mel-spectrogram, and chroma—from the speech audio samples. The proposed adaptive quantization (LAQ) method achieves classification accuracies of 99.6%, 75.9%, and 81.3% for the TESS, EMODB, and SAVEE datasets, respectively. Our approach not only provides comparable or superior results to existing models but also does so with significantly fewer parameters, totaling only 169K. The model maintains an average bit-width between 1.22 and 2.00, resulting in a maximum model size of 43 KB. While this model demonstrates high efficiency and accuracy, it has a limitation: the lack of cross-dataset validation, which is necessary to assess its generalizability and robustness. Future research will focus on extending the adaptive quantization technique to attention-based models for speech emotion recognition tasks across diverse datasets.

References

[1] Beard, R., Das, R., Ng, R.W., Gopalakrishnan, P.K., Eerens, L., Swietojanski, P. and Miksik, O., 2018, October. Multi-modal sequence fusion via recursive attention for emotion recognition. In Proceedings of the 22nd conference on computational natural language learning (pp. 251-259).

[2] Zhao, J., Mao, X. and Chen, L., 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical signal processing and control, 47, pp.312-323.

[3] Akinpelu, S., Viriri, S. and Adegun, A., 2024. An enhanced speech emotion recognition using vision transformer. Scientific Reports, 14(1), p.13126.

[4] Brown, T.B., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[5] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z. and Chu, E., 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.

[6] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A. and Millican, K., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

[7] Baevski, A., Zhou, Y., Mohamed, A. and Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, pp.12449-12460.

[8] Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A., 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing, 29, pp.3451-3460.

[9] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X. and Wu, J., 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), pp.1505-1518.

[10] Goron, E., Asai, L., Rut, E. and Dinov, M., 2024, April. Improving Domain Generalization in Speech Emotion Recognition with Whisper. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 11631-11635). IEEE.

[11] Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. and Pang, R., 2020. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100.

[12] Shor, J., Jansen, A., Han, W., Park, D. and Zhang, Y., 2022, May. Universal paralinguistic speech representations using self-supervised conformers. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3169-3173). IEEE.

[13] Li, Z., Li, H. and Meng, L., 2023. Model compression for deep neural networks: A survey. Computers, 12(3), p.60.

[14] Verma, D. and Mukhopadhyay, D., 2016, April. Age driven automatic speech emotion recognition system. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 1005-1010). IEEE.

[15] Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A.A., Alhadlaq, A. and Lee, H.N., 2022. Two-way feature extraction for speech emotion recognition using deep learning. Sensors, 22(6), p.2378.

[16] Bahdanau, D., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[17] Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y. and Li, C., 2018. Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition.

[18] Du, Q., Gu, W., Zhang, L. and Huang, S.L., 2018, November. Attention-based LSTM-CNNs for time-series classification. In Proceedings of the 16th ACM conference on embedded networked sensor systems (pp. 410-411).

[19] Chumachenko, K., Iosifidis, A. and Gabbouj, M., 2022, August. Self-attention fusion for audiovisual emotion recognition with incomplete data. In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 2822-2828). IEEE.

[20] Li, H., Ding, W., Wu, Z. and Liu, Z., 2020. Learning fine-grained cross modality excitement for speech emotion recognition. arXiv preprint arXiv:2010.12733.

[21] Sajjad, M. and Kwon, S., 2020. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE access, 8, pp.79861-79875.

[22] Zeng, Y., Mao, H., Peng, D. and Yi, Z., 2019. Spectrogram based multi-task audio classification. Multimedia Tools and Applications, 78, pp.3705-3722.

[23] Han, S., Pool, J., Tran, J. and Dally, W., 2015. Learning both weights and connections for efficient neural network. Advances in neural information processing systems, 28.

[24] Li, H., Kadav, A., Durdanovic, I., Samet, H. and Graf, H.P., 2016. Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710.

[25] Yin, M., Sui, Y., Liao, S. and Yuan, B., 2021. Towards efficient tensor decomposition-based dnn model compression with optimization framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10674-10683).

[26] Hinton, G., 2015. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

[27] Kim, N., Shin, D., Choi, W., Kim, G. and Park, J., 2020. Exploiting retraining-based mixed-precision quantization for low-cost DNN accelerator design. IEEE Transactions on Neural Networks and Learning Systems, 32(7), pp.2925-2938.

[28] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W. and Keutzer, K., 2022. A survey of quantization methods for efficient neural network inference. In Low-Power Computer Vision (pp. 291-326). Chapman and Hall/CRC.

[29] Deng, L., 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine, 29(6), pp.141-142.

[30] Krizhevsky, A. and Hinton, G., 2009. Learning multiple layers of features from tiny images.

[31] Aftab, A., Morsali, A., Ghaemmaghami, S. and Champagne, B., 2022, May. Light-sernet: A lightweight fully convolutional neural network for speech emotion recognition. In ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6912-6916). IEEE.

[32] You, C., Chen, N., Liu, F., Yang, D. and Zou, Y., 2020. Towards data distillation for end-to-end spoken conversational question answering. arXiv preprint arXiv:2010.08923.

[33] Pimentel, A., Guimarães, H., Avila, A.R., Rezagholizadeh, M. and Falk, T.H., 2023. On the Impact of Quantization and Pruning of Self-Supervised Speech Models for Downstream Speech Recognition Tasks" In-the-Wild". arXiv preprint arXiv:2309.14462.

[34] Chang, Y., Ren, Z., Nguyen, T.T., Qian, K. and Schuller, B.W., 2023, June. Knowledge transfer for on-device speech emotion recognition with neural structured learning. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[35] Zhao, H., Xiao, Y., Han, J. and Zhang, Z., 2019, May. Compact convolutional recurrent neural networks via binarization for speech emotion recognition. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6690-6694). IEEE.

[36] Yang, G., Yu, S., Yang, H., Nie, Z. and Wang, J., 2023. HMC: Hybrid model compression method based on layer sensitivity grouping. Plos one, 18(10), p.e0292517.

[37] Tang, C., Ouyang, K., Wang, Z., Zhu, Y., Ji, W., Wang, Y. and Zhu, W., 2022, October. Mixed-precision neural network quantization via learned layer-wise importance. In European Conference on Computer Vision (pp. 259-275). Cham: Springer Nature Switzerland.

[38] Mirsamadi, S., Barsoum, E. and Zhang, C., 2017, March. Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2227-2231). IEEE.

[39] Issa, D., Demirci, M.F. and Yazici, A., 2020. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, p.101894.

[40] Sadok, S., Leglaive, S. and Séguier, R., 2023, June. A vector quantized masked autoencoder for speech emotion recognition. In 2023 IEEE International conference on acoustics, speech, and signal processing workshops (ICASSPW) (pp. 1-5). IEEE.

[41] Jalal, M.A., Loweimi, E., Moore, R.K. and Hain, T., 2019, September. Learning temporal clusters using capsule routing for speech emotion recognition. In Proceedings of interspeech 2019 (pp. 1701-1705). ISCA.

[42] Gerczuk, M., Amiriparian, S., Ottl, S. and Schuller, B.W., 2021. Emonet: A transfer learning framework for multi-corpus speech emotion recognition. IEEE Transactions on Affective Computing, 14(2), pp.1472-1487.

[43] Shahin, I., Nassif, A.B. and Hamsa, S., 2019. Emotion recognition using hybrid Gaussian mixture model and deep neural network. IEEE access, 7, pp.26777-26787.

[44] Han, S., Mao, H. and Dally, W.J., 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.

[45] Molchanov, P., Tyree, S., Karras, T., Aila, T. and Kautz, J., 2016. Pruning convolutional neural networks for resource efficient inference. arXiv preprint arXiv:1611.06440.

[46] Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R. and Bengio, Y., 2016. Binarized neural networks. Advances in neural information processing systems, 29.

[47] Tian, Y., Krishnan, D. and Isola, P., 2019. Contrastive representation distillation. arXiv preprint arXiv:1910.10699.

[48] Elkerdawy, S., Elhoushi, M., Singh, A., Zhang, H. and Ray, N., 2020. To filter prune, or to layer prune, that is the question. In proceedings of the Asian conference on computer vision.

[49] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. and Weiss, B., 2005, September. A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

[50] Jackson, P. and Haq, S., 2014. Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK.

[51] Pichora-Fuller, M.K. and Dupuis, K., 2020. Toronto emotional speech set (TESS); 2020. URL: https://tspace. library. utoronto. ca/handle/1807/24487. DOI: https://doi. org/10.5683/SP2/E8H2MF.

[52] Islam, M.M., Kabir, M.A., Sheikh, A., Saiduzzaman, M., Hafid, A. and Abdullah, S., 2024, May. Enhancing Speech Emotion Recognition Using Deep Convolutional Neural Networks. In Proceedings of the 2024 9th International Conference on Machine Learning Technologies (pp. 95-100).

[53] Swain, M., Routray, A. and Kabisatpathy, P., 2018. Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21, pp.93-120.

[54] Assunção, G., Menezes, P. and Perdigão, F., 2020. Speaker Awareness for Speech Emotion Recognition. Int. J. Online Biomed. Eng., 16(4), pp.15-22.

[55] Ye, J., Wen, X.C., Wei, Y., Xu, Y., Liu, K. and Shan, H., 2023, June. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[56] Sidorov, M., Ultes, S. and Schmitt, A., 2014, May. Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4803-4807). IEEE.

[57] Tuncer, T., Dogan, S. and Acharya, U.R., 2021. Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowledge-Based Systems, 211, p.106547.

[58] Bayraktar, U., Kilimci, H., Kilinc, H.H. and Kilimci, Z.H., 2023, November. Assessing Audio-Based Transformer Models for Speech Emotion Recognition. In 2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS) (pp. 1-7). IEEE.

[59] Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G. and Schuller, B., 2011, May. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5688-5691). IEEE.

[60] Farooq, M., Hussain, F., Baloch, N.K., Raja, F.R., Yu, H. and Zikria, Y.B., 2020. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors, 20(21), p.6008.

[61] Krishnan, P.T., Joseph Raj, A.N. and Rajangam, V., 2021. Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition. Complex & Intelligent Systems, 7, pp.1919-1934.

[62] Latif, S., Rana, R., Younis, S., Qadir, J. and Epps, J., 2018. Transfer learning for improving speech emotion classification accuracy. arXiv preprint arXiv:1801.06353.

[63] Wen, X.C., Ye, J.X., Luo, Y., Xu, Y., Wang, X.Z., Wu, C.L. and Liu, K.H., 2022. Ctl-mtnet: A novel capsnet and transfer learning-based mixed task net for the single-corpus and cross-corpus speech emotion recognition. arXiv preprint arXiv:2207.10644.

[64] Sivanagaraja, T., Ho, M.K., Khong, A.W. and Wang, Y., 2017, December. End-to-end speech emotion recognition using multi-scale convolution networks. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 189-192). IEEE.

[65] Ye, J.X., Wen, X.C., Wang, X.Z., Xu, Y., Luo, Y., Wu, C.L., Chen, L.Y. and Liu, K.H., 2022. GM-TCNet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition. Speech Communication, 145, pp.21-35.

[66] Yogesh, C.K., Hariharan, M., Ngadiran, R., Adom, A.H., Yaacob, S., Berkai, C. and Polat, K., 2017. A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. Expert Systems with Applications, 69, pp.149-158.

[67] Guizzo, E., Weyde, T., Scardapane, S. and Comminiello, D., 2023. Learning speech emotion representations in the quaternion domain. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, pp.1200-1212.

[68] Wen, G., Li, H., Huang, J., Li, D. and Xun, E., 2017. Random deep belief networks for recognizing emotions from speech signals. Computational intelligence and neuroscience, 2017(1), p.1945630.