Layer-Importance guided Adaptive Quantization for Efficient Speech Emotion Recognition

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 crucial for enhancing human-machine interaction. Deep learning models have achieved significant success in SER without manually engineered features, but they require substantial computational resources, processing power, and hyper-parameter tuning, limiting their deployment on edge devices. To address these limitations, we propose an efficient and lightweight Multilayer Perceptron (MLP) classifier within a custom SER framework. Furthermore, we introduce a novel adaptive quantization scheme based on layer importance to reduce model size. This method balances model compression and performance by adaptively selecting bit-width precision for each layer based on its importance, ensuring the quantized model maintains accuracy within an acceptable threshold. Unlike previous mixed-precision methods, which are often complex and costly, our approach is both interpretable and efficient. Our model is evaluated on the benchmark SER datasets, focusing on features such as Mel-Frequency Cepstral Coefficient (MFCC), Chroma, and Mel-spectrogram. Our experiments show that our quantization scheme achieves performance comparable to state-of-the-art methods while significantly reducing model size, making it well-suited for lightweight devices.

1 Introduction

Humans are quite good in recognizing the emotions, whereas it is still a very challenging task for the machines. Speech emotion recognition (SER) focuses on understanding and identifying the emotional states embedded in human speech. The subjective nature and complexity of the human emotional state and expression is what makes SER difficult. SER has an extensive range of applications such as virtual assistants, social robots, lie-detection, call-center answering, mental health and fitness analysis, human-computer interaction, and so on (1). In recent years, SER have significantly improved the ability to detect and interpret human emotions. Researchers are increasingly leveraging deep learning techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), Long Short-Term Memory (LSTMs) (2), and transformers (3) to enhance the accuracy of emotion classification from speech. Additionally, the use of large annotated datasets has accelerated model training and evaluation. Also, with the advancements in the field of large language models (LLMs) (e.g., GPT-3, GPT-4, PALM, and Gemini) (8; 9; 10) and pre-trained speech models (e.g., wav2vec (11), HuBERT (12), wavLM (13), Whisper (14), and Conformer (15)), improving efficiency has become crucial. While computational resources have driven much of the success in these deep learning models, their increasing size, often encompassing billions of parameters, poses significant challenges for real-world applications, particularly in terms of computational complexity, deployment costs, and environmental impact (16). This concern extends to SER models as well, where highly over-parameterized networks may hinder practical usage. As a result, optimizing both model architectures and hardware-aware solutions is critical to achieve real-time performance in SER tasks.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The demanding task of SER is to identify suitable features that can capture maximum information from speech. Mel-frequency cepstral coefficients (MFCC) is one of the most prominent features that has been consistently reported in the literature (36). Among other audio features, Mel-spectrogram, Tonnetz, and Chroma are widely used (3), (55). When it comes to models, CNNs and LSTMs, or their combinations, are commonly used, often enhanced with attention-based networks applied to the basic versions to improve speech emotion recognition performance (7), (57),(56). Zhao et al. (56) applied attention-based model on Bidirectional LSTM, and fully Connected Networks for learning spatio-temporal emotional features and machine learning classifier for speech emotion classification. Another work along the same line (57) utilized attention-based model and 1D-CNN. A multimodal deep learning approach using text and speech with temporal alignment technique presented in (5) combined CNN, LSTM, and attention networks, obtaining notable accuracy in classifying different emotions on the benchmark datasets. Furthermore, a novel hybrid-based audio transformer, named Conformer-HuBERT (6) provided a significantly improved performance, learning from large-scale unannotated data.

Various methods have been proposed for reducing model sizes, including model pruning (17; 18), low-rank factorization (41), knowledge distillation (19), and quantization (20; 21). Most existing research in the field of model compression centers on image datasets, such as MNIST (22) and CIFAR (23). However, (24; 25; 26; 27; 28) have explored the application of model compression techniques in speech emotion recognition and question-answering systems, respectively. Building on these efforts, our work investigates an adaptive model quantization scheme to reduce the size of SER neural networks. The quantization scheme, which reduces the bit-precision of weights and activations, is widely used for model compression due to its storage and memory efficiency (21). Most quantization methods apply uniform bit-width precision across all layers, reducing model size but often sacrificing accuracy, especially for complex tasks. Since different layers contribute variably to a model's performance, uniform quantization fails to capture this variation (29). Mixed-precision quantization has been explored to address this issue, which introduces challenges in determining the optimal bit-width for each layer, requiring costly optimization, and may result in sub-optimal trade-offs between compression and accuracy (30).

In this paper, we introduce a simple and lightweight MLP model with only three hidden layers to recognize whether the spoken voice manifests anger, fear, disgust, happiness, surprise, sadness, or neutral, among others. Furthermore, we present a framework for adaptive layer-wise quantization in the pursuit of a lightweight model. The main contributions of this work are as follows:

- A lightweight model with limited number of parameters ($\sim 169K$) for SER task.
- A novel layer importance computation based on number of parameters and variance of weights within a layer.
- A novel adaptive layer-wise quantization method that assigns different bit-widths to individual layers as per their importance score.

The rest of this paper is organized as follows. Section 2 describes the proposed method in detail. The experimental setup is described in Section 3 and Section 4 presents experimental results. Finally, Section 5 concludes the paper.

2 Methodology

2.1 Proposed Framework

Fig. 1 illustrates a novel framework for efficient speech emotion recognition that integrates audio feature extraction, a lightweight multi-layer perceptron (MLP) model, and adaptive quantization to enhance performance while reducing model size. The framework begins with raw audio samples, from which three key features namely MFCC, Chroma, and Mel-spectrogram are extracted. These features are chosen because they often perform well in identifying emotions in speech signals. The features are processed by a compact MLP architecture with three hidden layers consisting of 256, 512, and 64 neurons, respectively. The framework also employs adaptive quantization, guided by layer importance, to optimize the precision of weights and reduce model size without sacrificing accuracy. This process involves (i) computing layer importance, (ii) performing iterative search optimization for bit precision, and (iii) applying layer-wise quantization. The resulting quantized

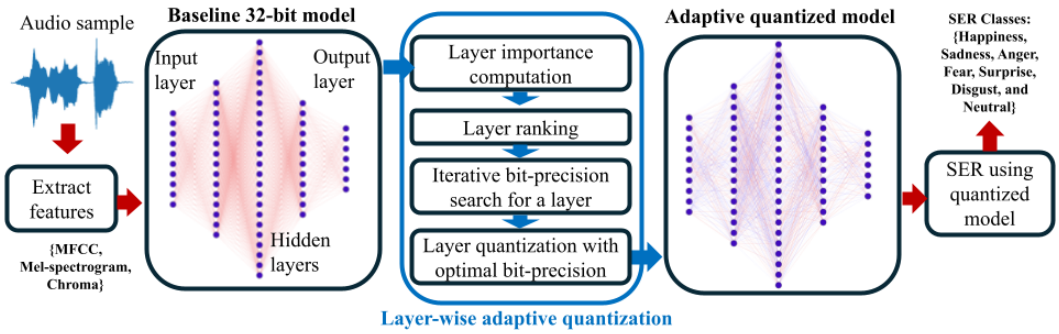


Figure 1: The framework of the proposed approach.

model is compressed, reducing its size and inference time while maintaining high performance. The final step involves classifying the speech sample into predefined emotion categories. Hence, by combining effective feature extraction with a scalable MLP model along with layer importance-guided adaptive quantization, our framework offers an efficient speech emotion recognition solution, suitable for resource-constrained environments and real-world applications.

2.2 Layer Importance Guided Adaptive Quantization Strategy

This section presents our adaptive layer-wise quantization strategy. The proposed method integrates a quantization framework and an iterative optimization process aimed at minimizing accuracy loss while reducing the model size. The strategy adapts the bit precision for each layer based on its importance. The main objective is to optimize the bit-width of each layer to strike a balance between model size and accuracy. Our approach is motivated by the observation that different layers of a DNN contribute unequally to the model's final accuracy (31). To explore this, we experimented by quantizing each layer at varying bit-widths while keeping the rest of the layers at 8-bit precision. We found that layers exhibit different sensitivities to quantization, leading to varying performance at the same bit precision. To leverage this observation, we propose quantizing each layer with a different bit-width based on its importance. However, determining the optimal quantization order is challenging, so we propose a layer ranking mechanism based on layer importance.

2.2.1 Layer Importance Computation and Layer Ranking

To compute the importance of each layer, we evaluate following metrics that balance between bit precision and model accuracy.

Number of parameters: The number of parameters in a layer is a key factor in model size. Layers with more parameters are prioritized for quantization to reduce the overall size. We define the parameter proportion as:

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

Variance: Parameter distribution in each layer affects its quantization. Layers with higher variance may benefit from quantization. We define normalized variance as:

$$N_V(l) = \log\left(e - 1 + \frac{\text{Variance of layer } l}{\max_k \left(\text{Variance of layer } k\right)}\right) \tag{2}$$

Layer Importance Calculation: The importance of a particular layer l can be quantified using a weighted sum of various criteria. Specifically, the importance of layer l is given by:

$$I(l) = \alpha \cdot N_P(l) + (1 - \alpha) \cdot N_V(l) \tag{3}$$

where α and $1 - \alpha$ are weights assigned to the two different criteria i.e., number of parameters, and variance. Each of these weights determine the importance of individual criterion in overall

importance score. $N_P(l)$ represents number of parameters in layer *l* as compared to the total number of parameters in the model. Whereas, $N_V(l)$ measures the relative variance of the layer *l*.

Layer Ranking: We propose a ranking system for the layers based on their importance. The layers are sorted in descending order of importance, ensuring that the most critical layers are quantized first. This prioritization is crucial, as it allows us to allocate computational resources effectively during the quantization process.

2.2.2 Iterative Bit-precision Search Algorithm

The optimal bit-precision for each layer is selected through a search process to minimize the overall bit-width while maintaining accuracy. Layers are ranked by importance, and quantization starts with the highest-ranked layers. A bit-precision search begins from the lowest possible value and stops when the performance degradation is within a threshold margin $T_{margin}(l)$, which adapts according to layer importance:

$$T_{margin}(l) = T_{margin} \times Importance(l) \tag{4}$$

where $T_{margin}(l)$ is the threshold margin for layer l, and $Importance(l)$ is the importance of the current layer l. This adaptive margin ensures efficient compression without significant performance loss, optimizing the bit precision of each layer iteratively.

2.2.3 Layer quantization

In this step, we proceed to quantize layer l using the optimal bit-precision identified in the previous search. The selected bit-precision for each layer is applied to convert the floating-point weights into lower bit-width representations. This quantization not only reduces the memory footprint of the model but also enhances inference speed, making it more suitable for deployment in resource-constrained environments. The careful selection of bit-precision ensures that while the model's size is minimized, its performance remains robust and reliable.

3 Experimental setup

Dataset Description: This work utilizes three benchmark SER datasets namely EMODB, SAVEE, and TESS. A detailed description of each of these datasets is presented in the Table 1.

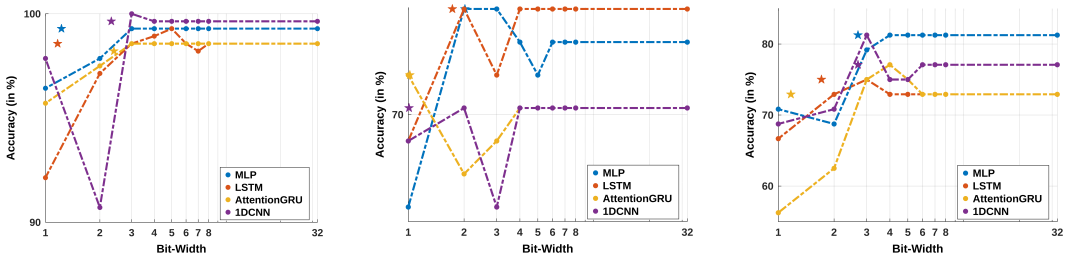
Dataset	#Samples	# Speakers	Gender (M/F)	Emotions
EMODB (39)	535	10	5/5	Anger, Boredom, Disgust, Fear, Happiness, Sadness, Neutral
SAVEE (54)	480	4	4/0	Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise
TESS (32)	2800	2	0/2	Anger, Disgust, Happiness, Sadness, Neutral, Pleasant Surprise, Fear

Table 1: Summary of the benchmark SER datasets considered in this work.

Implementation details: In this work, rather than utilizing raw audio signals as input to the Multi-Layer Perceptron (MLP) (33), we draw inspiration from existing studies (4), which focus on feature extraction. Specifically, we derive three audio features, namely Mel-frequency cepstral coefficients (MFCCs), Chroma, and Mel-spectrogram. The MLP architecture was implemented in PyTorch, with evaluation metrics generated using scikit-learn. All experiments were conducted on the Kaggle platform, utilizing CPUs for neural network training. The dataset is split into training (80%), validation (10%), and test (10%) sets. For training, the model used Adam optimizer with a learning rate of 0.001 and Cross-Entropy loss as the loss function. A batch size of 32 was employed, and regularization was implemented through early stopping with a patience of 5 epochs to prevent overfitting. Additionally, a dropout rate of 0.1 was applied to enhance generalization.

Evaluation metrics: The performance evaluation is presented in terms of accuracy and average bit-width of the model ($\bar{b}$). This metric measures overall bit-precision used across all layers of the model. It reflects how the bit-width varies from layer to layer and indicates the extent of compression applied to the model. For a given model with L layers, the average bit-width $\bar{b}$ can be calculated as:

$$\bar{b} = \sum_{l=1}^{L} N_P(l) \cdot b(l) \tag{5}$$



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

Figure 2: Comparison of model accuracy vs. bit-width for different models across benchmark datasets. Performance of the proposed method is highlighted with star markers.

Datasets	**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%
Fixed quantization 8-bit	169	99.29%	169	74.07%	169	81.25%
7-bit	147	99.29%	147	74.07%	147	81.25%
6-bit	126	99.29%	126	74.07%	126	81.25%
5-bit	105	99.29%	105	72.22%	105	81.25%
4-bit	84	99.29%	84	74.07%	84	81.25%
3-bit	63	99.29%	63	75.93%	63	79.17%
2-bit	42	97.86%	42	75.93%	42	68.75%
1-bit	21	96.43%	21	64.81%	21	70.83%
Proposed adaptive quantization	25	**99.29%**	43	**75.93%**	56	**81.25%**
Avg. Bit-Width ($\bar{b}$)	1.22		2.00		2.69	

Table 2: Comparison of model accuracy, size and average bit-width ($\bar{b}$) across different quantization variants applied to the baseline model.

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

Model architecture: We experimented with different lightweight MLP architectures, with smaller depth and width. We empirically selected the lightweight configuration as a simple three hidden layers network consisting of 256, 512, and 64 neurons respectively. With all fully connected (FC) layers, the SER model for benchmark datasets contains about 169K parameters.

4 Results

Three benchmark datasets i.e., TESS, EMODB, and SAVEE are used to verify the effectiveness of the proposed approach. This section presents the results of our adaptive layer-wise quantization method. We have evaluated the performance of the quantized models compared to their full-precision counterparts, focusing on accuracy and model size. Layers were ranked based on their importance as computed using equation (3), and each layer was sequentially quantized 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 $T_{margin}(l)$ corresponding to the layer l. Further, we also evaluated the performance of different models such as MLP, LSTM, AttentionGRU and 1DCNN for the SER task on all three datasets considering their baseline models and quantized versions (fixed as well as layer-wise adaptive). The results of performance comparison (in terms of accuracy and bit-width) are presented in Figure 2.

Table 2 compares the accuracy and model size, measured in terms of average bit-width ($\bar{b}$), for different quantization levels applied to the baseline model across the three datasets. The baseline model, which operates at full 32-bit precision, achieves high accuracy rates of 99.29%, 74.07%, and 81.25% on TESS, EMODB and SAVEE dataset, respectively. However, its large model size of 676 KB highlights the need for more efficient quantization techniques. Fixed-bit quantization was applied from 8-bit down to 1-bit, resulting in significant reduction in model size while maintaining comparable accuracy levels. On the TESS dataset, 8-bit quantization retained the baseline accuracy of 99.29% with a reduced model size of 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 pattern was observed on the EMODB

Model	Year	TESS	Model	Year	EMODB	Model	Year	SAVEE
CNN (33)	2024	98.0% (4M)	Logistic Model Tree (47)	2020	80.0% (58M)	TIM-Net (42)	2023	77.3% (10M)
LSTM (33)	2024	77.0% (2M)	ANN (48)	2014	74.6% (1M)	TSP+INCA (43)	2021	83.4% (75M)
Transformer (34)	2023	98.2% (100M)	DNN (50)	2011	79.1% (7M)	DCNN (45)	2020	82.1% (62M)
EMD+LDA (35)	2021	93.3% (-)	DBN (52)	2018	72.4% (3M)	CPAC (44)	2022	83.7% (7M)
Vision Transformer (3)	2024	98% (4M)	MCNN (51)	2017	50% (1.3M)	GM-TCN (46)	2022	83.9% (-)
SVM (36)	2016	96% (-)	DCNN-DTPM (36)	2017	76.3% (5.2M)	PSOBBO+ELM (49)	2017	62.5% (2M)
Quaternion CNN (37)	2023	97% (5M)	RDBN (53)	2017	82.3% (3M)	RDBN (53)	2017	53.6% (3M)
Proposed	-	**99.29**% (169K × **1.22**)	**Proposed**	-	**75.93**% (169K × **2.00**)	**Proposed**	-	**81.25**% (169K × **2.69**)

Table 3: Comparison of the proposed approach with existing studies on three benchmark datasets: TESS, EMODB, and SAVEE. *(The classification accuracy (in %) and the number of model parameters (in brackets) are provided for each dataset).*

dataset, where accuracy was maintained at 74.07% with 8-bit, 7-bit, and 6-bit quantization, with a slight decrease to 72.22% at 5-bit and further decreases to 64.81% at 1-bit, alongside 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, along with a reduction in model size. In contrast, the proposed adaptive quantization method, which assigns varying bit-widths to different layers based on their importance, achieves near-baseline accuracy while substantially reducing the average bit-width and model size. For the TESS dataset, the proposed method achieves an accuracy of 99.29% with an average bit-width of 1.22 and a model size of 25 KB. On the EMODB dataset, it achieves an improvement of about 2% with 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 2.69 and a model size of 56 KB. These results demonstrate the effectiveness of the proposed adaptive quantization approach in maintaining high accuracy while reducing model size, making it a more efficient alternative to fixed-bit quantization methods for neural networks.

A comparison of the proposed method with several existing studies on the TESS, EMODB, and SAVEE datasets are presented in Table 3. For TESS dataset, proposed approach achieves the highest accuracy of 99.29% with only 25K parameters, surpassing other models such as CNN (33) (98.0%) and Vision Transformer (3) (98.0%), both of which have significantly larger parameter sizes. Additionally, it outperforms methods like SVM (36) (96%) and Quaternion-valued CNN (37) (97%). For EMODB, the proposed model demonstrates strong performance with an accuracy of 75.93% and only 43K parameters. This surpasses several other methods, including 1D-CNN (38) (71.6%) and VQ Masked Autoencoder (40) (65.8%), while maintaining a much lower parameter count. For the SAVEE dataset, it achieves 81.25% accuracy with 56K parameters, outperforming various models such as TIM-Net (42) (77.3%) and DCNN (45) (82.1%), which have higher parameter counts. Moreover, the average bit-width ($\bar{b}$) for each parameter is reduced to 1.22 for TESS, 4.06 for EMODB, and 2.69 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. Overall, the results highlight the potential of our adaptive layer-wise quantization in deploying deep neural networks for SER task on resource-constrained edge devices. By optimizing the bit-width adaptively, our approach 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.

5 Conclusion

This study proposes a lightweight and adaptively quantized MLP neural network consisting of only three hidden layers for the classification of seven emotions on three benchmark datasets namely TESS, EMODB, and SAVEE. We have extracted three audio features namely MFCC, Mel-spectrogram, and Chroma from the speech audio samples. The proposed method achieves an accuracy of 99.6%, 75.9%, and 81.3% for the seven-class classification of emotions on the TESS, EMODB, and SAVEE datasets, respectively. Our method provides comparable or better results than the existing models but with far lesser parameters (about an average bit width of 2 and maximum model size of 56 KB). This efficient model is able to provide results at par with the existing studies. The major advantage of the proposed model is its simple architecture with very few parameters $(169K)$. Furthermore, an adaptive quantization strategy is employed to further reduce the model size. A limitation of this work is that we have not performed cross-dataset experiments to check the generalizability and robustness of the model.

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] 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.

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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).

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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).

[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] 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).

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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).