AVSE-Pruner: Filter Pruning of Audio-Visual Speech Enhancement System using Multi-objective Binary Particle Swarm Optimization

Rahma Fourati[1,2], Jihene Tmamna[1], Najwa Kouka[3], Tassadaq Hussain[4], Mandar Cogate[4], Kia Dashtipour[4], Tughrul Arslan[5],Amir Hussain[4]
[1]REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia
[2]Université de Jendouba, Faculté des Sciences Juridiques, Economiques et de Gestion de Jendouba, 8189 Jendouba, Tunisie
[3] Dipartimento di Informatica Università degli Studi via Celoria 18 20133
[4] School of Computing, Merchiston Campus, Edinburgh Napier University, Edinburgh, EH10 5DT, Scotland, UK.
[5] School of Engineering, The University of Edinburgh, Edinburgh, EH9 3FF, UK

***Abstract*—This paper optimizes filter pruning as a constrained multi-objective optimization problem using a new binary multi-objective particle swarm optimization with dynamic learning strategies (AVSE-Pruner). AVSE-Pruner aims to balance network performance and computational cost by incorporating dynamic learning strategies to adjust search behavior. Applying AVSE-Pruner, we pruned the baseline model for the Audio-Visual Speech Enhancement (AVSE) Challenge, which enhances speech intelligibility in noisy environments using audio and visual inputs. Our pruned model maintains high performance while significantly reducing computational burden, demonstrating its suitability for real-time embedded applications.**

I. Pruning process description

We propose a pruned model of the baseline using a multiobjective binary particle swarm optimization (BPSO) approach. Our method aims to optimize the balance between model complexity and performance, enhancing both the computational efficiency and the accuracy of the speech enhancement process. The results demonstrate that our pruned model achieves significant reductions in computational overhead while maintaining high levels of speech intelligibility and quality, as validated by established objective measures and human subject tests. This contribution not only advances the field of multimodal speech enhancement but also sets a new standard for efficient and effective AV model deployment in real-world noisy environments. The determination of the optimal pruned model to be deployed is illustrated in Fig.1. Starting with the original model $M(W, b)$, the BPSO-FPruner runs for a fixed number of iterations. Ultimately, the $Gbest$ represents the best solution found by the algorithm, assigning a value of "1" to filters to be retained and "0" to those to be removed. It's important to note that the pruning process exclusively pertains to the convolutional layers, excluding max-pooling and fully connected layers. Subsequently, the pruned model $M^{'}(W^{'}, b^{'})$ is constructed by extracting relevant filters along with their weight values and biases from the original model. In other words, the frozen weights of the original model are retained, while irrelevant filters are removed. The newly structured model does not initially achieve high performance. To enhance the pruned model's performance, it undergoes fine-tuning for several epochs using the training dataset, ultimately yielding the optimal model $M^*(W^*, b^*)$ for deployment. In summary, the pruning process involves the original model $M(W, b)$, the optimally pruned model $M^{'}(W^{'}, b^{'})$, and the enhanced optimal pruned model $M^*(W^*, b^*)$.

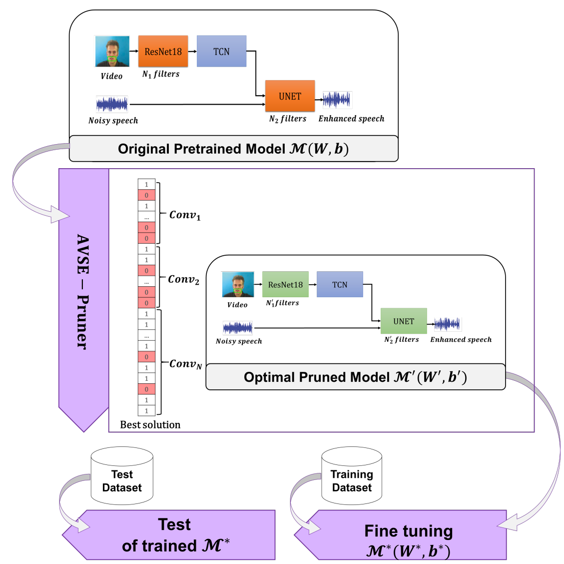

Fig. 1. The pruning process

II. Evaluation

TABLE I
Model performance before and after the pruning

Model	FLOPs	Params	PESQ	STOI	SISNR
Baseline model	96725.85M	76.01M			
Pruned model					