An Audio-Visual Speech Enhancement Approach Based on DCNN-RNN for the 2nd COG-MHEAR AVSE Challenge: System Description

Rami Ahmed[1], Kia Dashtipour[2], and Amir Hussain[2]

[1]Edinburgh Napier University; ramiscience@gmail.com
[2]School of Computing, Edinburgh Napier University; k.dashtipour, a.hussain@napier.ac.uk

July 21, 2023

1 AVSE-2 System Description

This report presents a comprehensive overview of our developed system, including the model architecture, training process, hardware specifications:

Table 1: Report Elements Description

1.	**Number of Parameters in the Model:**
	Trainable params: 2.0 M
	Non-trainable params:0
	Total params: 2.0 M
	Total estimated model params size: 7.948 (MB)
2.	**Floating-Point Operations per Second (FLOPS):**
	7.57 GFLOPS
3.	**Number of Training Steps:**
	135 steps
4.	**Latency (with Hardware Specifications):**
	CPU: Intel(R) Xeon(R) Platinum 8369B (2.9GHz, 4 cores, 30G RAM)
	GPU: NVIDIA A10 (Total available VRAM: 23.03GB)
	Latency: 144.43 milliseconds Per Sample.
5.	**Real-Time Factor (RTF):**
	Average of (0.15) RTF
6.	**Training Time (Time per Epoch):**
	1 min. – 43 secs.
7.	**Memory Footprint:**

	Training: GPU (22453 MiB) – RAM (11918 MiB)
	Inference: GPU (6464 MiB) – RAM (17607 MiB)
	Model Loading: GPU (302 MiB) – RAM (14031 MiB)
8.	**Hardware Specifications Used for Training and Inference:**
	CPU: Intel(R) Xeon(R) Platinum 8369B (2.9GHz, 4 cores, 30G RAM)
	GPU: NVIDIA A10 (Total available VRAM: 23.03GB)
9.	**Number and Type of GPUs Used:**
	1 x NVIDIA A10 GPU
10.	**Training Process:**
	Data Preprocessing: The images of the MP4 frames (images) have been resized to (128px) and enhanced by applying histogram equalization (cv2.equalizeHist) method.
	The ResNet module has been replaced by DCNN (08-Convs. Layers)
	Batch Normalization, Dropout (0.3), and MaxPooling functions have been applied.
	Only 02-LSTM Layers of Tensor [512, 32, 32] have been used with the Dropout (0.3).
	Optimization Algorithm: RMSprop
	Batch Size: 32
	Learning Rate: Exponential decay starting from 0.001
	Number of Training Epochs: 200
11.	**Reproducibility:**
	Code Availability: The source code is available on GitHub at https://github.com/RamiSaad/AVSE-2-Challenge.git.
	Instructions for Reproduction: To reproduce the system's performance, follow the steps provided in the GitHub repository's README file.
12.	**Known Limitations or Constraints of the Developed System:**
	Although the new model is smaller and faster, it still needs improvement to surpass the base model's accuracy.
	The visual model and LSTM model have undergone the most changes.
	The audio model has not undergone any changes.
13.	**Specific Hardware or Software Requirements:**
	Python version: 3.9.0
	PyTorch version: 2.1.0.dev20230622+cu121
	NVIDIA GPU compute capability: 8.6
	CUDA Toolkit version: 12.1
	cuDNN version: 8.0.1