Our entry is AVSE01. We just used the network structure of the baseline system which consists of a visual feature extraction net, an audio feature extraction net and a visual-audio fusion net. We did not use the pytorch-lightning framework in the baseline system, but rewrote the model and trainer in PyTorch and trained the model in distributed data parallel. We used 16 gpus to train the model; every gpu processes 4 input samples in one batch, so the actual batch size is 64 in total. Other hyperparameters and configurations were kept the same as of the baseline. We trained the model 15 epochs, and produced the enhanced sample audio.