Reference Plans:

[1] C. Barhoumi, and Y. BenAyed, “Real-time speech emotion recognition using deep learning and data augmentation,” Artificial Intelligence Review, vol. 58, no. 2, pp. 49, 2024/12/20, 2024.

- Method: Inject carefully controlled white noise during training (SNR set to 20 dB).
- Improve model robustness and performance for emotion recognition in real-world noisy environments.

[2] J. Wagner et al., "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10745-10759, 1 Sept. 2023, doi: 10.1109/TPAMI.2023.3263585.

- Method: Add natural soundscapes, human non-speech, and indoor/household sounds from the ESC-50 dataset at SNR levels of 0 dB, 10 dB, or 20 dB.

Initial Plan

Objectives

- Assess whether adding white/natural noise augmentation improves SER generalization (clean → noisy test).
- Evaluate whether Transformer (AST) is more robust than traditional architectures in noisy scenarios.
- Analyze the impact of different SNR levels and identify potential performance turning points.

Data and Splits

- Main dataset: RAVDESS (prioritize `Audio_Speech_Actors_01-24`).
- Split strategy (by actor to avoid speaker leakage):
  - Train: Actor 01–16
  - Validation: Actor 17–20
  - Test: Actor 21–24
- Optional extension: Add a subset of IEMOCAP for cross-validation in a later stage.

Noise Sources and SNR Settings

- Noise types:
  - White noise (baseline).
  - Three ESC-50 categories: natural soundscapes (environmental), human non-speech, indoor/household sounds.
- SNR set: {0 dB, 5 dB, 10 dB, 20 dB}; white noise fixed at 20 dB as a reference point.

Augmentation and Mixing Strategy (On-the-fly during training)

- Probability: use `p_aug = 0.7` to mix noise into training samples; keep the rest clean (`1 - p_aug = 0.3`).
- Sampling:
  - Uniform over noise categories (white noise / three ESC-50 categories).
  - Uniform over SNR set.
  - At most one noise type mixed per original sample.

- Duration alignment:
  - If the noise clip is shorter than speech, loop or randomly concatenate; if longer, randomly crop.
- Mixing and normalization (RMS alignment, avoid distortion):
  - Compute speech RMS: `R_s`, noise RMS: `R_n`.
  - Noise gain `g = R_s / (R_n × 10^(SNR/20))`.
  - Mixed signal `x_mix = x_speech + g × x_noise`.
  - Peak limiting and renormalization: peak ≤ −1 dBFS to avoid clipping.

Model and Features

- Model: AST (Audio Spectrogram Transformer).
- Features:
  - Sample rate 16 kHz.
  - Log-Mel spectrogram, Mel bins = 64, window = 25 ms, hop = 10 ms.
  - Disable other strong augmentations during training (e.g., SpecAugment) to isolate the effect of noise augmentation; add them back gradually in ablation.

Training Settings (Unified)

- Optimizer: AdamW, initial LR = 1e-4, weight decay = 1e-4, cosine or multi-step decay.
- Batch size: 32; epochs: 50; early stopping: patience = 7.
- Random seeds: {0, 1, 2} for three independent runs.
- Class imbalance: use weighted cross-entropy or balanced sampling.
- Logging: save training curves, best validation metrics, and corresponding checkpoints to `logs/` and `checkpoints/`.

Experiment Lineup (aligned with comparative design)

- E0 Baseline (clean training → clean/noisy test): no noise augmentation.
- E1 White noise at 20 dB: augment training set with 20 dB white noise only.
- E2 ESC-50 multi-class, multi-SNR: randomly choose category and SNR ∈ {0, 5, 10, 20} during training.
- E3 SNR ablation: fix noise category (natural soundscapes), train separate models at SNR = 0/5/10/20; plot performance vs. SNR.
- E4 Model robustness comparison: under identical settings, compare AST to a non-Transformer baseline (e.g., simple CNN or BiLSTM); report performance differences on noisy test sets.

Evaluation

- Test sets:
  - Clean test set (original RAVDESS).
  - Noisy test sets: for white noise and the three ESC-50 categories, generate offline mixtures at SNR ∈ {0, 5, 10, 20} and evaluate all models consistently.
- Metrics:
  - Weighted Accuracy (WA), Unweighted Accuracy (UA), Macro F1.
  - Confusion matrix (focus on confusable pairs such as Happy/Neutral, Sad/Calm).
- Visualization and artifacts:
  - Save training curves, confusion matrices, and performance–SNR curves to logs.
  - Tabulate key results (model × noise × SNR).

Risks and Controlled Variables

- Keep other augmentations (e.g., speed perturbation, SpecAugment) off or fixed to highlight the main effect of noise augmentation.
- Unify feature extraction, optimizer, and training epochs.
- Use fixed seeds and report mean ± standard deviation.
- Strictly avoid speaker leakage across splits.

Comparative Experiment Design

1. Verify: Adding speech mixed with specific noise categories can effectively improve SER accuracy.
2. Verify: Transformer models such as AST show stronger robustness than non-Transformer models on noisy data.
3. Evaluate: Performance variations under different SNR levels.