Project Supervision Briefing Document

# 🧩 项目题目（可修改）

中文：基于多特征与深度模型的语音情感识别系统研究

English: A Study on Speech Emotion Recognition Based on Multi-Feature and Deep Models

创新点 separate noise from voice

# 📌 1. 项目背景与意义 / Background & Relevance

中文：随着线上教育、语音助手、智能客服等场景日益普及，理解用户的情绪状态成为提升交互体验的重要途径。当前主流语音 API 多侧重语音识别（ASR），而情感识别准确率普遍低于 65%，无法满足实际应用需求。因此，构建一个高效、可解释的语音情感识别系统具有明确的现实价值。

English: With the rise of online education, virtual assistants, and intelligent customer services, understanding users’ emotions is critical to enhancing interaction quality. Existing speech APIs mostly focus on ASR, but emotion recognition accuracy remains below 65%, which limits practical deployment. Hence, this project aims to build an accurate and interpretable Speech Emotion Recognition (SER) system.

# 🌟 2. 研究创新点 / Project Novelty

- 中文：  
1. 提出“多特征 × 多模型”组合的系统性对比策略；  
2. 首次将 Wav2Vec 2.0 等自监督语音嵌入与传统 MFCC、Mel Spectrogram 融合；  
3. 引入新型 Transformer 架构（如 Conformer、MIT AST），探索其在 SER 任务中的表现；  
4. 利用 Grad-CAM 热图可视化分析模型关注区域，提升系统可解释性。

- English:  
1. Proposes a systematic comparison across multiple features × multiple models;  
2. Integrates self-supervised embeddings (Wav2Vec 2.0, HuBERT) with traditional MFCC and Mel Spectrogram for the first time in this task;  
3. Introduces novel Transformer-based models like Conformer and MIT AST to test their capability in SER;  
4. Uses Grad-CAM visualization to identify key attention regions, enhancing model interpretability.

# 🎧 3. 数据集与特征选择 / Datasets & Features

Datasets:  
RAVDESS: ≈1.4 小时，8 类情感，音质高  
IEMOCAP: ≈12 小时，标注精细，交互丰富  
CREMA-D: ≈7 小时，多说话人、真实场景

Feature Types (select 2–3):  
1. MFCC + Δ + ΔΔ  
2. Log-Mel Spectrogram  
3. Wav2Vec 2.0 / HuBERT Embeddings

# 🧠 4. 模型设置 / Model Setup

Baseline Models:  
- MLP（处理 MFCC）  
- Shallow CNN（处理 2D Mel Spec）  
- ResNet-18（2D CNN）

Transformer 系列（可选，若时间允许）:  
- Conformer-Tiny  
- MIT AST（Audio Spectrogram Transformer）

# 📏 5. 评估指标 / Evaluation Metrics

Accuracy, Precision, Recall, F1-score (Macro), Confusion Matrix（混淆矩阵）

# 📊 6. 可视化与表格输出 / Visualization & Reporting

- 性能对比图（不同特征 × 模型）  
- Grad-CAM 热图  
- Pandas + Matplotlib 图表

# 🛠️ 7. 技术工具 / Tools & Libraries

- PyTorch + torchaudio  
- Pandas, Matplotlib  
- MIT AST 官方脚本

# 🧠 8. 为什么不能由生成式 AI 替代？ / Why can’t Generative AI do it?

中文：生成式 AI（如 GPT）只能处理文字输入，无法感知音频或分析声学特征。语音情感识别涉及模型训练、语音特征提取、错误率评估与可视化，这些都依赖于真实音频数据与 GPU 加速，当前 LLM 并不具备直接替代能力。

English: Generative AI models like GPT can only process text. They lack the ability to “listen” to audio, train on acoustic features, or generate performance matrices. SER requires real speech input, GPU-based training, and interpretability tools that LLMs currently cannot replicate.

# 📅 9. 时间计划 / Timeline

第1周（7.1 - 7.7）：确定任务、数据准备、特征提取  
第2–3周（7.8 - 7.21）：训练基础模型（MLP、CNN、ResNet）  
第4周（7.22 - 7.28）：性能评估 + 混淆矩阵  
第5周（7.29 - 8.4）：实现 Transformer 与 AST，集成 Grad-CAM  
第6周（8.5 - 8.11）：可视化分析与中期总结，最终代码，  
第7周（8.12 - 8.18）：模型优化与结果整理  
第8周（8.19 - 8.25）：撰写论文 + 制作答辩材料  
第9周（8.26 - 8.31）：模拟答辩 + 提交最终材料，8.10 前提交论文

English:  
Week 1 (Jul 1–7): Finalize scope, prepare datasets, feature extraction  
Week 2–3 (Jul 8–21): Train baseline models (MLP, CNN, ResNet)  
Week 4 (Jul 22–28): Initial evaluations + confusion matrix  
Week 5 (Jul 29–Aug 4): Implement Transformer & AST, integrate Grad-CAM  
Week 6 (Aug 5–11): Visualization and interim summary  
Week 7 (Aug 12–18): Optimization and results  
Week 8 (Aug 19–25): Write-up and prepare presentation  
Week 9 (Aug 26–31): Mock defense and final submission

 could you please let me know the exact requirements and final submission deadline for the MSc dissertation?

must have some innovation

next meeting online？25/07 basic code

finalcode 10/08