TaskSummaryDoc

Week 5 – “Audio Is All You Need” Sprint (5-day edition)

0 · Purpose

This report condenses the full lecture transcript into a focused five-day execution plan. Keep it open as your living reference while you sprint through the tasks.

---

1 · Five-day timeline at a glance

Day	Key outcome
Mon (today)	• Set up repo, poetry/conda env & data folders.• Read this report + skim slides.• Download **UrbanSound 8K**& script log-mel extractor.
Tue	**Task 1 complete** – CNN baseline ≥ 70 % validation accuracy; confusion matrix saved.
Wed	**Task 2.1 complete** – Whisper tiny fine-tuned; at least **three experimental variants** reported (see §4.2).
Thu	**Task 2.2 complete** – first note-transcription model trained & demo MIDI renders.
Fri	**Extra Task** – reproduce minimal text-to-speech (TTS) pipeline from Xie et al. 2025; slide deck draft & open questions.

> **Tip:** Each evening push code & artefacts; a class-mate must be able to reproduce the current milestone before you sleep.

---

2 · Task catalogue

2.1 Task 1 – UrbanSound 8K sound-event classification

- **Why**: warms you up on audio I/O, spectrograms and CNNs.
- **Pipeline**: 4 s WAV → *log-mel spectrogram* (128 bands, 50 ms window, 25 ms hop) → small ResNet-style CNN → linear classifier.
- **Deliverables**: `train.py`, weights, metrics JSON, model card.
- **Common pitfalls**: class imbalance, overly aggressive time-masking. Use focal loss or class weighting; listen to false-positives.

2.2 Task 2.1 – Whisper pronunciation fine-tune

Baseline code is provided for “Hello, my name is Bess”. Your goal is to design **experiments** that probe catastrophic forgetting and token mechanics.

Required experiments (choose at least three; automate with Hydra or W&B):

1. **Encoder freeze vs full-model** – compare WER and memory usage.
2. **LoRA adapters vs vanilla SGD** on tiny-Whisper (40 MB) for rapid iterations.
3. **Prompt engineering** – explicitly include `<|transcribe|>` to avoid unintended translation.
4. **Custom vocabulary tokens** – inject the rare name into the system prompt to bias decoding.
5. **Accent or emotion token** – add an auxiliary classification head or extra token as in last year’s accent-classifier.
6. **Weight-drift audit** – log the mean and variance of each layer before and after updates.

Evaluation: before/after WER on `{hello, Bess, catastrophic, Italian accent}` sentence set plus five sentences of your own.

2.3 Task 2.2 – Music note transcription

- Collect 200–500 `<wav, MIDI>` pairs (MusicNet, MAESTRO, or self-recorded).
- Transform to log-mel → encoder-decoder → note tokens.
- Metrics: note-on F1 score + 30 s audio reconstruction. Time-pool with convolution to keep < 400 tokens.

2.4 Extra Task – Text-to-Speech survey & reproduction

- Read Xie et al. 2025 for the landscape of options.

3 · Technical primer (fast-track)

- **Wave → log-mel**: STFT over 30 ms windows, convert to Mel, then log-scale; gives a time-frequency “image”.
- **1-D time-pooling convolution**: a stride-2 kernel halves the 3 000-step spectrogram to 1 500 tokens before attention.
- **Whisper token protocol**: `<|startoftranscript|> <language>` `<|transcribe|>/<|translate|>` then predicted tokens.
- **Bug-fix trick**: force `<|transcribe|>` to stop unwanted auto-translation.

4 · Experiment design crib sheet

Theme	Hypothesis	Method
Catastrophic forgetting	Small LR + frozen encoder preserves base accuracy.	Train variants with learning rates {1e-5, 1e-4}; encoder {frozen, LoRA, full}. Measure WER on original sentence set.
Prompt control	Explicit `<	transcribe
Custom vocabulary token	Adding a rare name to the system prompt boosts recall.	Edit prompt: “ The following text uses … ‘Bess’ ...”. Measure name error-rate.
Accent head	Auxiliary classifier for accent retains speaker traits.	Add a one-layer linear head on encoder and use multitask loss.
Weight-drift stats	Large drift correlates with forgetting.	Log mean and variance per layer after each experiment.

5 · Risks & mitigations

- **Imbalanced classes (Task 1)** → use focal loss or oversample minority classes.
- **GPU OOM (music ≥ 120 s)** → chunk & overlap decoding; downsample to 22 kHz.
- **Scope creep into productising** → prioritise research questions over shiny demos.

6 · Definition of done

1. Reproducible commands (`make train_task1`, etc.).
2. Metrics JSON + model cards for each task.
3. Peer can validate in ≤ 1 h.
4. Push to main by **Friday 18:00**.

7 · Quick reference links

- UrbanSound 8K
- TorchAudio transforms docs
- OpenAI Whisper repo + tiny checkpoint

- Xie et al. 2025 (link in slides, appendix C)

Happy sprinting—now hit the **Mon-AM checklist** and get spectrograms flowing! 🎧
