Speaker Identification & Confidence Calibration – Home Assignment

1. Background

SpeakCare keeps a roster of **N** nurses.

- **Roster audio.** For each nurse we currently give you **one** enrollment utterance (about 15 seconds, 16 kHz, `.m4a`).
  Design your code so it can later accept multiple enrollment utterances per nurse without major changes (e.g. 10 utterance)
- **Session audio.** You receive **one treatment-session recording** (≈ 30 min, 16 kHz, *m4a*). Exactly **one nurse from the roster speaks for the majority of the session**, but other people (patient, family, staff) may also speak briefly.

This file is for **inference only -** do not use it for training or calibration.

2. Task

Build a system that, given the enrollment set and the session recording, outputs:

Output	Description
Predicted nurse ID	The roster nurse you conclude is speaking most of the time
Probability vector	P(nurse_i | session) for all N nurses, summing to 1
Confidence / quality / credibility statement	A quantitative indication of how certain you are in the top prediction

You may choose to segment the session or apply any preprocessing techniques to handle silence or the presence of other speakers. The specific approach is up to you to determine and justify.

3 Documentation (`README` ≤ 2 pages)

Describe:

- Model / library choices.
- How you decode *.m4a* (and resample, if needed).
- Any segmentation, embedding, scoring, calibration, or confidence-estimation methods you implemented.
- Instructions for installing and running your code.

4 Submission

1. **Source code** (Python ≥ 3.10) with reproducible instructions in your `README`
2. `README` (≤ 2 pages) containing:
   - Overall top decision for the provided session (and, if you tested more sessions on your own, a confusion matrix or accuracy table, or any other metrics you calculated).
   - The probability vector and your chosen confidence / credibility measure, or any other quality metric you calculated.
3. **Command line** we should run to score a new session, for example:

Shell

```
python infer.py --enroll_dir ENROLL/ --session SESSION.m4a --out results.json
```

5 Evaluation Criteria

Criterion	**How we judge it**
Correctness	Whether the top prediction matches the true dominant nurse on a hidden test set.
Probability quality	Likelihood (or other principled metric) of the true nurse under your probability vector, and plausibility of your confidence measure.
Clarity & reproducibility	We must be able to run your code end-to-end, on CPU, without modification.

5 Data

You can find a roster of nurses with 6 recordings, and a single treatment session here - Link