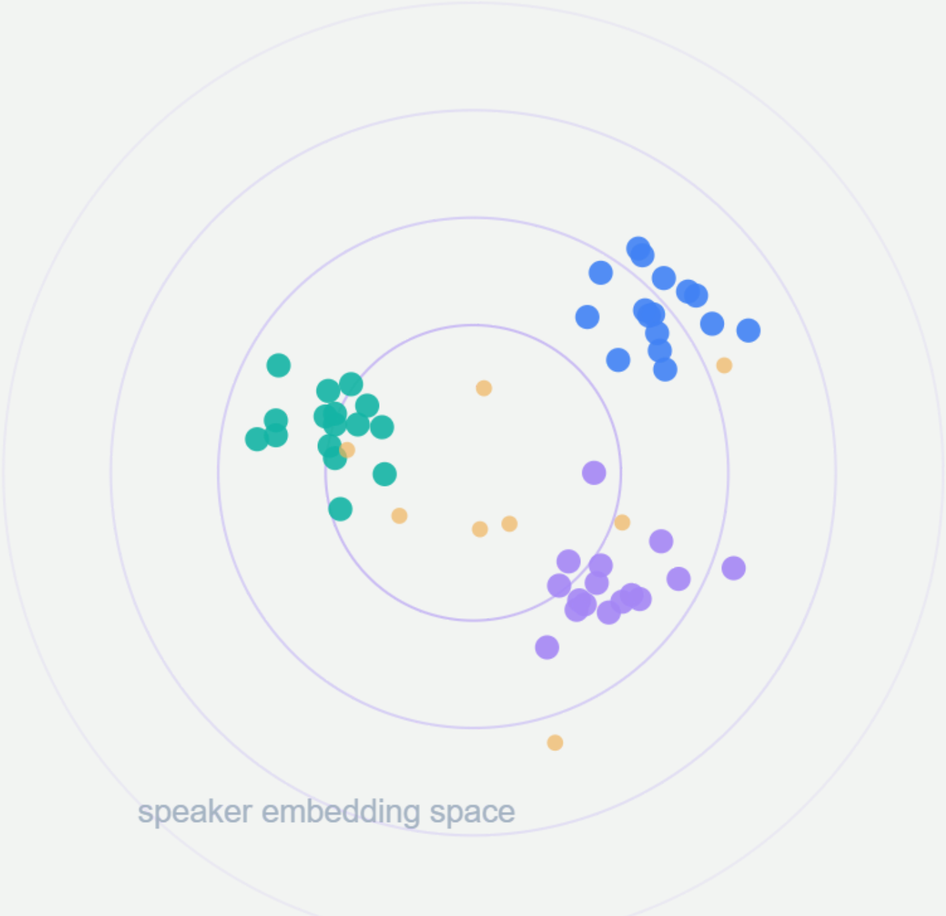
Master DATA SCIENCE AND ANALYTICS ·
SPEECH PROCESSING

المدرسة العليا للتكنولوجيا–اسفي
ECOLE SUPÉRIEURE DE TECHNOLOGIE -SAFI
MARRAKECH
جامعة القاضي عياض
UNIVERSITÉ CADI AYYAD

Speaker recognition



speech waveform



speaker embedding space

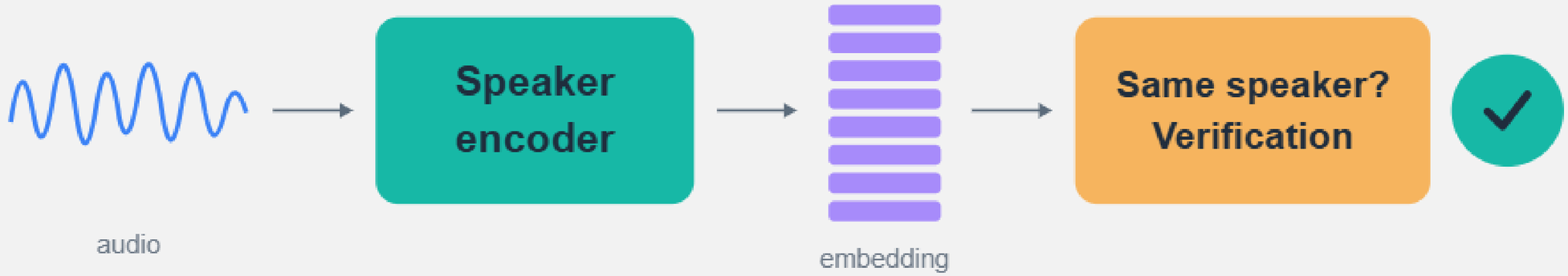
Presented by : NMISSI Nadia, JAIFRA Ichrak,Siham EL KHASSAL & OUANNANI Manal

Supervised by : Pr .HOURRI Soufiane

OUTLINE

1. CONTEXT AND ROADMAP
2. RELATED WORK AND PAPER CLASSIFICATION
3. SEED: METHODOLOGY AND EXPERIMENTAL SETUP
4. RESULTS AND CRITICAL ANALYSIS
5. OUR WORK AND CONCLUSION

Context and Roadmap



Speaker recognition

Who is speaking?

Access control, forensics, verification, voice interfaces.

ASR

What is being said?

Transcription, regardless of the speaker.

Part I

Classification, comparison and contributions of Eight related papers.

Part II

Detailed analysis of SEED (Interspeech 2025).

Eight Papers, Four Categories

1 · ASR and speaker adaptation

Adaptive AVSR

Pseudo-labeling, elderly speech

2 · Diarization

SDR-LLM

Global-Local Speaker Classification

Speaker recognition

3 · Federated learning

Dual Classification Heads

Fisher Information

4 · Speaker recognition enhancement

Pitch raising study

SEED (selected paper)

Category 1: ASR and Speaker Adaptation

THE GENERAL IDEA:



noisy or elderly speech

Speaker / environment adaptation or pseudo-labeling



transcription

TWO RESEARCH DIRECTIONS :

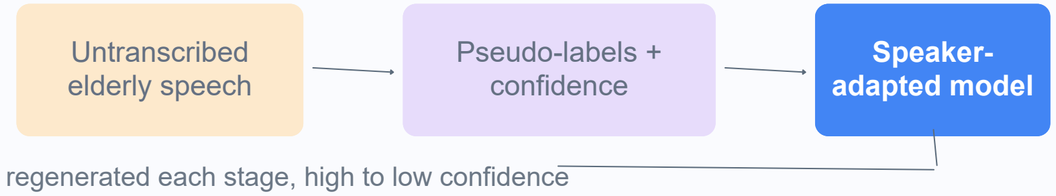
Adaptive AVSR



Main approach: Speaker and environmental embeddings are integrated into an audio-visual speech recognizer.

Objective: Keep recognition robust in noisy conditions.

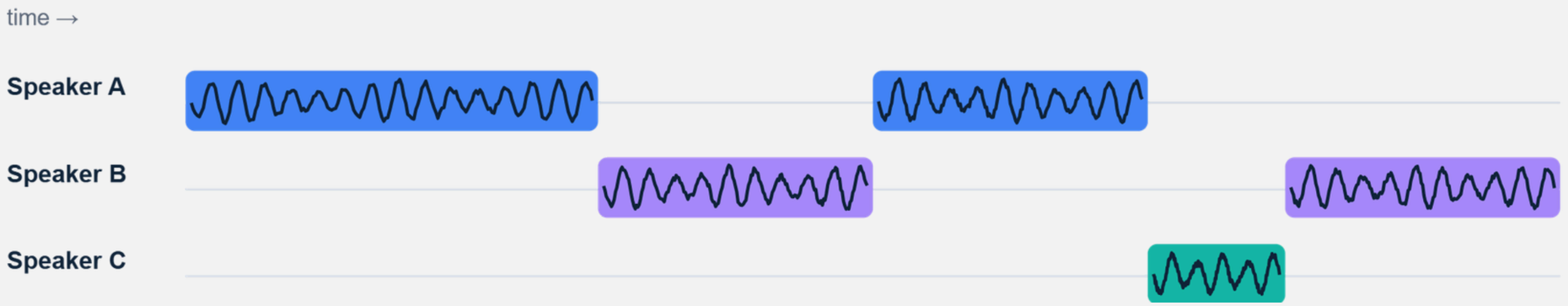
Elderly-speech pseudo-labeling



Main approach: Incremental pseudo-labeling guided by confidence scores (high to low), with speaker adaptation.

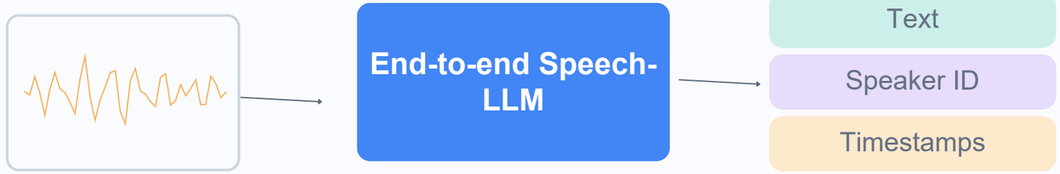
Objective: Use untranscribed data to improve elderly speech recognition.

Category 2: Diarization: Who Spoke When?



TWO RESEARCH DIRECTIONS :

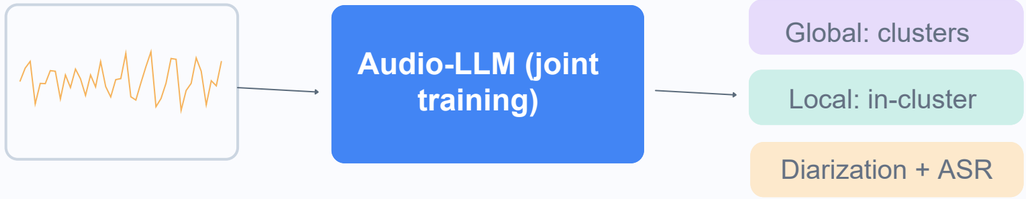
SDR-LLM



Main approach: One end-to-end speech-LLM predicts text, speaker identity and timestamps.

Objective: Replace separate diarization and ASR modules with one model.

Global-Local Speaker Classification

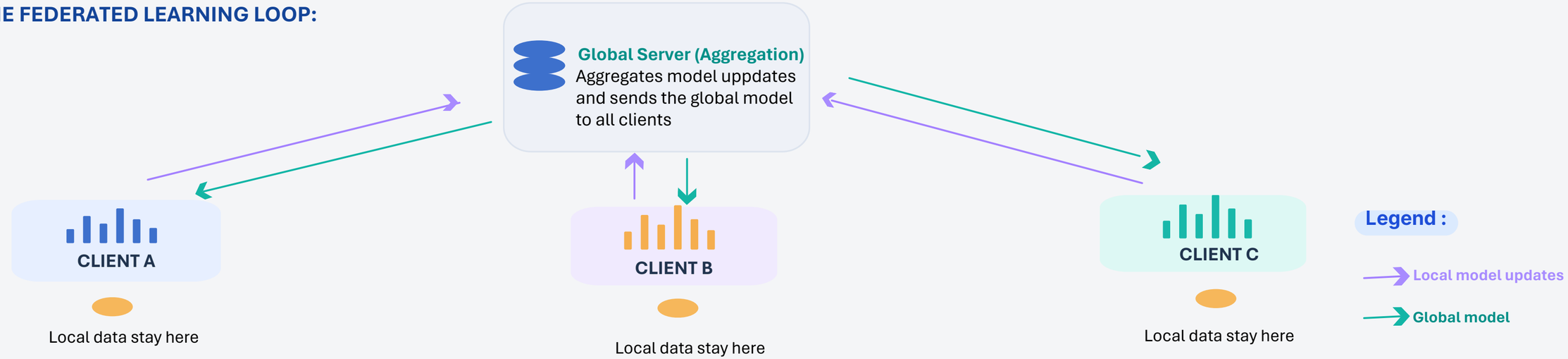


Main approach: Speaker classification is trained jointly with diarization and recognition, using global clusters and local speakers.

Objective: Better speaker discrimination without large real conversation data.

Category 3 : Federated Speaker Recognition

THE FEDERATED LEARNING LOOP:



TWO RESEARCH DIRECTIONS :

Dual Classification Heads

A Federated Learning-Based Speaker Recognition Method with Dual Classification Heads



Main approach: Uses two classification heads: one client-specific and one global

Objective: Handle differences between clients while preserving shared knowledge

Fisher Information

Learning Global Key Knowledge for Federated Speaker Recognition via Fisher Information

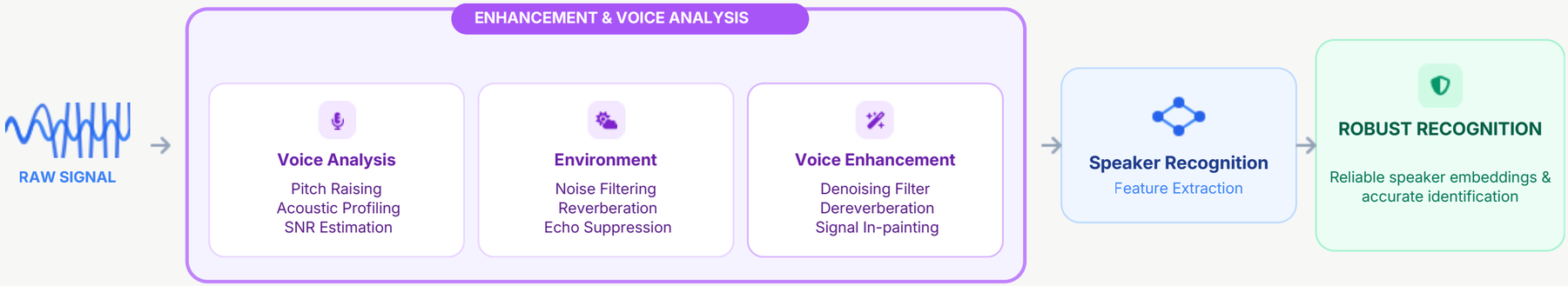


Main approach: Uses Fisher Information to identify important dimensions of speaker embeddings.

Objective: Preserve useful speaker information and improve collaboration between clients.

Category 4 : Speaker recognition enhancement

FROM VOICE ANALYSIS TO ROBUST SPEAKER RECOGNITION



TWO RESEARCH DIRECTIONS :

Pitch Raising Study

Examines voluntary and synthetic pitch raising in female speakers and its effect on automatic speaker recognition.

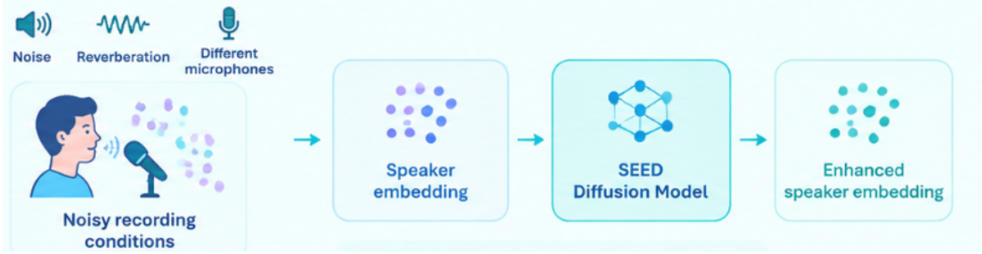


Main approach: Studies the effect of voluntarily and synthetically raising the pitch of female voices on automatic speaker recognition.

Objective: Determine how pitch modifications affect the system's ability to recognize a speaker.

SEED

Speaker Embedding Enhancement Diffusion Model



Main approach: Uses a diffusion model to enhance speaker embeddings rather than directly denoising the audio signal.

Objective: Improve the robustness of speaker representations under changing recording conditions.

From Related Work to the Selected Study

Key Insights from the Literature Review



1. Speaker Adaptation

Objective Adapting systems to speakers and speaking conditions.



2. Speaker Diarization

Objective Identifying who speaks and when.



3. Federated Speaker Recognition

Objective Collaborative training across distributed datasets.



4. Enhancement and Voice Analysis

Objective Studying voice changes and improving speaker representations.

The Common Challenge

How can we maintain reliable speaker recognition when voices or recording conditions change?

SELECTED PAPER

SEED: Speaker Embedding Enhancement Diffusion Model

Problem Environmental mismatch can degrade speaker recognition performance.

Approach A diffusion model refines speaker embeddings instead of enhancing the raw audio.

Goal Improve recognition robustness without modifying the pretrained speaker recognition model.

The Problem: Environment Mismatch

1. The Basic Principle

A speaker recognition system learns to represent a person's voice using a **speaker embedding** :a numerical representation that captures speaker-related characteristics.



2. What Goes Wrong?

Environmental changes can alter the acoustic characteristics of a recording and make the speaker's representation less reliable.

Consequence :

The system may find it harder to distinguish the speaker from other speakers, leading to more recognition errors.

3. The Key Research Question

How can we make speaker embeddings more robust to changes in recording conditions?

Core Idea: Diffusion in Embedding Space



Clean + noisy audio → Frozen speaker encoder → Speaker embeddings → SEED diffusion model → Refined embedding

Key point: SEED refines the speaker vector — it does not denoise the waveform or replace the pretrained encoder.

Method and Architecture: Equation 9

TRAINING



$$L_SEED = E\left[\|x_0 - f\theta(x_t, t)\| + \Sigma_{k=0}^{N-1} \|x_0 - f\theta(y_t^k, t)\| \right]$$

Clean-embedding term + N noisy-embedding terms (N = 3)

TARGET

x_0 = clean speaker embedding
The same target is used for every noisy variant.

INPUTS TO SEED

x_t = diffused clean embedding
y_t^k = diffused embedding from noisy version k

NETWORK

fθ predicts the clean embedding directly.
3 residual fully connected blocks; LayerNorm + SiLU + Linear.

Why the second term matters: it trains SEED to map noisy-condition embeddings back toward the clean target x_0.

Experimental Protocol

TRAINING SPEECH

LibriTTS-R · 460 h
Libri-Light · 577 h

→

AUGMENTATION

MUSAN: noise + music
Simulated RIR: reverberation

→

PRETRAINED ENCODERS

ResNet34 · 512-D
ECAPA-TDNN · 256-D
WavLM-ECAPA · 192-D

→

SEED TRAINING

AdamW · LR 0.0005 · ~60 epochs
DDIM: 1,000 training steps
Inference: one step at t = 50

EVALUATION DATASETS

Generalisation: Vox1-O, Vox1-E, Vox1-H
Environment mismatch: VoxSRC23, VC-Mix

Purpose: test both standard performance and robustness to recording-condition changes.

METRICS

EER — Equal Error Rate: false-accept and false-reject rates are equal at this point.

minDCF — minimum Detection Cost Function, with Cmiss = 1, Cfa = 1, Ptarget = 0.05.

Lower values are better.

Training needs clean speech but no speaker labels. Each clean utterance is paired with three simulated noisy versions.

Results: Gains Depend on Model and Condition

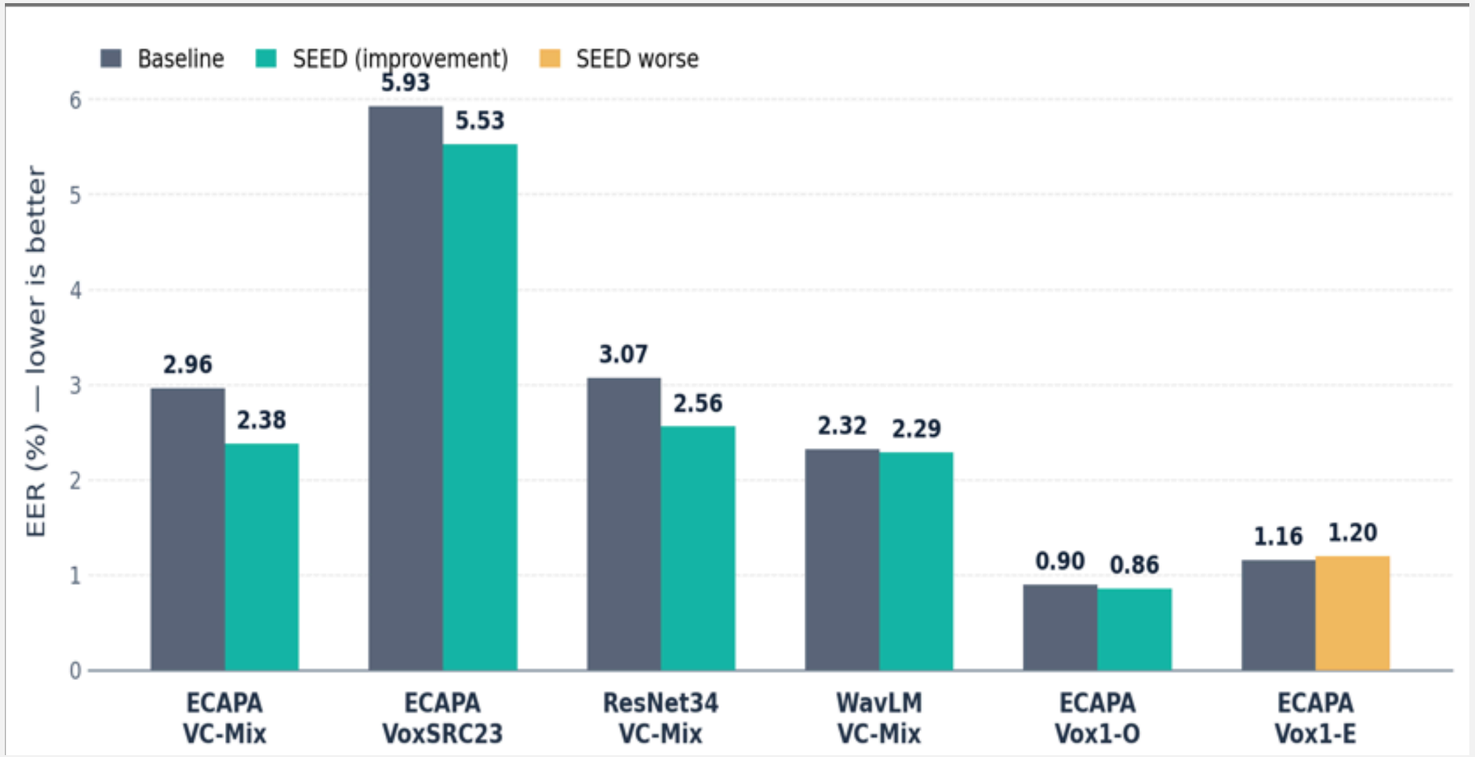


Table 2 · cost (ECAPA)

ECAPA →	SE+ECAPA	SEED
Params (M)	46.6	18.8
RTF	0.0038	0.0020
Memory (GiB)	0.88	0.59
EER VC-Mix (%)	4.63	2.38

EER (%), lower is better. Largest gain: ECAPA on VC-Mix, 2.96 → 2.38 (–0.58 points, –19.6% relative). Small gain for WavLM-ECAPA; slight loss for ECAPA on Vox1-E.

Critical Analysis

Strengths

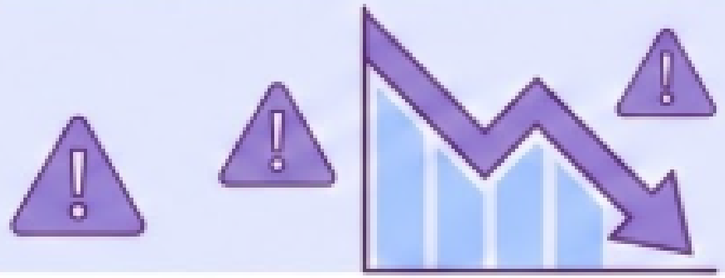


Strengths

- No speaker labels needed
- Lightweight: fewer parameters, less memory
- No change to the pretrained extractor
- Best under environment mismatch



Limitations



- Gains vary across models (authors' tables)
- Slight degradation, e.g. ECAPA on Vox1-E (our observation)
- Training instability (authors)
- Depends on extractor quality (ours)
- Reproduction needs data and compute (ours)

Our Research and Reproduction Plan

Starting point: reproduce SEED (full or reduced) with a pretrained ECAPA-TDNN baseline.

Candidate contribution directions · one or several, to be selected

Non-diffusion ablation
Replace diffusion with a simple enhancer (MLP or denoising autoencoder) to test what diffusion adds.

Inference steps
Study the diffusion time step and one-step versus multi-step sampling.

Adaptive fusion
Combine original and enhanced embeddings, with a justified weighting.

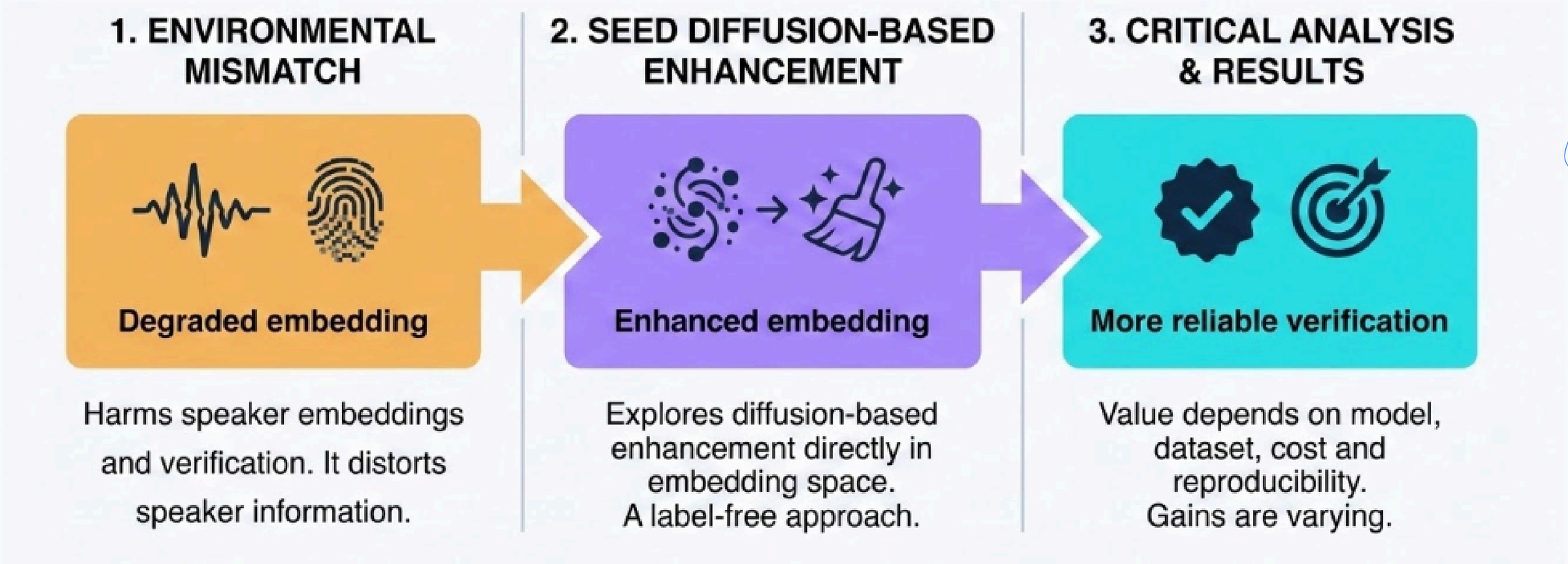
Objective and stability
Explore alternative losses (e.g. angular) and ways to stabilize training.

Broader degradations
Test other noise types, codecs, languages and embedding extractors.

Subgroup analysis
Compare gains across speaker groups (e.g. gender) and conditions.

Any direction is evaluated with EER and minDCF, using disjoint training and test speakers. Proposed work, not results.

Conclusion


1. ENVIRONMENTAL MISMATCH
Degraded embedding
Harms speaker embeddings and verification. It distorts speaker information.
2. SEED DIFFUSION-BASED ENHANCEMENT
Enhanced embedding
Explores diffusion-based enhancement directly in embedding space.
A label-free approach.
3. CRITICAL ANALYSIS & RESULTS
More reliable verification
Value depends on model, dataset, cost and reproducibility.
Gains are varying.

Thank You