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



Speaker Recognition

Summary of 6 papers by problem + in-depth study of cross-domain retrieval

AUDIO → EMBEDDING → MATCH /SEARCH

Presented by : Ihssane LABIBI - Houda ELBASIT - Ibrahim RZIQI

Presentation Plan

1 PART 1 : Sorting by problem

- Recording conditions: codec, noise, duration
- Continuous, real-time verification
- Voice imitation and pitch change
- Goals and solutions from the papers

2 PART 2 : Chosen paper

On the Robustness of Speaker Embeddings for Cross-Domain Speaker Retrieval

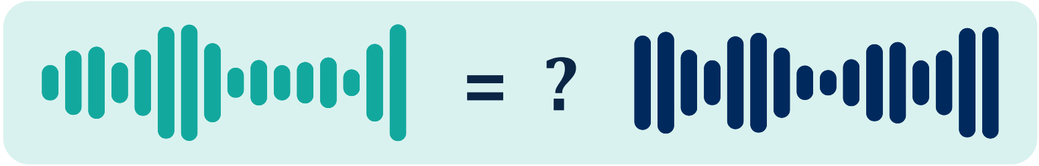
- Problem and motivation
- Main idea of the solution
- Concept architecture

BACKGROUND

Before comparing the papers: two different tasks

This difference helps us avoid mixing up the goals of the systems.

Speaker verification : “Is it the same person?”



Input: two voice recordings.

The system computes how similar their representations are, then accepts or rejects the claimed identity.

Example: compare a voice recorded today with a reference voice.

Speaker retrieval : “Who sounds like this voice?”



Input: one query voice and a collection of recordings.

The system ranks the recordings from most similar to least similar.

Example: find clips of the same speaker in a database.

Voice embedding: a list of numbers that sums up voice features that help tell speakers apart.

PART 1

Sorting by problem

The seven papers do not all deal with the same difficulty.

A. Signal quality

2) Codec + duration

4) Background noise

B. Duration and continuity

3) Continuous verification

5) Verification on short segments

C. Voice variation

6) Voice imitation

7) Voluntary or synthetic pitch

Common question

How can we avoid mistaking a change of channel, duration, noise or speaking style for a change of identity?



PART 1

Classification of Speaker Recognition Problems

Six papers · Different challenges · Proposed solutions

PROBLEM 1

Codec, duration and speaker-specific effect

Paper: Codec-induced Mismatch, Speech Duration, and Speaker-dependent Effect in a DNN-based Forensic Speaker Recognition System

PROBLEM

- The codec compresses or changes the signal.
- Audio compression and speech length can make recognition less reliable, with different effects on different people.
- Short clips contain less voice information.

GOAL

Measure how codec mismatch, speech length and speaker identity affect the scores and the recognition errors.

SOLUTION / STUDY

Compare the scores in several conditions: original or coded audio, segments of different lengths, and several speakers. Keep the other factors under control.



Different clip lengths, different speakers

PROBLEM 2

Continuous and real-time verification

Paper: A Lightweight Continuous Speaker Verification System for Real-time Monitoring

PROBLEM

During a call (bank, insurance), someone else can take the person’s place.

GOAL

Detect whether the voice we hear still matches the authorized speaker, with low delay.

SOLUTION / MAIN IDEA

A fast program compares the voice with the saved voice every half second and warns the employee if it changes.



PROBLEM 3

Noise and robustness

Paper : Mixture Consistency Learning for Robust Speaker Verification in Noisy Environments

PROBLEM

Street noise, background conversations or a fan can hide voice features and make verification worse.

GOAL

Recognize the voice despite the noise.

SOLUTION / MAIN IDEA

The program learns by itself to separate the voice from the noise, without needing a perfect recording as a model.



PROBLEM 4

Verification on very short speech

Paper : Stabilizing Short Duration Speaker Verification through Neural Re-scoring with Hybrid Enrollment

PROBLEM

Over a few hundred milliseconds or a few seconds, embeddings can be unstable and scores can be unreliable.

GOAL

Make decisions more stable when the test recording is short and the speaker's enrollment uses a hybrid strategy.

SOLUTION / MAIN IDEA

We use two types of recordings of the person (the same phrase and free speech), then a second program refines the decision.



PROBLEM 5

Assessing voice imitation

Paper : ECAPA-TDNN-based Speaker Embedding Framework for Voice Mimicry Assessment

PROBLEM

A person can imitate the tone, rhythm or intonation of another. A strong sound similarity does not always mean the two voices have the same identity.

GOAL

Measure how similar an original voice and an imitation are, using representations learned to tell speakers apart.

SOLUTION / MAIN IDEA

The program listens to the sound and the way of speaking, then ranks the imitators. It finds the best one 3 times out of 4.





3 / 4 **best imitator found**

PROBLEM 6

Pitch change

Paper : Assessing the effect of volitional and synthetic pitch raising in female speakers on automatic speaker recognition

PROBLEM

A person can use a higher voice to hide.

GOAL

Check whether raising the pitch causes recognition errors, and whether the effect is different between a natural and a synthetic change.

SOLUTION / STUDY

Compare original recordings, voices with a voluntarily raised pitch, and digitally changed versions, using the same recognition system.



Variables to control: **speaker, duration, spoken content, sound level and method of change.**



PART 2

Study of the paper : On the Robustness of Speaker Embeddings for Cross-Domain Speaker Retrieval

PART 2

Why choose this paper and not another one?

Choice criteria for a Master's project that can really be done.



RELEVANCE

- Real problem: change of domain.
- Directly linked to embeddings and voice recognition.
- Lets us discuss generalization.



FEASIBILITY

- A prototype is possible with a pre-trained model.
- No need to retrain a large network at the start.
- Results with numbers.



WHY NOT THE OTHERS?

- Codec / pitch: a narrower problem.
- Noise / short segments: sometimes tied to a specific method that is hard to reproduce.
- Real time: adds delay constraints.

Still, the choice depends on the code and data that are available.

PART 2

Chosen paper : Problem and goal

Working title: On the Robustness of Speaker Embeddings for Cross-Domain Speaker Retrieval

PROBLEM

While voice recognition systems achieve excellent results during clean lab tests, they often fail in real-world applications when the recording conditions change.

The paper studies four types of differences:

Recording channel: Phone or network.

Acoustic environment: quiet room, echo or noise.

Language: recordings in different languages.

Age: Voices recorded across life stages, testing resilience to aging.

MAIN IDEA OF THE PAPER:

Check whether a system can still find the right people when voices are recorded in different conditions, then test a strategy that improves the results without retraining the model.
The paper compares several voice representation models and leverages Adaptive Symmetric Normalization (ASN) as a training-free backend method to improve results when recording conditions are different.

PART 2

Concept architecture for the study

1

Voice recording

A person’s voice

→

2

Pre-trained model

Extracts the voice features

→

3

Speaker embedding

A list of numbers that represents the voice



→

4

Comparing the vectors

Cosine similarity calculation and ranking of the voices

Speaker embedding — Made-up example: [0.32 ; 0.71 ; −0.18 ; …]

The six models compared are:

CAM++ | ECAPA-TDNN | ERes2Net | x-vector | RDINO | SDPN

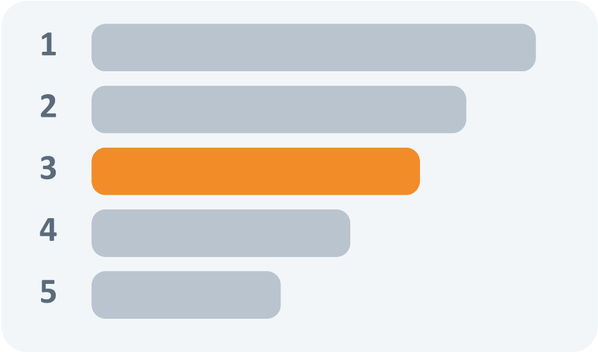
They are pre-trained on VoxCeleb2 and then tested on different scenarios, without retraining on the target domains

THE PROPOSED SOLUTION

Adaptive Symmetric Normalization (ASN)

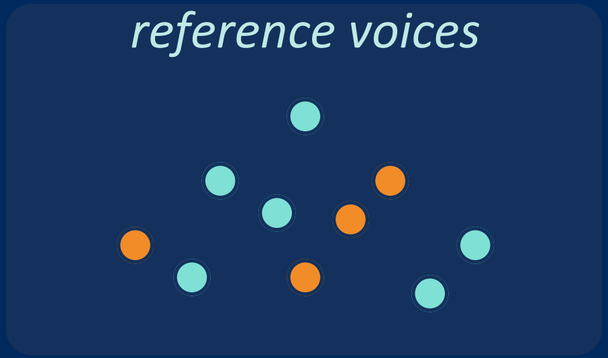
Fix the score differences that appear when recordings come from different conditions, without changing the neural models or retraining their parameters.

1 Initial similarity scores



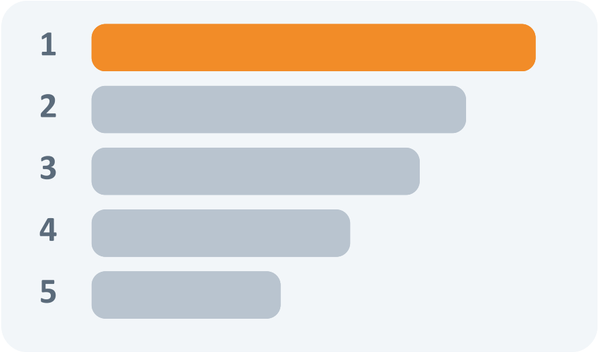
Domain differences can disturb the ranking.

2 ASN



Picks reference voices and adjusts the scores based on their distribution.

3 Improved ranking



The recordings most likely to belong to the wanted person move up in the ranking.



Thank you for your attention