Speaker Recognition

Thematic review of nine articles and in-depth study of a system used in police investigations

Literature Review

October 2026

Presented by:

Badreddine BENHILA
Issam ELBOUAZOULI
Mohamed Khalil SIDQI
Ouassim DERJA

Supervised by:
Dr . Soufiane hourri

Outline

01

Part 1 · Overview by theme

Nine articles on speaker recognition, grouped into four themes according to their main objective.

02

Part 2 · In-depth study

How audio compression, speech duration and differences between speakers affect a forensic voice system (Deng, Wang & Hughes, Interspeech 2026). Steps: Question, Literature, Method, Results, Conclusion.

03

Conclusion

Key results, limitations of the study and outlook.



Part.1

Overview by theme

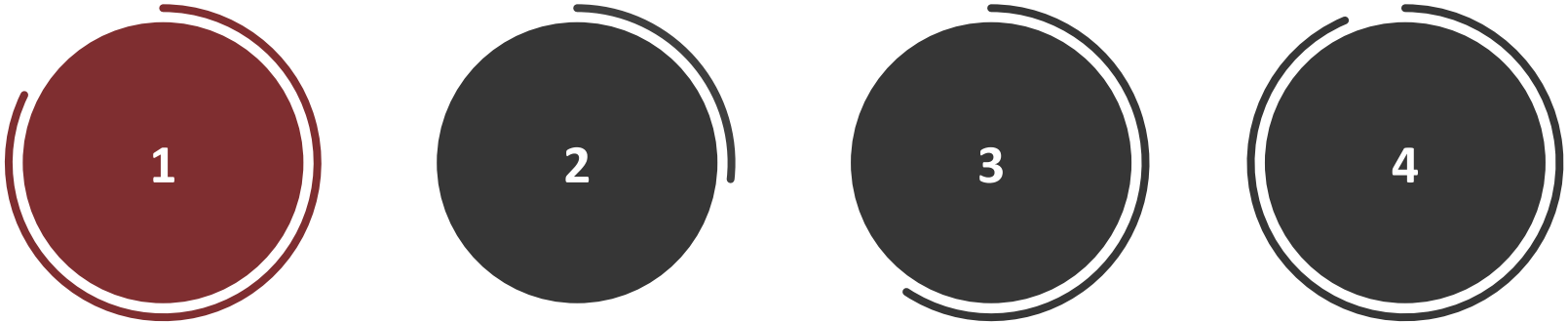
Nine articles, four themes

The nine articles chosen

1. Assessing the effect of volitional and synthetic pitch raising in female speakers on automatic speaker recognition
2. Codec-induced Mismatch, Speech Duration, and Speaker-dependent Effect in a DNN-based Forensic Speaker Recognition System
3. Vocal Tract Disparity and Potential Implications for Speaker Recognition
4. A Federated Learning-Based Speaker Recognition Method with Dual Classification Heads
5. Learning Global Key Knowledge for Federated Speaker Recognition via Fisher Information
6. VieSpeaker: A Large-Scale Vietnamese Speaker Recognition Dataset Beyond Visual Dependency
7. Balancing ASR and diarization in end-to-end LLMs for multi-talker speech recognition
8. Joint Learning Global-Local Speaker Classification to Enhance End-to-End Speaker Diarization and Recognition
9. SDR-LLM: Speech-LLM Based End-to-End Speaker Diarization and Recognition with Sentence-Level Temporal Modeling

Classification: four themes, nine articles

Criterion: the main objective of each article: the problem studied, the method used and the main result.



Reliability and voice characteristics

3 articles

Voice pitch, codecs, vocal tract

Private (federated) learning

2 articles

Local and shared models

Voice databases

1 article

Large voice corpus

Diarization and conversations

3 articles

Who speaks when, and what

Groups may overlap: each article is placed according to its main contribution.

Theme 1 · Reliability and voice characteristics

Which factors make speaker recognition more or less reliable?

VOICE

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

Does raising the voice pitch, deliberately or with software, change recognition quality for women?

CHANNEL

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

How audio compression (phone calls) and recording duration affect a forensic voice system.

Chosen for Part 2

BODY

Vocal Tract Disparity and Potential Implications for Speaker Recognition

How the shape of the throat and mouth (vocal tract) differs from one person to another, and why it matters.

Common point: each article finds a factor that can make recognition less accurate.

Theme 1 in plain words · Voice reliability

1

The voice itself changes

Article 1: does raising the voice pitch, deliberately or with software, change recognition quality for women?

2

The phone line changes the sound

Article 2 (our article): compression in calls and apps, and short recordings, make a voice harder to recognize.

3

Every body is different

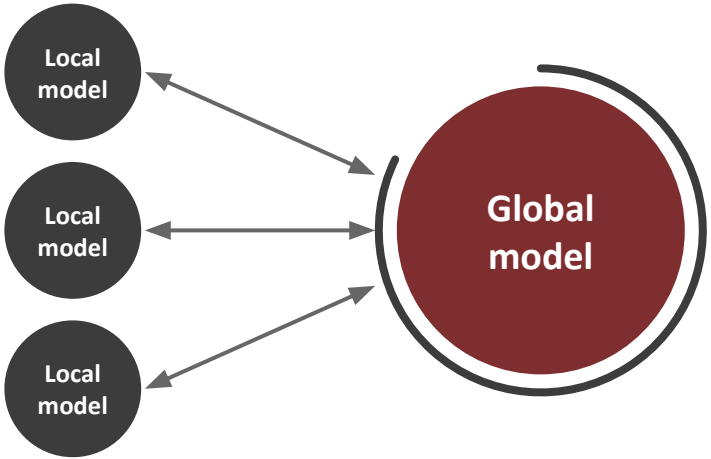
Article 3: the shape of each person’s vocal tract (throat, mouth, nose) differs, and this may matter for recognition.

IN ONE SENTENCE

Recognition is not always equally reliable: the voice, the channel and the body can change the result.

Themes 2 and 3 · Private learning and voice databases

Theme 2 · Federated learning = private learning



knowledge shared in both directions

Voice data never leaves each client.

Challenge: each client's data is different.

Article 4 · two classifiers working together to make private learning work.

Article 5 · chooses the most useful knowledge to share (Fisher information).

Theme 3 · Voice databases

Text descriptions and transcripts

↓

No need to see the speaker on screen

↓

Large Vietnamese voice database

Article 6 · VieSpeaker, a database for training and testing systems.

Theme 2 in plain words · Federated learning

1

The problem

Training a model takes a lot of voices, but voices are private. Sending them all to a single server is risky.

2

The idea

Each client trains a local model with its own voices. Only what was learned is sent to a shared (global) model.

3

The difficulty

Each client has different voices, so the local models drift apart. Articles 4 (two classifiers) and 5 (Fisher information) help.



Theme 3 in plain words · Voice databases

1

The problem

AI learns from examples. Without a large collection of labeled voices, a good system cannot be trained or tested.

2

What a database is

A large organized set of recordings, each labeled with the person speaking. Article 6 builds one: VieSpeaker (Vietnamese).

3

The trick

No need to see the speaker on video: text information (descriptions and transcripts) is used to find and label speakers.

IN ONE SENTENCE

No data, no good system: a database makes it possible to learn and to test.

Theme 4 · Diarization (who speaks when) and conversations

One conversation, three questions



1

Diarization
who speaks when

2

Transcription
what is said

3

Speaker Recognition
who the person is

A single model learns the three tasks together, instead of separate steps.

Three articles

Article 7

Balances transcription and “who speaks when” in a large language model, for several speakers.

Article 8

Looks at the whole recording and short segments together to find and identify speakers.

Article 9

SDR-LLM: a language model for speech, which tracks time sentence by sentence.

Theme 4 in plain words · Diarization and conversations

1

The situation

In a meeting, several people speak. We want to know who speaks when, what is said, and who each person is.

2

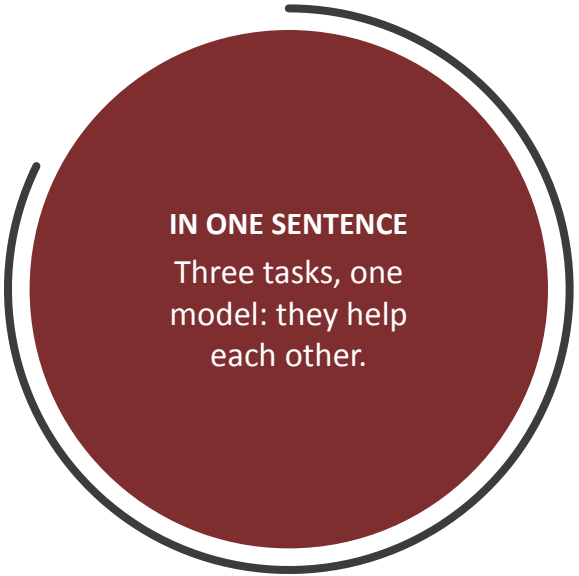
The old method

Three separate tools, one after the other. An error in the first step spreads to the next ones.

3

The new method

A single model learns the three tasks together (articles 7 to 9). Two use large language models, AIs like ChatGPT.



Part.2

In-depth study of the article

Question · Literature · Method · Results · Conclusion

Article · The principle: comparing two voices

QUESTIONED RECORDING

Voice of the unknown author (call, message). Often goes through a phone or an app, and is often short.

KNOWN RECORDING

Voice of the suspect, recorded at the police station (interview). Long and of good quality.

The system compares the two voices

Result: a likelihood ratio (LR)

LR above 1: more likely the same person. LR below 1: more likely two different people. The further from 1, the stronger the evidence.

Article · Question: what problem to solve?

1

Phone calls damage the voice

In a real case, the recording of the unknown author often goes through a phone or an app. Compression (the “codec”) erases details of the voice.

2

Sometimes there is very little speech

The questioned recording is often short. The suspect’s, made during an interview, is long and of good quality.

3

A good average can hide bad cases

Systems are judged on an overall score. But a court needs a reliable result for this specific person.

THE QUESTION

Does a neural network system remain reliable for each speaker when the sound is compressed and the recording is short?

Article · Literature: what is known, and the gap

WHAT IS ALREADY KNOWN

1. Neural network (DNN) systems are now the best, even when recording channels differ.
2. Compression and lack of speech lower accuracy; low-bitrate codecs do more damage (Alsalihi & Sztahó, 2025).
3. These systems are increasingly used in real cases, in many countries.

THE GAP

1. A single codec, or identical conditions: no side-by-side comparison of several codecs.
2. Codec and speech duration are studied separately: we do not know how they influence each other.
3. Only overall scores: nobody checks whether each speaker is treated well.

This study: 3 codecs (4 settings) × speech durations from 5 to 90 s, looking at each speaker one by one.

Article · Method (1): the data

102 young Cantonese men from Hong Kong (average age: 23)

4 recordings per person, in 2 sessions at least 3 weeks apart

30 cm from the microphone, in a soundproof room, for very good quality

TWO SESSIONS

Session 1 → “known” recording

Mock interview + free discussion (topic drawn from a card). Long recordings, over 90 s.

Session 2 → “questioned” recording

Conversation with an accomplice + free discussion. This is the one that will be degraded and then shortened.

Preparation: original high-quality files (WAV 44.1 kHz). Manual cutting: only the speaker’s speech is kept, so file duration = speech duration.

Article · Method (2): the codecs

A codec is a program that compresses sound so it can be sent quickly. It loses small details of the voice. The “bitrate” (kb/s) is the amount of information kept per second: the lower it is, the more is lost.

G.711 A-law	Landline phone, 64 kb/s. The sound is compressed while keeping its waveform (sampled at 8 kHz).
AMR-NB	Mobile phone. Instead of the sound, voice parameters are sent. Two settings: 12.2 kb/s (good) and 6.7 kb/s (heavily compressed).
Opus	Internet calls and voice messages (WhatsApp), 128 kb/s: very good quality.
HQ	Original recording, without compression. Used as the reference.

Simulation: FFmpeg for G.711 and AMR-NB, libopus for Opus.

Article · Method (3): the test protocol

68
speakers to train the comparator

This statistical model (LDA-PLDA) learns what changes from one person to another, and what changes within the same person.

34
other speakers for testing

Known = session 1 (intact). Questioned = session 2, degraded by the codec, then shortened to 5, 10 … 90 s (the first seconds are kept).

130 | 4,292
same-speaker pairs | different-speaker pairs

Each questioned recording is compared with all known ones; calibration and testing by leave-one/two-out cross-validation. (3 recordings shorter than 90 s are discarded.)

What we vary: 5 codec conditions (HQ, G.711, AMR-NB 12.2, AMR-NB 6.7, Opus) × 18 speech durations (from 5 s to 90 s, in 5 s steps). Only the questioned recording is degraded: the known one stays of good quality, as in a real case.

Article · Method (4): from sound to voiceprint

1

Prepare the sound

The sound is brought down to 8 kHz (phone quality). A filter boosts high frequencies (pre-emphasis).

2

Image of the sound

Every 10 ms, the energy is measured in 64 frequency bands: this gives a “64 × time” image (spectrogram), normalized (CMVN).

3

ResNet34 network

A 34-layer network, known for images, reads this image and picks out voice traits at each moment.

4

Attention + summary

The network (8 attention heads) gives more weight to useful moments and summarizes everything in 512 numbers: the voiceprint.

NETWORK TRAINING

Trained on VoxCeleb 1 and 2 (7,205 people). To withstand difficult conditions, noise (MUSAN), room echo (RIR) and masked regions (SpecAugment) are added to the sound. Network tuned with the AAM-softmax loss for 105 epochs.

Article · Method (5): from voiceprint to LR

1

Simplify

LDA reduces the 512 numbers to 50: it keeps what distinguishes people and discards part of the channel effect.

2

Compare

PLDA computes the probability of the two voiceprints if it is the same person, divided by that if they are two different people. It accounts for similarity and for how rare the voice is.

3

Calibrate

A logistic regression (bi-Gaussian calibration) corrects the scores so that the LRs are honest: an LR of 100 must really be worth 100.

4

Result

One LR per pair of voices, usable to express the strength of evidence.

LR =

probability of the voices if same person

÷

probability of the voices if different people

The "cosine" used elsewhere only measures similarity. The LR adds how rare the voice is, which is essential in court.

Article · Method (6): performance measures

Cllr
whole system

Measures how far LRs go in the wrong direction (high LR for two different people, or low for the same one). 0 = perfect; less than 1 = useful.

Cllr / speaker
a single person

Same calculation, but for one speaker: their “same person” pairs and their pairs against all the others. Few pairs: more variable value.

Regression
sensitivity to codecs

At 90 s, for each speaker: mean Cllr (overall level) and standard deviation across codecs (sensitivity). We test whether the two are linked.

Also used: Tippett curves, which show how LRs are distributed for “same speaker” and “different speakers” pairs.

Article · Results (1): the whole system

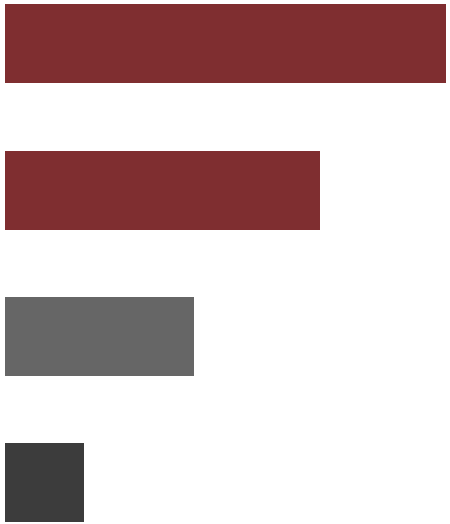
Which codec does the most damage?

AMR-NB 6.7 kb/s
Strongest degradation

AMR-NB 12.2 kb/s
Strong degradation

G.711 A-law
Only a moderate increase

Opus
Almost like high quality



Order from figure 1 (Cllr error). Bar lengths: schematic.

30 s

After 30 s of speech, adding more brings only small gains.

< 0.3

Even in the worst case (5 s, AMR-NB 6.7), the Cllr stays below 0.3.

Compared with earlier work: same finding, low-bitrate codecs do more damage.

Article · Results (2): speaker by speaker

2
speakers misleading at short durations

Nos. 80 and 93 have a per-speaker Cllr above 1 at short durations with AMR-NB 6.7 kb/s, then stabilize.

2
speakers still unstable

After 60 s, no. 91 (all codecs) and no. 85 (AMR-NB 6.7 only) stay above 0.3.

32
speakers stable

Of the 34 speakers tested, the 32 other than nos. 85 and 91 (including nos. 80 and 93) stay below 0.2 after 60 s.

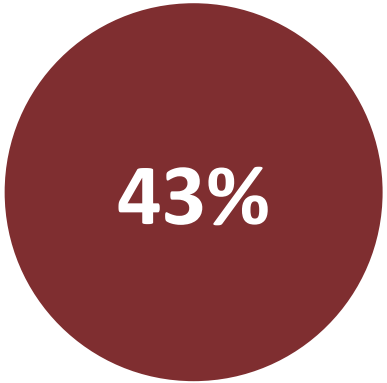
THE SYSTEM LOOKS GOOD… ON AVERAGE

Even in the worst case, the overall Cllr is below 0.3. Yet some speakers get misleading or unstable LRs: overall scores hide these cases. For some, longer speech would be needed for a stable result.

Article · Results (3): who suffers the most?

“EASY” SPEAKER

Good result with all codecs. Their score changes little from one codec to another.



of the sensitivity explained ($R^2 = 0.43$)

“DIFFICULT” SPEAKER

Already a weaker result, which degrades much more with a heavily compressed codec.

Possible explanation: Speakers who are already difficult for the system are also the most sensitive to codecs ($\beta = 0.23$; $p < 0.001$; 34 speakers). The article sees this as an indication, not proof: the models are trained on data with little variety, so robustness is not the same for everyone.

Article · Conclusion

KEY MESSAGES

1. Good overall results, especially beyond 30 s; AMR-NB (low bitrate) harms the most, Opus hardly at all.
2. Individual results hide problems: some speakers remain poor or unstable.
3. Speakers who are already difficult suffer the most from compression: a limitation of current models, trained on data with little diversity.

LIMITATIONS AND NEXT STEPS

1. Only young Cantonese men from Hong Kong, 34 test speakers and a small homogeneous training set: per-person scores vary.
2. One codec at a time; real cases sometimes chain several (landline, then mobile).
3. Next: identify fragile speakers and choose the right reference group for each case.

To remember: good on average does not mean good for everyone. Let’s test speaker by speaker.