ANALYTICAL REPORT

Voice match: source recordings vs. generated file

09/10/2026 · SpeechBrain ECAPA-TDNN model (spkrec-ecapa-voxceleb) · comparison of speaker embeddings, not speech content

Conclusion: very strong match — the recordings in the source folder and the voice used to generate generated/generated.wav most likely belong to the same person. The calibrated P(same speaker) at a 50% prior is above 99%. Cosine similarity of averaged embeddings: 0.76 (a typical different speaker in this model is < 0.20; EER threshold ≈ 0.25).

>99%	0.76	0.96	79%
P(same speaker), prior 50%	cosine generated vs source	baseline source/1 vs source/2	generated relative to source baseline

1. Research question

Do the recordings in the source folder come from the same voice whose recordings (not necessarily the same files) were used to generate the voice model that produced the file generated/generated.wav?

The test concerns speaker identity, not whether the model was trained specifically on the files source/1.wav and source/2.wav. Other recordings of the same person would give a similar result.

2. Materials

File	Duration	Format	4 s speech chunks
source/1.wav	7 min 37 s	48 kHz, 16-bit, mono	70
source/2.wav	6 min 07 s	48 kHz, 24-bit, mono	70
generated/generated.wav	1 min 17 s	48 kHz, 16-bit, mono	36

3. Method

4-second speech segments were extracted from each recording (energy-based detection), resampled to 16 kHz, and speaker vectors were computed using the ECAPA-TDNN model trained on VoxCeleb (SpeechBrain spkrec-ecapa-voxceleb). Similarity is the cosine between vectors: 1.0 means an identical speaker profile. Both averaged embeddings of whole files and pairs of short chunks were compared.

The probability was estimated using a likelihood ratio based on typical cosine distributions for this model: same speaker ≈ N(0.48, 0.13), different speaker ≈ N(0.08, 0.10), 50% prior. The tails of these distributions are an approximation, which is why the report states “above 99%” rather than a value read off the tail of a Gaussian.

4. Results

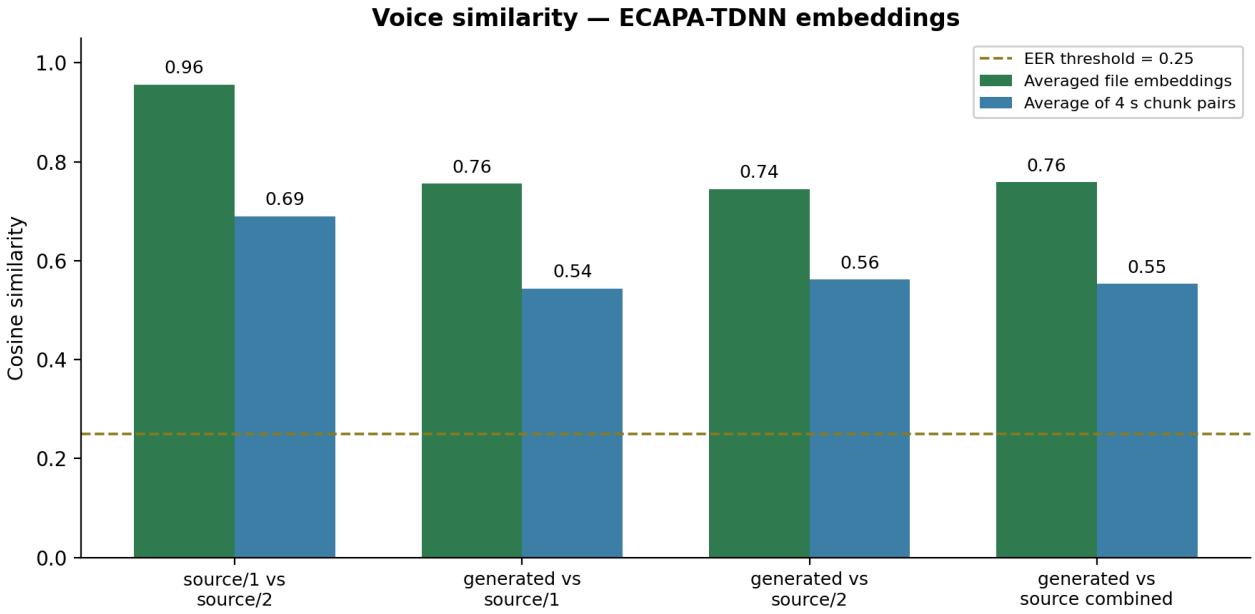

Fig. 1. Cosine similarity. Dashed line: model's typical EER threshold (~0.25).

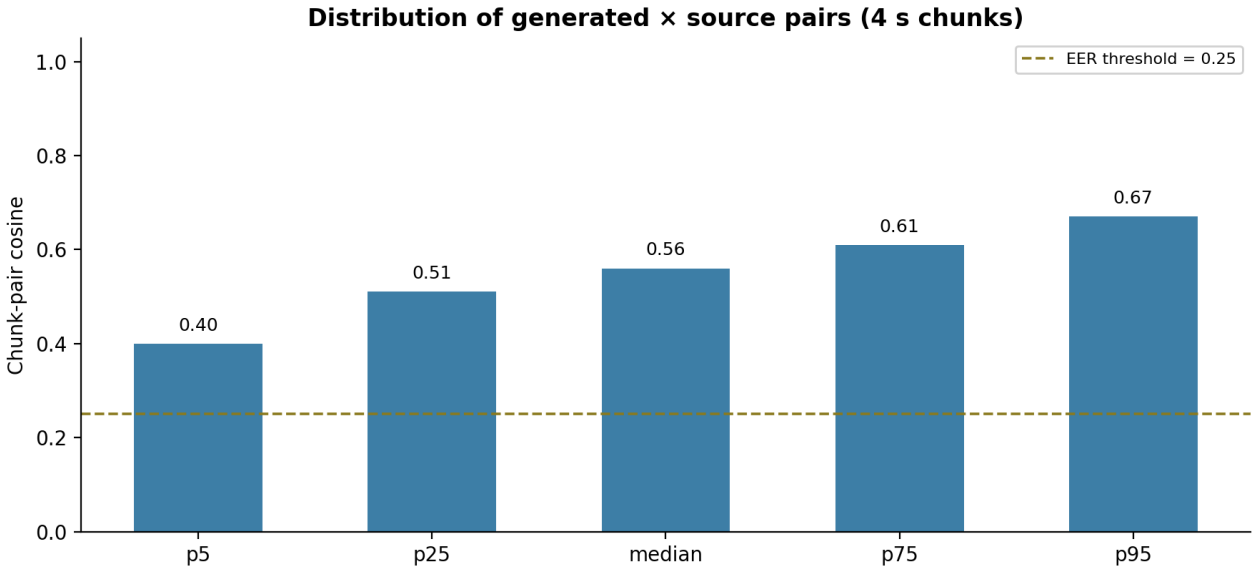

Fig. 2. Distribution of cosine values for all generated × source chunk pairs. 95% of pairs lie above 0.40.

Comparison table

Comparison	Cosine (averaged embeddings)	Cosine (chunk pairs)	Interpretation
source/1 vs source/2	0.956	0.690	Same speaker, natural recordings
generated vs source/1	0.756	0.544	Very strong match
generated vs source/2	0.745	0.562	Very strong match
generated vs source combined	0.759	0.553	Same voice as the model

Within-recording consistency (average chunk-pair cosine)

Recording	Mean	Std. dev.	p5	p95
source/1.wav	0.684	0.101	0.486	0.832
source/2.wav	0.753	0.068	0.636	0.857
generated.wav	0.745	0.060	0.648	0.844

5. Interpretation

source/1 and source/2 are almost certainly the same person (cosine of averaged embeddings 0.956). This serves as the “live recording vs. live recording” reference point.

generated vs source (0.759) is lower than this baseline — typical when comparing a synthesized or cloned voice with a natural recording (different channel, vocoder, content). It still lies far above the different-speaker zone. Relative to the source baseline, the generated result is about 79%.

The average of short chunk pairs (0.553) is always lower than the comparison of averaged vectors, because 4 seconds carry phonetic content. Source vs. source on chunks is 0.690, so generated stays within a similar order of magnitude.

6. Limitations

• The result does not confirm that the TTS/cloning model was trained specifically on the files in the source folder.

• There was no local cohort of other speakers (same gender, language, microphone). With similar voices the threshold can be higher, but a cosine of 0.76 is still outside the typical impostor range for ECAPA-TDNN.

• This report is a technical analysis of embeddings, not a forensic phonetic expertise for judicial purposes.

7. Numerical summary

Measure	Value
Model	speechbrain/spkrec-ecapa-voxceleb
Cosine generated vs source (averaged embeddings)	0.759
Cosine generated vs source (average of chunk pairs)	0.553
Baseline source/1 vs source/2	0.956
P(same speaker), 50% prior	> 99%
Interpretation band	very strong match