ANALYSEBERICHT

Stimmenübereinstimmung: Quellaufnahmen und generierte Datei

10.09.2026 · Modell SpeechBrain ECAPA-TDNN (spkrec-ecapa-voxceleb) · Vergleich von Sprecher-Embeddings, nicht des Gesprochenen

Ergebnis: sehr starke Übereinstimmung — die Aufnahmen im Ordner source und die für die Erzeugung von generated/generated.wav verwendete Stimme gehören höchstwahrscheinlich derselben Person an. Die kalibrierte P(gleicher Sprecher) liegt bei einem Prior von 50 % über 99 %. Cosine-Ähnlichkeit der gemittelten Embeddings: 0,76 (ein typischer anderer Sprecher liegt bei diesem Modell < 0,20; EER-Schwelle ≈ 0,25).

>99%	0,76	0,96	79%
P(gleicher Sprecher), Prior 50 %	Cosine generated vs. source	Baseline source/1 vs. source/2	generated bezogen auf source-Baseline

1. Forschungsfrage

Stammen die Aufnahmen im Ordner source von derselben Stimme, deren Aufnahmen (nicht notwendigerweise dieselben Dateien) zur Erzeugung des Stimmmodells verwendet wurden, aus dem die Datei generated/generated.wav entstanden ist?

Der Test betrifft die Sprecheridentität, nicht die Frage, ob das Modell genau auf den Dateien source/1.wav und source/2.wav trainiert wurde. Andere Aufnahmen derselben Person würden ein ähnliches Ergebnis liefern.

2. Material

Datei	Dauer	Format	4-s-Sprachsegmente
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. Methode

Aus jeder Aufnahme wurden 4-Sekunden-Sprachsegmente extrahiert (energiebasierte Erkennung), auf 16 kHz resampelt, und mit dem auf VoxCeleb trainierten ECAPA-TDNN-Modell (SpeechBrain spkrec-ecapa-voxceleb) wurden Sprechervektoren berechnet. Die Ähnlichkeit ist die Cosine-Ähnlichkeit zwischen den Vektoren: 1,0 bedeutet ein identisches Sprecherprofil. Verglichen wurden sowohl die gemittelten Embeddings ganzer Dateien als auch Paare kurzer Segmente.

Die Wahrscheinlichkeit wurde über ein Likelihood-Verhältnis auf Basis typischer Cosine-Verteilungen dieses Modells geschätzt: gleicher Sprecher ≈ N(0,48; 0,13), anderer Sprecher ≈ N(0,08; 0,10), Prior 50 %. Die Ränder dieser Verteilungen sind Näherungswerte, weshalb im Bericht „über 99 %“ angegeben wird und nicht ein Wert vom Rand einer Gaußverteilung.

4. Ergebnisse

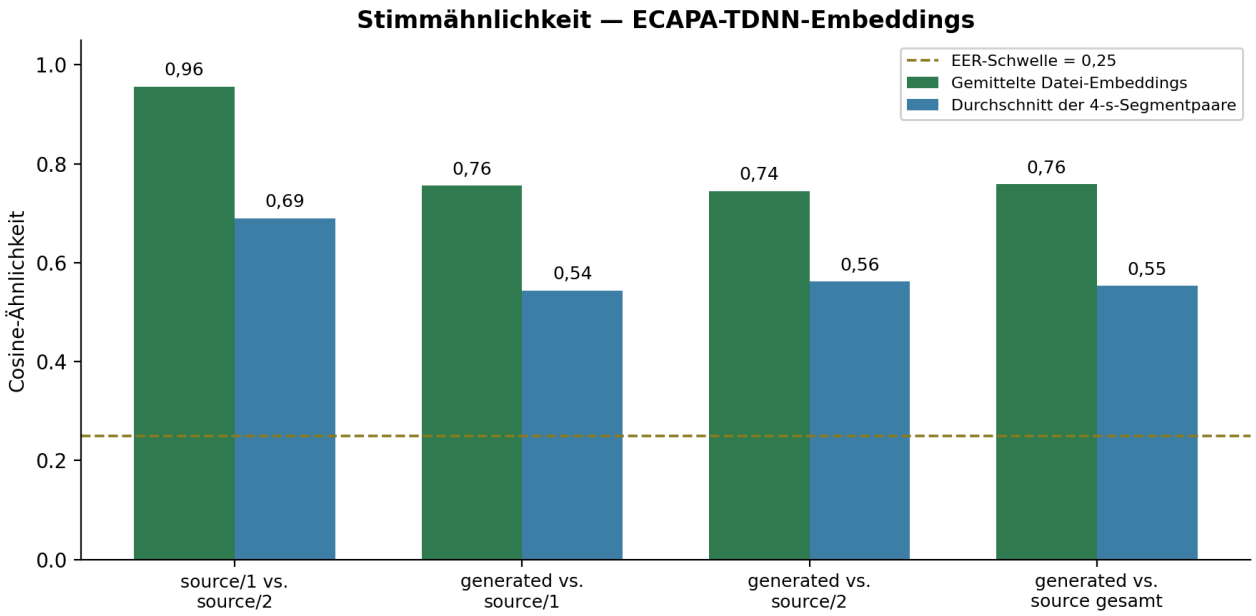

Abb. 1. Cosine-Ähnlichkeit. Gestrichelte Linie: typische EER-Schwelle des Modells (~0,25).

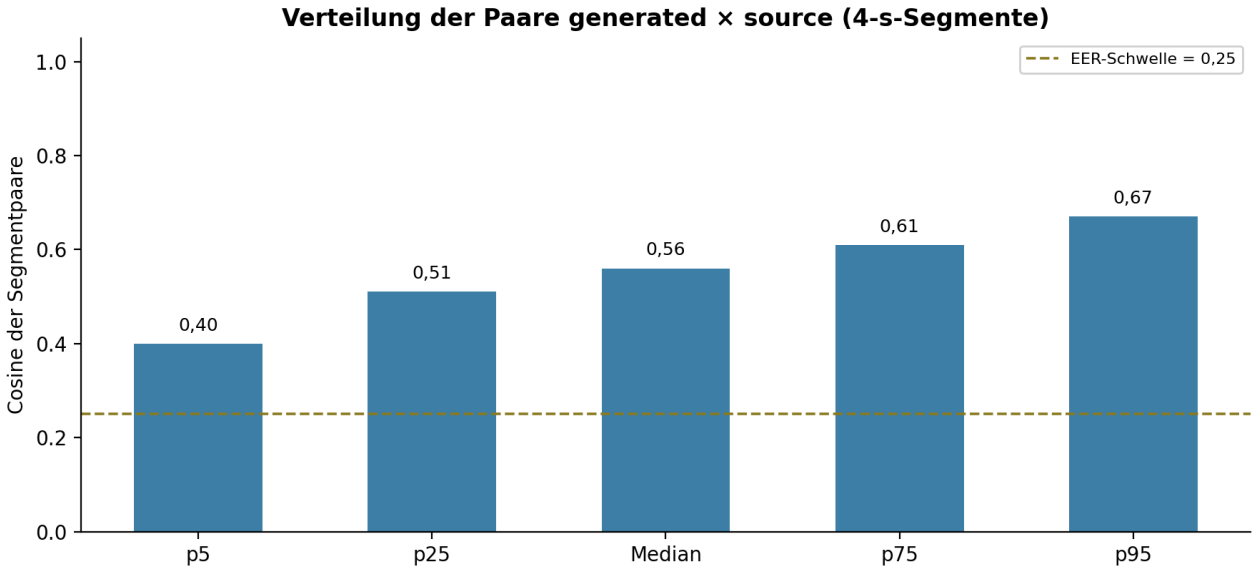

Abb. 2. Verteilung der Cosine-Werte aller generated × source Segmentpaare. 95 % der Paare liegen über 0,40.

Vergleichstabelle

Vergleich	Cosine (gemittelte Embeddings)	Cosine (Segmentpaare)	Interpretation
source/1 vs. source/2	0,956	0,690	Gleicher Sprecher, natürliche Aufnahmen
generated vs. source/1	0,756	0,544	Sehr starke Übereinstimmung
generated vs. source/2	0,745	0,562	Sehr starke Übereinstimmung
generated vs. source gesamt	0,759	0,553	Gleiche Stimme wie das Modell

Konsistenz innerhalb der Aufnahme (durchschnittliche Cosine-Ähnlichkeit der Segmentpaare)

Aufnahme	Mittelwert	Std.-Abw.	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 und source/2 sind fast sicher dieselbe Person (Cosine der gemittelten Embeddings 0,956). Dies dient als Referenzpunkt „Live-Aufnahme vs. Live-Aufnahme“.

generated vs. source (0,759) liegt unter dieser Baseline — typisch beim Vergleich einer synthetisierten oder geklonten Stimme mit einer natürlichen Aufnahme (anderer Kanal, Vocoder, Inhalt). Der Wert liegt weiterhin weit über dem Bereich unterschiedlicher Sprecher. Bezogen auf die source-Baseline erreicht generated etwa 79 %.

Der Durchschnitt kurzer Segmentpaare (0,553) ist stets niedriger als der Vergleich gemittelter Vektoren, da 4 Sekunden phonetischen Inhalt tragen. source vs. source liegt bei Segmenten bei 0,690, sodass generated in einer ähnlichen Größenordnung bleibt.

6. Einschränkungen

• Das Ergebnis bestätigt nicht, dass das TTS-/Klon-Modell exakt auf den Dateien im Ordner source trainiert wurde.

• Es gab keine lokale Kohorte anderer Sprecher (gleiches Geschlecht, gleiche Sprache, gleiches Mikrofon). Bei ähnlichen Stimmen kann die Schwelle höher liegen, doch ein Cosine-Wert von 0,76 liegt weiterhin außerhalb des typischen Impostor-Bereichs für ECAPA-TDNN.

• Der Bericht ist eine technische Analyse von Embeddings, kein phonoskopisches Gutachten für gerichtliche Zwecke.

7. Zahlenübersicht

Kennzahl	Wert
Modell	speechbrain/spkrec-ecapa-voxceleb
Cosine generated vs. source (gemittelte Embeddings)	0,759
Cosine generated vs. source (Durchschnitt der Segmentpaare)	0,553
Baseline source/1 vs. source/2	0,956
P(gleicher Sprecher), Prior 50 %	> 99%
Interpretationsbereich	sehr starke Übereinstimmung