RAPORT ANALITYCZNY

Zgodność głosu: nagrania źródłowe a plik wygenerowany

10.09.2026 · model SpeechBrain ECAPA-TDNN (spkrec-ecapa-voxceleb) · porównanie embeddingów mówcy, nie treści wypowiedzi

Wniosek: bardzo silne dopasowanie — nagrania z folderu source i głos użyty do wygenerowania generated/generated.wav należą najpewniej do tej samej osoby. Kalibrowane P(ten sam mówca) przy priorze 50% wynosi powyżej 99%. Cosine uśrednionych embeddingów: 0,76 (typowy inny mówca w tym modelu < 0,20; próg EER ≈ 0,25).

>99%	0,76	0,96	79%
P(ten sam mówca) prior 50%	cosine generated vs source	baseline source/1 vs source/2	generated względem baseline source

1. Pytanie badawcze

Czy nagrania w folderze source pochodzą od tego samego głosu, którego nagrania (niekoniecznie te same pliki) wykorzystano do wygenerowania modelu głosu, z którego powstał plik generated/generated.wav?

Test dotyczy tożsamości mówcy, nie tego, czy model uczono dokładnie plików source/1.wav i source/2.wav. Inne nagrania tej samej osoby dałyby podobny wynik.

2. Materiały

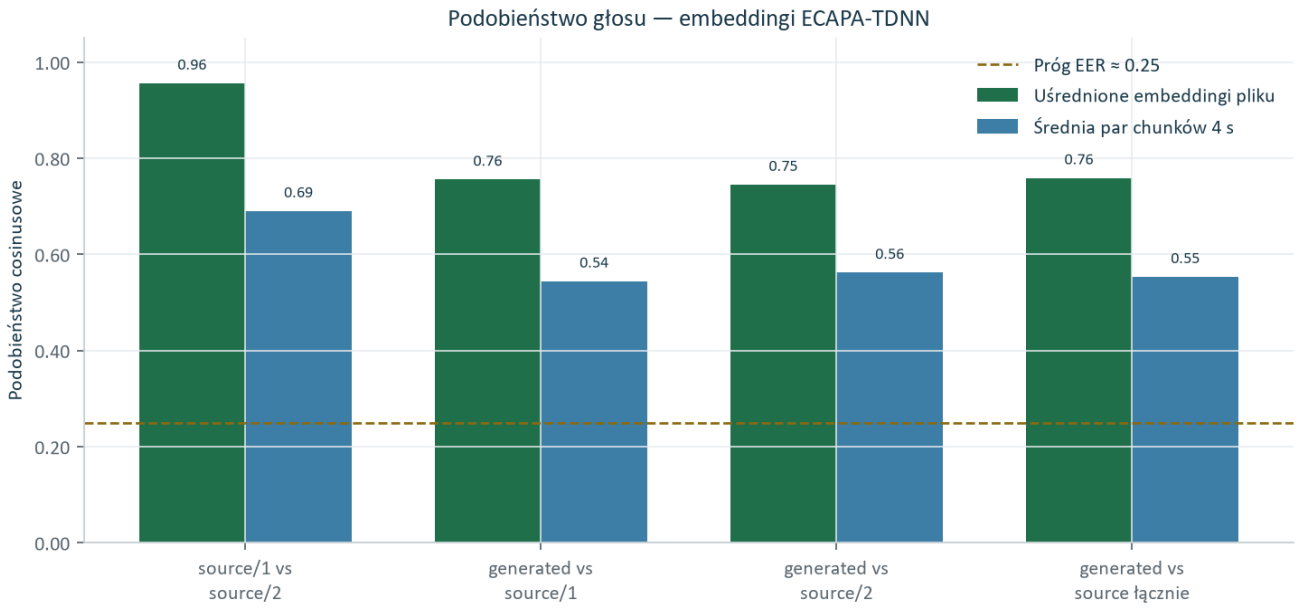
Plik	Czas	Format	Chunki mowy 4 s
source/1.wav	7 min 37 s	48 kHz, 16-bit, mono	70
source/2.wav	6 min 07 s	48 kHz, 24-bit, mono	70
generated/generated.wav	1 min 17 s	48 kHz, 16-bit, mono	36

3. Metoda

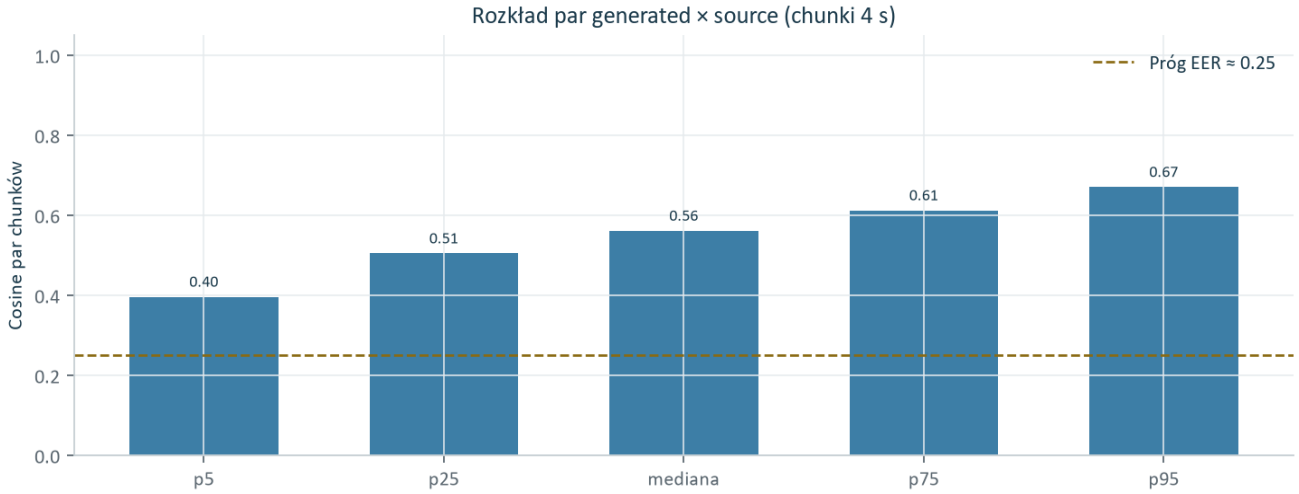
Z każdego nagrania wyodrębniono fragmenty mowy po 4 sekundy (detekcja energii), przesamplowano do 16 kHz i policzono wektory mówcy modelem ECAPA-TDNN wytrenowanym na VoxCeleb (SpeechBrain spkrec-ecapa-voxceleb). Podobieństwo to cosine między wektorami: 1,0 oznacza identyczny profil mówcy. Porównano zarówno uśrednione embeddingi całych plików, jak i pary krótkich chunków.

Prawdopodobieństwo oszacowano ilorazem wiarygodności na typowych rozkładach cosine tego modelu: ten sam mówca ≈ N(0,48; 0,13), inny mówca ≈ N(0,08; 0,10), prior 50%. Końce tych rozkładów są przybliżeniem, dlatego w raporcie podawane jest powyżej 99%, a nie wartość z ogona gaussjana.

4. Wyniki



Rys. 1. Podobieństwo cosinusowe. Linia przerywana: typowy próg EER modelu (~0,25).



Rys. 2. Rozkład cosine wszystkich par chunków generated × source. 95% par leży powyżej 0,40.

Tabela porównań

Porównanie	Cosine (średnie embeddingi)	Cosine (pary chunków)	Interpretacja
source/1 vs source/2	0,956	0,690	Ten sam mówca, nagrania naturalne
generated vs source/1	0,756	0,544	Bardzo silne dopasowanie
generated vs source/2	0,745	0,562	Bardzo silne dopasowanie
generated vs source łącznie	0,759	0,553	Ten sam głos co model

Spójność wewnątrz nagrania (średni cosine par chunków)

Nagranie	Średnia	Odch. std.	p5	p95
source/1.wav	0,684	0,101	0,486	0,832
source/2.wav	0,753	0,068	0,636	0,857

Nagranie	Średnia	Odch. std.	p5	p95
generated.wav	0,745	0,060	0,648	0,844

5. Interpretacja

source/1 i source/2 to prawie na pewno ta sama osoba (cosine uśrednionych embeddingów 0,956). To punkt odniesienia „żywe nagranie vs żywe nagranie”.

generated vs source (0,759) jest niższe niż ten baseline — typowe przy porównaniu syntezy lub klonu z nagraniem naturalnym (inny kanał, vocoder, treść). Nadal leży daleko powyżej strefy innego mówcy. Względem baseline source wynik generated to ok. 79%.

Średnia par krótkich chunków (0,553) jest zawsze niższa niż porównanie uśrednionych wektorów, bo 4 s niosą treść fonetyczną. source vs source na chunkach to 0,690, więc generated trzyma zbliżony rząd wielkości.

6. Ograniczenia

• Wynik nie potwierdza, że model TTS/klonu trenowano dokładnie na plikach z folderu source.

• Nie było lokalnej kohorty innych mówców (ta sama płeć, język, mikrofon). Przy podobnych głosach próg bywa wyższy, ale cosine 0,76 nadal jest poza typowym zakresem impostora dla ECAPA-TDNN.

• Raport jest analizą techniczną embeddingów, nie ekspertyzą fonoskopijną do celów sądowych.

7. Podsumowanie liczbowe

Miara	Wartość
Model	speechbrain/spkrec-ecapa-voxceleb
Cosine generated vs source (średnie embeddingi)	0,759
Cosine generated vs source (średnia par chunków)	0,553
Baseline source/1 vs source/2	0,956
P(ten sam mówca), prior 50%	> 99%
Pasmo interpretacji	bardzo silne dopasowanie