

RAPORT ANALITYCZNY

Zgodność głosu: files/source.wav a files/generated.wav

11.09.2026 · model SpeechBrain ECAPA-TDNN (spkrec-ecapa-voxceleb) · materiały: folder files · porównanie embeddingów mówcy, nie treści wypowiedzi

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

>99%	0,76	0,68	80%
P(ten sam mówca) prior 50%	cosine generated vs source	baseline spójności source (chunki)	pary chunków względem baseline source

1. Pytanie badawcze

Czy nagranie files/source.wav pochodzi od tego samego głosu, którego nagrania (niekoniecznie ten sam plik) wykorzystano do wygenerowania modelu głosu, z którego powstał plik files/generated.wav?

Test dotyczy tożsamości mówcy, nie tego, czy model uczono dokładnie tego pliku źródłowego. Inne nagrania tej samej osoby dałyby podobny wynik.

2. Materiały

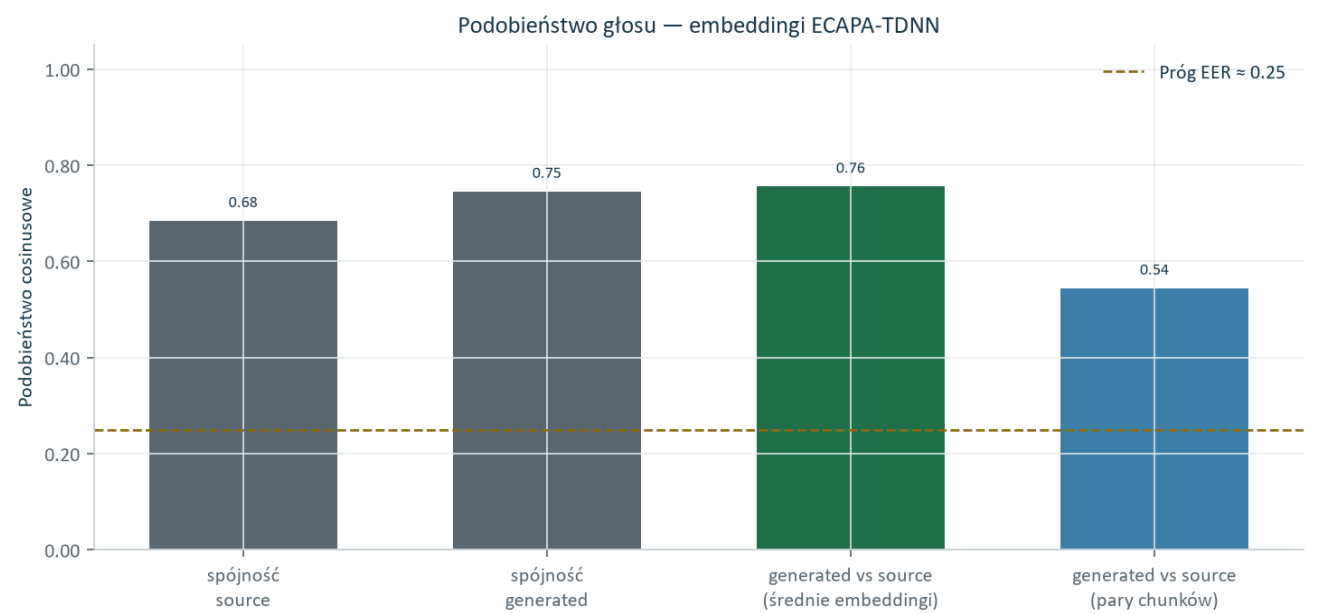
Plik	Czas	Format	Chunki mowy 4 s
files/source.wav	7 min 37 s	48 kHz, 16-bit, mono	70
files/generated.wav	1 min 17 s	48 kHz, 16-bit, mono	36

3. Metoda

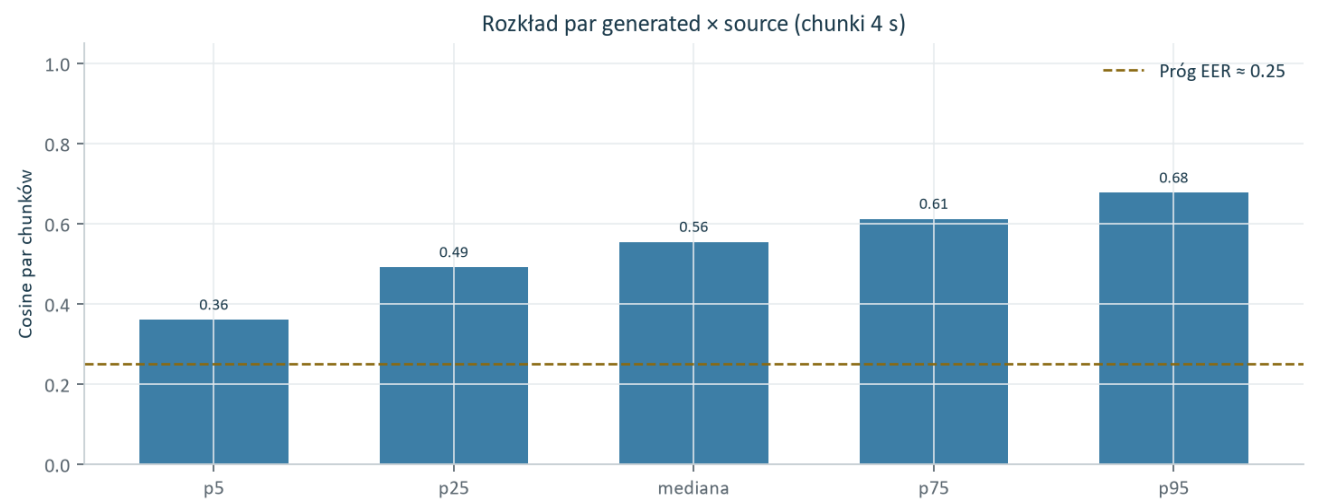
Z każdego nagrania wyodrębniono fragmenty mowy po 4 sekundy (detekcja energii), przesampłowano do 16 kHz i policzono wektory mówcy modelem ECAPA-TDNN wytrenowanym na VoxCeleb (SpeechBrain spkrec-ecapa-voxceleb). Podobieństwo to cosine między wektorami: 1,0 oznacza identyczny profil mówcy. Porównano uśrednione embeddingi całych plików oraz 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 >99%, a nie wartość z ogona gaussjana. Przy jednym pliku źródłowym baseline to spójność chunków wewnątrz source.wav (brak drugiego nagrania naturalnego tej samej osoby).

4. Wyniki



Rys. 1. Podobieństwo cosinusowe. Linia przerywana: typowy próg EER modelu (~0,25). Spójność source / generated to średni cosine par chunków wewnątrz nagrania.



Rys. 2. Rozkład cosine wszystkich par chunków generated × source. 5. percentyl = 0,36, 95. percentyl = 0,68.

Tabela porównań

Porównanie	Cosine	Interpretacja
generated vs source (średnie embeddingi pliku)	0,756	bardzo silne dopasowanie
generated vs source (średnia par chunków 4 s)	0,544	silne dopasowanie na krótkich fragmentach
spójność wewnątrz source.wav (pary chunków)	0,684	baseline tego samego nagrania naturalnego
spójność wewnątrz generated.wav (pary chunków)	0,745	wewnętrzna spójność syntezy

Spójność wewnątrz nagrania (średni cosine par chunków)

Nagranie	Średnia	Odch. std.	p5	p95
files/source.wav	0,684	0,101	0,486	0,832
files/generated.wav	0,745	0,060	0,648	0,844

5. Interpretacja

Uśrednione embeddingi generated vs source dają cosine 0,76. To wynik z zakresu tej samej osoby: typowy inny mówca w ECAPA-TDNN ma cosine poniżej ok. 0,20, a próg EER modelu wynosi ok. 0,25.

Średnia par krótkich chunków (0,54) jest niższa niż spójność wewnątrz source (0,68) — ok. 80% tego baseline. To oczekiwane: 4 s niosą treść fonetyczną, a synteza różni się kanałem i vocoderem od nagrania naturalnego. Nadal daleko powyżej strefy innego mówcy (5. percentyl par = 0,36).

6. Ograniczenia

- Wynik nie potwierdza, że model TTS/klonu trenowano dokładnie na files/source.wav.
- Jest tylko jedno nagranie źródłowe, więc baseline to spójność wewnątrz tego pliku, a nie porównanie dwóch niezależnych nagrań naturalnych tej samej osoby.
- 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
Folder wejściowy	files/
Model	speechbrain/spkrec-ecapa-voxceleb
Cosine generated vs source (średnie embeddingi)	0,756
Cosine generated vs source (średnia par chunków)	0,544
Baseline spójności source (pary chunków)	0,684
P(ten sam mówca), prior 50%	>99%
Pasmo interpretacji	bardzo silne dopasowanie