

АНАЛІТИЧНИЙ ЗВІТ

Відповідність голосу: вихідні записи та згенерований файл

10.09.2026 · модель SpeechBrain ECAPA-TDNN (spkrec-ecapa-voxceleb) · порівняння ембеддингів мовця, а не змісту мовлення

Висновок: дуже сильна відповідність — записи з папки source та голос, використаний для генерації generated/generated.wav, найімовірніше належать одній і тій самій особі. Калібрована P(той самий мовець) за пріору 50% перевищує 99%. Косинусна подібність усереднених ембеддингів: 0,76 (типовий інший мовець у цій моделі < 0,20; поріг EER $\approx$ 0,25).

>99% P(той самий мовець), пріор 50%	0,76 косинус generated vs source	0,96 базовий рівень source/1 vs source/2	79% generated відносно базового рівня source
---	---	---	---

1. Дослідницьке питання

Чи походять записи в папці source від того самого голосу, записи якого (не обов'язково ті самі файли) було використано для створення голосової моделі, з якої отримано файл generated/generated.wav?

Тест стосується ідентичності мовця, а не того, чи навчалася модель саме на файлах source/1.wav та source/2.wav. Інші записи тієї самої особи дали б подібний результат.

2. Матеріали

Файл	Тривалість	Формат	Фрагменти мовлення 4 с
source/1.wav	7 хв 37 с	48 кГц, 16 біт, моно	70
source/2.wav	6 хв 07 с	48 кГц, 24 біт, моно	70
generated/generated.wav	1 хв 17 с	48 кГц, 16 біт, моно	36

3. Метод

З кожного запису було виділено 4-секундні фрагменти мовлення (виявлення на основі енергії), пересемпльовано до 16 кГц, і за допомогою моделі ECAPA-TDNN, навченої на VoxCeleb (SpeechBrain spkrec-ecapa-voxceleb), обчислено вектори мовця. Подібність — це косинус між векторами: 1,0 означає ідентичний профіль мовця. Порівняно як усереднені ембеддинги цілих файлів, так і пари коротких фрагментів.

Ймовірність оцінено за допомогою відношення правдоподібності на типових розподілах косинуса цієї моделі: той самий мовець $\approx N(0,48; 0,13)$, інший мовець $\approx N(0,08; 0,10)$, пріор 50%. Хвости цих розподілів є наближенням, тому в звіті наведено «понад 99%», а не значення з хвоста гаусіани.

4. Результати

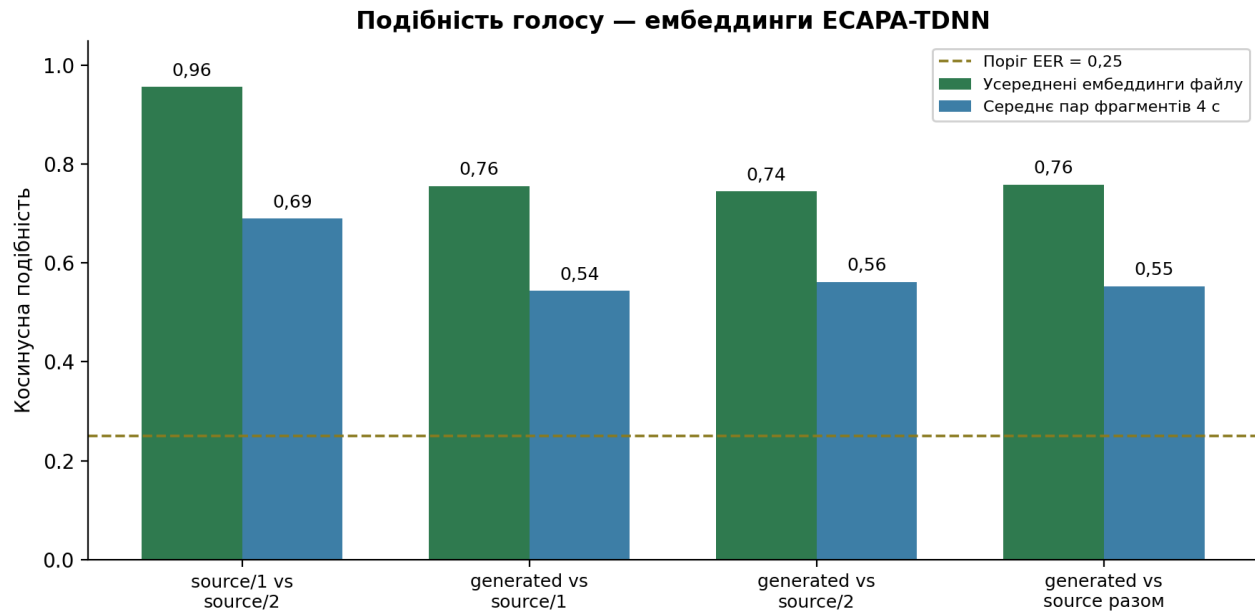


Рис. 1. Косинусна подібність. Пунктирна лінія: типовий поріг EER моделі (~0,25).

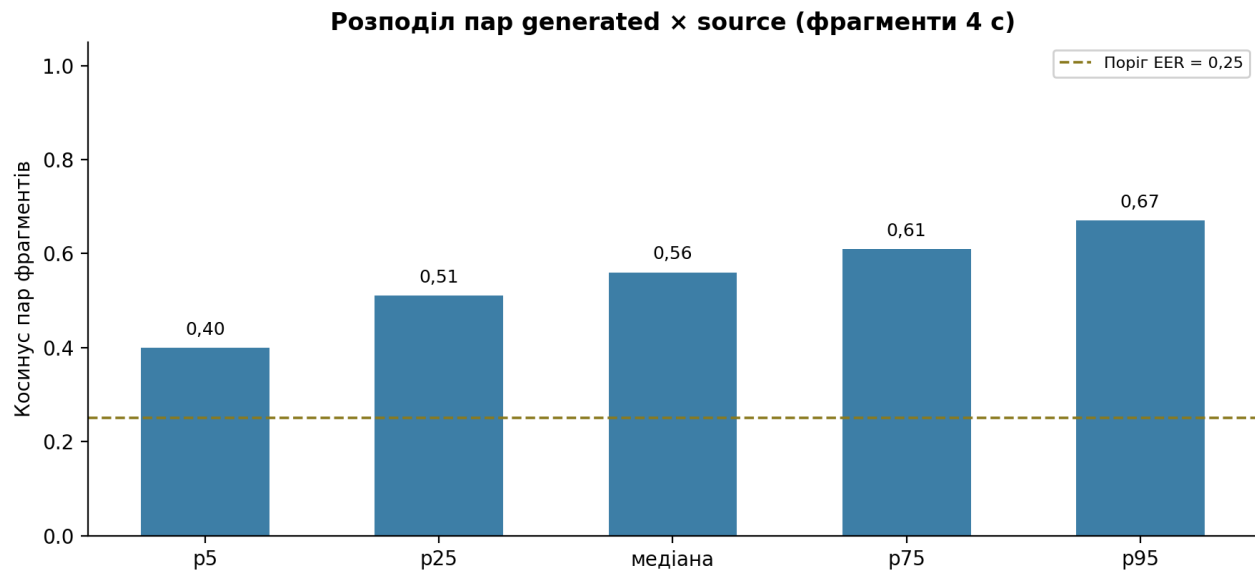


Рис. 2. Розподіл косинуса всіх пар фрагментів generated x source. 95% пар лежать вище 0,40.

Таблиця порівнянь

Порівняння	Косинус (усереднені ембеддинги)	Косинус (пари фрагментів)	Інтерпретація
source/1 vs source/2	0,956	0,690	Той самий мовець, природні записи
generated vs source/1	0,756	0,544	Дуже сильна відповідність
generated vs source/2	0,745	0,562	Дуже сильна відповідність
generated vs source разом	0,759	0,553	Той самий голос, що й модель

Згуртованість всередині запису (середній косинус пар фрагментів)

Запис	Середнє	Станд. відхил.	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. Інтерпретація

source/1 і source/2 майже напевно є однією і тією самою особою (косинус усереднених ембеддингів 0,956). Це слугує точкою відліку «живий запис проти живого запису».

generated vs source (0,759) нижче за цей базовий рівень — типово при порівнянні синтезованого або клонованого голосу з природним записом (інший канал, вокодер, зміст). Значення все ще лежить далеко вище зони іншого мовця. Відносно базового рівня source результат generated становить приблизно 79%.

Середнє коротких пар фрагментів (0,553) завжди нижче за порівняння усереднених векторів, оскільки 4 секунди несуть фонетичний зміст. source vs source на фрагментах становить 0,690, тож generated залишається в подібному порядку величини.

6. Обмеження

- Результат не підтверджує, що модель TTS/клонування навчалася саме на файлах з папки source.
- Не було локальної когорти інших мовців (та сама стаття, мова, мікрофон). За схожих голосів поріг може бути вищим, але косинус 0,76 усе одно перебуває поза типовим діапазоном самозванця для ECAPA-TDNN.
- Цей звіт є технічним аналізом ембеддингів, а не фоноскопичною експертизою для судових цілей.

7. Числове підсумовування

Показник	Значення
Модель	speechbrain/spkrec-ecapa-voxceleb
Косинус generated vs source (усереднені ембеддинги)	0,759
Косинус generated vs source (середнє пар фрагментів)	0,553
Базовий рівень source/1 vs source/2	0,956
P(той самий мовець), пріор 50%	> 99%
Смуга інтерпретації	дуже сильна відповідність