Wielojęzykowe podejście do audio DeepFake'ów

Autor: Bartłomiej Marek
Opiekun: dr inż. Piotr Syga

Motywacje

Rosnące zagrożenia społeczno-ekonomiczne:

- Manipulacja opinią publiczną poprzez sfałszowane nagrania polityków (np. manipulacja wypowiedzi Spiker Izby Reprezentantów Stanów Zjednoczonych [1], czy wygenerowane przemówienia polskich polityków [2];
- Straty finansowe wynikające z oszustw głosowych (np. oszustwo na kwotę 25,6 mln USD w Hongkongu z wykorzystaniem podszywania się pod kierownictwo [3];

Wyzwania technologiczne:

- Postęp w dziedzinie generatywnej AI umożliwia tworzenie realistycznych nagrań audio w różnych językach. Według najlepszej wiedzy autora, w obecnym stanie wiedzy istnieje luka badawcza nad zbadaniem skuteczności metod detekcji audio DeepFake dla różnych języków. Ze względu na dominację języka angielskiego, najnowocześniejsze modele wykrywające są trenowane i oceniane przy wykorzystaniu zbiorów referencyjnych, które są anglojęzyczne.

Cele pracy

- Dokonać oceny obecnie wykorzystywanych modeli detekcji audio DeepFake w kontekście wielojęzykowym, ze szczególnym uwzględnieniem faktu, że większość z nich została wytrenowana i zwalidowana na zbiorach referencyjnych w języku angielskim.
- Zbadać wpływ dostrajania (ang. fine-tuning) modeli na skuteczność detekcji audio DeepFake w konkretnym języku oraz w obrębie szerszej rodziny językowej.

W trakcie ich realizacji w obliczu niejednoznacznych wyników i podejrzenia uczenia się cech modeli użytych do generowania, zamiast cech języka, dodano:

- Zbadanie wpływu generatora na detekcję audio DeepFake dla różnych modeli (zastosowanie zbiorów niezależnych pod kątem architektur generatorów, a nie tylko próbek).

Metodologia badań

Badania zostały wykonane dla 8 języków z 3 różnych rodzin językowych:

- **języki germańskie:** niemiecki, angielski;
- **języki romańskie:** francuski, hiszpański, włoski;
- **języki słowiańskie**: polski, ukraiński, rosyjski;

Próbki pochodziły z zbioru MLAAD (The Multi-Language Audio Anti-Spoofing Dataset)[4].

Wykorzystano modele wytrenowane w języku angielskim (na zbiorach ASVspoof[5]) oraz implementację architektur sieci przedstawionych w ramach repozytorium do publikacji [6] oraz [7].

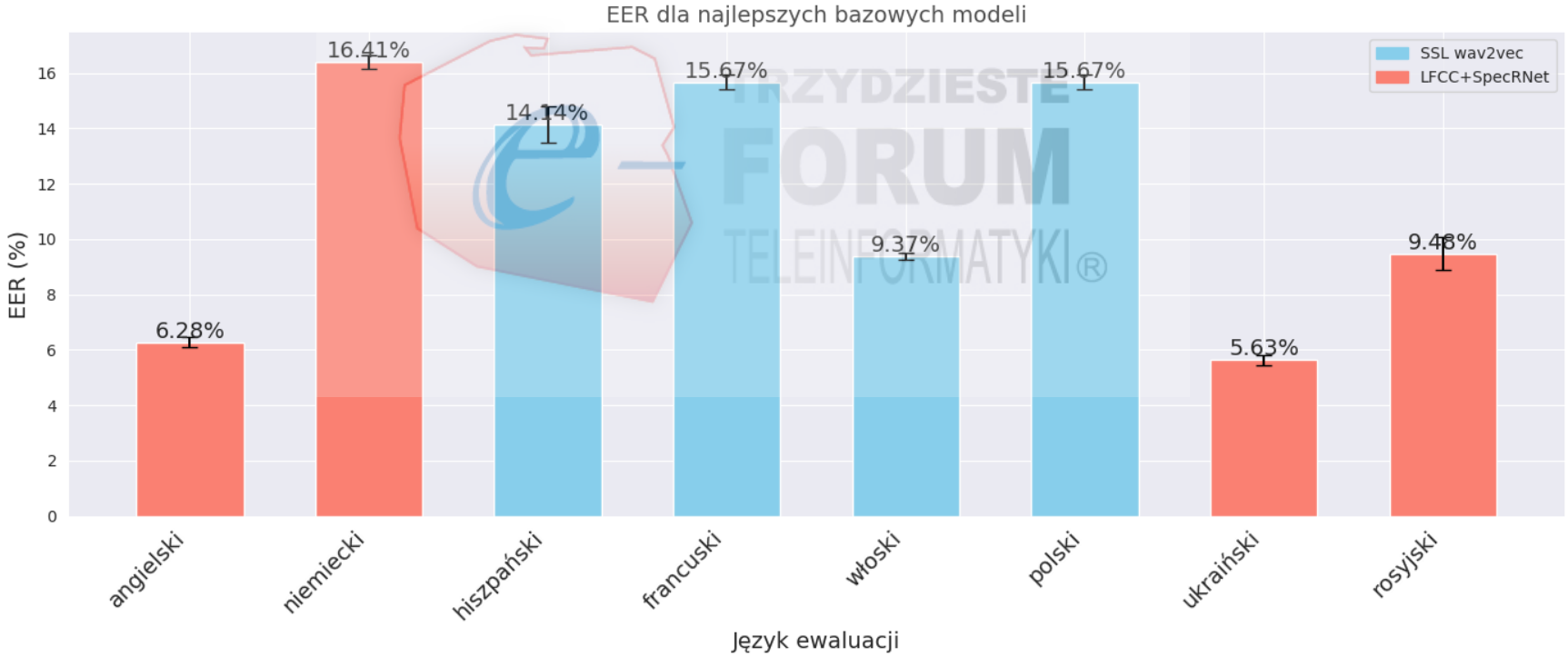
Podczas ewaluacji, zbiór testowy był podzielony na 5 częsci, z których każda kolejno nie była uwzględniona do testowania. Do oceny modelu wykorzystano metrykę Equal Error Rate (ERR) w skali procentowej.

Faza I - wyznaczenie baseline

- **Modele**: SSL-wav2vec (state-of-the-art), FrontEnd+SpecRNet, FrontEnd+MesoNet

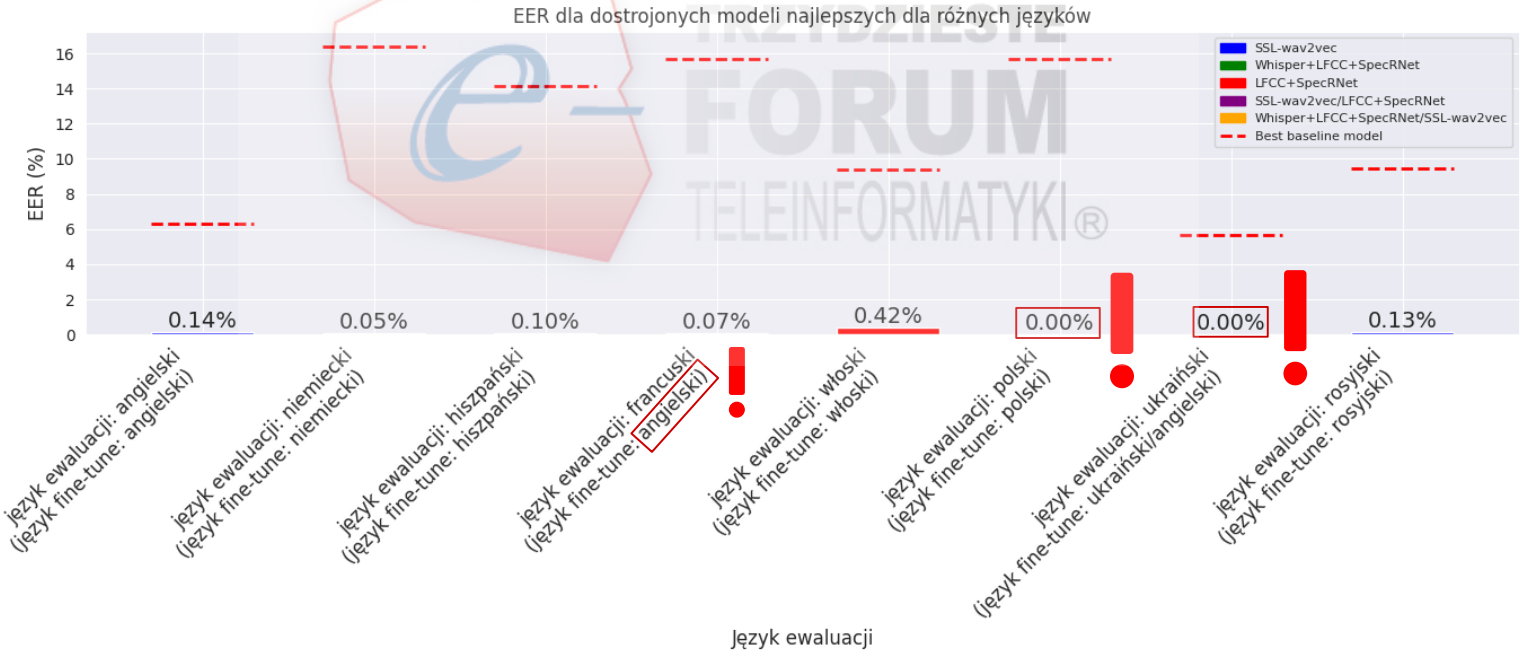
  FrontEnd = {Whisper; LFCC; Whisper+LFCC; MFCC; Whisper+ MFCC}

- **Zbiór danych do ewaluacji:** 25% losowo wybranych próbek dla danego języka;



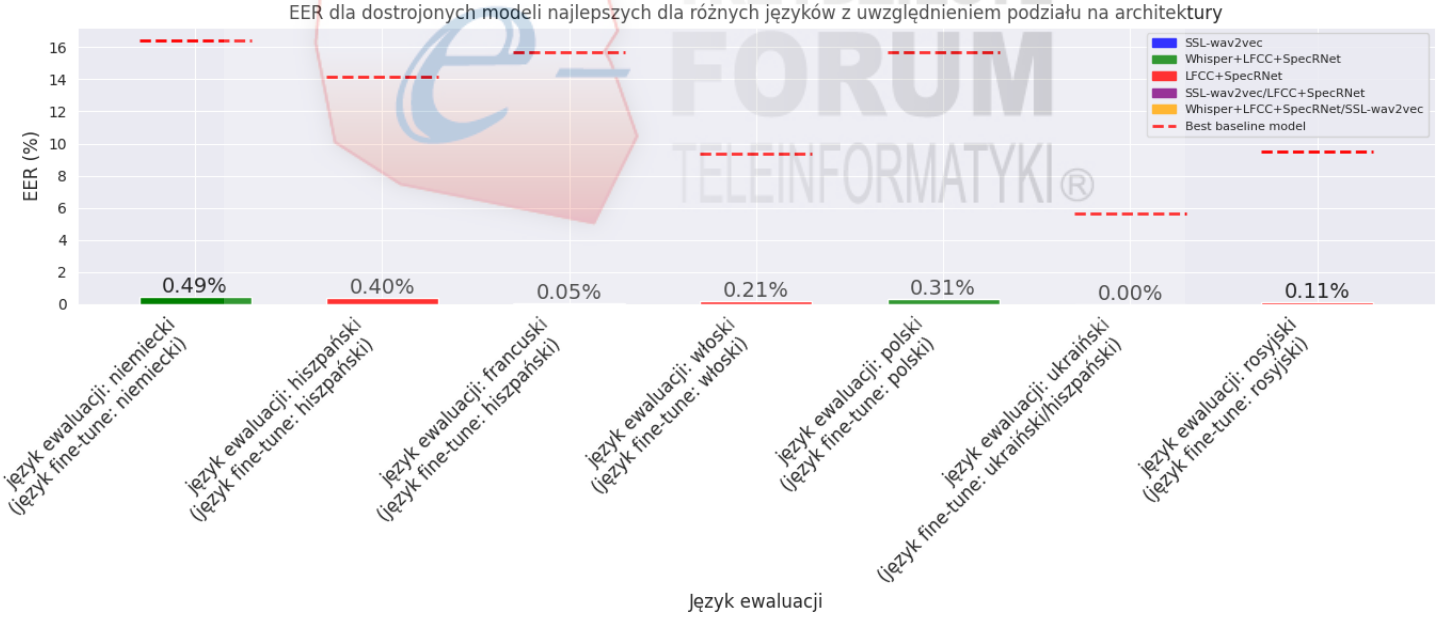
Faza II - najlepsze modele dla różnych języków

- **Modele:** SSL-wav2vec, LFCC+SpecRNet, MFCC+MesoNet, Whisper+LFCC+MesoNet
- **Parametry dostrajania:** 25 epok, learning-rate: 10e-6, zastosowanie early stopping (patience=5)
- **Zbiór danych do ewaluacji:** 25% losowo wybranych próbek dla danego języka (ten sam zbiór co w fazie 1)
- **Zbiór danych do dostrajania:** pozostałe 75% próbek (z których 10% to zbiór walidacyjny)

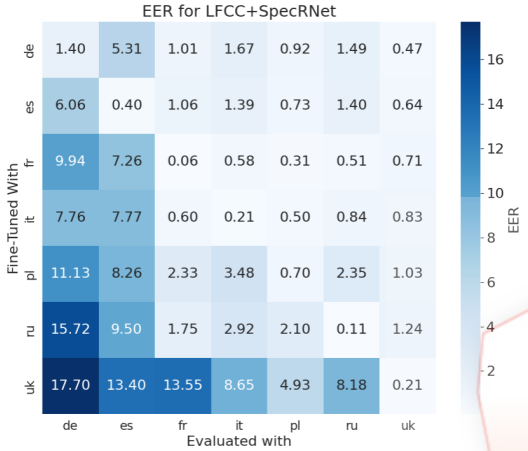


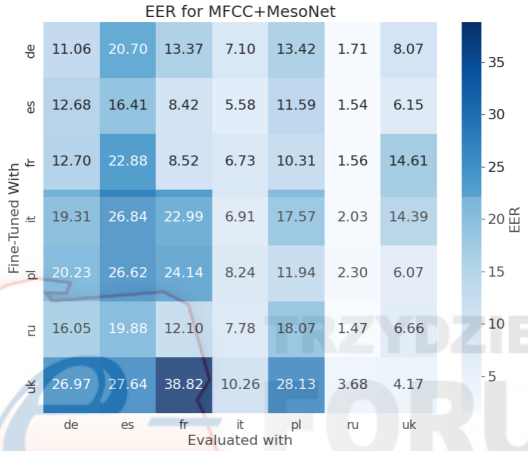
Faza III - najlepsze modele

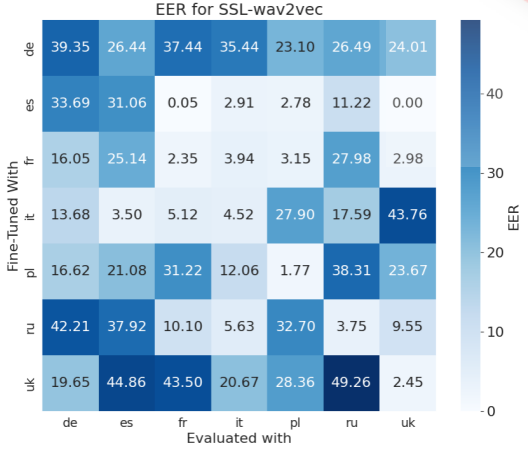
- **Modele:** SSL-wav2vec, LFCC+SpecRNet, MFCC+MesoNet, Whisper+LFCC+SpecRNet
- **Parametry dostrajania:** 25 epok, learning-rate: 10e-6, zastosowanie early stopping (patience=5)
- **Zbiór danych do ewaluacji:** próbki wygenerowane przez architektury: griffin_lim i xtts_v2
- **Zbiór danych do dostrajania:** próbki wygenerowane przez architektury: vits i xtts_v1.1

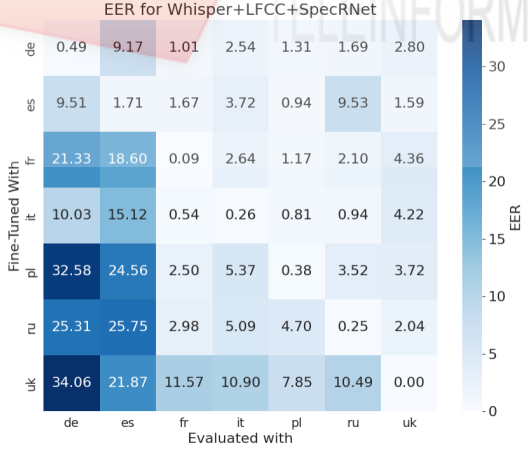


Faza III - wpływ innych języków









Dla modeli bazujących na SpecRNet nie zaobserwowano znaczących różnic między fazą II a fazą III. Dla obu z nich dla większości języków, najefektywniej jest użyć język docelowy (najlepsze wyniki na diagonalnej).

Znaczna degradacja wyników detekcji dla modelu SSL-wav2vec, co wskazuje na uczenie się charakterystyki generatora.

Wnioski końcowe

- Efektywność najnowocześniejszych modeli wytrenowanych przy użyciu języka angielskiego zależy od języka;
- dostrajanie znacząco poprawia wyniki;
- Modele oparte na SpecRNecie wykazują małą zależność od generatora;
- SSL-wav2vec traci skuteczność przy separacji zbiorów wg architektury generatora;
- Brak korelacji między językami w obrębie rodzin językowych;
- Wysoka skuteczność dla ukraińskiego wynika z niskiej jakości próbek [4]

Kierunki dalszych badań:

- Testowanie innych architektur sieci i front-end'ów;
- Badanie wpływu augmentacji danych;
- Analiza przyczyn degradacji SSL-wav2vec w fazie III;
- Balansowanie zbioru danych;
- Współpraca z lingwistami;

Bibliografia

[1] Reuters, "'Drunk' Nancy Pelosi video is manipulated," Reuters, Aug. 03, 2020. [Online]. Dostępne: https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated-idUSKCN24Z2BI

[2] Notes from Poland, "Opposition criticised for using AI-generated deepfake voice of PM in Polish election ad," Notes from Poland, Aug. 25, 2023. [Online]. Dostępne: https://notesfrompoland.com/2023/08/25/opposition-criticised-for-using-ai-generated-deepfake-voice-of-pm-in-polish-election-ad/

[3] CNN World, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'," CNN World, Feb. 04, 2024. [Online]. Dostępne: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[4] N. M. Müller et al., "MLAAD: The Multi-Language Audio Anti-Spoofing Dataset," arXiv preprint arXiv:2401.09512, 2024.

[5] A. Nautsch et al., "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 2, pp. 252-265, 2021, doi: 10.1109/TBIOM.2021.3059479.

[6] TAK, Hemlata, et al. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. (2022).

[7] Kawa, Piotr, et al. "Improved DeepFake Detection Using Whisper Features." (2023).