AKADEMIA GÓRNICZO-HUTNICZA
im. Stanisława Staszica w Krakowie

Technologia Mowy

Anonimizacja mowy

Sposoby anonimizowanej resyntezy mowy wraz z porównaniem ich efektywności

Jakub Czernecki
Wojciech Sabała
Bartosz Wąsik

1. Cel projektu

Celem projektu było opracowanie systemu anonimizacji mowy, którego produktem jest resyntezowany sygnał wejściowy o zmienionej charakterystyce. Takowe systemy pozwalają na zachowanie informacji na temat treści wypowiedzi, jednocześnie uniemożliwiając zidentyfikowanie oryginalnego mówcy.

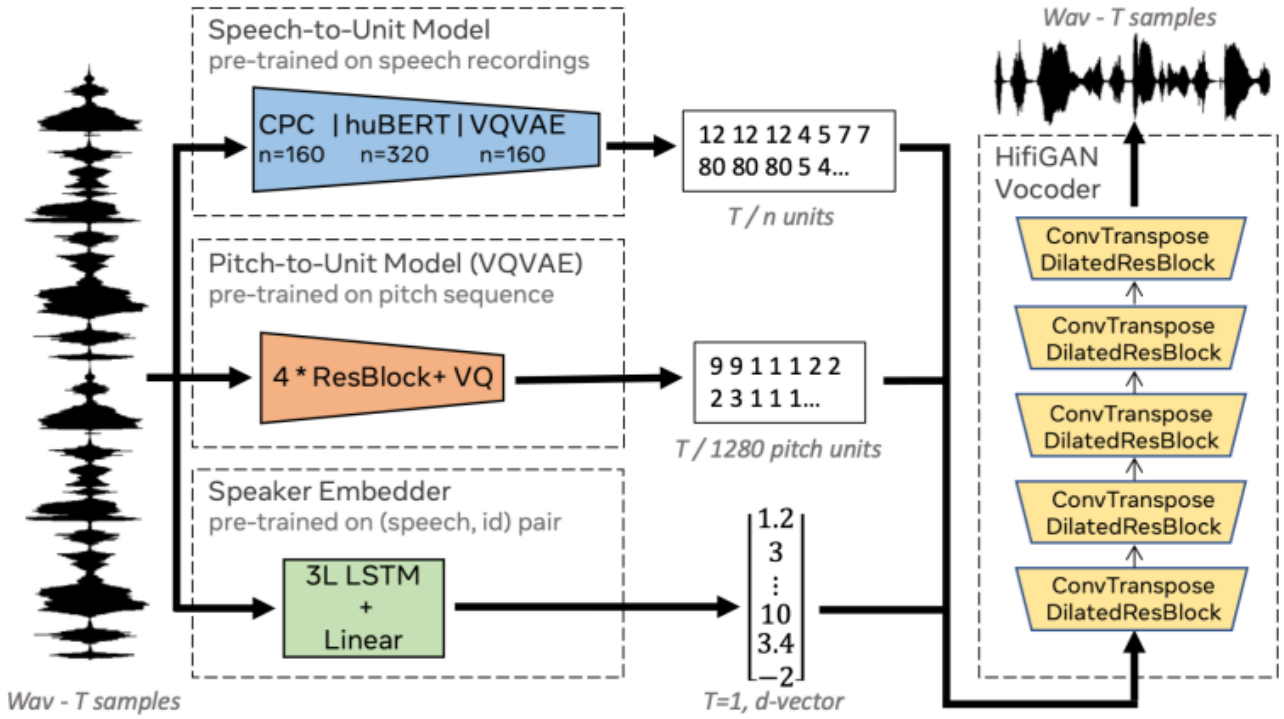
2. Założenia projektowe

W celu utworzenia pożądanego systemu, w ograniczonym czasie przewidzianym w planie zajęć, istotnym elementem okazało się dobranie możliwego do osiągnięcia celu oraz jasnego zarysu drogi doń. Jako kluczową funkcjonalność przyjęliśmy stworzenie systemu resyntezującego mowę, który w przypadku problemów uniemożliwiających wdrożenie, początkowo, teoretyzowanego sposobu anonimizacji, przeprowadzałby embedding oraz decoding informacji. W ramach uproszczeń związanych z wysoką złożonością owych systemów zdecydowano się na używanie przetrenowanych modeli syntezy mowy, czy ekstrakcji poszczególnych jej cech (jak x-vectory czy zawartość fonetyczną – o czym mowa poniżej). Z tego też powodu największym ograniczeniem okazała się możliwość znalezienia systemów, które były możliwe do wykorzystania w realizacji pierwotnych planów. Trudność ta okazała się kluczowa w sposobie rozwoju projektu. Spowodowała rozdzielenie pracy na jednym z etapów w celu szukania alternatyw z powodu braku możliwości ukończenia przyjętych pierwotnie założeń. Końcowo, pomimo wątpliwości co do realnej możliwości zrealizowania pierwszego z przyjętych systemów udało się zrealizować zarówno jego, jak i drugie zupełnie odrębne podejście. Podejście, które początkowo miało być alternatywą, jednakże zostało zakończone otrzymaniem innego w sposobie resyntezy, lecz równowartościowego produktu.

3. Anonimizacja poprzez ekstrakcje wektorów informacji

3.1 Architektura systemu

Zaplanowany w początkowej fazie projektu system anonimizacji miał polegać na resyntezie mowy z wyekstrahowanych wektorów reprezentujących trzy informacje zawarte w sygnale mowy: zdarzenia fonetyczne, ton podstawowy oraz x-vector (wektor charakteryzujący mówcę). Następnie x-vector miał zostać zmodyfikowany według przyjętego sposobu jego anonimizacji. Aby uzyskać sygnał wyjściowy, należałoby więc wykorzystać vocoder pozwalający na odtworzenie informacji zawartych w wektorach.



Rys. 1 Architektura systemu z HiFi-GAN, bez bloku anonimizacji [1].

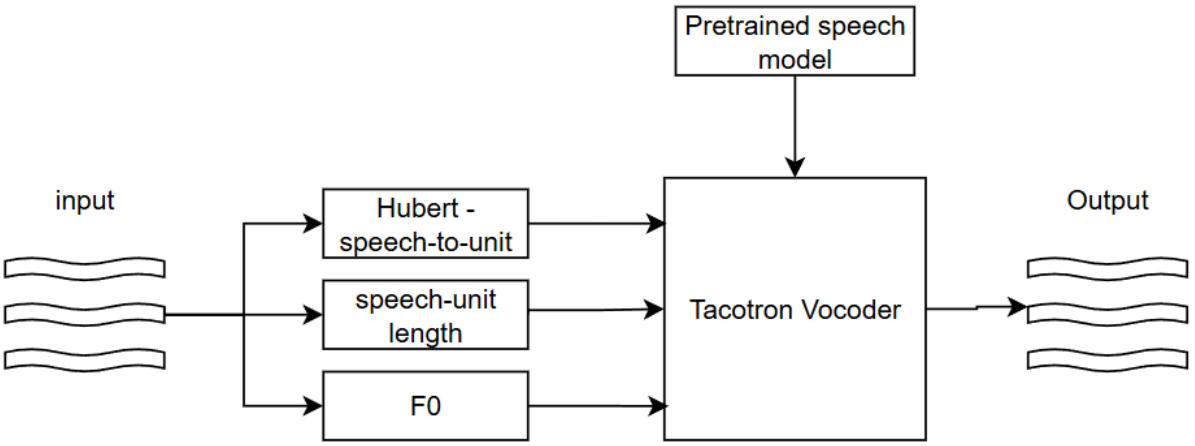
Podejście to, choć łatwe do zrozumienia, stwarza problem w modyfikacji wybranego vocodera, gdyż pretrenowany model obróbki danych z wektorów nie był dostępny. Postanowiono więc porzucić konkretny model resyntezy w poszukiwaniu innego, ale wciąż opartego na ekstrahowaniu wektorów informacji.

Znaleziono odpowiednio zmodyfikowany model vocodera – Tacotron 2, który również bazując na publikacji [1] był w stanie generować sygnał mowy na podstawie wektorów informacji. Jednakże w tym przypadku wykluczona została resynteza z użyciem x-vectorów, gdyż vocoder ten bazuje na przetrenowanych modelach głosowych takich jak użyty finalnie Hubert. Co to oznacza? Utracona zostanie możliwość modyfikacji głosu mówcy poprzez modyfikacje wektora osadzenia. Wytrenowanie modelu podobnego do Huberta używanego przez vocoder graniczy z niemożliwością, biorąc uwagę czas trwania projektu oraz dostępność danych do trenowania. Nie zastosujemy również żadnego sposobu losowego doboru głosu resyntezy, gdyż łączyłoby się to z tworzeniem nowego obiektu klasy vocodera w każdej instancji, co wiąże się z ładowaniem dużych modeli językowych.

Z drugiej strony, jeśli przyjmiemy za cel projektu jedynie wysoki stopień anonimizacji, to tego typu model zapewnia kompletną empirycznie zmianę charakterystyki mówcy na charakterystykę neutralną zaciągniętą z dużego modelu.

Znaleziony vocoder podobnie jak w poprzednim podejściu odtwarza sygnał wejściowy na podstawie wektora zjawisk fonetycznych oraz wektora częstotliwości podstawowych, pomijając jednak x-vector. Dodatkowo przyjmowany jest wektor długości zdarzeń fonetycznych, który redukuje powtarzające się wartości pierwotnego wektora. Oczywiście zagłębiając się w implementacje vocodera, jesteśmy również w stanie zmienić używany do resyntezy model językowy.

Najważniejszą zaletą tego podejścia jest uzyskanie anonimizacji, w której nie zatraca się informacja na temat prozodii mowy. W ostatecznej wersji pojedyncze okno analizy obejmuje 20 ms sygnału wejściowego, co pokrywa się z literaturową wartością długości stacjonarności sygnału mowy. Dzięki oknu tej długości jesteśmy w stanie również ograniczyć nienaturalne brzmienie resyntezowanego sygnału spowodowane błędami kategoryzacji w wektorze zdarzeń fonetycznych.



Rys. 2 Ostateczna architektura podejścia resyntezy wektorami informacji.

3.2 Sukcesy podejścia

Stosując przedstawione podejście, udało się uzyskać działający system, który pozwala na całkowitą anonimizacje sygnału wejściowego, jednocześnie zachowując przejrzystość i zrozumiałość mowy, szczególnie w przypadku języka angielskiego. Jako wynik otrzymany zostaje plik .wav o identycznym co do długości pojedynczego okna analizy czasie trwania, który jednocześnie zachowuje długość wymowy poszczególnych wyrazów oraz kierunek tonalny. W ostatecznej ewaluacji warto również zwrócić uwagę na dodatkowy pozytyw użycia pretrenowanego modelu językowego – zrozumiałość i względna naturalność resyntezowanego sygnału. Okazuje się bowiem, że pomimo „komputerowego" brzmienia, które ujawnia się przy niektórych z głosek, sygnał wyjściowy jest pozbawiony szumów, okresów niezrozumiałości czy nietypowych sposobów wypowiadania zlepków fonemów.

3.3 Jak należałoby dalej rozwijać to podejście?

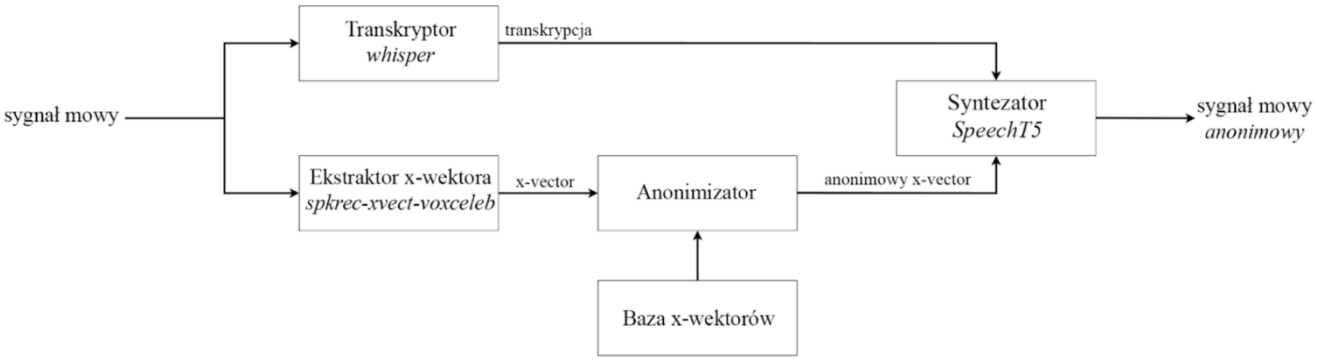
Głównym sposobem na ulepszenie pracy systemu byłaby dalsza modyfikacja użytego vocodera, tak aby móc łatwiej i swobodniej kontrolować głos resyntezy. Użycie x-vectorów, które tworzone byłyby z gotowych baz danych (co pokazano w dalszej części niniejszego raportu), niewątpliwie rozwiązałoby ten problem, jednakże implementacja manipulacji embedingiem w vocoderze bazującym na dużych pretrenowanych modelach wydaje się karkołomne.

W tym przypadku może wypadałoby się zastanowić nad zmianą użytego vocodera na wymieniany w pierwotnym planie HiFi-GAN i zdecydować się na wytrenowanie modelu, który jest w stanie dekodować informacje zawarte w wektorach na gotowe dane na wejście syntezatora mowy.

4. Anonimizacja poprzez manipulację wektorem mówcy.

4.1 Architektura systemu

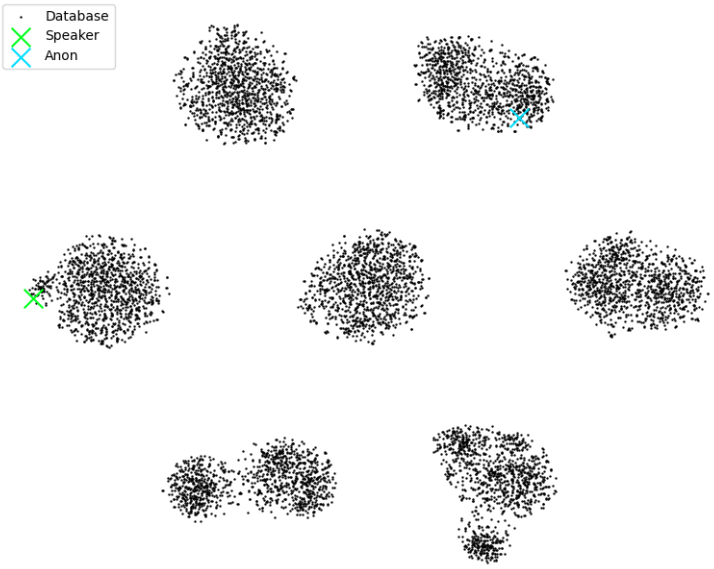
Chcąc kontynuować rozwój projektu na podstawie manipulacji wektorami osadzenia, podjęto dalsze poszukiwania kombinacji przetrenowanych modeli przyjmujących na wejściu x-vector jako jedną ze zmiennych. Wiedząc, że największy problem stanowi synteza, postanowiono porzucić pozostałe z pierwotnie zakładanych zmiennych. Rezultatem poszukiwań był model SpeechT5 stworzony przez firmę Microsoft. Jest to model text to speech, a więc oprócz x-vectora przyjmował tekst, w odróżnieniu od systemu opisanego w punkcie trzecim, co wiązało się z utratą prozodii mowy. Mając kluczowy element systemu adaptowano pod nią resztę architektury układu.



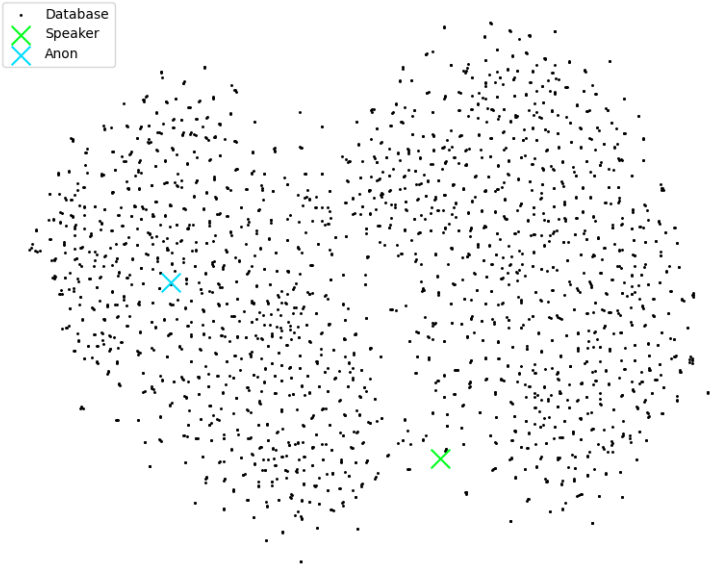
Rys. 3 Architektura dla anonimizatora z wykorzystaniem x-vectorów.

W tym celu system oparto o mechanizm transkrypcji z wykorzystaniem modelu Whisper odpowiedzialnego za dyskretyzację treści językowej, a charakterystykę oparto o x-vector otrzymywany za pomocą jednego z modeli dostępnych w pakiecie speechbrain – spkrec-xvect-voxceleb. W celach badawczych anonimizację przeprowadzono, korzystając z dwóch baz danych.

Pierwsza z nich – oparta o korpus CMU ARCTIC – składała się z siedmiu mówców w ponad siedmiu tysiącach nagrań, a druga – wykorzystująca korpus libritts – składała się z dwóch i pół tysiąca nagrań pomiędzy mnogimi mówcami. Postanowiono dodatkowo przeprowadzić fine-tuning wykorzystywanego syntezatora, aby system działający w języku angielskim zaadaptować do przetwarzania mowy polskiej.



Rys. 4 Wykres t-SNE wektorów mówców dla systemu z bazą CMU ARCTIC.



Rys. 5 Wykres t-SNE wektorów mówców dla systemu z bazą libritts.

4.2 Anonimizacja

Dla opisywanego modelu anonimizacja przebiega na podstawie manipulacji wektorem osadzenia mówcy [2]. Wymyślono, oraz zaimplementowano więc trzy sposoby przekształceń. Pierwszy opierał się na losowym doborze wektora z bazy danych, który następnie stawał się embeddingiem anonimowego pseudomówcy. Drugi z nich , swoją bazę miał w odległości euklidesowej. W tym przypadku wektorem anonimowym stawała się średnia arytmetyczna spośród trzech wektorów o największej odległości, od mówcy którego sygnał jest przetwarzany. Trzeci sposób, również bazujący na odległości euklidesowej zakładał wybranie wektora o największej odległości.

4.3 Osiągnięcia metody

Rozwijając ów model, stworzono działający system resyntezujący mowę zarówno w języku angielskim i polskim. Implementacja metod anonimizacji umożliwiła skuteczne zamaskowanie tożsamości oryginalnego nagrania. System wykorzystujący bazę CMU ARCTIC charakteryzuje się większą przejrzystością dźwięku wyjściowego, podczas gdy ten z libritts charakteryzował się dużą różnorodnością głosów wyjściowych.

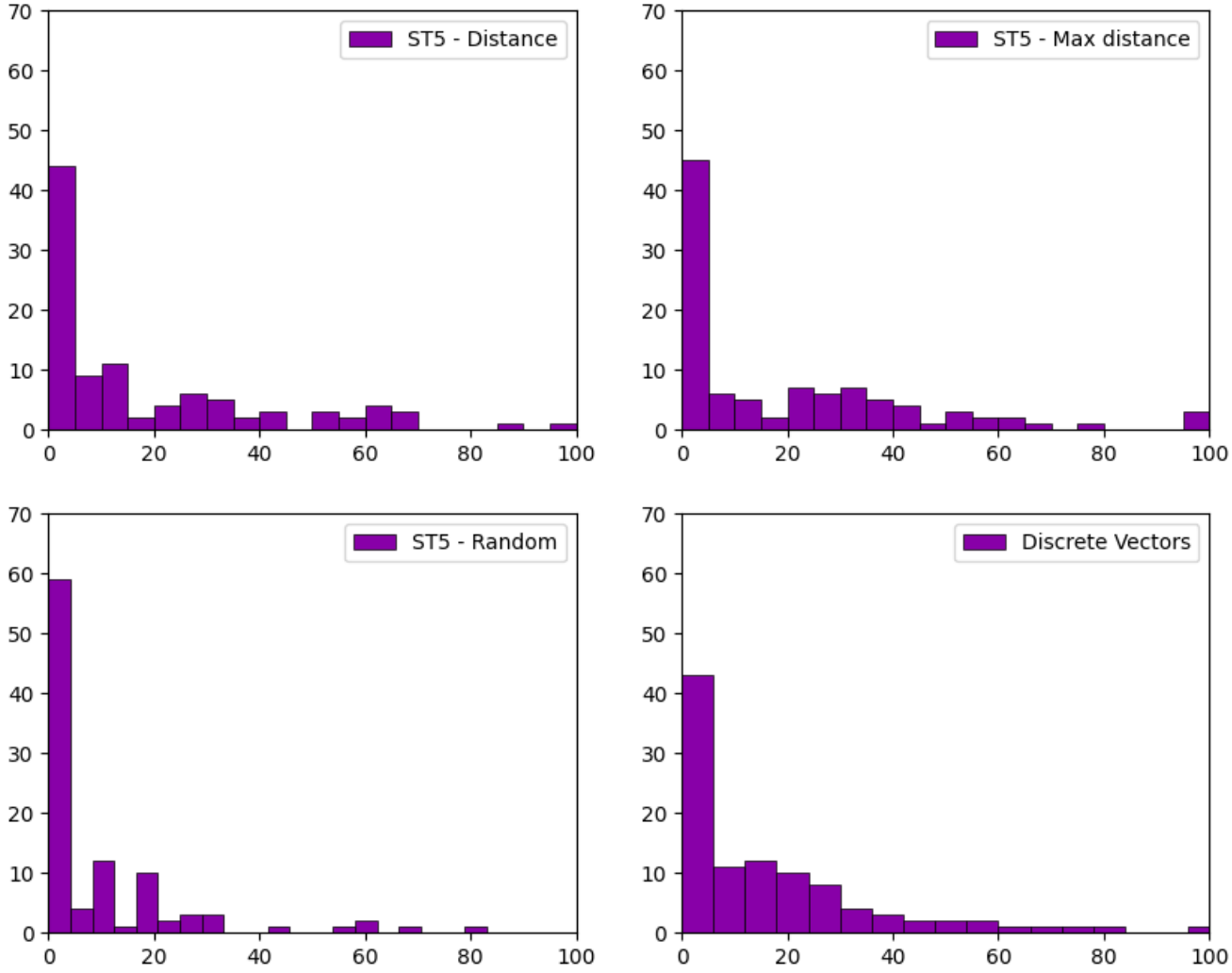
4.4 Kierunek rozwoju podejścia

Z racji na dowolność przekształceń wektorów osadzeń istnieje wiele sposobów, na które można byłoby przeprowadzić anonimizację. Dodatkowo dobrym pomysłem jest dalszy rozwój polskiego systemu. Wymagałoby to wykorzystania innego zbioru danych. Obecny system trenowany był na polskim podzbiorze zbioru VoxPopuli. Udało się zawrzeć tokenizację każdego z polskich znaków diakrytycznych, jednak aby uzyskać ich poprawną reprezentację w mowie resyntezowanej, należałoby zwiększyć ilość danych, lub reprezentatywność rzadko wykorzystywanych liter, poprzez zmianę bazy danych. Ową bazę wykorzystać można byłoby również do anonimizacji x-vectorów, by w pełni zlokalizować system. Angielska wersja zwraca mowę w pełni zrozumiałą i zgodną z zadanym tekstem. Zdarzają się jednak niekiedy w każdej wersji modelu halucynacje zniekształcające sygnał wyjściowy. Ostatecznie, można byłoby podjąć próbę inkorporacji modelu, w system opisany w punkcie trzecim, aby augmentować go o możliwość anonimizacji.

5. Ewaluacja modeli

5.1 Word error rate

Jako metrykę analizującą wierność odwzorowania i jakość akustyczną obraliśmy word error rate. Zrobiono to, porównując transkrypcje modelem Whisper sygnałów przed i po anonimizacją.



Rys. 6 Rozkłady WER [%] dla każdego z badanych modeli.

Tab. 1 Miary rozkładów WER [%].

	Distance anonymization	**Max distance speaker**	**Random speaker**	**Discrete Vectors**
Średnia	17,34	18,98	9,40	18,84
Odchylenie standardowe	22,73	24,02	16,11	26,31
Mediana	9,00	9,00	0,00	11,00

Do błędów występujących najczęściej zaliczyć można błędną interpretację obcojęzycznych nazw własnych, a także wyrazach złożonych. Dodatkowo wystąpiły błędy w rozumieniu gramatyki tekstu, gdzie pod uwagę brane były różnice pomiędzy znakami interpunkcyjnymi – kropki, przecinki i apostrofy przyległe do słów. Ostatnim z głównych archetypów błędów były cyfry, liczby i daty, który występował wyłącznie w modelach wykorzystujących SpeechT5 z powodu nieobsługiwania przezeń tokenów numerycznych. Odpowiednia manipulacja transkrypcją oraz preparacja pozwoliłaby na uniknięcie tego problemu przez substytucję liczb tekstem, jednak w dostępnym czasie, osiągniecie pełnej efektywności tokenizacji, nie było możliwe.

5.2 Cosine distance

W celu ewaluacji anonimizacji samej w sobie warto zadać sobie pytanie, czym tak naprawdę jest anonimowość głosu. Jeśli głos traktujemy jako część naszej biometrii, skorzystać można wówczas z miar wykorzystywanych w systemach biometrycznych, np. cosinusa kąta pomiędzy embeddingiem mówcy oryginalnego kontra anonimowego (1)

$$cos(x_1, x_2) = \frac{x_1 \cdot x_2}{||x_1|| \cdot ||x_2||} \quad (1)$$

gdzie:

- x_1 – x-vector oryginalnego mówcy,
- x_2 – x-vector pseudomówcy anonimowego,
- $||x_1||$ – norma euklidesowa sygnału oryginalnego,
- $||x_2||$ – norma euklidesowa sygnału anonimowego.

Tab. 2 Wartości odległości cosinusowej dla badanych modeli.

	Distance anonymization	**Max distance speaker**	**Random speaker**	**Discrete Vectors**
Średnia	-0,08	-0,08	-0,03	0,85
Odchylenie standardowe	0,01	0,01	0,02	0,03
Kąt [°]	94,59	94,59	91,72	31,79

6. Wnioski

W ciągu czterech tygodni prac udało się nam stworzyć systemy anonimizacji mowy działające na podstawie dwóch różnych systemów. Wynikiem każdego z nich jest zrozumiała mowa w języku angielskim. Dodatkowo przeprowadzono fine-tuning syntezatora mowy SpeechT5 w celu stworzenia systemu działającego w języku polskim. Modele ewaluowano zarówno pod kątem poprawności przenoszenia treści językowej, jak i treści biometrycznej. Całościowo trudno jest stwierdzić, który system, a także która z metod anonimizacji jest najlepsza. Oba układy posiadają swoje atuty i wady. Analogicznie, każda z metod anonimizacji jest na swój sposób unikatowa. Metoda odległościowa pozwala na stworzenie pseudomówcy nieznajdującego się w zbiorze danych. Metoda maksymalnej odległości pozwala oddalić się w bazie danych najbardziej od mówcy oryginalnego. Metoda losowa jest nieprzewidywalna, co sprawia, że odległość wektora mówcy od wektora anonimowego jest losowa z każdym wywołaniem. Podobnie działa metoda anonimizacji wektorami dyskretnymi, gdzie odległość wektora mówcy od anonimowego jest różna, w zależności od mówcy, co zapobiega możliwości wykonania operacji deanonimizacji głosu. Podsumowując, temat prywatyzacji głosu jest bardzo obszerny i trudno go dokładnie zbadać, jednak podczas tego projektu udało się nam poznać jego podstawy i stworzyć działające modele odwzorowujące mowę ze zmienioną charakterystyką mówcy.

Wyrażamy zgodę na prezentację wyników projektu, którego dotyczy ten raport na stronie Zespołu Przetwarzania Sygnałów oraz w ramach wydarzeń promujących uczelnię, takich jak Noc Naukowców, Dzień Otwarty AGH, Festiwal Nauki.

Wyrażamy zgodę na publikację naszych imion i nazwisk podczas prezentacji wyników projektu, którego dotyczy ten raport.

Bibliografia

[1] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux: „Speech Resynthesis from Discrete Disentangled Self-Supervised Representations”

[2] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, “Speaker Anonymization Using X-vector and Neural Waveform Models,” in 10th ISCA Speech Synthesis Workshop, 2019.