Ewelina Bartuzi-Trokielewicz*

Analiza potencjału ataków wykorzystujących syntetycznie wygenerowane głosy na systemy biometrycznej weryfikacji mówców oraz wykrywania materiałów deepfake

Streszczenie

Postęp technologiczny w dziedzinie głębokiego uczenia znacząco przyczynił się do rozwoju syntezowania głosu, umożliwił tworzenie realistycznych nagrań audio, które mogą naśladować indywidualne cechy głosów ludzkich. Chociaż ta innowacja otwiera nowe możliwości w dziedzinie technologii mowy, niesie ze sobą również poważne obawy dotyczące bezpieczeństwa, zwłaszcza w kontekście potencjalnego wykorzystania technologii deepfake do celów przestępczych. Przeprowadzone badanie koncentrowało się na ocenie wpływu syntetycznych głosów na systemy biometrycznej weryfikacji mówców w języku polskim oraz skuteczności wykrywania deepfake'ów narzędziami dostępnymi publicznie, z wykorzystaniem dwóch głównych metod generowania głosu, tj. przekształcenia tekstu na mowę oraz konwersji mowy.

Jednym z głównych wniosków analizy jest potwierdzenie zdolności syntetycznych głosów do zachowania charakterystycznych cech biometrycznych i otwierania drogi przestępcom do nieautoryzowanego dostępu do zabezpieczonych systemów lub danych. To podkreśla potencjalne zagrożenia dla indywidualnych użytkowników oraz instytucji, które polegają na technologiach rozpoznawania mówcy jako metodzie uwierzytelniania

* Dr inż. Ewelina Bartuzi-Trokielewicz, adiunkt, Zespół Złożonych Systemów, Instytut Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska, e-mail: ewelina.trokielewicz@pw.edu.pl, ORCID: 0000-0001-6245-2908.

i wskazuje na konieczność wdrażania modułów wykrywania ataków. Badanie ponadto pokazało, że deepfaki odnalezione w polskiej części internetu dotyczące promowania fałszywych inwestycji lub kierowane w celach dezinformacji najczęściej wykorzystują popularne i łatwo dostępne narzędzia do syntezy głosu.

Badanie przyniosło również nowe spojrzenie na różnice w skuteczności metod konwersji tekstu na mowę i klonowania mowy. Okazuje się, że metody klonowania mowy mogą być bardziej skuteczne w przekazywaniu biometrycznych cech osobniczych niż metody konwersji tekstu na mowę, co stanowi szczególny problem z punktu widzenia bezpieczeństwa systemów weryfikacji.

Wyniki eksperymentów podkreślają potrzebę dalszych badań i rozwoju w dziedzinie bezpieczeństwa biometrycznego, żeby skutecznie przeciwdziałać wykorzystywaniu syntetycznych głosów do nielegalnych działań. Wzrost świadomości o potencjalnych zagrożeniach i kontynuacja pracy nad ulepszaniem technologii weryfikacji mówców są ważne dla ochrony przed coraz bardziej wyrafinowanymi atakami wykorzystującymi technologię deepfake.

Słowa kluczowe: cyberbezpieczeństwo, biometria, weryfikacja mówców, audio deepfake, syntetyczne głosy, metody sztucznej inteligencji, uczenie maszynowe, uczenie głębokie

Wstęp

W erze cyfrowej, gdy technologia ewoluuje w zawrotnym tempie, pojawiają się nowe wyzwania związane z bezpieczeństwem cybernetycznym. Jednym z najbardziej niepokojących zjawisk ostatnich lat jest rozwój i upowszechnienie technologii deepfake audio. Technologie te, wykorzystujące zaawansowane algorytmy sztucznej inteligencji do syntetycznego generowania głosów, zyskują na popularności, dlatego że oferują możliwości, które jeszcze niedawno wydawały się niemożliwe. Jednakże tak samo szybko jak rosną możliwości technologiczne pojawiają się też metody ich nadużycia, które stwarzają realne zagrożenia dla indywidualnych i zbiorowych aspektów bezpieczeństwa. W tym kontekście systemy biometrycznej weryfikacji mówców, które są coraz powszechniej stosowane w różnych dziedzinach życia, od bankowości po systemy bezpieczeństwa, stają przed wyjątkowymi wyzwaniami.

Rozwój technologii audio deepfake, począwszy od wczesnych lat dwudziestych XXI wieku, otworzył nowe fronty w cyberprzestępczości. Ponieważ technologia ta umożliwia tworzenie niemal nierozróżnialnych od oryginału syntetycznych kopii głosów, więc stwarza potencjalne zagrożenia dla systemów weryfikacji tożsamości opartych na głosie. Ataki te mogą mieć różne cele, od oszustw finansowych po manipulowanie opinią publiczną i interwencje w procesy demokratyczne. Przykłady wykorzystania tej technologii przez oszustów, którzy podszywają się pod głosy osób na wysokich stanowiskach w celu wyłudzenia ogromnych sum pieniędzy, uwydatniają realne ryzyko

związane z tymi narzędziami. Ponadto zdolność do szerzenia dezinformacji poprzez manipulację audio może podważać zaufanie do mediów, instytucji oraz indywidualnych osób, stawać się narzędziem w rękach propagandystów i terrorystów[1]. Syntetycznie wygenerowane głosy są też wykorzystywane w materiałach wideo. Coraz częściej spotykamy się z manipulowaniem głosami i wizerunkami osób publicznych w celu tworzenia fałszywych wiadomości lub wypowiedzi, które nigdy nie miały miejsca[2]. Ta praktyka nie tylko wprowadza w błąd odbiorców, lecz także zagraża reputacji osób, których dotyczy i może prowadzić do niesprawiedliwych osądów publicznych. W konsekwencji rośnie zapotrzebowanie na zaawansowane technologie wykrywające deepfaki oraz na świadomość społeczną krytycznego odbioru treści cyfrowych.

Oprócz generowania syntetycznych głosów w celach rozrywkowych, gdzie technologia ta pozwala na ożywienie historycznych postaci w dokumentach czy tworzenie bardziej realistycznych postaci w grach wideo, warto zwrócić uwagę na przypadki, które uwydatniają ciemniejszą stronę tej technologii. Przykładowo, w 2019 roku oszuści wykorzystali technologię deepfake do naśladowania głosu dyrektora generalnego dużej firmy w celu nakłonienia pracownika do przelania kilkuset tysięcy dolarów na fałszywe konto bankowe[3]. Innym razem deepfake audio użyto do próby wymuszenia okupu[4]. Mieszkanka Arizony usłyszała w telefonie rzekomo porwaną córkę, a później porywaczy żądających okupu i grożących przy tym konsekwencjami.

Ponadto zdolność do szerzenia dezinformacji poprzez manipulację audio może podważać zaufanie do mediów, instytucji oraz indywidualnych osób, stać się narzędziem w rękach propagandystów i terrorystów. Syntetycznie wygenerowane głosy wykorzystywane są również w materiałach wideo, co coraz

1 *New Orleans magician says he made AI Biden robocall for aide to challenger*, https://www.theguardian.com/us-news/2024/feb/23/ai-biden-robocall-magician-new-orleans [dostęp: 15.02.2024].

2 *Podrobiony szef InPostu zachęca do inwestycji. Po chwili zaczyna mówić w obcym języku*, https://spidersweb.pl/2024/03/inpost-rafal-brzoska-oszustwo-deepfake.html [dostęp: 15.02.2024]; *Zostań inwestorem projektu Baltic Pipe? Uwaga na scam*, https://demagog.org.pl/fake_news/zostan-inwestorem-projektu-baltic-pipe-uwaga-na-scam/ [dostęp: 15.02.2024].

3 *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [dostęp: 15.02.2024].

4 *„Mom, these bad men have me": She believes scammers cloned her daughter's voice in a fake kidnapping*, https://amp.cnn.com/cnn/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html [dostęp: 15.02.2024].

częściej prowadzi do tworzenia przekonujących fałszywych wypowiedzi polityków lub celebrytów, które mogą być rozpowszechniane w mediach społecznościowych i tym samym wprowadzać w błąd opinię publiczną i wpływać na procesy polityczne i społeczne na całym świecie.

Jednocześnie globalne raporty wskazują na rosnącą świadomość i doświadczenie osób, które padły ofiarą oszustw wykorzystujących klonowanie głosu[5]. Pomimo rosnącej świadomości zagrożeń wykrycie manipulacji audio i im przeciwdziałanie pozostaje wyzwaniem. Normatywne akty prawne takie jak zakazy używania sztucznej inteligencji do fałszowania głosów w połączeniach telefonicznych stanowią krok w dobrym kierunku, ale technologia nadal wyprzedza regulacje.

Generowanie syntetycznych głosów

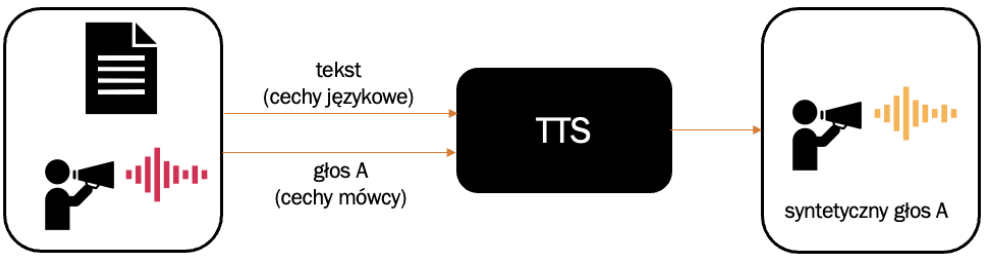
Audio deepfake to technologia wykorzystująca algorytmy uczenia maszynowego do tworzenia fałszywych, ale realistycznie brzmiących nagrań głosowych. Jest to forma syntetycznej generacji mowy, która może naśladować głos konkretnej osoby na podstawie dostępnych próbek audio. Chociaż technologia ta ma potencjał do pozytywnych zastosowań, np. przywracanie zdolności mówienia osobom po utracie głosu czy tworzenie bardziej naturalnych interfejsów głosowych, niesie ze sobą również znaczne ryzyko nadużyć.

Początkowo stosowano ataki odtworzeniowe polegające na wykorzystaniu nagranych i odtwarzanych próbek głosowych, miały one jednak swoje ograniczenia. Obecnie rozwój technik sztucznej inteligencji pozwolił na realistyczne klonowanie głosu ludzi. Wyróżniamy dwa główne nurty generowania syntetycznych głosów, znanych również jako audio deepfaki, tj. przekształcenie tekstu na mowę (ang. *Text-To-Speech* – TTS) oraz konwersji głosu (ang. *Voice Conwersion* – VC, lub *Speech-To-Speech* – STS).

Technologia TTS wykorzystuje sieci neuronowe do zamiany tekstu na mowę syntetyczną przez szereg zaawansowanych procesów. Na wstępie tekst jest analizowany i rozbijany na podstawowe jednostki fonetyczne, które są dalej przetwarzane przez zaawansowane mechanizmy lingwistyczne w celu odtworzenia naturalnych modulacji głosu takich jak ton czy akcent. Następnie

5 *Kobieta pokazała, jak złodzieje sklonowali jej głos i na Facebooku próbowali wyłudzić kod BLIK*, https://natemat.pl/546320,oszustwo-na-blika-zlodzieje-podrobili-glos-z-ai [dostęp: 15.02.2024].

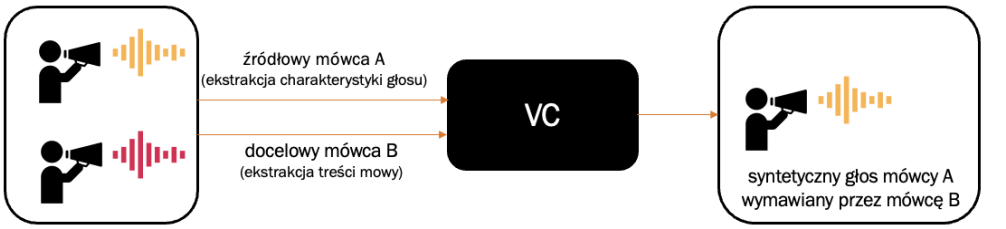
technologia ta skupia się na naśladowaniu ludzkich cech akustycznych, które są wynikiem działania różnorodnych elementów anatomicznych, w tym jamy nosowej, podniebienia, języka, warg, żuchwy oraz strumienia powietrza wydychanego z płuc. Rezonans, który pojawia się w trakcie przepływu powietrza przez aparat mowy, tworzy dźwięki o unikalnych właściwościach dla każdej osoby. W technologii TTS te unikalne cechy dźwięku są przekształcane w cyfrowe reprezentacje, tzw. embeddingi akustyczne. Ostatni etap łączy te cechy akustyczne z fonetycznymi aspektami tekstu, co umożliwia wygenerowanie głosu brzmiącego realistycznie. Centralną część tego procesu stanowi *vocoder*, czyli moduł kodowania głosu (ang. *voice coder*), który syntezuje fale dźwiękowe. Te cyfrowe zapisy audio, odtwarzane później jako mowa, wiernie oddają charakterystyczne cechy głosu danej osoby.



Źródło: opracowanie własne.

Schemat 1. Uproszczony schemat działania systemu zamiany tekstu na mowę syntetyczną

Konwersja głosu to zaawansowana metoda pozwalająca na tworzenie cyfrowych replik ludzkich głosów. Techniki VC są rozwiniętymi metodami obróbki dźwięku umożliwiającymi przeniesienie specyficznych cech głosowych z głosu jednej osoby – źródłowego mówcy „A", do tworzenia mowy syntetycznej przez inną osobę – docelowego mówcy „B", który zapewnia treść i intonację. Procedura ta rozpoczyna się od wydobycia unikalnych właściwości głosu z nagranych próbek dźwiękowych mówcy „A". Te unikalne właściwości określane jako embeddingi odzwierciedlają to, co sprawia, że każdy głos jest niepowtarzalny. Następnie technologia VC wykorzystuje te embeddingi do tworzenia mowy, która zachowuje charakterystyczne cechy głosu źródłowego, lecz przekazuje treść mówcy „B". Dzięki tym systemom możliwe jest nie tylko odwzorowanie tonu i sposobu artykulacji konkretnego człowieka, lecz także imitowania jego emocji i stylu mówienia.



Źródło: opracowanie własne.

Schemat 2. Uproszczony schemat działania systemu VC.
Moduł ten tworzy syntetyczny głos na podstawie cech akustycznych mówcy źródłowego oraz treści i sposobu mówienia docelowego mówcy (fałszerza)

Narzędzia do syntezy głosu można podzielić na rozwiązania komercyjne oraz narzędzia o otwartym kodzie źródłowym (ang. *open-source*). Popularne narzędzia komercyjne to rozwiązania firm: Google Cloud Text-to-Speech[6], Microsoft Azure Text to Speech[7], Amazon Polly[8], IBM Watson Text to Speech[9], ElevenLabs[10]. Charakteryzują się one niezwykłą jakością i realizmem.

W kontekście tych wyzwań niniejszy artykuł bada jak technologie audio deepfake wpływają na skuteczność i wiarygodność systemów biometrycznej weryfikacji mówców oraz metody wykrywania audio deepfake'ów. Analiza ta ma na celu nie tylko zrozumienie obecnego stanu zagrożeń, ale również zidentyfikowanie potencjalnych ścieżek rozwoju technologicznego i regulacyjnego mogących pomóc w ochronie przed negatywnymi skutkami audio deepfake'ów.

Biometryczne systemy rozpoznawania mówców

Systemy biometryczne to zaawansowane technologie identyfikacji i weryfikacji tożsamości oparte na unikalnych cechach fizycznych lub behawioralnych człowieka. W przeciwieństwie do tradycyjnych metod takich jak hasła czy karty dostępu, biometria wykorzystuje indywidualne atrybuty, żeby zapewnić wysoki poziom bezpieczeństwa i wygody użytkowania. Te cechy mogą obejmować

6 *Google Cloud Text-to-Speech*, https://cloud.google.com/text-to-speech [dostęp: 15.02.2024].
7 *Microsoft Azure Text to Speech*, https://azure.microsoft.com/en-us/products/ai-services/text-to-speech [dostęp: 15.02.2024].
8 *Amazon Polly*, https://aws.amazon.com/polly/ [dostęp: 15.02.2024].
9 *IBM Watson Text to Speech*, https://www.ibm.com/products/text-to-speech [dostęp: 15.02.2024].
10 *ElevenLabs*, https://elevenlabs.io [dostęp: 15.02.2024].

odciski palców, geometrię dłoni, wzory tęczówki lub siatkówki oka, a także charakterystyki behawioralne takie jak podpis czy sposób chodzenia.

Wśród różnorodnych systemów biometrycznych systemy rozpoznawania mówców wyróżniają się unikalną modalnością, która identyfikuje i weryfikuje osoby na podstawie ich cech głosu. Głos jako modalność biometryczna jest szczególnie atrakcyjny ze względu na jego nieinwazyjność i zdolność do zdalnego stosowania, co umożliwia zastosowanie w wielu scenariuszach, od autoryzacji dostępu do systemów teleinformatycznych, przez telefoniczne systemy bankowe, po inteligentne domy.

Głos jest wyjątkowo złożoną cechą, która jest kształtowana przez kombinację czynników fizjologicznych i behawioralnych, w tym kształt i rozmiar struktur anatomicznych takich, jak: jama ustna, gardło, język, a także styl mówienia i akcent. To sprawia, że każdy głos jest unikatowy, co czyni go skutecznym środkiem do identyfikacji osobowej, Systemy rozpoznawania mówców wykorzystują te unikalne cechy, analizują różne aspekty sygnału głosowego, takie, jak: barwa, tonacja, siła głosu, szybkość mówienia, a także bardziej subtelne cechy akustyczne.

Systemy te działają na zasadzie porównywania cech głosu próbki z cechami głosu zapisanymi w bazie danych. Proces ten można podzielić na dwie główne fazy – rejestrację (ang. *enrollment*) i weryfikację lub identyfikację. Podczas rejestracji system analizuje głos użytkownika, ekstrahuje charakterystyczne cechy i tworzy wzorzec biometryczny, który jest następnie zapisywany. W fazie weryfikacji porównuje się prezentowane próbki głosu z wcześniej zapisanymi wzorcami, żeby stwierdzić, czy próbka należy do zadeklarowanej osoby.

Pomimo licznych zalet systemy rozpoznawania mówców napotykają na wyzwania takie, jak: wrażliwość na zmiany w środowisku akustycznym, choroby wpływające na głos, a także manipulacje technologiczne, np. syntetyczne generowanie głosów (audio deepfake). Rozwój technologii deepfake stawia przed tymi systemami nowe wyzwania związane z zapewnieniem bezpieczeństwa i wiarygodności weryfikacji.

Wykrywanie audio deepfake'ów

Wykrywanie audio deepfake'ów staje się coraz ważniejszym elementem obrony przed manipulacją cyfrową, zwraca uwagę na różnorodne aspekty nienaturalności, które mogą ujawnić fałszerstwo. Jednym z podstawowych wskaźników są nienaturalne pauzy w wypowiedzi, które mogą sugerować syntetyczne

generowanie mowy, nieodpowiadające naturalnemu rytmowi rozmowy. Nietypowe intonacje, zmiany głosu lub akcentu w trakcie wypowiedzi, a także nagłe zmiany tonu i wysokości dźwięku mogą również wskazywać na manipulację. Dziwna wymowa słów lub nietypowy dla danej osoby styl mówienia, np. niecharakterystyczny ton czy tempo mowy, dodatkowo podkreślają potencjalną nienaturalność nagrania. Ponadto brak naturalnych odgłosów tła lub niespójności w hałasie w tle mogą sygnalizować, że nagranie zostało zmodyfikowane lub stworzone w kontrolowanym środowisku. Mogą pojawić się także subtelne artefakty dźwiękowe, które są związane z obróbką cyfrową i stanowią kolejny element, na który zwracają uwagę eksperci podczas analizy nagrań pod kątem autentyczności. Zrozumienie i rozpoznawanie tych wskaźników jest ważne dla skutecznego identyfikowania manipulacji audio i ochrony przed ich potencjalnie szkodliwym wpływem.

Podatność systemów na audio deepfaki

Pojawia się coraz więcej badań związanych ze zdolnością wykrywania deepfake'ów przez ludzi. Zespół z University College London badał, na ile efektywnie ludzie potrafią rozróżnić prawdziwą mowę od tej generowanej sztucznie z użyciem metod deepfake[11]. Wyniki pokazały, że uczestnicy zdołali prawidłowo zidentyfikować mowę stworzoną przez sztuczną inteligencję jedynie w 73% przypadków. Uczestnicy, którzy odpowiedzieli poprawnie, wskazywali, że zwrócili uwagę na takie elementy, jak: sposób oddychania, przerwy między słowami i ogólna płynność mowy. Interesujące jest to, że nawet po edukacji uczestników na temat charakterystycznych cech mowy generowanej sztucznie i przedstawieniu im przykładowych nagrań deepfake, ogólna zdolność do ich identyfikacji poprawiła się nieznacznie.

W pracach badawczych pojawiają się analizy wykorzystujące moduły wykrywania ataków odtworzeniowych oraz syntetycznie wygenerowane głosy. Postanowiliśmy przeanalizować wpływ audiodeepfake'ów tworzonych różnymi metodami na działanie przykładowego systemu rozpoznawania mówców oraz dostępnego online narzędzia do wykrywania deepfake'ów na fałszywe nagrania krążące w polskim internecie.

11 K.T. Mai, S. Bray, T. Davies, L.D. Griffin, *Warning: Humans cannot reliably detect speech deepfakes*, „PLoS ONE" 2023, nr 18(8), https://doi.org/10.1371/journal.pone.0285333 [dostęp: 12.02.2024].

W tym celu utworzono zbiory zawierające próbki prawdziwe i syntetyczne:

– **Fake-Public** – audio deepfaki osób publicznych opublikowane w mediach społecznościowych, o nieokreślonych metodach generowania głosu, ale prawdopodobnie wykorzystujące popularne metody syntezy głosu TTS oraz VC – 45 nagrań przedstawiających głos polskich polityków, dziennikarzy, influencerów oraz sportowców. Próbki mogą być w większości łatwo rozpoznawalne przez człowieka jako dane syntetyczne przez cechy takie, jak: robotyczny wydźwięk, błędy w odmianie liczb, inny sposób mowy oraz barwy głosu w porównaniu z oryginalnymi materiałami,

– **Real-Public** – referencyjne autentyczne nagrania osób ze zbioru **Fake--Public,**

– **Fake-Private** – 45 fałszywych próbek wykorzystujących popularne i niszowe metody syntezy głosu:

a) 5 próbek wygenerowanych narzędziem Coqui-TTS[12];

b) 5 próbek wytworzonych z wykorzystaniem ElevenLabs-TTS;

c) 10 próbek wygenerowanych SVC-VC[13], w tym 5 nagrań zawierających szum;

d) 15 próbek wygenerowanych RVC-VC[14], w tym 5 nagrań śpiewu i 5 nagrań zawierających szum;

e) 5 próbek wygenerowanych narzędziem ElevenLabs-VC;

f) 5 próbek zawierających śpiew – VC,

– **Real-Private** – referencyjny zbiór 45 prawdziwych nagrań użytych do wytworzenia zbioru Fake-Private, w tym 5 zawierających śpiew.

Do analizy biometrycznej wykorzystano dwa open-source'owe narzędzia weryfikacji mówców – Resambyzer[15] oraz SpeechBrain[16]. Testy rozpoznawania syntetycznych głosów przeprowadzono z wykorzystaniem popularnych, dostępnych online narzędzi Deepfake-TotaL[17] i AIorNOT[18]. Do oceny skuteczności wyznaczano procent poprawnie rozpoznanych materiałów syntetycznych jako deepfaki.

12 https://github.com/coqui-ai/TTS [dostęp: 28.01.2024].
13 https://github.com/svc-develop-team/so-vits-svc [dostęp: 28.01.2024].
14 https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI [dostęp: 28.01.2024].
15 https://github.com/resemble-ai/Resemblyzer [dostęp: 28.01.2024].
16 https://github.com/speechbrain/speechbrain [dostęp: 28.01.2024].
17 https://deepfake-total.com [dostęp: 28.01.2024].
18 https://www.aiornot.com [dostęp: 28.01.2024].

Wyniki przeprowadzonych testów zaprezentowano w tabeli 1. Zauważa się, że syntetyczne nagrania mimo słyszalnych nieścisłości są zbliżone do oryginalnych głosów i akceptowane przez systemy biometryczne. Lepszą rozróżnialnością charakteryzuje się system SpeechBrain, który osiąga skuteczność rozpoznawania audio deepfake'ów na poziomie 35,56% dla zbioru materiałów publicznych oraz 31,11% dla materiałów przygotowanych różnymi metodami syntezy głosu. Druga metoda wykrywa jedynie 33,33% w przypadku deepfake'ów ze zbioru Fake-Public i 24,44% ze zbioru Fake-Private. W przypadku metod celowanych na wykrywanie sklonowanych głosów zdecydowanie lepiej działa narzędzie Deepfake-Total, pozwala bowiem na wykrywanie 93,33% materiałów ze zbioru publicznych. Może to wskazywać, że spreparowane materiały zostały stworzone popularnymi metodami syntezy głosu. W przypadku zbioru przygotowanego na potrzeby pracy badawczej narzędzie to miało 53,33% skuteczności, z czego wykrywało wszystkie ataki przeprowadzone metodami TTS. Najwięcej problemów miało z deepfake'ami zawierającymi śpiew i dźwięki otoczenia, a także z metodami mniej popularnymi – SVC i RVC.

Tabela 1. Skuteczność rozpoznawania mowy syntetycznej systemami biometrycznej weryfikacji mówców oraz narzędziami do wykrywania deepfake'ów

Zbiór	Weryfikacja mówców (%)		Wykrywanie deepfake'ów (%)	
	Resablyzer	SpeechBrain	Deepfake-Total	AIorNOT
Fake-Public	33,33	35,56	93,33	8,89
Fake-Private	24,44	31,11	53,33	6,67
TTS	80,00	90,00	100,0	13,33
VC – bez szumu	20,00	20,00	73,33	6,67
VC – z szumem	0,00	6,67	13,33	0,00
VC – śpiew	0,00	6,67	6,67	0,00

Źródło: Opracowanie własne.

Zakończenie

Artykuł dotyczył analizy audio deepfake'ów osób publicznych dostępnych w internecie oraz wygenerowanych próbek za pomocą narzędzi do syntezy głosu, tj. dwóch narzędzi zamiany tekstu na mowę (TTS) oraz trzech narzędzi konwersji głosu (VC). Celem badania było poznanie, w jaki sposób te syntetyczne próbki głosu są rozpoznawane przez systemy biometrycznej weryfikacji mówców oraz narzędzia wykrywające audio deepfake.

Eksperymenty wykazały, że próbki wygenerowane przez popularne metody TTS i VC są łatwo rozpoznawalne zarówno przez narzędzie do wykrywania audio deepfake'ów, jak i przez system weryfikacji tożsamości. To sugeruje, że obecne algorytmy i technologie mają zdolność do efektywnej identyfikacji próbek syntetycznych, co jest ważnym krokiem w walce z potencjalnym nadużyciem technologii deepfake.

Badanie pokazało, że w przypadku bardziej niszowych metod generowania głosu oraz próbek zawierających szum otoczenia lub śpiew skuteczność narzędzi w rozróżnianiu próbek syntetycznych od naturalnych znacząco spada. Mimo słyszalnych niedoskonałośc, te bardziej specyficzne lub złożone typy audio są trudniejsze do zidentyfikowania przez obecne systemy.

To potwierdza potrzebę dalszego rozwoju i dostosowywania narzędzi wykrywających deepfaki oraz systemów weryfikacji tożsamości, żeby mogły one skutecznie radzić sobie z rosnącą różnorodnością i zaawansowaniem technik generowania syntetycznego audio. Wnioski z tego badania wskazują na konieczność ciągłej adaptacji i ulepszania technologii weryfikacyjnych w związku z ewolucją metod tworzenia deepfake'ów w celu zapewnienia bezpieczeństwa i autentyczności komunikacji cyfrowej.

Bibliografia

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [dostęp: 15.02.2024].

Kobieta pokazała, jak złodzieje sklonowali jej głos i na Facebooku próbowali wyłudzić kod BLIK, https://natemat.pl/546320,oszustwo-na-blika-zlodzieje-podrobili-glos-z-ai [dostęp: 15.02.2024].

Mai K.T., Bray S., Davies T., Griffin L.D., *Warning: Humans cannot reliably detect speech deepfakes*, „PLoS ONE" 2023, nr 18(8), https://doi.org/10.1371/journal.pone.0285333 [dostęp: 12.02.2024].

„Mom, these bad men have me": She believes scammers cloned her daughter's voice in a fake kidnapping, https://amp.cnn.com/cnn/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html [dostęp: 15.02.2024].

New Orleans magician says he made AI Biden robocall for aide to challenger, https://www.theguardian.com/us-news/2024/feb/23/ai-biden-robocall-magician-new-orleans [dostęp: 15.02.2024].

Podrobiony szef InPostu zachęca do inwestycji. Po chwili zaczyna mówić w obcym języku, https://spidersweb.pl/2024/03/inpost-rafal-brzoska-oszustwo-deepfake.html [dostęp: 15.02.2024].

Zostań inwestorem projektu Baltic Pipe? Uwaga na scam, https://demagog.org.pl/fake_news/zostan-inwestorem-projektu-baltic-pipe-uwaga-na-scam/ [dostęp: 15.02.2024].

Analysis of the Potential for Attacks Utilizing Synthetically Generated Voices on Biometric Speaker Verification Systems

Abstract

Technological advancements in the field of deep learning have significantly contributed to the development of voice synthesis, enabling the creation of realistic audio recordings that can mimic the individual characteristics of human voices. While this innovation opens up new possibilities in the field of speech technology, it also raises serious security concerns, especially in the context of the potential use of deepfake technology for criminal purposes. Our study focuses on assessing the impact of synthetic voices on biometric speaker verification systems in Polish and the effectiveness of detecting deepfakes with publicly available tools, considering two main approaches to voice generation: text-to-speech conversion and speech conversion.

One of the main findings of our research is the confirmation that synthetic voices are capable of retaining biometric characteristics, which could allow criminals unauthorized access to protected systems or data. The analysis showed that the greater the biometric similarity between the „victim's" voice and the „criminal's" synthetic voice, the more difficult it is for verification systems to distinguish between real and fake voices. This highlights the potential threats to individual users and institutions that rely on speaker recognition technologies as a method of authentication.

Our study also provides a new perspective on the differences in the effectiveness of text-to-speech conversion methods versus speech cloning. It turns out that speech cloning methods may be more effective in conveying individual biometric characteristics than text-to-speech conversion methods, posing a particular problem from the security perspective of verification systems.

The results of the experiments underscore the need for further research and development in the field of biometric security to effectively counteract the use of synthetic voices for illegal activities. Increasing awareness of potential threats and continuing work on improving speaker verification technologies are crucial for protecting against increasingly sophisticated attacks utilizing deepfake technology.

Key words: cybersecurity, biometrics, speaker verification, audio deepfakes, synthetic voices, artificial intelligence methods, machine learning, deep learning