Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Informatyka Stosowana (IST)**
Specjalność: **Projektowanie Systemów Informatycznych (PSI)**

PRACA DYPLOMOWA
MAGISTERSKA

Analiza możliwości wykorzystania metod uczenia maszynowego w rekonstrukcji nagrań dźwiękowych

Sebastian Łakomy

Opiekun pracy
Dr inż. Maciej Walczyński

Uczenie Maszynowe, sieci GAN, Rekonstrukcja Audio

WROCŁAW (2024)

Spis treści

1) Wstęp

1.1) Tło historyczne nagrań dźwiękowych

Historia rejestracji dźwięku sięga połowy XIX wieku, kiedy to w 1857 roku Édouard-Léon Scott de Martinville skonstruował fonoautograf - pierwsze urządzenie zdolne do zapisywania dźwięku [1]. Choć fonoautograf nie umożliwiał odtwarzania zarejestrowanych dźwięków, stanowił przełom w dziedzinie akustyki i zapoczątkował erę nagrań dźwiękowych. Pierwszym nagraniem uznawanym za możliwe do odtworzenia była francuska piosenka ludowa „Au Clair de la Lune", zarejestrowana przez Scotta w 1860 roku [1].

Kolejnym kamieniem milowym w historii rejestracji dźwięku było wynalezienie fonografu przez Thomasa Edisona w 1877 roku. Urządzenie to nie tylko zapisywało dźwięk, ale również umożliwiało jego odtwarzanie, co otworzyło drogę do komercjalizacji nagrań dźwiękowych [2]. W następnych dekadach technologia nagrywania ewoluowała, przechodząc przez etapy takie jak płyty gramofonowe, taśmy magnetyczne, aż po cyfrowe nośniki dźwięku [3].

Kluczowe momenty w historii rejestracji dźwięku obejmują:

1. 1888 - Wprowadzenie płyt gramofonowych przez Emile'a Berlinera [4]
2. 1920-1930 - Rozwój nagrań elektrycznych, znacząco poprawiających jakość dźwięku [5]
3. 1948 - Pojawienie się płyt długogrających (LP) [6]
4. 1963 - Wprowadzenie kaset kompaktowych przez Philips [7]
5. 1982 - Komercjalizacja płyt CD, rozpoczynająca erę cyfrową w muzyce [8]

Rozwój technologii nagrywania miał ogromny wpływ na jakość i dostępność nagrań muzycznych. Wczesne nagrania charakteryzowały się ograniczonym pasmem częstotliwości, wysokim poziomem szumów i zniekształceń [9]. Wraz z postępem technologicznym, jakość dźwięku stopniowo się poprawiała, osiągając szczyt w erze cyfrowej. Jednocześnie, ewolucja nośników dźwięku od płyt gramofonowych przez kasety magnetyczne po pliki cyfrowe, znacząco zwiększyła dostępność muzyki dla szerokiego grona odbiorców [10].

Warto zauważyć, że mimo znacznego postępu technologicznego, wiele historycznych nagrań o ogromnej wartości kulturowej i artystycznej wciąż pozostaje w formie, która nie oddaje pełni ich oryginalnego brzmienia. Stwarza to potrzebę rozwoju zaawansowanych technik rekonstrukcji i restauracji nagrań, co stanowi jedno z głównych wyzwań współczesnej inżynierii dźwięku i muzykologii [11], [12].

1.2) Problematyka jakości historycznych nagrań

Historyczne nagrania dźwiękowe, mimo ich ogromnej wartości kulturowej i artystycznej, często charakteryzują się niską jakością dźwięku, co stanowi istotne wyzwanie dla współczesnych słuchaczy i badaczy. Problematyka ta wynika z kilku kluczowych czynników.

Ograniczenia wczesnych technologii nagrywania stanowiły główną przeszkodę w uzyskiwaniu wysokiej jakości dźwięku. Wczesne urządzenia rejestrujące charakteryzowały się wąskim pasmem przenoszenia, co skutkowało utratą zarówno niskich, jak i wysokich częstotliwości [9]. Typowe pasmo przenoszenia dla nagrań z początku XX wieku wynosiło zaledwie 250-2500 Hz, co znacząco ograniczało pełnię brzmienia instrumentów i głosu ludzkiego [13]. Ponadto, pierwsze systemy nagrywające wprowadzały znaczne szumy i zniekształcenia do rejestrowanego materiału, co było spowodowane niedoskonałościami mechanicznymi i elektrycznymi ówczesnych urządzeń [14].

Wpływ warunków przechowywania na degradację nośników jest kolejnym istotnym czynnikiem wpływającym na jakość historycznych nagrań. Nośniki analogowe, takie jak płyty gramofonowe czy taśmy magnetyczne, są szczególnie podatne na uszkodzenia fizyczne i chemiczne [15]. Ekspozycja na wilgoć, ekstremalne temperatury czy zanieczyszczenia powietrza może prowadzić do nieodwracalnych zmian w strukturze nośnika, co przekłada się na pogorszenie jakości odtwarzanego dźwięku. W przypadku taśm magnetycznych, zjawisko print-through, polegające na przenoszeniu sygnału magnetycznego między sąsiednimi warstwami taśmy, może wprowadzać dodatkowe zniekształcenia [16].

Przykłady znaczących nagrań historycznych o niskiej jakości dźwięku są liczne i obejmują wiele kluczowych momentów w historii muzyki. Jednym z najbardziej znanych jest nagranie Johannesa Brahmsa wykonującego fragment swojego „Tańca węgierskiego nr 1" z 1889 roku, które jest najstarszym znanym nagraniem muzyki poważnej [17]. Nagranie to, mimo swojej ogromnej wartości historycznej, charakteryzuje się wysokim poziomem szumów i zniekształceń. Innym przykładem są wczesne nagrania bluesa, takie jak „Crazy Blues" Mamie Smith z 1920 roku, które pomimo przełomowego znaczenia dla rozwoju gatunku, cechują się ograniczonym pasmem częstotliwości i obecnością szumów tła [18].

Wyzwania związane z odtwarzaniem i konserwacją starych nagrań są złożone i wymagają interdyscyplinarnego podejścia. Odtwarzanie historycznych nośników często wymaga specjalistycznego sprzętu, który sam w sobie może być trudny do utrzymania w dobrym stanie [19]. Proces digitalizacji, choć kluczowy dla zachowania dziedzictwa audio, niesie ze sobą ryzyko wprowadzenia nowych zniekształceń lub utraty subtelnych niuansów oryginalnego nagrania [20]. Ponadto, konserwacja fizycznych nośników wymaga stworzenia odpowiednich warunków przechowywania, co może być kosztowne i logistycznie skomplikowane [21].

Dodatkowo, etyczne aspekty restauracji nagrań historycznych stanowią przedmiot debaty w środowisku muzycznym i konserwatorskim. Pytanie o to, jak daleko można posunąć się w procesie cyfrowej rekonstrukcji bez naruszenia integralności oryginalnego

dzieła, pozostaje otwarte [22]. Problematyka jakości historycznych nagrań stanowi zatem nie tylko wyzwanie techniczne, ale również kulturowe i etyczne. Rozwój zaawansowanych technik rekonstrukcji audio, w tym metod opartych na sztucznej inteligencji, otwiera nowe możliwości w zakresie przywracania i zachowania dziedzictwa dźwiękowego, jednocześnie stawiając przed badaczami i konserwatorami nowe pytania dotyczące granic ingerencji w historyczny materiał [23].

1.3) Znaczenie rekonstrukcji nagrań muzycznych

Rekonstrukcja historycznych nagrań muzycznych odgrywa kluczową rolę w zachowaniu i promowaniu dziedzictwa kulturowego, oferując szereg korzyści zarówno dla badaczy, jak i dla szerszej publiczności.

Wartość kulturowa i historyczna archiwów muzycznych jest nieoceniona. Nagrania dźwiękowe stanowią unikalne świadectwo rozwoju muzyki, technik wykonawczych i zmian w stylistyce muzycznej na przestrzeni lat [24]. Rekonstrukcja tych nagrań pozwala na zachowanie i udostępnienie szerszemu gronu odbiorców dzieł, które w przeciwnym razie mogłyby zostać zapomniane lub stać się niedostępne ze względu na degradację nośników [25]. Ponadto, zrekonstruowane nagrania umożliwiają współczesnym słuchaczom doświadczenie wykonań legendarnych artystów w jakości zbliżonej do oryginalnej, co ma ogromne znaczenie dla zrozumienia historii muzyki i ewolucji stylów wykonawczych [26].

Rola zrekonstruowanych nagrań w badaniach muzykologicznych jest fundamentalna. Wysokiej jakości rekonstrukcje pozwalają naukowcom na dokładną analizę technik wykonawczych, interpretacji i stylów muzycznych z przeszłości [27]. Umożliwiają one również badanie ewolucji praktyk wykonawczych oraz porównywanie różnych interpretacji tego samego utworu na przestrzeni lat [28]. W przypadku kompozytorów, którzy sami wykonywali swoje dzieła, zrekonstruowane nagrania stanowią bezcenne źródło informacji o intencjach twórców [29].

Wpływ jakości nagrań na percepcję i popularność utworów muzycznych jest znaczący. Badania wskazują, że słuchacze są bardziej skłonni do pozytywnego odbioru i częstszego słuchania nagrań o wyższej jakości dźwięku [30]. Rekonstrukcja historycznych nagrań może przyczynić się do zwiększenia ich dostępności i atrakcyjności dla współczesnych odbiorców, potencjalnie prowadząc do odkrycia na nowo zapomnianych artystów lub utworów [31]. Ponadto, poprawa jakości dźwięku może pomóc w lepszym zrozumieniu i docenieniu niuansów wykonania, które mogły być wcześniej niezauważalne ze względu na ograniczenia techniczne oryginalnych nagrań [32].

Ekonomiczne aspekty rekonstrukcji nagrań są również istotne. Rynek remasterów i zrekonstruowanych nagrań historycznych stanowi znaczący segment przemysłu muzycznego [33]. Wydawnictwa specjalizujące się w tego typu projektach, takie jak „Deutsche Grammophon" czy „Naxos Historical", odnoszą sukcesy komercyjne, co świadczy o istnieniu popytu na wysokiej jakości wersje klasycznych nagrań [34]. Ponadto, rekonstrukcja nagrań może prowadzić do powstania nowych źródeł przychodów dla artystów lub ich spadkobierców, a także instytucji kulturalnych posiadających prawa do historycznych nagrań [33], [35].

Warto również zauważyć, że rekonstrukcja nagrań muzycznych ma istotne znaczenie dla edukacji muzycznej. Zrekonstruowane nagrania historyczne mogą służyć jako cenne narzędzie dydaktyczne, umożliwiając studentom muzyki bezpośredni kontakt z wykonaniami wybitnych artystów z przeszłości i pomagając w zrozumieniu ewolucji stylów muzycznych.

Podsumowując, znaczenie rekonstrukcji nagrań muzycznych wykracza daleko poza samą poprawę jakości dźwięku. Jest to proces o fundamentalnym znaczeniu dla zachowania dziedzictwa kulturowego, wspierania badań naukowych, edukacji muzycznej oraz rozwoju przemysłu muzycznego. W miarę jak technologie rekonstrukcji dźwięku, w tym metody oparte na sztucznej inteligencji, stają się coraz bardziej zaawansowane, można oczekiwać, że ich rola w przywracaniu i promowaniu historycznych nagrań będzie nadal rosła, przynosząc korzyści zarówno dla świata nauki, jak i dla miłośników muzyki na całym świecie [36].

1.4) Cel i zakres pracy

Głównym celem pracy jest dogłębna analiza i ocena potencjału metod uczenia maszynowego w dziedzinie rekonstrukcji nagrań dźwiękowych. Szczególny nacisk położony zostanie na zbadanie efektywności zaawansowanych technik sztucznej inteligencji w przywracaniu jakości historycznym nagraniom muzycznym, koncentrując się na wyzwaniach, które tradycyjne metody przetwarzania sygnałów audio często pozostawiają nierozwiązane [23].

W ramach pracy zostanie położony nacisk na trzy kluczowe zagadnienia:

1. Eliminacja szumów i zakłóceń typowych dla historycznych nagrań [37].
2. Poszerzanie pasma częstotliwościowego w celu wzbogacenia brzmienia nagrań o ograniczonym paśmie [38].
3. Rekonstrukcja uszkodzonych fragmentów audio, co jest szczególnie istotne w przypadku wielu historycznych nagrań [39].

Podejście badawcze opiera się na implementacji i analizie wybranych metod uczenia maszynowego, skupiając się na architekturze **Generatywnych Sieci Przeciwstawnych** (Generative Adversarial Networks - GAN) [40]. Wybór tej architektury wynika z jej udokumentowanej skuteczności w zadaniach generatywnych i rekonstrukcyjnych w innych powiązanych dziedzinach, takich jak przetwarzanie obrazów [41].

W ramach badań zostanie podjęta próba opracowania zaawansowanego modelu GAN, który będzie wykorzystywał strukturę enkoder-dekoder z połączeniami skip dla generatora oraz architekturę konwolucyjną dla dyskryminatora. Zastosowany zostanie szereg technik normalizacji i regularyzacji, takich jak Batch Normalization, Spectral Normalization czy Gradient Penalty, celem poprawy stabilności i wydajności treningu. Kluczowym elementem będzie wykorzystanie kompleksowego zestawu funkcji strat, obejmującego Adversarial Loss, Content Loss, oraz specjalistyczne funkcje straty dla domen audio, jak Spectral Convergence Loss czy Phase-Aware Loss. Zaimplemenotwane zostaną zaawansowane techniki optymalizacji, w tym Adam Optimizer z niestandardowymi parametrami oraz dynamiczne dostosowywanie współczynnika uczenia.

Metodologia badawcza obejmuje kilka kluczowych etapów. Na początku przygotowany zostanie obszerny zestaw danych treningowych, składający się z par nagrań oryginalnych i ich zdegradowanych wersji. Następnie zaimplementowane zostaną różnorodne warianty architektury GAN, dostosowane do specyfiki przetwarzania sygnałów audio. Proces treningu będzie wykorzystywał zaawansowane techniki, takie jak augmentacja danych czy przetwarzanie na spektrogramach STFT. Ostatnim etapem będzie wszechstronna ewaluacja wyników, łącząca wiele obiektywnych metryk jakości audio.

Oczekiwane rezultaty pracy obejmują kompleksową ocenę skuteczności proponowanych metod uczenia maszynowego w zadaniach rekonstrukcji nagrań audio, w zestawieniu z tradycyjnymi technikami przetwarzania sygnałów. Przeprowadzona zostanie szczegółowa analizę wpływu różnych komponentów architektury i parametrów modeli na jakość rekonstrukcji. Istotnym elementem będzie identyfikacja mocnych stron i ograniczeń metod opartych na AI w kontekście specyficznych wyzwań związanych z restauracją historycznych nagrań muzycznych. Na podstawie uzyskanych wyników, zostaną sformułowane wnioski dotyczące potencjału sztucznej inteligencji w dziedzinie rekonstrukcji nagrań, wraz z rekomendacjami dla przyszłych badań i zastosowań praktycznych [42].

Realizacja powyższych celów ma potencjał nie tylko do znaczącego wkładu w dziedzinę przetwarzania sygnałów audio i uczenia maszynowego, ale również do praktycznego zastosowania w procesach restauracji i zachowania dziedzictwa muzycznego [25]. Wyniki pracy mogą znaleźć zastosowanie w instytucjach kulturalnych, archiwach dźwiękowych oraz w przemyśle muzycznym, przyczyniając się do lepszego zachowania i udostępnienia cennych nagrań historycznych szerokiej publiczności.

2) Zagadnienie poprawy jakości sygnałów dźwiękowych

2.1) Charakterystyka zniekształceń w nagraniach muzycznych

Zagadnienie poprawy jakości sygnałów dźwiękowych jest ściśle związane z charakterystyką zniekształceń występujących w nagraniach muzycznych. Zrozumienie natury tych zniekształceń jest kluczowe dla opracowania skutecznych metod ich redukcji lub eliminacji.

Główne typy zniekształceń występujących w nagraniach audio obejmują szereg problemów, które Szczotka [43] identyfikuje w swojej pracy, w tym szumy, brakujące dane, intermodulację i flutter. Szczególnie istotnym problemem, zwłaszcza w przypadku historycznych nagrań, jest ograniczenie pasma częstotliwościowego, co stanowi główny przedmiot badań w pracy nad BEHM-GAN [44]. Wczesne systemy rejestracji dźwięku często były w stanie uchwycić jedynie wąski zakres częstotliwości, co prowadziło do utraty wielu detali dźwiękowych, szczególnie w zakresie wysokich i niskich tonów. Ponadto, jak wskazują badania nad rekonstrukcją mocno skompresowanych plików audio [45], kompresja może wprowadzać dodatkowe zniekształcenia, które znacząco wpływają na jakość dźwięku.

Zniekształcenia nieliniowe stanowią kolejną kategorię problemów, które mogą poważnie wpłynąć na jakość nagrania. Mogą one wynikać z niedoskonałości w procesie nagrywania, odtwarzania lub konwersji sygnału. Efektem tych zniekształceń może być wprowadzenie niepożądanych harmonicznych składowych lub intermodulacji, co prowadzi do zmiany charakteru dźwięku [15].

W przypadku historycznych nagrań na nośnikach analogowych, takich jak płyty winylowe czy taśmy magnetyczne, często występują specyficzne rodzaje zniekształceń. Na przykład, efekt print-through w taśmach magnetycznych może prowadzić do pojawienia się echa lub przesłuchów między sąsiednimi warstwami taśmy [16]. Z kolei w przypadku płyt winylowych, charakterystyczne trzaski i szumy powierzchniowe są nieodłącznym elementem, który może znacząco wpływać na odbiór muzyki.

Wpływ tych zniekształceń na percepcję muzyki i jej wartość artystyczną jest znaczący. Badania pokazują, że jakość dźwięku ma istotny wpływ na to, jak słuchacze odbierają i oceniają muzykę [46]. Zniekształcenia mogą maskować subtelne niuanse wykonania, zmieniać barwę instrumentów czy głosu, a w skrajnych przypadkach całkowicie zniekształcać intencje artystyczne twórców.

W kontekście historycznych nagrań, zniekształcenia mogą stanowić barierę w pełnym docenieniu wartości artystycznej i kulturowej danego dzieła. Nawet niewielkie poprawy w jakości dźwięku mogą znacząco wpłynąć na odbiór i interpretację wykonania.

Jednocześnie warto zauważyć, że niektóre rodzaje zniekształceń, szczególnie te charakterystyczne dla określonych epok czy technologii nagrywania, mogą być postrzegane jako element autentyczności nagrania. To stawia przed procesem rekonstrukcji dźwięku wyzwanie znalezienia równowagi między poprawą jakości a zachowaniem historycznego charakteru nagrania [22].

Zrozumienie charakterystyki zniekształceń w nagraniach muzycznych jest kluczowym krokiem w opracowaniu skutecznych metod ich redukcji. W kolejnych częściach pracy skupię się na tym, jak zaawansowane techniki uczenia maszynowego, w szczególności sieci GAN, mogą być wykorzystane do adresowania tych problemów, jednocześnie starając się zachować artystyczną integralność oryginalnych nagrań.

2.1.1) Szumy i trzaski

Szumy i trzaski stanowią jeden z najbardziej powszechnych problemów w historycznych nagraniach. Źródła szumów są różnorodne i obejmują ograniczenia sprzętowe, takie jak szum termiczny w elektronice, oraz zakłócenia elektromagnetyczne pochodzące z otoczenia lub samego sprzętu nagrywającego [14]. Charakterystyka trzasków jest często związana z przyczynami mechanicznymi, takimi jak uszkodzenia powierzchni płyt winylowych, lub elektronicznymi, wynikającymi z niedoskonałości w procesie zapisu lub odtwarzania. Wpływ szumów i trzasków na jakość odsłuchu jest znaczący. Mogą one maskować subtelne detale muzyczne, zmniejszać dynamikę nagrania oraz powodować zmęczenie słuchacza. W skrajnych przypadkach, intensywne szumy lub częste trzaski mogą całkowicie zaburzyć odbiór muzyki, czyniąc nagranie trudnym lub niemożliwym do słuchania [46].

2.1.2) Ograniczenia pasma częstotliwościowego

Historyczne ograniczenia w rejestrowaniu pełnego spektrum częstotliwości są jednym z kluczowych wyzwań w rekonstrukcji nagrań. Wczesne systemy nagrywania często były w stanie zarejestrować jedynie wąski zakres częstotliwości, typowo między 250 Hz a 2500 Hz [13]. To ograniczenie miało poważne konsekwencje dla brzmienia instrumentów i wokalu, prowadząc do utraty zarówno niskich tonów, nadających muzyce głębię i ciepło, jak i wysokich częstotliwości, odpowiedzialnych za klarowność i przestrzenność dźwięku. Znaczenie szerokiego pasma dla naturalności i pełni dźwięku jest trudne do przecenienia. Współczesne badania pokazują, że ludzkie ucho jest zdolne do percepcji dźwięków w zakresie od około 20 Hz do 20 kHz, choć z wiekiem górna granica często się obniża. Pełne odtworzenie tego zakresu jest kluczowe dla realistycznego oddania brzmienia instrumentów i głosu ludzkiego. Rekonstrukcja szerokiego pasma częstotliwościowego w historycznych nagraniach stanowi zatem jedno z głównych zadań w procesie ich restauracji, co odzwierciedlają badania nad technikami takimi jak BEHM-GAN [44].

2.1.3) Zniekształcenia nieliniowe

Zniekształcenia nieliniowe stanowią szczególnie złożoną kategorię problemów w rekonstrukcji nagrań audio. Definiuje się je jako odstępstwa od idealnej, liniowej relacji między sygnałem wejściowym a wyjściowym w systemie audio. Przyczyny tych zniekształceń mogą być różnorodne, obejmując między innymi nasycenie magnetyczne w taśmach analogowych, nieliniową charakterystykę lamp elektronowych w starszym sprzęcie nagrywającym, czy też ograniczenia mechaniczne w przetwornikach [15]. Wpływ zniekształceń nieliniowych na nagrania jest znaczący i często subtelny. Prowadzą one do powstania dodatkowych harmonicznych składowych dźwięku, które nie były obecne w oryginalnym sygnale, oraz do zjawiska intermodulacji, gdzie różne częstotliwości wejściowe generują nowe, niepożądane tony. W rezultacie, brzmienie instrumentów może ulec zmianie, a czystość i przejrzystość nagrania zostaje zaburzona. W niektórych przypadkach, zwłaszcza w muzyce elektronicznej czy rockowej, pewne formy zniekształceń nieliniowych mogą być celowo wprowadzane dla uzyskania pożądanego efektu artystycznego. Korekcja zniekształceń nieliniowych stanowi jedno z największych wyzwań w procesie rekonstrukcji audio. W przeciwieństwie do zniekształceń liniowych, które można stosunkowo łatwo skorygować za pomocą filtrów, zniekształcenia nieliniowe wymagają bardziej zaawansowanych technik. Tradycyjne metody często okazują się niewystarczające, co skłania badaczy do poszukiwania rozwiązań opartych na uczeniu maszynowym, takich jak adaptacyjne modelowanie nieliniowości czy zastosowanie głębokich sieci neuronowych [45]. Trudność polega na tym, że korekta tych zniekształceń wymaga precyzyjnego odtworzenia oryginalnego sygnału, co jest szczególnie skomplikowane w przypadku historycznych nagrań, gdzie brakuje referencyjnego materiału wysokiej jakości.

2.2) Tradycyjne metody poprawy jakości nagrań

Ewolucja technik restauracji nagrań audio przeszła znaczącą transformację od prostych metod analogowych do zaawansowanych technik cyfrowych. Początkowo, restauracja nagrań opierała się głównie na fizycznej konserwacji nośników i optymalizacji sprzętu odtwarzającego. Wraz z rozwojem technologii cyfrowej, pojawiły się nowe możliwości manipulacji sygnałem audio, co znacząco rozszerzyło arsenał narzędzi dostępnych dla inżynierów dźwięku [47]. Nogales i inni w swojej pracy porównują efektywność klasycznych metod filtracji, takich jak filtr Wienera, z nowoczesnymi technikami głębokiego uczenia, ilustrując tę ewolucję.

Jednak tradycyjne metody, mimo swojej skuteczności w wielu przypadkach, mają pewne ograniczenia. Głównym problemem jest trudność w selektywnym usuwaniu szumów bez wpływu na oryginalny sygnał muzyczny. Ponadto, rekonstrukcja utraconych lub zniekształconych częstotliwości często prowadzi do artefaktów dźwiękowych, które mogą być równie niepożądane jak oryginalne zniekształcenia. Cheddad i Cheddad [48] w swoich badaniach nad aktywną rekonstrukcją utraconych sygnałów audio podkreślają te ograniczenia, proponując jednocześnie nowe podejścia uzupełniające klasyczne techniki restauracji.

2.2.1) Filtracja cyfrowa

Filtracja cyfrowa stanowi podstawę wielu technik restauracji audio. Wyróżniamy trzy podstawowe typy filtrów: dolnoprzepustowe, górnoprzepustowe i pasmowe. Dai i inni [49] w swoich badaniach nad super-rozdzielczością sygnałów muzycznych pokazują, jak tradycyjne metody filtracji mogą być rozszerzone i ulepszone dzięki zastosowaniu uczenia maszynowego.

Zastosowanie filtracji w redukcji szumów polega na identyfikacji i selektywnym tłumieniu częstotliwości, w których dominuje szum. W korekcji częstotliwościowej, filtry są używane do wzmacniania lub osłabiania określonych zakresów częstotliwości, co pozwala na poprawę balansu tonalnego nagrania.

Wady filtracji cyfrowej obejmują ryzyko wprowadzenia artefaktów dźwiękowych, zwłaszcza przy agresywnym filtrowaniu, oraz potencjalną utratę subtelnych detali muzycznych. Zaletą jest natomiast precyzja i powtarzalność procesu, a także możliwość niedestrukcyjnej edycji.

2.2.2) Remasterowanie

Remasterowanie to proces poprawy jakości istniejącego nagrania, często z wykorzystaniem nowoczesnych technologii cyfrowych. Celem remasteringu jest poprawa ogólnej jakości dźwięku, zwiększenie głośności do współczesnych standardów oraz dostosowanie brzmienia do współczesnych systemów odtwarzania.

Typowe etapy remasteringu obejmują normalizację, kompresję i korekcję EQ. Moliner i Välimäki [49] w swojej pracy nad BEHM-GAN pokazują, jak nowoczesne techniki mogą być wykorzystane do przezwyciężenia ograniczeń tradycyjnego remasteringu, szczególnie w kontekście rekonstrukcji wysokich częstotliwości w historycznych nagraniach muzycznych.

Kontrowersje wokół remasteringu często dotyczą konfliktu między zachowaniem autentyczności oryginalnego nagrania a dążeniem do poprawy jakości dźwięku. Lattner i Nistal [45] w swoich badaniach nad stochastyczną restauracją mocno skompresowanych plików audio pokazują, jak zaawansowane techniki mogą być wykorzystane do poprawy jakości nagrań bez utraty ich oryginalnego charakteru, co stanowi istotny głos w debacie o autentyczności vs. jakość dźwięku.

Mimo swoich ograniczeń, tradycyjne metody poprawy jakości nagrań wciąż odgrywają istotną rolę w procesie restauracji audio. Jednakże, rosnąca złożoność wyzwań związanych z restauracją historycznych nagrań skłania badaczy do poszukiwania bardziej zaawansowanych rozwiązań, w tym metod opartych na sztucznej inteligencji, które mogą przezwyciężyć niektóre z ograniczeń tradycyjnych technik.

2.3) Wyzwania w rekonstrukcji historycznych nagrań muzycznych

Proces rekonstrukcji historycznych nagrań muzycznych stawia przed badaczami szereg złożonych wyzwań, wymagających interdyscyplinarnego podejścia i zaawansowanych technik przetwarzania sygnałów.

Fundamentalnym problemem jest brak oryginalnych, wysokiej jakości źródeł dźwięku. Wiele historycznych nagrań przetrwało jedynie w formie znacznie zdegradowanej, często na nośnikach analogowych, które same uległy deterioracji [15]. Szczotka [43] zwraca uwagę, że niedobór niezakłóconych sygnałów referencyjnych komplikuje proces uczenia modeli rekonstrukcyjnych, zmuszając do opracowywania zaawansowanych metod symulacji degradacji dźwięku.

Identyfikacja i separacja poszczególnych instrumentów w nagraniach historycznych stanowi kolejne istotne wyzwanie. Dai i współpracownicy [49] podkreślają znaczenie tego aspektu, szczególnie w kontekście rekonstrukcji złożonych utworów orkiestrowych, gdzie ograniczenia wczesnych systemów nagrywania często prowadziły do nakładania się ścieżek instrumentalnych.

Kluczowym dylematem jest zachowanie autentyczności brzmienia przy jednoczesnej poprawie jakości. Moliner i Välimäki [44] akcentują potrzebę znalezienia równowagi między poprawą technicznej jakości dźwięku a utrzymaniem charakterystycznego, historycznego brzmienia nagrania. Zbyt agresywna ingerencja może prowadzić do utraty autentyczności i kontekstu historycznego.

Etyczne aspekty ingerencji w historyczne nagrania budzą kontrowersje w środowisku muzycznym i konserwatorskim. Lattner i Nistal [45] poruszają kwestię granic dopuszczalnej modyfikacji oryginalnego nagrania, argumentując za ostrożnym stosowaniem zaawansowanych technik rekonstrukcji.

Techniczne ograniczenia w odtwarzaniu oryginalnego brzmienia wynikają z fundamentalnych różnic między historycznymi a współczesnymi technologiami audio. Cheddad [48] zwracają uwagę na trudności w wiernym odtworzeniu charakterystyki akustycznej dawnych sal koncertowych czy specyfiki historycznych instrumentów.

Złożoność wyzwań związanych z rekonstrukcją historycznych nagrań muzycznych wymaga kompleksowego podejścia. Integracja zaawansowanych technik przetwarzania sygnałów, metod uczenia maszynowego, wiedzy muzykologicznej oraz refleksji etycznej jest kluczowa dla skutecznego rozwiązywania napotkanych problemów. Badania prowadzone przez Nogalesa i in. [47] wskazują na potrzebę ciągłego doskonalenia istniejących metod oraz opracowywania nowych rozwiązań. Przyszłość rekonstrukcji nagrań historycznych zależy od zdolności naukowców do tworzenia innowacyjnych technik, które będą w stanie sprostać unikalnym wymaganiom każdego historycznego dzieła muzycznego, zachowując jednocześnie jego autentyczność i wartość artystyczną.

3) Metody sztucznej inteligencji w poprawie jakości nagrań dźwiękowych

3.1) Przegląd technik uczenia maszynowego w przetwarzaniu dźwięku

Rozwój metod uczenia maszynowego w ostatnich latach przyniósł znaczący postęp w dziedzinie przetwarzania i analizy sygnałów dźwiękowych. Techniki te znajdują coraz szersze zastosowanie w poprawie jakości nagrań, rekonstrukcji uszkodzonych fragmentów oraz ekstrakcji informacji z sygnałów audio.

3.1.1) Ewolucja zastosowań uczenia maszynowego w dziedzinie audio

Początki wykorzystania uczenia maszynowego w przetwarzaniu dźwięku sięgają lat 90. XX wieku, kiedy to zaczęto stosować proste modele statystyczne do klasyfikacji gatunków muzycznych czy rozpoznawania mowy [50]. Wraz z rozwojem mocy obliczeniowej komputerów oraz postępem w dziedzinie sztucznych sieci neuronowych, nastąpił gwałtowny wzrost zainteresowania tymi technikami w kontekście analizy i syntezy dźwięku.

Przełomowym momentem było zastosowanie głębokich sieci neuronowych, które umożliwiły modelowanie złożonych zależności w sygnałach audio. Badania wykazały, że głębokie sieci konwolucyjne potrafią skutecznie wyodrębniać cechy charakterystyczne dźwięków, co otworzyło drogę do bardziej zaawansowanych zastosowań, takich jak separacja źródeł dźwięku czy poprawa jakości nagrań.

W ostatnich latach coraz większą popularność zyskują modele generatywne, takie jak sieci GAN (Generative Adversarial Networks) czy modele dyfuzyjne, które umożliwiają nie tylko analizę, ale także syntezę wysokiej jakości sygnałów audio [49]. Te zaawansowane techniki znajdują zastosowanie w rekonstrukcji uszkodzonych nagrań oraz rozszerzaniu pasma częstotliwości starych rejestracji dźwiękowych.

3.1.2) Klasyfikacja głównych podejść: nadzorowane, nienadzorowane, półnadzorowane

W kontekście przetwarzania sygnałów audio można wyróżnić trzy główne podejścia do uczenia maszynowego:

a) Uczenie nadzorowane: W tym podejściu model uczy się na podstawie par danych wejściowych i oczekiwanych wyników. W dziedzinie audio może to obejmować uczenie się mapowania między zaszumionymi a czystymi nagraniami w celu usuwania szumów, czy też klasyfikację instrumentów na podstawie oznaczonych próbek dźwiękowych. Przykładem zastosowania uczenia nadzorowanego jest praca Nogales A. i innych [47], w której autorzy wykorzystali konwolucyjne sieci neuronowe do rekonstrukcji uszkodzonych nagrań audio.

b) Uczenie nienadzorowane: Techniki nienadzorowane skupiają się na odkrywaniu ukrytych struktur w danych bez korzystania z etykiet. W kontekście audio może to obejmować grupowanie podobnych dźwięków czy wyodrębnianie cech charakterystycznych bez uprzedniej wiedzy o ich znaczeniu.

c) Uczenie półnadzorowane: To podejście łączy elementy uczenia nadzorowanego i nienadzorowanego, wykorzystując zarówno oznaczone, jak i nieoznaczone dane. Jest szczególnie przydatne w sytuacjach, gdy dostępna jest ograniczona ilość oznaczonych próbek, co często ma miejsce w przypadku historycznych nagrań audio.

3.1.3) Rola reprezentacji dźwięku w uczeniu maszynowym: spektrogramy, cechy MFCC, surowe próbki

Wybór odpowiedniej reprezentacji dźwięku ma kluczowe znaczenie dla skuteczności modeli uczenia maszynowego w zadaniach przetwarzania audio.

a) Spektrogramy: Przedstawiają rozkład częstotliwości sygnału w czasie, co pozwala na analizę zarówno cech czasowych, jak i częstotliwościowych. Spektrogramy są szczególnie przydatne w zadaniach takich jak separacja źródeł czy poprawa jakości nagrań. W pracy [49] autorzy wykorzystali spektrogramy logarytmiczne jako wejście do modelu GAN, osiągając dobre wyniki w zadaniu rozszerzania pasma częstotliwości nagrań muzycznych.

b) Cechy MFCC (Mel-Frequency Cepstral Coefficients): Reprezentują charakterystykę widmową dźwięku w sposób zbliżony do ludzkiego systemu słuchowego. MFCC są często stosowane w zadaniach klasyfikacji i rozpoznawania mowy. Badania wykazały, że cechy MFCC mogą być skutecznie wykorzystywane w ocenie jakości rekonstrukcji nagrań historycznych.

c) Surowe próbki: Niektóre modele, szczególnie te oparte na sieciach konwolucyjnych, mogą pracować bezpośrednio na surowych próbkach audio. Podejście to eliminuje potrzebę ręcznego projektowania cech, pozwalając modelowi na samodzielne odkrywanie istotnych wzorców w sygnale.

Wybór odpowiedniej reprezentacji zależy od specyfiki zadania oraz architektury modelu. Coraz częściej stosuje się też podejścia hybrydowe, łączące różne reprezentacje w celu uzyskania lepszych wyników.

Techniki uczenia maszynowego oferują szerokie spektrum możliwości w dziedzinie przetwarzania i poprawy jakości sygnałów audio. Ewolucja tych metod, od prostych modeli statystycznych po zaawansowane sieci generatywne, umożliwia rozwiązywanie coraz bardziej złożonych problemów związanych z rekonstrukcją i poprawą jakości nagrań dźwiękowych. W kontekście przetwarzania sygnałów audio kluczowe znaczenie ma odpowiedni dobór podejścia (nadzorowane, nienadzorowane lub półnadzorowane) oraz reprezentacji dźwięku. Właściwe decyzje w tym zakresie pozwalają na optymalne wykorzystanie potencjału uczenia maszynowego, co przekłada się na skuteczność i efektywność opracowywanych rozwiązań. Postęp w tej dziedzinie otwiera nowe możliwości w zakresie zachowania i odtwarzania dziedzictwa kulturowego, jakim są historyczne nagrania dźwiękowe.

3.2) Sieci neuronowe w zadaniach audio

Sieci neuronowe stały się fundamentalnym narzędziem w przetwarzaniu sygnałów dźwiękowych, oferując niezrównaną elastyczność i zdolność do modelowania złożonych zależności. Ich adaptacyjna natura pozwala na automatyczne wyodrębnianie istotnych cech z surowych danych audio, co czyni je niezwykle skutecznymi w szerokiej gamie zastosowań - od klasyfikacji dźwięków po zaawansowaną syntezę mowy.

Różnorodność architektur sieci neuronowych pozwala na dobór optymalnego rozwiązania do specyfiki danego zadania audio. Konwolucyjne sieci neuronowe (CNN) wykazują szczególną skuteczność w analizie lokalnych wzorców w spektrogramach, podczas gdy rekurencyjne sieci neuronowe (RNN) doskonale radzą sobie z modelowaniem długoterminowych zależności czasowych. Autoenkodery z kolei znajdują zastosowanie w kompresji i odszumianiu sygnałów, oferując możliwość redukcji wymiarowości przy zachowaniu kluczowych cech dźwięku.

Efektywność poszczególnych architektur może się znacząco różnić w zależności od konkretnego zadania. Badania empiryczne wskazują, że hybrydowe podejścia, łączące zalety różnych typów sieci, często prowadzą do najlepszych rezultatów w złożonych scenariuszach przetwarzania audio.

3.2.1) Konwolucyjne sieci neuronowe (CNN)

Konwolucyjne sieci neuronowe zrewolucjonizowały sposób, w jaki analizujemy sygnały audio. Ich unikalna architektura, inspirowana biologicznym systemem wzrokowym, okazała się niezwykle skuteczna w wyodrębnianiu hierarchicznych cech z reprezentacji czasowo-częstotliwościowych dźwięku.

W kontekście analizy audio, CNN operują najczęściej na spektrogramach, traktując je jako dwuwymiarowe „obrazy" dźwięku. Warstwy konwolucyjne działają jak filtry, wyodrębniając lokalne wzorce spektralne, które mogą odpowiadać konkretnym cechom akustycznym, takim jak akordy, formanty czy charakterystyki instrumentów.

Klasyfikacja dźwięków i rozpoznawanie mowy to obszary, w których sieci CNN wykazują szczególną skuteczność. W zadaniach identyfikacji gatunków muzycznych czy detekcji słów kluczowych, sieci te potrafią automatycznie nauczyć się rozpoznawać istotne cechy spektralne, często przewyższając tradycyjne metody oparte na ręcznie projektowanych cechach.

Adaptacje CNN do specyfiki danych dźwiękowych obejmują m.in. zastosowanie dilated convolutions. Ta technika pozwala na zwiększenie pola recepcyjnego sieci bez zwiększania liczby parametrów, co jest szczególnie przydatne w modelowaniu długoterminowych zależności czasowych w sygnałach audio. Dilated CNN znalazły zastosowanie m.in. w generowaniu dźwięku w czasie rzeczywistym.

3.2.2) Rekurencyjne sieci neuronowe (RNN)

Rekurencyjne sieci neuronowe wyróżniają się zdolnością do przetwarzania sekwencji danych, co czyni je naturalnym wyborem do analizy sygnałów audio. Ich architektura, oparta na pętlach sprzężenia zwrotnego, pozwala na uwzględnienie kontekstu czasowego w przetwarzaniu dźwięku, co jest kluczowe w wielu zadaniach, takich jak modelowanie muzyki czy rozpoznawanie mowy ciągłej.

LSTM (Long Short-Term Memory) i GRU (Gated Recurrent Unit) to popularni „następcy" klasycznych RNN, którzy rozwiązują problem zanikającego gradientu. Te zaawansowane jednostki rekurencyjne potrafią efektywnie przetwarzać długie sekwencje audio, zachowując informacje o odległych zależnościach czasowych.

W syntezie mowy, modele oparte na LSTM wykazały się zdolnością do generowania naturalnie brzmiących wypowiedzi, uwzględniających niuanse prozodyczne. W dziedzinie modelowania muzyki, sieci rekurencyjne znalazły zastosowanie w generowaniu sekwencji akordów czy komponowaniu melodii, potrafiąc uchwycić złożone struktury harmoniczne i rytmiczne.

3.2.3) Autoenkodery

Autoenkodery to fascynująca klasa sieci neuronowych, której głównym zadaniem jest nauczenie się efektywnej, skompresowanej reprezentacji danych wejściowych. W kontekście audio, ta zdolność do redukcji wymiarowości otwiera szereg możliwości - od kompresji sygnałów po zaawansowane techniki odszumiania.

Klasyczny autoenkoder składa się z enkodera, który „ściska" dane wejściowe do niższego wymiaru, oraz dekodera, który próbuje odtworzyć oryginalne dane z tej skompresowanej reprezentacji. W zastosowaniach audio, autoenkodery mogą nauczyć się reprezentacji, które zachowują kluczowe cechy dźwięku, jednocześnie eliminując szum czy niepożądane artefakty.

Wariacyjne autoenkodery (VAE) idą o krok dalej, wprowadzając element losowości do procesu kodowania. Ta cecha czyni je szczególnie przydatnymi w generowaniu nowych, unikalnych dźwięków, zachowujących charakterystykę danych treningowych. VAE znalazły zastosowanie m.in. w syntezie mowy i efektów dźwiękowych.

Splotowe autoenkodery (CAE) łączą zalety autoenkoderów i CNN, co czyni je skutecznymi w zadaniach związanych z przetwarzaniem spektrogramów. Ich zdolność do wyodrębniania lokalnych cech spektralnych przy jednoczesnej redukcji wymiarowości sprawia, że są cennym narzędziem w odszumianiu i restauracji nagrań audio.

3.3) Generatywne sieci przeciwstawne (GAN) w kontekście audio

Generatywne sieci przeciwstawne (GAN) to innowacyjna architektura uczenia maszynowego, która zrewolucjonizowała podejście do generacji i przetwarzania danych, w tym sygnałów audio. Podstawowa idea GAN opiera się na „rywalizacji" dwóch sieci neuronowych: generatora, który tworzy nowe dane, oraz dyskryminatora, który ocenia ich autentyczność. Ta koncepcja, początkowo opracowana dla obrazów, została z powodzeniem zaadaptowana do domeny audio, otwierając nowe możliwości w syntezie i manipulacji dźwiękiem.

W kontekście danych dźwiękowych, architektura GAN wymaga specyficznego podejścia. Generator często pracuje na reprezentacjach czasowo-częstotliwościowych, takich jak spektrogramy, tworząc nowe „obrazy" dźwięku. Dyskryminator z kolei analizuje te reprezentacje, ucząc się rozróżniać między autentycznymi a wygenerowanymi próbkami. Kluczowym wyzwaniem jest zapewnienie, aby wygenerowane spektrogramy były nie tylko realistyczne wizualnie, ale także przekładały się na spójne i naturalne brzmienia po konwersji z powrotem do domeny czasowej.

Zastosowania GAN w dziedzinie audio są niezwykle różnorodne. W syntezie dźwięku, sieci te potrafią generować realistyczne efekty dźwiękowe czy nawet całe utwory muzyczne, naśladując style konkretnych artystów. W zadaniach super-rozdzielczości audio, sieci GAN wykazują imponującą zdolność do rekonstrukcji wysokich częstotliwości w nagraniach o ograniczonym paśmie, co znajduje zastosowanie w restauracji historycznych nagrań. Transfer stylu audio, inspirowany podobnymi technikami w przetwarzaniu obrazów, pozwala na przenoszenie charakterystyk brzmieniowych między różnymi nagraniami, otwierając fascynujące możliwości w produkcji muzycznej.

Trening GAN dla sygnałów audio niesie ze sobą specyficzne wyzwania. Niestabilność treningu, charakterystyczna dla GAN, jest szczególnie problematyczna w domenie audio, gdzie nawet drobne artefakty mogą znacząco wpłynąć na jakość percepcyjną. Projektowanie odpowiednich funkcji straty, które uwzględniają specyfikę ludzkiego słuchu, stanowi kolejne wyzwanie. Ponadto, zapewnienie spójności fazowej w generowanych spektrogramach wymaga dodatkowych technik, takich jak wykorzystanie informacji o fazie lub bezpośrednie generowanie w domenie czasowej.

3.4) Modele dyfuzyjne w rekonstrukcji dźwięku

Modele dyfuzyjne reprezentują nowatorskie podejście do generacji danych, które w ostatnich latach zyskało ogromną popularność w dziedzinie przetwarzania dźwięku. U podstaw tej koncepcji leży idea stopniowego dodawania szumu do danych, a następnie uczenia się procesu odwrotnego - usuwania szumu, co prowadzi do generacji nowych, wysokiej jakości próbek.

Proces generacji dźwięku w modelach dyfuzyjnych można podzielić na dwa etapy. W pierwszym, zwanym procesem forward, do oryginalnego sygnału audio stopniowo dodawany jest szum gaussowski, aż do otrzymania czystego szumu. W drugim etapie, zwanym procesem reverse, model uczy się krok po kroku usuwać ten szum, rozpoczynając od losowej próbki szumu i stopniowo przekształcając ją w realistyczny sygnał audio. Ta unikalna architektura pozwala na generację dźwięku o wysokiej jakości i szczegółowości.

Zastosowania modeli dyfuzyjnych w rekonstrukcji i syntezie audio są obiecujące. W zadaniach rekonstrukcji uszkodzonych nagrań, modele te wykazują zdolność do „wypełniania" brakujących fragmentów w sposób spójny z resztą nagrania. W syntezie mowy, modele dyfuzyjne potrafią generować niezwykle naturalne i ekspresyjne wypowiedzi, uwzględniając subtelne niuanse prozodyczne.

W porównaniu z GAN, modele dyfuzyjne oferują kilka istotnych zalet w kontekście zadań audio. Przede wszystkim, ich trening jest bardziej stabilny i przewidywalny, co przekłada się na konsekwentnie wysoką jakość generowanych próbek. Modele dyfuzyjne wykazują również lepszą zdolność do modelowania różnorodności danych, unikając problemu „mode collapse" charakterystycznego dla GAN. Jednakże, kosztem tych zalet jest zazwyczaj dłuższy czas generacji, co może ograniczać ich zastosowanie w aplikacjach czasu rzeczywistego.

Aktualne osiągnięcia w dziedzinie modeli dyfuzyjnych dla dźwięku są imponujące. Modele takie jak WaveGrad czy DiffWave demonstrują wysoką jakość w syntezie mowy, często przewyższając modele autoregresyjne. W dziedzinie muzyki, modele dyfuzyjne pokazują rezultaty w generacji instrumentalnej i wokalnej, zachowując niezwykłą szczegółowość brzmienia.

Eksplorowane są techniki łączenia modeli dyfuzyjnych z innymi architekturami, takimi jak transformery, w celu lepszego modelowania długoterminowych zależności w sygnałach audio. Rosnące zainteresowanie multimodalnych modeli dyfuzyjnych otwiera możliwości syntezy audio skorelowanej z innymi modalnościami, takimi jak obraz czy tekst.

Zarówno GAN, jak i modele dyfuzyjne reprezentują przełomowe podejścia w dziedzinie generacji i rekonstrukcji dźwięku. Każda z tych technik oferuje unikalne zalety. Dalszy rozwój tych metod niewątpliwie przyczyni się do postępu w takich dziedzinach jak restauracja historycznych nagrań, synteza mowy czy produkcja muzyczna, otwierając nowe horyzonty w przetwarzaniu i generacji sygnałów audio.

4) Zastosowania metod sztucznej inteligencji w rekonstrukcji nagrań muzycznych

4.1) Ogólny przegląd praktycznych zastosowań AI w restauracji nagrań

Zastosowanie sztucznej inteligencji (AI) w rekonstrukcji nagrań muzycznych stale zyskuje na popularności. Tradycyjne techniki restauracji, jak filtry analogowe i cyfrowe, miały swoje ograniczenia, szczególnie w kontekście skomplikowanych sygnałów muzycznych. Nowoczesne metody AI, w tym głębokie uczenie i generatywne sieci przeciwstawne (GAN), oferują nowe możliwości w przywracaniu uszkodzonych i zdegradowanych nagrań muzycznych, poprawiając ich jakość w sposób, który wcześniej nie był możliwy.

Przykładowo, praca Dai et al. pokazuje, jak sieci GAN mogą być wykorzystane do poprawy rozdzielczości sygnałów muzycznych, co prowadzi do bardziej precyzyjnej i szczegółowej rekonstrukcji dźwięku [49]. Z kolei badania przedstawione przez Nogales et al. wykorzystują głębokie autoenkodery do przywracania jakości nagrań, przewyższając tradycyjne metody, takie jak filtracja Wienera [47].

4.2) Porównanie skuteczności metod AI z tradycyjnymi technikami

Tradycyjne metody rekonstrukcji nagrań muzycznych, takie jak filtry Wienera czy metody interpolacji oparte na DSP, są powszechnie stosowane, ale ich skuteczność jest ograniczona. Wprowadzenie technik AI, w szczególności głębokich sieci neuronowych, znacząco poprawiło jakość odtwarzania i rekonstrukcji nagrań.

Przykładem jest zastosowanie GAN do poprawy jakości mocno skompresowanych plików MP3. Artykuł z MDPI pokazuje, jak stochastyczne generatory oparte na GAN są w stanie wytworzyć próbki bliższe oryginałowi niż tradycyjne metody DSP, szczególnie w przypadku dźwięków perkusyjnych i wysokich częstotliwości [45]. Dodatkowo, metody takie jak nienegatywna faktoryzacja macierzy (NMF) i głębokie sieci neuronowe (DNN) zostały zastosowane do odrestaurowania historycznych nagrań fortepianowych, jak pokazuje praca na temat rekonstrukcji nagrania Johannesa Brahmsa z 1889 roku [50].

4.3) Wpływ postępu w dziedzinie AI na możliwości rekonstrukcji nagrań

Postęp w dziedzinie AI, a zwłaszcza rozwój modeli dyfuzyjnych i GAN, otworzył nowe możliwości w rekonstrukcji nagrań muzycznych. Modele te pozwalają na generowanie dźwięku o wysokiej jakości, nawet z uszkodzonych i silnie skompresowanych źródeł. Artykuł na temat modeli dyfuzyjnych dla restauracji dźwięku przedstawia kompleksowe omówienie tego tematu, podkreślając ich zdolność do generowania naturalnie brzmiących próbek dźwiękowych [11].

4.4) Usuwanie szumów i zakłóceń

Usuwanie szumów i zakłóceń z nagrań muzycznych stanowi kluczowe wyzwanie w procesie rekonstrukcji dźwięku, szczególnie w kontekście metod opartych na sztucznej inteligencji. Nagrania muzyczne mogą być narażone na różnorodne typy szumów, takie jak szumy tła, impulsowe zakłócenia oraz artefakty powstałe podczas konwersji analogowo-cyfrowej. W celu ich skutecznego usunięcia, konieczne jest zrozumienie charakterystyki każdego z tych szumów, a także ich wpływu na jakość odbioru dźwięku przez słuchacza.

W ostatnich latach, metody oparte na sztucznej inteligencji, w tym sieci neuronowe, zyskały na popularności w kontekście identyfikacji i separacji szumów od sygnału muzycznego. Zastosowanie autoenkoderów oraz sieci GAN okazało się szczególnie efektywne w odszumianiu nagrań, co potwierdzają liczne badania [51]. Autoenkodery, ze względu na swoją zdolność do kompresji danych i ich rekonstrukcji, umożliwiają wyodrębnienie istotnych cech sygnału, a jednocześnie eliminację niepożądanych szumów. Z kolei sieci GAN, które składają się z generatora i dyskryminatora, pozwalają na generowanie bardziej realistycznych rekonstrukcji sygnału dźwiękowego, dzięki czemu możliwe jest zachowanie większej ilości detali muzycznych podczas usuwania szumów [51].

Porównanie efektywności różnych architektur sieci neuronowych w zadaniu usuwania szumów wykazało, że tradycyjne metody oparte na filtracji spektralnej ustępują nowoczesnym podejściom opartym na głębokim uczeniu się. Przykładem może być zastosowanie bloków rezydualnych oraz technik normalizacji w architekturze sieci, co prowadzi do znaczącej poprawy jakości odszumionego dźwięku [51].

Niemniej jednak, wyzwania związane z zachowaniem detali muzycznych podczas usuwania szumów pozostają istotnym problemem. Głębokie sieci uczące się często mają tendencję do usuwania nie tylko szumów, ale również subtelnych niuansów muzycznych, co może prowadzić do utraty pierwotnego charakteru nagrania. Aby zminimalizować ten efekt, stosowane są zaawansowane funkcje strat, takie jak Perceptual Loss czy Signal-to-Noise Ratio Loss, które pomagają w zachowaniu jak największej ilości oryginalnych detali dźwiękowych [52].

4.5) Rozszerzanie pasma częstotliwościowego

Rozszerzanie pasma częstotliwościowego w historycznych nagraniach stanowi istotne wyzwanie technologiczne i badawcze, mające na celu poprawę jakości dźwięku przy zachowaniu integralności oryginalnego materiału.

4.5.1) Problematyka ograniczonego pasma w historycznych nagraniach

Historyczne nagrania, z uwagi na ograniczenia technologiczne ówczesnych systemów rejestracji dźwięku, często charakteryzują się ograniczonym pasmem przenoszenia, co prowadzi do utraty wyższych częstotliwości i w rezultacie zubożenia jakości dźwięku. Tego typu nagrania są zwykle poddawane cyfryzacji, a następnie obróbce mającej na celu odzyskanie jak największej ilości utraconej informacji. Rozszerzanie pasma częstotliwościowego staje się tutaj kluczowym narzędziem, które umożliwia przywrócenie pełniejszego brzmienia nagrania, a co za tym idzie, zbliżenie się do oryginalnego zamysłu artystycznego twórcy.

4.5.2) Techniki AI do estymacji i syntezy brakujących wysokich częstotliwości

Zastosowanie sztucznej inteligencji, w szczególności technik uczenia maszynowego, przyniosło nowe możliwości w zakresie rekonstrukcji brakujących informacji w historycznych nagraniach. Przykładem tego jest metoda Blind Audio Bandwidth Extension (BABE), która wykorzystuje model dyfuzyjny do estymacji brakujących wysokich częstotliwości w nagraniach o ograniczonym paśmie przenoszenia. Model ten, działający w tzw. trybie zero-shot, pozwala na realistyczne odtworzenie utraconych części spektrum częstotliwości bez konieczności znajomości szczegółów degradacji sygnału [53]. Testy subiektywne potwierdziły, że zastosowanie BABE znacząco poprawia jakość dźwięku w nagraniach historycznych [53].

4.5.3) Zastosowanie sieci GAN w super-rozdzielczości spektralnej

Sieci generatywne (GAN) znalazły szerokie zastosowanie w przetwarzaniu dźwięku, w tym w rozszerzaniu pasma częstotliwościowego. Metoda BEHM-GAN wykorzystuje sieci GAN do rozszerzania pasma częstotliwościowego w nagraniach muzycznych z początku XX wieku. Zastosowanie GAN pozwala na realistyczną syntezę brakujących wysokich częstotliwości, co przekłada się na znaczną poprawę percepcyjnej jakości dźwięku [44].

4.5.4) Metody oceny jakości rozszerzonego pasma częstotliwościowego

Ocena jakości dźwięku po zastosowaniu technik rozszerzania pasma częstotliwościowego jest kluczowym etapem procesu. W przypadku historycznych nagrań ocena ta jest szczególnie istotna, ponieważ dodanie nowych informacji może wpłynąć na oryginalny charakter nagrania. W związku z tym stosuje się zarówno metody obiektywne, jak i subiektywne. Przykładem są testy preferencyjne, w których słuchacze oceniają jakość dźwięku pod kątem jego spójności i naturalności [53].

4.5.5) Etyczne aspekty dodawania nowych informacji do historycznych nagrań

Dodawanie nowych informacji do historycznych nagrań rodzi szereg pytań etycznych. Główna kwestia dotyczy tego, na ile możemy modyfikować oryginalny materiał, by nie zatracić jego autentyczności. Rozszerzanie pasma częstotliwościowego za pomocą AI i GAN musi być prowadzone z poszanowaniem dla oryginalnego dzieła, aby zachować jego integralność i nie wprowadzać zmian, które mogłyby zostać odebrane jako manipulacje oryginałem [54], [55].

4.6) Uzupełnianie brakujących fragmentów

4.6.1) Przyczyny i charakterystyka ubytków

Braki w nagraniach muzycznych mogą mieć różnorodne przyczyny, takie jak uszkodzenia fizyczne nośników, błędy w digitalizacji, czy celowe wycięcia fragmentów w procesie edycji. Charakterystyka tych ubytków jest równie zróżnicowana — od krótkich, niemal niezauważalnych przerw, po dłuższe fragmenty, które znacząco wpływają na integralność utworu muzycznego. W związku z tym, rekonstrukcja brakujących fragmentów stała się kluczowym zadaniem w konserwacji i restauracji nagrań dźwiękowych.

4.6.2) Metody AI do interpolacji brakujących fragmentów

W ostatnich latach znaczący postęp dokonał się w dziedzinie sztucznej inteligencji, szczególnie w kontekście interpolacji brakujących danych audio. Metody te wykorzystują zaawansowane modele uczenia maszynowego, które są zdolne do odtwarzania brakujących próbek dźwiękowych w sposób, który jest trudny do odróżnienia od oryginału. Na przykład, techniki oparte na modelach autoregresyjnych, takich jak Rekurencyjne Sieci Neuronowe (RNN) i Long Short-Term Memory (LSTM), umożliwiają przewidywanie brakujących próbek na podstawie istniejącego kontekstu dźwiękowego, co prowadzi do bardziej naturalnej rekonstrukcji [56].

4.6.3) Wykorzystanie kontekstu muzycznego w rekonstrukcji ubytków

Modele te mogą efektywnie wykorzystywać kontekst muzyczny, analizując struktury melodyczne, rytmiczne i harmoniczne, co pozwala na precyzyjne wypełnienie braków w sposób, który zachowuje spójność i naturalność nagrania. Ważnym aspektem jest tutaj także ocena spójności muzycznej rekonstruowanych fragmentów, która może być przeprowadzona zarówno subiektywnie, poprzez testy odsłuchowe, jak i obiektywnie, z wykorzystaniem narzędzi analitycznych [43].

4.7) Poprawa jakości mocno skompresowanych plików audio

Kompresja stratna, taka jak MP3, AAC, czy OGG, jest powszechnie stosowana w celu redukcji rozmiaru plików audio. Jednak proces ten nieodłącznie wiąże się z utratą pewnych informacji, co wpływa na jakość dźwięku. W szczególności mogą pojawić się artefakty, takie jak brakujące detale w wyższych częstotliwościach czy zniekształcenia perkusji, które negatywnie wpływają na odbiór muzyczny [49].

Aby przeciwdziałać tym problemom, rozwijane są techniki oparte na sztucznej inteligencji (AI). Jednym z obiecujących podejść jest zastosowanie głębokich sieci neuronowych, które mogą identyfikować i redukować artefakty kompresji. Przykładowo, modele oparte na architekturze U-Net czy Wave-U-Net są w stanie skutecznie poprawiać jakość dźwięku, szczególnie w przypadku nagrań mocno skompresowanych [47].

Zastosowanie Generatywnych Sieci Przeciwstawnych (GAN) otwiera nowe możliwości w odtwarzaniu detali utraconych podczas kompresji. GAN-y potrafią generować brakujące fragmenty sygnału audio w sposób realistyczny, co pozwala na znaczną poprawę jakości muzyki. Badania wykazują, że sieci GAN są szczególnie skuteczne w zwiększaniu rozdzielczości częstotliwościowej nagrań oraz w poprawie jakości dźwięku w skompresowanych plikach MP3 [45].

Istotną częścią tych procesów jest odpowiednie szkolenie modeli AI. Trening odbywa się na parach nagrań przed i po kompresji, co umożliwia modelom nauczenie się odtwarzania utraconych detali. Wyzwanie stanowi jednak generalizacja tych modeli na różne formaty kompresji, gdyż algorytmy mogą wykazywać różną skuteczność w zależności od typu kompresji. Dalsze badania są konieczne, aby zapewnić efektywne działanie tych technologii w szerokim spektrum formatów [57], [58].

5) Charakterystyka wybranej metody - sieci GAN w rekonstrukcji nagrań muzycznych

Generatywne Sieci Przeciwstawne (GAN) to potężne narzędzie w przetwarzaniu i rekonstrukcji sygnałów dźwiękowych. Składają się z generatora, który tworzy nowe próbki, oraz dyskryminatora, który ocenia ich jakość. W kontekście audio, sieci GAN umożliwiają przywracanie zniekształconych sygnałów poprzez identyfikację i korekcję utraconych detali, wykorzystując techniki takie jak szybka transformacja Fouriera (STFT) [59]. Wybór GAN jako głównej metody do analizy w tej pracy wynika z jej ponad przeciętnej zdolności do odtwarzania cech sygnału, które zostały utracone podczas kompresji oraz z powodu innych zniekształceń [45].

5.1) Architektura sieci GAN dla zadań audio

W kontekście zadań przetwarzania audio, architektura Generatywnych Sieci Przeciwstawnych (GAN) składa się z dwóch głównych komponentów: generatora i dyskryminatora. Generator jest odpowiedzialny za tworzenie nowych próbek danych, które mają naśladować rzeczywiste dane, podczas gdy dyskryminator ocenia te próbki, starając się odróżnić generowane dane od prawdziwych. Proces ten polega na iteracyjnym szkoleniu obu komponentów, gdzie generator uczy się tworzyć coraz bardziej realistyczne dane, a dyskryminator staje się coraz bardziej wyrafinowany w wykrywaniu sztucznie wygenerowanych danych [47].

Adaptacje architektury GAN do specyfiki danych audio często obejmują zastosowanie konwolucji jednokierunkowych (1D), które są bardziej odpowiednie do przetwarzania sygnałów dźwiękowych niż tradycyjne konwolucje dwuwymiarowe. W modelach audio GAN konwolucje 1D pozwalają na skuteczniejsze modelowanie ciągłości czasowej sygnału dźwiękowego, co jest kluczowe dla zachowania naturalnego brzmienia [60].

Znaczącą rolę w architekturze GAN dla zadań audio odgrywają spektrogramy oraz reprezentacje czasowo-częstotliwościowe. Spektrogramy, które reprezentują sygnał audio w domenie czasowo-częstotliwościowej, są często używane jako wejście do generatora lub dyskryminatora. Tego typu reprezentacje pozwalają modelom GAN na lepsze wychwycenie charakterystycznych wzorców akustycznych, co przekłada się na wyższą jakość generowanych sygnałów dźwiękowych [61].

Przykłady konkretnych implementacji GAN dla rekonstrukcji nagrań muzycznych obejmują takie podejścia jak Dual-Step-U-Net, które łączą konwencjonalne techniki głębokiego uczenia z GAN. W tego typu modelach, szczególnie ważne jest eliminowanie zakłóceń obecnych w nagraniach analogowych, co osiąga się poprzez głębokie uczenie na rzeczywistych, zaszumionych danych audio [62].

5.2) Proces uczenia sieci GAN

Generatywne Sieci Przeciwstawne (GAN) opierają się na zasadzie przeciwstawnego uczenia, gdzie dwie sieci neuronowe — generator i dyskryminator — rywalizują ze sobą. Generator stara się tworzyć próbki danych, które mają naśladować rzeczywiste próbki, podczas gdy dyskryminator ocenia, czy próbka pochodzi od generatora, czy jest oryginalnym danymi. Proces ten prowadzi do stopniowego udoskonalania obu modeli: generator staje się coraz lepszy w tworzeniu realistycznych danych, a dyskryminator w ich rozróżnianiu [50].

Trening GAN na danych audio niesie ze sobą specyficzne wyzwania. Ze względu na różnorodność sygnałów dźwiękowych oraz ich reprezentacji, takie jak spektrogramy czy bezpośrednie fale dźwiękowe, trening wymaga precyzyjnego dopasowania architektury sieci oraz funkcji straty. Jednym z problemów jest zachowanie ciągłości sygnału oraz uniknięcie problemu zaniku gradientu, który może prowadzić do niestabilności procesu uczenia [47]. Aby stabilizować proces uczenia, w kontekście przetwarzania dźwięku często stosuje się techniki takie jak normalizacja spektralna. Pozwala ona na lepsze zarządzanie wartościami wag sieci i zapobiega eksplozji gradientów, co jest kluczowe dla zachowania stabilności modelu w dłuższej perspektywie [61].

Strategie doboru hiperparametrów, takie jak rozmiar wsadu (batch size), stopa uczenia się, czy wybór optymalizatora, są kluczowe dla efektywnego treningu sieci GAN. W kontekście przetwarzania dźwięku istotne jest również zarządzanie procesem treningowym poprzez monitorowanie postępów modelu i dynamiczne dostosowywanie parametrów, aby unikać problemów takich jak nadmierne dopasowanie (overfitting) [43].

5.3) Funkcje straty i metryki oceny jakości

W kontekście GAN stosowanych do rekonstrukcji audio, funkcje straty odgrywają kluczową rolę w kształtowaniu jakości generowanych danych. Najczęściej stosowane są adversarial loss, która odpowiada za przeciwstawne uczenie generatora i dyskryminatora, oraz reconstruction loss, która pomaga w precyzyjnym odtwarzaniu szczegółów sygnału dźwiękowego [49].

Do obiektywnej oceny jakości rekonstrukcji audio stosuje się metryki takie jak PESQ (Perceptual Evaluation of Speech Quality) i STOI (Short-Time Objective Intelligibility). Metryki te pozwalają na ilościowe porównanie jakości sygnałów oryginalnych i odtworzonych, co jest kluczowe dla oceny efektywności modeli GAN w zadaniach rekonstrukcji dźwięku [63]. Oprócz obiektywnych metryk, subiektywne metody ewaluacji, takie jak testy odsłuchowe przeprowadzone przez ekspertów, są często używane do oceny jakości generowanych próbek dźwiękowych. Metody te pozwalają na uwzględnienie percepcji ludzkiego słuchu, co jest niezbędne przy ocenie jakości nagrań muzycznych [58].

5.4) Modyfikacje i rozszerzenia standardowej architektury GAN

W standardowej architekturze GAN, mimo jej licznych zalet, istnieje szereg ograniczeń, które mogą wpływać na skuteczność i stabilność modeli, zwłaszcza w kontekście zadań związanych z przetwarzaniem dźwięku. Motywacja do wprowadzania modyfikacji w standardowej architekturze GAN wynika z potrzeby przezwyciężenia tych ograniczeń, takich jak problem zaniku gradientów, wolna konwergencja, czy trudności z generowaniem wysokiej jakości próbek w złożonych przestrzeniach danych.

W ciągu ostatnich lat powstało wiele wariantów GAN, które zostały dostosowane do specyficznych wymagań zadań związanych z przetwarzaniem dźwięku. Do najważniejszych wariantów stosowanych w audio należą m.in. Conditional GAN, który pozwala na kontrolowane generowanie danych na podstawie dodatkowych informacji, oraz WaveGAN, który jest szczególnie przydatny do generowania surowych sygnałów audio. Warianty te wprowadzają unikalne modyfikacje zwiększające ich efektywność w określonych zastosowaniach [64].

5.4.1) Conditional GAN

Conditional GAN (cGAN) wprowadza koncepcję warunkowego generowania, gdzie proces generowania danych jest kontrolowany przez dodatkowe informacje, takie jak etykiety klas czy inne cechy danych. W ten sposób możliwe jest uzyskanie bardziej precyzyjnych wyników, dostosowanych do specyficznych wymagań zadania. W kontekście rekonstrukcji nagrań audio, cGAN może być wykorzystany do kontrolowania parametrów rekonstrukcji, co pozwala na lepsze odwzorowanie oryginalnego sygnału lub dostosowanie go do określonych warunków [65].

Zastosowania Conditional GAN w rekonstrukcji nagrań obejmują m.in. rozszerzenie pasma częstotliwości dźwięku, co jest szczególnie istotne w przypadku nagrań o ograniczonej jakości. Modele cGAN są wykorzystywane do augmentacji danych audio, co pozwala na zwiększenie ilości danych treningowych i poprawę jakości wyników w zadaniach takich jak diagnostyka oparta na dźwięku [66].

5.4.2) CycleGAN

CycleGAN jest modelem generatywnym, który umożliwia uczenie się transformacji między różnymi domenami danych bez potrzeby posiadania sparowanych próbek treningowych. W przeciwieństwie do tradycyjnych metod nadzorowanych, CycleGAN wykorzystuje mechanizm dwóch cykli generacyjnych (cykl do przodu i cykl do tyłu), co pozwala na uczenie bez nadzoru. Główna idea tego podejścia polega na tym, że model uczy się odwzorowywać dane z jednej domeny na drugą w taki sposób, aby możliwe było odzyskanie oryginalnych danych przy odwrotnym procesie transformacji.

CycleGAN znalazł szerokie zastosowanie w transferze stylu audio i konwersji głosu. Dzięki zdolności do uczenia się z danych nieparowanych, model ten jest wykorzystywany do zmiany stylu muzycznego utworów czy konwersji głosu pomiędzy różnymi mówcami. CycleGAN został użyty w celu transformacji dźwięków silników okrętowych [67].

CycleGAN oferuje ogromny potencjał w rekonstrukcji nagrań bez par treningowych. Dzięki swojej zdolności do pracy z nieparowanymi danymi, model ten może być używany do odtwarzania sygnałów dźwiękowych w przypadkach, gdy brak jest sparowanych próbek treningowych, co czyni go wyjątkowo użytecznym w rekonstrukcji historycznych nagrań lub innych złożonych zadań dźwiękowych [67].

5.4.3) Progressive GAN

Progressive GAN (ProGAN) wprowadza innowacyjną koncepcję stopniowego zwiększania rozdzielczości generowanych danych. Zamiast trenować model na danych o docelowej rozdzielczości od samego początku, ProGAN zaczyna od niskiej rozdzielczości i stopniowo dodaje nowe warstwy, które zwiększają poziom szczegółowości generowanych danych. Takie podejście pozwala na uzyskanie bardziej stabilnych i realistycznych wyników, co jest szczególnie korzystne dla zastosowań związanych z generowaniem złożonych danych, takich jak dźwięki [68].

Adaptacje Progressive GAN do domeny audio opierają się właśnie na koncepcji stopniowego zwiększania rozdzielczości, ale zastosowanej do sygnałów dźwiękowych. Dzięki temu podejściu możliwe jest generowanie próbek audio o coraz wyższej jakości, co jest istotne w kontekście syntezy dźwięku, gdzie precyzja i jakość są kluczowe [69].

ProGAN jest wykorzystywany w wielu zadaniach związanych z generowaniem wysokiej jakości próbek dźwiękowych. Przykładem może być zastosowanie tego modelu do syntezy mowy, gdzie stopniowe zwiększanie jakości generowanych sygnałów pozwala na uzyskanie bardziej naturalnych i realistycznych próbek dźwiękowych [69].

6) Implementacja i eksperymenty

6.1) Opis zestawu danych

W ramach przeprowadzonych badań nad zastosowaniem sieci GAN w rekonstrukcji nagrań dźwiękowych, wykorzystano zestaw danych **MusicNet Dataset** [70], składający się z wysokiej jakości nagrań muzyki klasycznej. Wybór tego gatunku muzycznego podyktowany był jego złożonością harmoniczną i dynamiczną, co stanowi istotne wyzwanie dla algorytmów rekonstrukcyjnych. Brak wokalu w nagraniach muzyki klasycznej pozwolił na skupienie się na czystej rekonstrukcji sygnałów muzycznych.

Proces przygotowania danych składał się z kilku kluczowych etapów, zaimplementowanych w skryptach `to_mp3.py`, `to_vinyl_crackle.py`, `generate_stfts.py` oraz `data_preparation.py`. Początkowo, pliki źródłowe w formacie WAV zostały przekonwertowane na format MP3 z przepływnością 320 kbit/s, co pozwoliło na zachowanie wysokiej jakości dźwięku przy jednoczesnej redukcji rozmiaru plików. Następnie, długie nagrania zostały podzielone na dokładnie 10-sekundowe fragmenty, co zapewniło jednolitą strukturę wejściową dla sieci neuronowej.

```
def process_file(file_info):
    input_path, output_dir = file_info
    filename = os.path.basename(input_path)
    file_id = os.path.splitext(filename)[0]

    audio = AudioSegment.from_wav(input_path)

    segment_length = 10 * 1000  # 10 seconds in milliseconds
    num_full_segments = len(audio)

    segments = []
    for i in range(num_full_segments):
        start = i * segment_length
        segment = audio[start:start + segment_length]
        output_filename = f"{file_id}-{i}.mp3"
        output_path = os.path.join(output_dir, output_filename)
        segment.export(output_path, format="mp3", bitrate="320k")
        segments.append(output_filename)

    # Check if there's a remaining segment and if it's exactly 10 seconds
    remaining_audio = audio[num_full_segments * segment_length:]
    if len(remaining_audio) == segment_length:
        output_filename = f"{file_id}-{num_full_segments}.mp3"
        output_path = os.path.join(output_dir, output_filename)
        remaining_audio.export(output_path, format="mp3", bitrate="320k")
        segments.append(output_filename)

    return filename, segments
```

Program 1: Procedura przygotowania danych

Istotnym elementem przygotowania danych była augmentacja, mająca na celu symulację efektów charakterystycznych dla historycznych nagrań. W tym celu opracowano algorytm generujący szum o charakterystyce spektralnej zbliżonej do autentycznych płyt winylowych. Proces ten, zaimplementowany w skrypcie `to_vinyl_crackle.py`, obejmował generowanie różnorodnych efektów, takich jak trzaski, pęknięcia i zadrapania, z wykorzystaniem technik przetwarzania sygnałów, w tym filtracji pasmowo-przepustowej.

```
def generate_vinyl_crackle(duration_ms, sample_rate):
    num_samples = int(duration_ms * sample_rate / 1000)
    samples = np.zeros(num_samples)

    event_density = 0.0001
    event_positions = np.random.randint(0, num_samples, int(num_samples *
event_density))

    for pos in event_positions:
        event_type = np.random.choice(['pop', 'crackle', 'scratch'])

        if event_type == 'pop':
            duration = np.random.randint(5, 15)
            event = np.random.exponential(0.01, duration)
            event = event * np.hanning(duration)
        elif event_type == 'crackle':
            duration = np.random.randint(20, 50)
            event = np.random.normal(0, 0.01, duration)
            event = event * (np.random.random(duration) > 0.7)
        else:  # scratch
            duration = np.random.randint(50, 200)
            event = np.random.normal(0, 0.05, duration)
            event = event * np.hanning(duration)

        end_pos = min(pos + len(event), num_samples)
        samples[pos:end_pos] += event[:end_pos - pos]

    nyquist = sample_rate / 2
    low = 500 / nyquist
    high = 7000 / nyquist
    b, a = signal.butter(3, [low, high], btype='band')
    samples = signal.lfilter(b, a, samples)

    samples = samples / np.max(np.abs(samples))

    return samples
```

Program 2: Procedura generowania trzasków winylowych

Kolejnym etapem było przekształcenie sygnałów audio do reprezentacji czasowo-częstotliwościowej przy użyciu **krótkoczasowej transformaty Fouriera (STFT)**. Skrypt `generate_stfts.py` realizował to zadanie, stosując funkcję `librosa.stft()` z oknem analizy o długości 2048 próbek i przeskokiem 512 próbek. Dodatkowo, zastosowano technikę skalowania `signed square root`, która pozwoliła na redukcję dynamiki sygnału przy jednoczesnym zachowaniu informacji o fazie.

```
def process_audio_file(file_path, output_dir, window_size=2048, hop_size=512):
    try:
        base_name = os.path.splitext(os.path.basename(file_path))[0]
        output_file = os.path.join(output_dir, f"{base_name}_stft.npz")

        if os.path.exists(output_file):
            return False, (0, 0)

        audio, sr = librosa.load(file_path, sr=None)
        stft = librosa.stft(audio, n_fft=window_size, hop_length=hop_size)
        stft_scaled = signed_sqrt(stft.real) + 1j * signed_sqrt(stft.imag)

                    np.savez_compressed(output_file,  stft=stft_scaled,  sr=sr,
window_size=window_size, hop_size=hop_size)

        return True, stft_scaled.shape
    except Exception as e:
        logging.error(f"Error processing {file_path}: {str(e)}")
        return False, (0, 0)
```

Program 3: Procedura tworzenia krótkoczasowych trasnformat Fouriera

W procesie normalizacji i skalowania danych, zaimplementowanym w klasie `STFTDataset`, zastosowano normalizację amplitudy do zakresu $[-1, 1]$. Proces ten obejmował oddzielne przetwarzanie magnitudy i fazy spektrogramów, co umożliwiło zachowanie pełnej informacji o strukturze czasowo-częstotliwościowej sygnału.

```
class STFTDataset(Dataset):
    def __init__(self, clean_files, noisy_files):
        self.clean_files = clean_files
        self.noisy_files = noisy_files

    def __getitem__(self, idx):
        clean_file = self.clean_files[idx]
        noisy_file = self.noisy_files[idx]

        clean_stft = np.load(clean_file)['stft']
        noisy_stft = np.load(noisy_file)['stft']

        clean_mag, clean_phase = np.abs(clean_stft), np.angle(clean_stft)
        noisy_mag, noisy_phase = np.abs(noisy_stft), np.angle(noisy_stft)

        clean_mag_original = clean_mag.copy()
        noisy_mag_original = noisy_mag.copy()

        clean_mag_norm = (clean_mag - np.min(clean_mag)) / (np.max(clean_mag) -
np.min(clean_mag)) * 2 - 1
        noisy_mag_norm = (noisy_mag - np.min(noisy_mag)) / (np.max(noisy_mag) -
np.min(noisy_mag)) * 2 - 1

        clean_data_norm = np.stack([clean_mag_norm, clean_phase], axis=0)
        noisy_data_norm = np.stack([noisy_mag_norm, noisy_phase], axis=0)

        clean_data_original = np.stack([clean_mag_original, clean_phase], axis=0)
        noisy_data_original = np.stack([noisy_mag_original, noisy_phase], axis=0)

        return (torch.from_numpy(noisy_data_norm).float(),
                torch.from_numpy(clean_data_norm).float(),
                torch.from_numpy(noisy_data_original).float(),
                torch.from_numpy(clean_data_original).float())
```

Program 4: Klasa przechowująca zbiory danych STFT

Finalny zestaw danych, przygotowany przez funkcję `prepare_data()` w skrypcie `data_preparation.py`, składał się z par nagrań: oryginalnych, wysokiej jakości fragmentów oraz ich odpowiedników z symulowanymi zniekształceniami winylowymi. Dane zostały podzielone na zbiory treningowy i walidacyjny z wykorzystaniem funkcji `train_test_split()` z biblioteki scikit-learn, zapewniając reprezentatywność obu zbiorów.

Warto podkreślić, że cały proces przygotowania danych został zoptymalizowany pod kątem wydajności obliczeniowej. Wykorzystano przetwarzanie równoległe z użyciem `ProcessPoolExecutor`, co znacząco przyspieszyło obróbkę dużej ilości plików audio. Dodatkowo, zastosowano techniki zarządzania pamięcią, przetwarzając dane w mniejszych porcjach, co umożliwiło jak najefektywniejsze wykorzystanie dostępnych zasobów sprzętowych.

6.2) Architektura proponowanego modelu

W ramach badań nad rekonstrukcją nagrań dźwiękowych opracowano architekturę Generatywnej Sieci Przeciwstawnej (GAN), składającą się z generatora i dyskryminatora. Model ten został zaprojektowany z myślą o efektywnym przetwarzaniu spektrogramów STFT, które stanowią reprezentację danych wejściowych.

6.2.1) Generator

Generator, będący kluczowym elementem architektury, wykorzystuje **zmodyfikowaną strukturę U-Net**. Wybór tej architektury podyktowany był jej skutecznością w zadaniach przetwarzania obrazów, które można zaadaptować do analizy spektrogramów. Wprowadzono jednak szereg modyfikacji dostosowujących model do specyfiki rekonstrukcji nagrań audio:

W przeciwieństwie do klasycznego U-Net, zastosowano normalizację spektralną w warstwach konwolucyjnych, co pomaga w stabilizacji treningu GAN i poprawia jakość generowanych wyników.

```
    def encoder_block(self, in_channels, out_channels):
        return nn.Sequential(
            spectral_norm(nn.Conv2d(in_channels, out_channels, 3, stride=2,
    padding=1)),
            nn.BatchNorm2d(out_channels),
            nn.LeakyReLU(0.2)
        )
```

Program 5: Blok składowy enkodera architektury Generatora

Centralną część generatora stanowi bottleneck składający się z trzech bloków rezydualnych. Ta modyfikacja pozwala na lepsze przetwarzanie cech na wysokim poziomie abstrakcji.

```
class ResidualBlock(nn.Module):
    def __init__(self, channels):
        super().__init__()
        self.conv1 = spectral_norm(nn.Conv2d(channels, channels, 3, padding=1))
        self.conv2 = spectral_norm(nn.Conv2d(channels, channels, 3, padding=1))
        self.norm1 = nn.BatchNorm2d(channels)
        self.norm2 = nn.BatchNorm2d(channels)
        self.relu = nn.LeakyReLU(0.2)

    def forward(self, x):
        residual = x
        x = self.relu(self.norm1(self.conv1(x)))
        x = self.norm2(self.conv2(x))
        return x + residual
```

Program 6: Rezydualny blok składowy architektury Generatora

Zamiast standardowej funkcji aktywacji ReLU, zastosowano LeakyReLU, co pomaga w uniknięciu problemu „umierających neuronów” i poprawia gradient przepływ w sieci.

Dekoder został dostosowany do specyfiki danych audio poprzez zastosowanie warstw dekonwolucyjnych z normalizacją spektralną:

```
def decoder_block(self, in_channels, out_channels):
    return nn.Sequential(
       spectral_norm(nn.ConvTranspose2d(in_channels, out_channels, 4, stride=2,
padding=1)),
        nn.BatchNorm2d(out_channels),
        nn.LeakyReLU(0.2)
    )
```

Program 7: Blok składowy dekodera architektury Generatora

Podobnie jak w klasycznym U-Net, zastosowano połączenia skip między odpowiadającymi sobie warstwami enkodera i dekodera. Jest to kluczowe dla zachowania drobnych detali w rekonstruowanych nagraniach.

Te modyfikacje pozwoliły na stworzenie architektury, która łączy zalety U-Net z specyficznymi wymaganiami rekonstrukcji nagrań audio w kontekście GAN.

6.2.2) Dyskryminator

Dyskryminator, drugi kluczowy komponent architektury GAN, wykorzystuje strukturę konwolucyjną. Składa się z pięciu bloków dyskryminatora, z których każdy zawiera warstwę konwolucyjną z normalizacją spektralną oraz funkcję aktywacji LeakyReLU:

```
    def discriminator_block(self, in_channels, out_channels):
        return nn.Sequential(
            spectral_norm(nn.Conv2d(in_channels, out_channels, 4, stride=2,
    padding=1)),
            nn.LeakyReLU(0.2)
        )
```

Program 8: Blok składowy architektury Dyskryminatora

W celu poprawy stabilności treningu i jakości generowanych wyników, w architekturze zastosowano szereg technik normalizacji. Normalizacja spektralna została wykorzystana we wszystkich warstwach konwolucyjnych, zarówno w generatorze, jak i dyskryminatorze. Technika ta efektywnie kontroluje dynamikę gradientów, co przyczynia się do bardziej stabilnego procesu uczenia.

Ponadto, w generatorze zastosowano normalizację wsadową (Batch Normalization) po każdej warstwie konwolucyjnej. Metoda ta normalizuje aktywacje w obrębie mini-batcha, co pomaga w redukcji wewnętrznego przesunięcia kowariancyjnego i przyspiesza konwergencję modelu.

Wykorzystanie spektrogramów STFT jako reprezentacji danych wejściowych stanowi kluczowy element proponowanej architektury. Spektrogramy te, obliczane z użyciem krótkoczasowej transformaty Fouriera, dostarczają bogatej reprezentacji czasowo-częstotliwościowej sygnału audio. Taka reprezentacja umożliwia modelowi efektywne przetwarzanie zarówno informacji o amplitudzie, jak i fazie sygnału, co jest kluczowe dla zadań rekonstrukcji nagrań dźwiękowych.

6.3) Proces treningu i optymalizacji

W ramach procesu treningu i optymalizacji opracowanego modelu GAN zastosowano szereg technik mających na celu poprawę stabilności uczenia i jakości generowanych wyników.

Implementacja funkcji strat stanowiła kluczowy element procesu optymalizacji. W modelu wykorzystano kombinację różnych funkcji strat, każda z przypisaną wagą, co pozwoliło na precyzyjne kierowanie procesem uczenia:

```
self.loss_weights = {
    'adversarial': 2.5,
    'content': 10.0,
    'spectral_convergence': 0.1,
    'spectral_flatness': 0.1,
    'phase_aware': 0.1,
    'multi_resolution_stft': 1.0,
    'perceptual': 0.1,
    'time_frequency': 1.0,
    'snr': 1.0
}
```

Program 9: Struktura wag funkcji strat

Adversarial Loss (Hinge Loss): Funkcja ta stanowi podstawę treningu przeciwstawnego w architekturze GAN. W implementacji wykorzystano wariant Hinge Loss, który promuje bardziej stabilne uczenie się generatora i dyskryminatora. Dla generatora, strata ta dąży do maksymalizacji prawdopodobieństwa, że dyskryminator sklasyfikuje wygenerowane próbki jako prawdziwe. Dla dyskryminatora, celem jest maksymalizacja marginesu między prawdziwymi a fałszywymi próbkami.

Content Loss (L1 lub L2): Ta funkcja straty mierzy bezpośrednią różnicę między wygenerowanym sygnałem a sygnałem docelowym. Implementacja umożliwia wybór między normą L1 (średnia wartość bezwzględna różnic) a normą L2 (średnia kwadratowa różnic). Norma L1 jest często preferowana w zadaniach związanych z przetwarzaniem sygnałów, gdyż jest mniej wrażliwa na ekstremalne wartości.

Spectral Convergence Loss: Funkcja ta mierzy podobieństwo widmowe między sygnałem wygenerowanym a docelowym. Jest szczególnie istotna w kontekście rekonstrukcji nagrań audio, gdyż koncentruje się na zachowaniu charakterystyki częstotliwościowej sygnału. Obliczana jest jako stosunek normy różnicy widm do normy widma oryginalnego.

Spectral Flatness Loss: Ta funkcja straty ocenia różnicę w „płaskości" widmowej między sygnałem wygenerowanym a docelowym. Płaskość widmowa jest miarą tego, jak równomiernie rozłożona jest energia sygnału w dziedzinie częstotliwości. Jest szczególnie przydatna w zachowaniu ogólnej charakterystyki tonalnej rekonstruowanego dźwięku.

Phase-Aware Loss: Funkcja ta uwzględnia zarówno informacje o amplitudzie, jak i fazie sygnału. Jest to kluczowe w rekonstrukcji dźwięku, gdzie zachowanie prawidłowych relacji fazowych jest niezbędne dla uzyskania naturalnie brzmiącego rezultatu. Składa się z dwóch komponentów: straty amplitudy i straty fazy.

Multi-Resolution STFT Loss: Funkcja analizuje sygnał w różnych skalach czasowo-częstotliwościowych. Wykorzystuje krótkoczasową transformatę Fouriera (STFT) o różnych rozmiarach okna, co pozwala na uchwycenie zarówno krótko- jak i długoczasowych struktur w sygnale audio.

Time-Frequency Loss: Funkcja ta łączy w sobie stratę w dziedzinie czasu i częstotliwości. Uwzględnia zarówno bezpośrednie różnice w próbkach czasowych, jak i różnice w reprezentacji częstotliwościowej sygnału.

Signal-to-Noise Ratio (SNR) Loss: Ta funkcja straty opiera się na klasycznej mierze jakości sygnału - stosunku sygnału do szumu. W kontekście rekonstrukcji audio, „szumem" jest różnica między sygnałem wygenerowanym a docelowym. Funkcja ta promuje generowanie sygnałów o wysokim SNR, co przekłada się na lepszą jakość percepcyjną.

Perceptual Loss: Wykorzystując wstępnie nauczony model ekstrakcji cech **VGGish** [71], funkcja ta porównuje wysokopoziomowe reprezentacje sygnału wygenerowanego i docelowego. Pozwala to na uwzględnienie bardziej abstrakcyjnych i percepcyjnie istotnych cech dźwięku, wykraczających poza proste porównania amplitud czy widm.

Optymalizacja modelu opierała się na algorytmie Adam z niestandardowymi parametrami beta:

```
g_optimizer = optim.Adam(gan.generator.parameters(), lr=g_lr, betas=(0.0, 0.9))
d_optimizer = optim.Adam(gan.discriminator.parameters(), lr=d_lr, betas=(0.0,
0.9))
```

Program 10: Parametry optymizatorów Adam

Z powodu rozmiarów modelu oraz dużych plików treningowych, napotkano na problemy z rozmiarem batcha wynikające z przekroczenia limitu pamięci wirtualnej karty graficznej. Z tego powodu zastosowano technikę akumulacji gradientów, co pozwoliło na efektywne zwiększenie rozmiaru batcha bez zwiększania zużycia pamięci:

```
g_loss = g_loss / self.accumulation_steps
g_loss.backward()

if (self.current_step + 1) % self.accumulation_steps == 0:
    torch.nn.utils.clip_grad_norm_(self.generator.parameters(), max_norm=1.0)
    self.g_optimizer.step()
```

Program 11: Technika akumulacji gradientów

Dynamiczne dostosowywanie współczynnika uczenia (learning rate) zostało zaimplementowane w oparciu o wartości funkcji straty:

```
if self.g_loss_ma > self.g_loss_threshold:
    for param_group in self.g_optimizer.param_groups:
        param_group['lr'] *= 1.01
```

Program 12: Dynamiczny współczynnik uczenia

W celu poprawy stabilności treningu zastosowano techniki regularyzacji. Gradient Penalty został wprowadzony do funkcji straty dyskryminatora:

```
gp = self.gradient_penalty(real_target_norm, generated_audio_norm.detach())
d_loss = d_loss + 10 * gp
```

Program 13: Gradient Penalty na podstawie Wasserstein GAN

Instance Noise z mechanizmem annealing został użyty do stopniowego zmniejszania szumu dodawanego do danych wejściowych w trakcie treningu:

```
self.instance_noise *= self.instance_noise_anneal_rate
```

Program 14: Mechanizm stopniowego zmniejszania szumu danych wejściowych Dyskryminatora

Monitorowanie i wizualizacja procesu treningu zostały zrealizowane za pomocą dedykowanych metod Callback. LossVisualizationCallback umożliwiał śledzenie i wizualizację różnych komponentów funkcji straty w czasie rzeczywistym:

```
loss_visualization_callback = LossVisualizationCallback(log_dir=run_log_dir)
loss_visualization_callback.on_epoch_end(epoch, combined_losses)
```

Program 15: Metody wizualizacji strat

Implementacja Early Stopping pozwoliła na automatyczne przerwanie treningu w przypadku braku poprawy wyników:

```
early_stopping_callback = EarlyStoppingCallback(patience=5, verbose=True,
delta=0.01,
                                                path=os.path.join(run_log_dir,
'best_model.pt'))
if early_stopping_callback(epoch, val_loss, gan):
    print("Early stopping triggered")
    break
```

Program 16: Mechanizm wczesnego przerwania treningu

Zapisywanie checkpointów umożliwiło zachowanie stanu modelu w regularnych odstępach czasu, co pozwoliło na wznowienie treningu w przypadku nagłego przerwania:

```
checkpoint_callback = CheckpointCallback(checkpoint_dir)
checkpoint_callback(epoch, gan)
```

Program 17: Mechanizm zachowania stanu modelu

6.4) Charakterystyka kodu źródłowego

Struktura projektu i organizacja modułów w opracowanym systemie rekonstrukcji nagrań dźwiękowych odzwierciedla modułowe i funkcjonalne podejście do rozwiązania problemu. Projekt został podzielony na kilka kluczowych modułów, każdy odpowiedzialny za specyficzny aspekt przetwarzania i analizy danych audio.

Główne moduły projektu obejmują:

1. `models.py`: Zawiera definicje klas `Generator`, `Discriminator` i `AudioEnhancementGAN`, które stanowią trzon architektury sieci GAN.
2. `losses.py`: Implementuje różnorodne funkcje straty wykorzystywane w procesie treningu, w tym `adversarial_loss`, `content_loss`, `spectral_convergence_loss` i inne.
3. `data_preparation.py`: Odpowiada za przygotowanie i przetwarzanie danych wejściowych, zawierając klasę `STFTDataset` do obsługi spektrogramów STFT.
4. `callbacks.py`: Implementuje mechanizmy monitorowania i wizualizacji procesu treningu, w tym `LossVisualizationCallback` i `CheckpointCallback`.
5. `main.py`: Stanowi punkt wejścia do aplikacji, integrując wszystkie komponenty i implementując logikę treningu.

Kluczowe klasy i funkcje w implementacji sieci GAN obejmują:

- Klasa `AudioEnhancementGAN`: Centralna klasa projektu, integrująca generator i dyskryminator oraz implementująca logikę treningu.
- Klasy `Generator` i `Discriminator`: Implementują odpowiednio architekturę generatora i dyskryminatora.
- Funkcja `generator_loss`: Implementuje funkcję straty dla generatora, łączącą różne komponenty straty.

Projekt w szerokim zakresie wykorzystuje biblioteki PyTorch, librosa i pydub:

- PyTorch służy jako podstawowy framework implementacji i treningu sieci neuronowych.
- Librosa jest wykorzystywana do zaawansowanego przetwarzania sygnałów audio, w szczególności do obliczania i manipulacji spektrogramami STFT.
- Pydub znajduje zastosowanie w procesie przygotowania danych, umożliwiając konwersję i manipulację plikami audio.

Mechanizmy przetwarzania równoległego i zarządzania pamięcią zostały zaimplementowane w celu optymalizacji wydajności:

- Wykorzystanie `ProcessPoolExecutor` i `ThreadPoolExecutor` do równoległego przetwarzania plików audio:
- Dynamiczne dostosowywanie liczby procesów do dostępnych zasobów CPU:
- Ograniczanie użycia pamięci RAM poprzez przetwarzanie danych w mniejszych porcjach:

Implementacja interfejsu wiersza poleceń (CLI) do obsługi skryptów została zrealizowana z wykorzystaniem modułu `argparse`, co umożliwia elastyczne konfigurowanie parametrów treningu:

```
parser = argparse.ArgumentParser(description='Audio Enhancement GAN')
parser.add_argument('--no-cuda', action='store_true', default=False,
                    help='disables CUDA training')
parser.add_argument('--batch-size', type=int, default=16, metavar='N',
                    help='input batch size for training (default: 16)')
parser.add_argument('--epochs', type=int, default=50, metavar='N',
                    help='number of epochs to train (default: 50)')
args = parser.parse_args()
```

Program 18: Mechanizm argumentów wierza poleceń

6.5) Metodologia eksperymentów

Przeprowadzone eksperymenty miały na celu ocenę skuteczności opracowanego modelu GAN w rekonstrukcji nagrań dźwiękowych, ze szczególnym uwzględnieniem usuwania szumów charakterystycznych dla płyt winylowych. Głównym celem było zbadanie zdolności modelu do odwzorowania oryginalnych sygnałów audio poprzez poprawę jakości oraz usunięcie artefaktów z próbek.

Opis przeprowadzonych eksperymentów:

1. Trening modelu na zbiorze danych zawierającym pary nagrań: oryginalne (czyste) oraz z dodanymi szumami winylowymi.
2. Walidacja modelu na oddzielnym zbiorze danych, niedostępnym podczas treningu.
3. Generowanie rekonstrukcji dla wybranych próbek testowych i analiza wyników.

Metody ewaluacji obejmowały obiektywne metryki jakości audio oraz walidację na oddzielnym zbiorze danych:

1. Obiektywne metryki:
   - Signal-to-Noise Ratio (SNR): Mierzy stosunek mocy sygnału do mocy szumu.
   - Spectral Convergence: Ocenia podobieństwo widmowe między sygnałem oryginalnym a zrekonstruowanym.
   - Perceptual Evaluation of Audio Quality (PEAQ): Symuluje subiektywną ocenę jakości dźwięku.
2. Walidacja na oddzielnym zbiorze danych:
   - Wykorzystano cross-walidację, dzieląc zbiór danych na część treningową i walidacyjną.
   - Monitorowano straty walidacyjne w trakcie treningu, aby uniknąć przeuczenia.

Ze względu na **niepowodzenie badania**, subiektywna **ocena przez odsłuch nie była możliwa**. W normalnych warunkach proces ten obejmowałby:
- Próbkę badaczy z możliwie różnych kohort oceniających jakość rekonstrukcji.
- Ślepe testy AB porównujące oryginalne nagrania z rekonstrukcjami.
- Ocenę parametrów takich jak czystość dźwięku, zachowanie detali muzycznych i ogólna jakość.

W ramach badań przeprowadzono eksperymenty mające na celu optymalizację architektury i hiperparametrów modelu GAN. W przypadku generatora, testowano różne konfiguracje sieci, modyfikując liczbę i rozmiar warstw konwolucyjnych oraz eksperymentując z różnymi typami bloków rezydualnych. Podobne modyfikacje wprowadzano w architekturze dyskryminatora, gdzie zweryfikowano wpływ głębokości sieci na jakość wyników oraz efektywność różnych technik normalizacji, ze szczególnym uwzględnieniem normalizacji spektralnej. Proces optymalizacji obejmował również dostrajanie kluczowych hiperparametrów, takich jak współczynniki uczenia się dla generatora i dyskryminatora. Badano także wpływ rozmiaru batcha oraz liczby kroków akumulacji gradientu na stabilność i efektywność procesu uczenia. Celem tych kompleksowych eksperymentów było znalezienie optymalnej konfiguracji modelu, zapewniającej najlepsze wyniki w zadaniu rekonstrukcji nagrań audio przy jednoczesnym zachowaniu stabilności treningu i efektywności obliczeniowej.

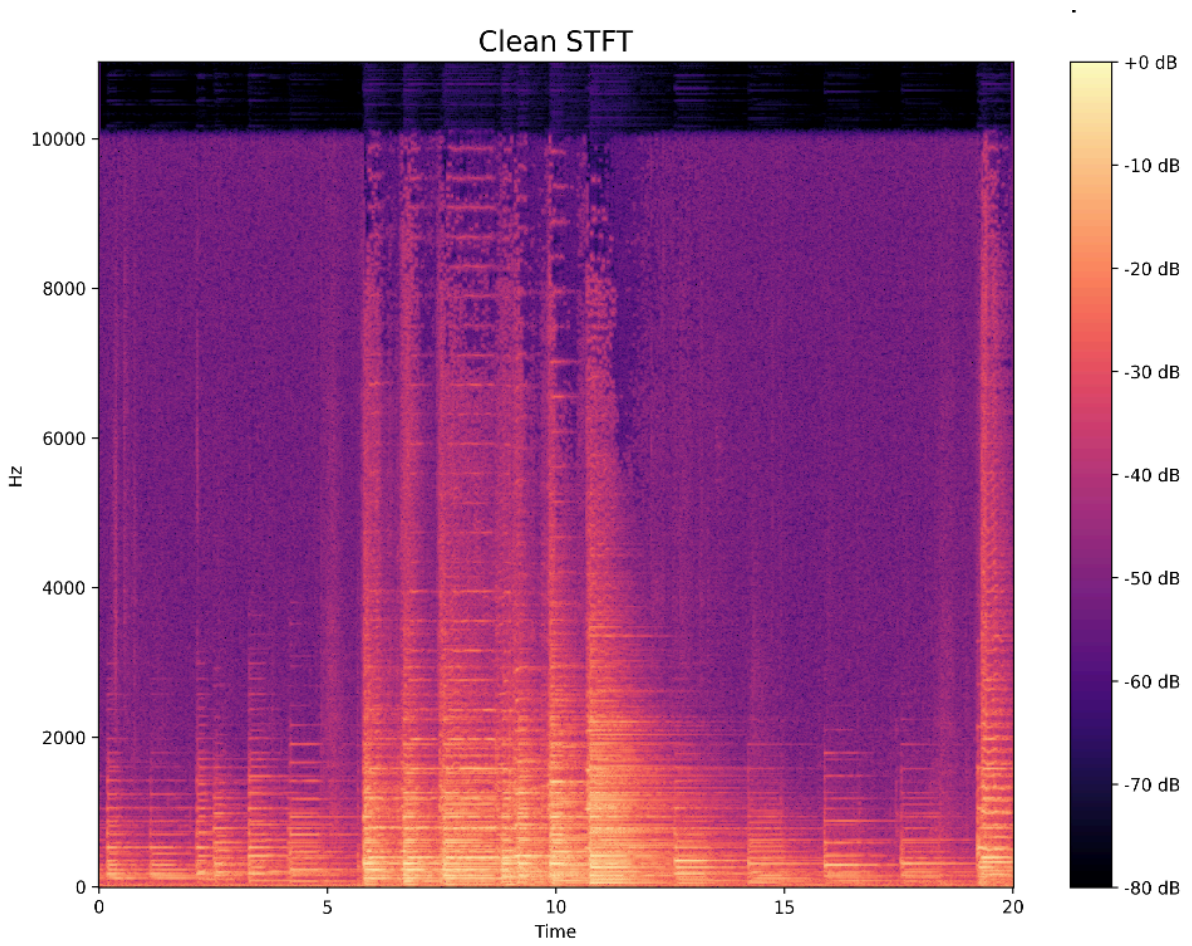
W ramach analizy wpływu poszczególnych komponentów na jakość rekonstrukcji, przeprowadzono badania ukierunkowane na różne aspekty modelu. W obszarze funkcji strat dokonano eksperymentów dotyczących wpływu różnych wag przypisanych poszczególnym komponentom oraz analizie efektywności funkcji takich jak spectral flatness loss czy phase-aware loss. Zweryfikowano także skuteczność różnych technik normalizacji, w tym batch normalization i instance normalization w generatorze.

Wyniki eksperymentów były analizowane poprzez wizualizację spektrogramów STFT, monitorowanie krzywych między epokami, oraz analizę metryk algorytmu podczas procesu uczenia. Mimo że badanie nie przyniosło oczekiwanych rezultatów w zakresie jakości rekonstrukcji audio, dostarczyło cennych informacji na temat zachowania modelu GAN w kontekście przetwarzania sygnałów dźwiękowych oraz wskazało potencjalne kierunki dalszych badań i ulepszeń.

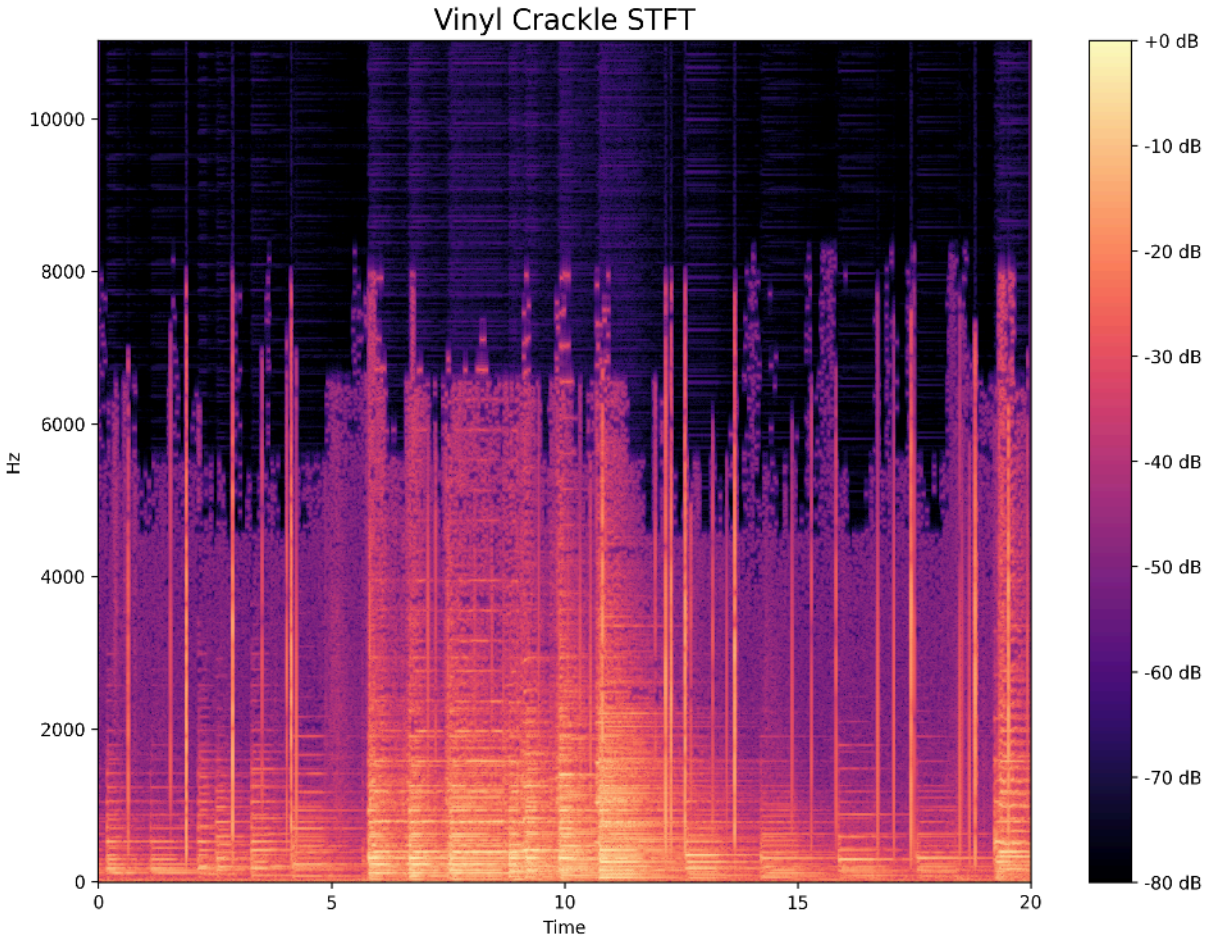
7) Analiza wyników

W ramach przeprowadzonego badania nad rekonstrukcją nagrań dźwiękowych z wykorzystaniem sieci GAN zastosowano obiektywne metody analizy celem oceny skuteczności opracowanego modelu. Proces ewaluacji koncentrował się na trzech głównych aspektach: **analizie wizualizacji spektrogramów STFT**, **obserwacji funkcji strat** w trakcie treningu oraz **testach odsłuchowych z wykorzystaniem odwrotnej transformaty STFT**.

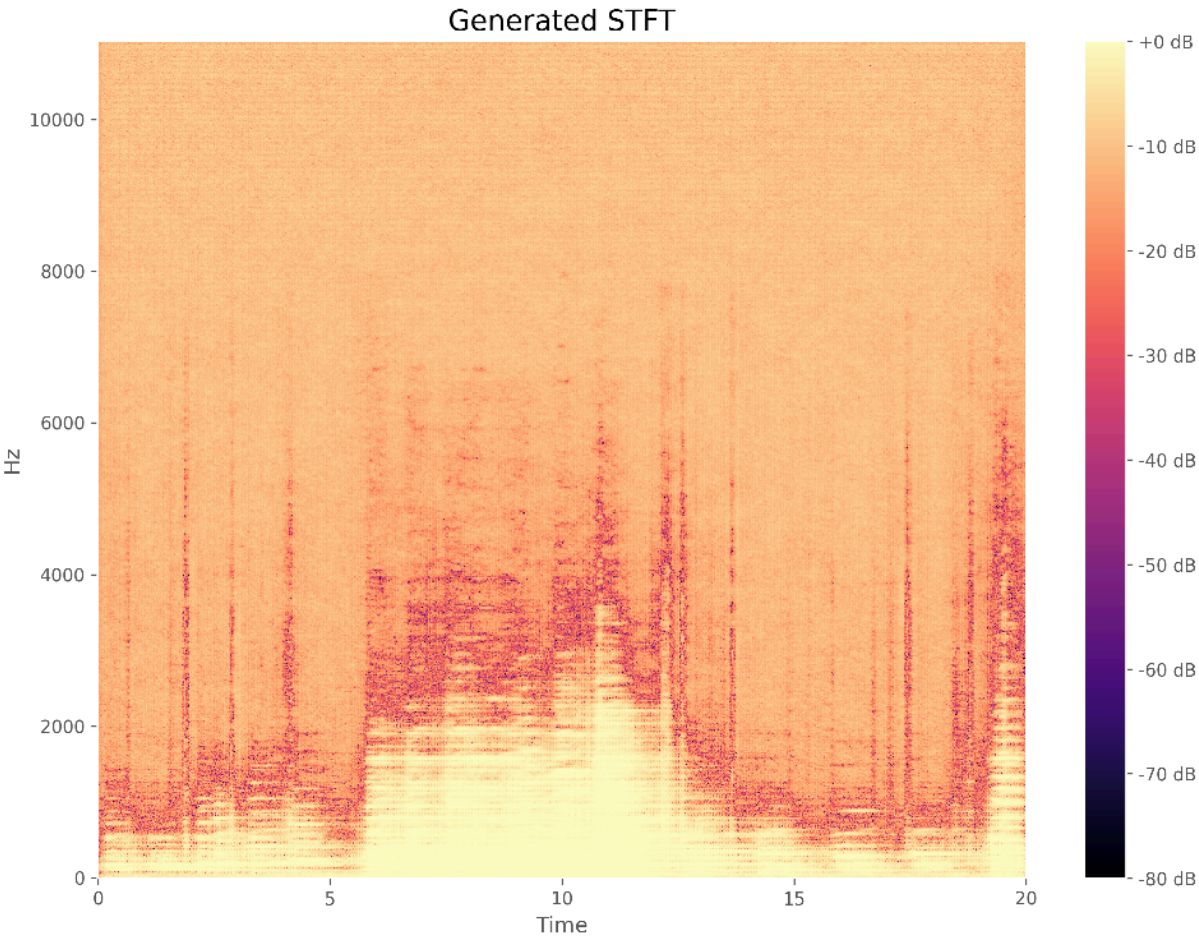
Analiza wizualizacji spektrogramów STFT stanowiła kluczowy element oceny obiektywnej. Porównanie spektrogramów oryginalnych nagrań, ich wersji z dodanymi szumami winylowymi oraz rekonstrukcji generowanych przez sieć GAN pozwoliło na bezpośrednią obserwację skuteczności modelu w usuwaniu charakterystycznych zniekształceń.



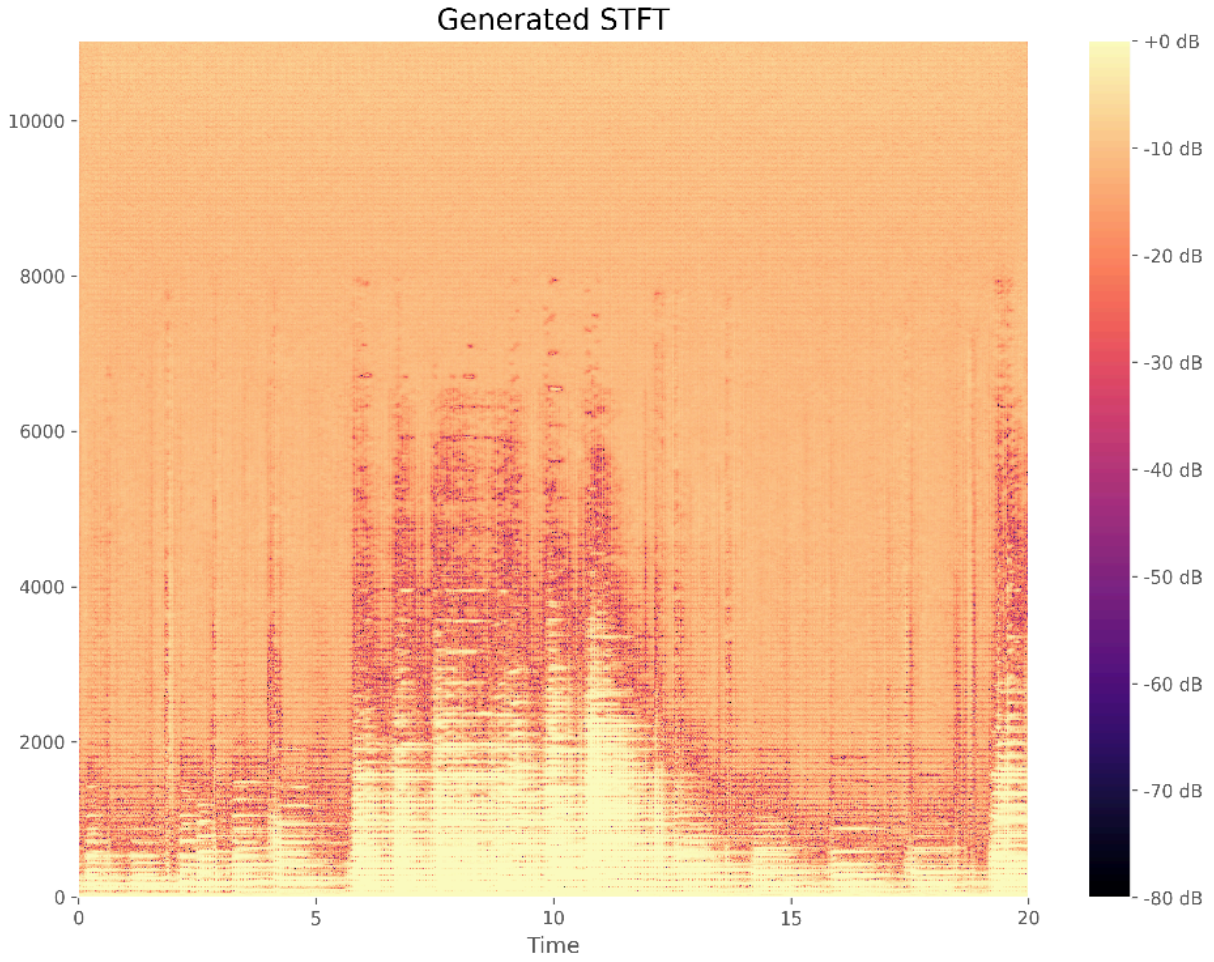
Rysunek 1: Spektogram oryginalnego fragmentu



Rysunek 2: Spektogram fragmentu z dodanym szumem winylowym



Rysunek 3: Spektogram fragmentu rekonstrukcji w początkowej fazie treningu



Rysunek 4: Spektogram fragmentu rekonstrukcji w końcowej fazie treningu

Analiza tych wizualizacji ujawniła, że model był wstanie w pewnym stopniu zniwelować trzaski charakterystyczne dla płyt winylowych, co widoczne było jako redukcja pionowych linii na spektrogramach reprezentujących nagłe, krótkotrwałe zakłócenia. Jednocześnie model nauczył się uwydatnić fale oznaczające wysokie dźwięki znajdujące się w oryginalnej próbce. **Model uczył się poprawnie i dążył w kierunku oryginalnych próbek, jednak nie był w stanie zniwelować szumów które wygenerował w początkowych fazach uczenia.**

Obserwacja funkcji strat w trakcie procesu uczenia stanowiła drugi istotny aspekt oceny obiektywnej. Analiza ta pozwoliła na śledzenie postępów w zdolności modelu do rekonstrukcji nagrań w miarę upływu epok treningu. Poniżej przedstawiono przykładowe wartości strat uzyskane podczas procesu uczenia:

epoka	0	1	2	...	14	15	16	...	20
wartość funkcji straty	249	197	169	...	44	49	45	...	46

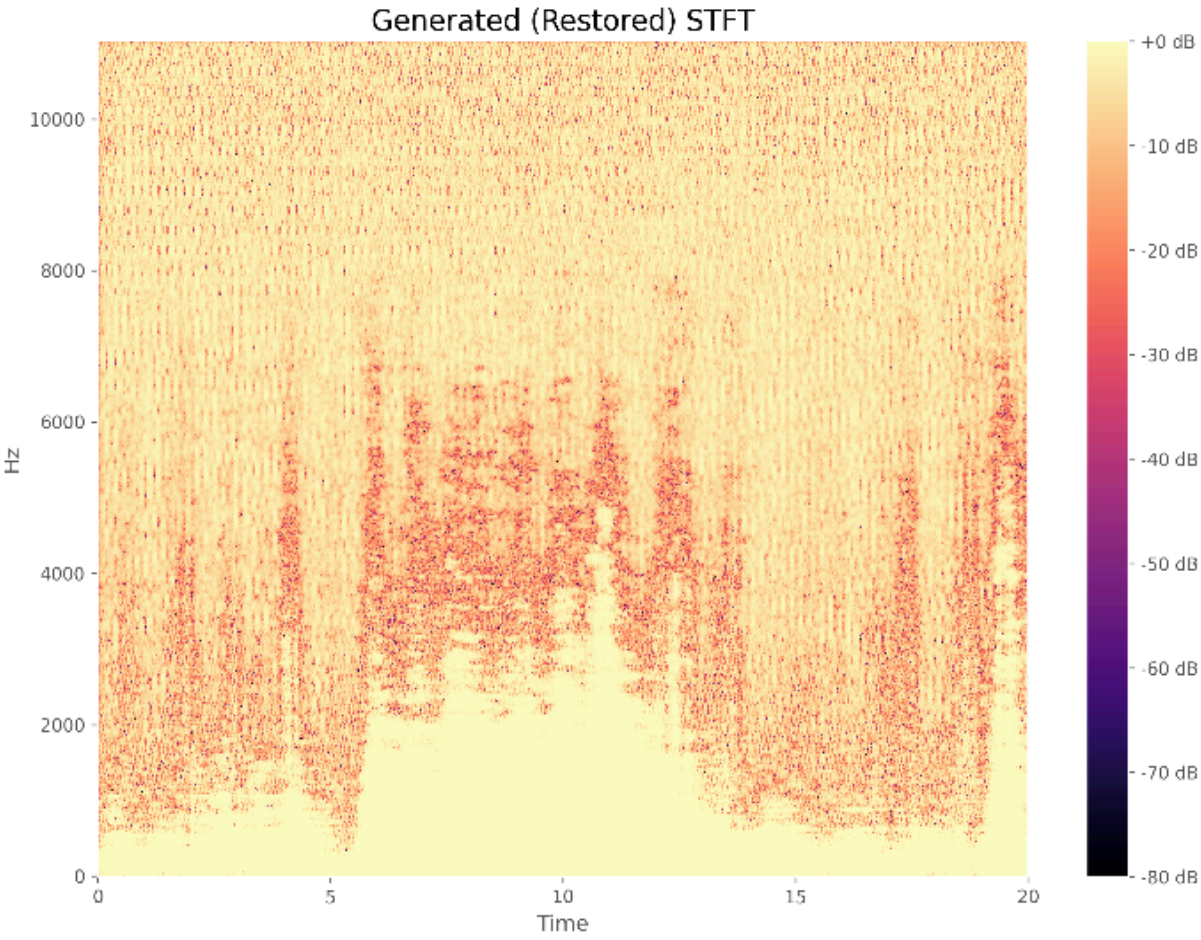
Obserwacja tych wartości ujawnia interesującą tendencję. W początkowych fazach treningu widoczny jest znaczący spadek wartości funkcji straty, co sugeruje, że model uczył się efektywnie redukować błędy rekonstrukcji. Jednakże, około 14-16 epoki wartość funkcji straty osiągnęła minimum (około 44-45) i przestała znacząco spadać, utrzymując się na podobnym poziomie w kolejnych epokach.

To zjawisko jest niepokojące, biorąc pod uwagę, że jakość rekonstrukcji audio pozostawała niezadowalająca. Sugeruje to, że model osiągnął **minimum lokalne**, które nie odpowiadało satysfakcjonującemu rozwiązaniu problemu rekonstrukcji. Innymi słowy, funkcja straty przestała dostarczać użytecznych informacji dla dalszej optymalizacji modelu, mimo że nie był on jeszcze w stanie generować wysokiej jakości rekonstrukcji audio.

Trzecim elementem oceny obiektywnej były testy odsłuchowe z wykorzystaniem odwrotnej transformaty STFT. W tym procesie, spektrogramy wygenerowane przez model były przekształcane z powrotem na sygnały audio w formacie MP3. Ta metoda miała na celu umożliwienie bezpośredniej oceny słuchowej jakości rekonstrukcji, stanowiąc istotne uzupełnienie analizy wizualnej i numerycznej.

Wyniki tych testów odsłuchowych okazały się jednak **skrajnie niezadowalające**. We wszystkich próbach, niezależnie od etapu treningu czy konfiguracji modelu, uzyskane sygnały audio składały się wyłącznie z niezrozumiałych szumów. Żadna z wygenerowanych próbek nie wykazywała cech przypominających muzykę czy jakiekolwiek rozpoznawalne dźwięki. Ta obserwacja stanowi najbardziej dobitne świadectwo nieefektywności modelu w zadaniu rekonstrukcji nagrań muzycznych, podkreślając znaczącą rozbieżność między częściowymi poprawami widocznymi na spektrogramach a faktyczną jakością dźwięku percypowaną przez ludzkie ucho.

Obserwacja zmian w spektrogramach w trakcie procesu uczenia ujawniła pewne interesujące tendencje. W początkowych fazach treningu, model wykazywał zdolność do częściowej redukcji najbardziej widocznych artefaktów, takich jak pionowe linie reprezentujące trzaski charakterystyczne dla płyt winylowych. Jednakże, w wielu podejściach do uczenia, w miarę postępu treningu, pojawiały się niepokojące zjawiska:



Rysunek 5: Spektogram fragmentu rekonstrukcji na początku błędów w nauce

Rysunek 6: Spektogram fragmentu rekonstrukcji pod koniec błędów w nauce

Na powyższej wizualizacji można zaobserwować, że model po kilkunastu epokach uczenia omylnie nauczył się podmieniać wartości zerami, odpowiadające dźwiękom na poziomie 0dB.

Dyskusja na temat niedostatecznej poprawy jakości do uzyskania użytecznych wyników audio musi uwzględnić kilka kluczowych aspektów:

1. Złożoność zadania: Rekonstrukcja pełnych nagrań muzycznych okazała się znacznie bardziej skomplikowana niż początkowo zakładano. Model musiał nie tylko usunąć szumy, ale także odtworzyć subtelne detale muzyczne, co stanowiło wyzwanie wykraczające poza możliwości obecnej architektury.
2. Nieadekwatność funkcji strat: Mimo zastosowania różnorodnych funkcji strat, mogły one nie w pełni odzwierciedlać percepcyjne aspekty jakości dźwięku. To mogło prowadzić do optymalizacji modelu w kierunku, który nie przekładał się bezpośrednio na poprawę słyszalnej jakości.
3. Ograniczenia architektury: Zastosowana architektura GAN, mimo swojej złożoności, mogła nie być wystarczająco dostosowana do specyfiki rekonstrukcji sygnałów audio. W szczególności, model mógł mieć trudności z zachowaniem spójności fazowej, co jest kluczowe dla naturalnego brzmienia dźwięku.

4. Problemy z danymi treningowymi: Jakość i reprezentatywność danych treningowych mogły nie być wystarczające do nauczenia modelu efektywnej rekonstrukcji. W szczególności, symulowane szumy winylowe mogły nie w pełni oddawać złożoność rzeczywistych zniekształceń występujących w historycznych nagraniach.
5. Niestabilność treningu GAN: Charakterystyczna dla architektury GAN niestabilność procesu uczenia mogła prowadzić do problemów z konwergencją, co objawiało się niekonsekwentną jakością generowanych spektrogramów w różnych fazach treningu.

Podsumowując, obiektywna ocena wyników wykazała, że opracowany model GAN, mimo pewnych obiecujących aspektów widocznych w analizie spektrogramów, nie był w stanie osiągnąć zadowalającego poziomu rekonstrukcji nagrań dźwiękowych. Całkowity brak rozpoznawalnych elementów muzycznych w wygenerowanych próbkach audio podkreśla głęboką rozbieżność między częściowymi poprawami obserwowanymi w domenie częstotliwościowej a faktyczną percepcją dźwięku.

Ta sytuacja uwypukla złożoność zadania rekonstrukcji nagrań muzycznych i wskazuje na potrzebę dalszych, pogłębionych badań w tym obszarze. Przyszłe prace powinny skupić się na:

1. Udoskonaleniu architektury modelu, ze szczególnym uwzględnieniem zachowania spójności fazowej sygnału.
2. Opracowaniu bardziej zaawansowanych funkcji strat, które lepiej odzwierciedlałyby percepcyjne aspekty jakości dźwięku.
3. Zwiększeniu rozmiaru i różnorodności zbioru danych treningowych, z możliwym uwzględnieniem rzeczywistych, a nie tylko symulowanych, zniekształceń.
4. Eksploracji alternatywnych podejść do generatywnego modelowania dźwięku, takich jak modele autoregresyjne czy modele dyfuzyjne.

Mimo że obecne wyniki nie spełniły oczekiwań w zakresie praktycznej użyteczności, stanowią one cenny wkład w zrozumienie wyzwań związanych z zastosowaniem technik uczenia maszynowego do rekonstrukcji nagrań dźwiękowych i otwierają nowe ścieżki dla przyszłych badań w tej dziedzinie.

8) Wnioski i perspektywy

8.1) Podsumowanie osiągniętych rezultatów

Niniejsza praca miała na celu zbadanie możliwości wykorzystania metod uczenia maszynowego, ze szczególnym uwzględnieniem Generatywnych Sieci Przeciwstawnych (GAN), w procesie rekonstrukcji nagrań dźwiękowych. Głównym celem było opracowanie i implementacja modelu GAN zdolnego do poprawy jakości historycznych nagrań muzycznych, ze szczególnym naciskiem na usuwanie szumów charakterystycznych dla płyt winylowych oraz rozszerzanie pasma częstotliwościowego.

Przeprowadzone eksperymenty dostarczyły cennych informacji na temat potencjału i ograniczeń zastosowania sieci GAN w tym kontekście. Kluczowe wyniki obejmują częściowe sukcesy w redukcji trzasków charakterystycznych dla płyt winylowych, co było widoczne na spektrogramach STFT generowanych próbek. Model wykazał zdolność do uczenia się pewnych aspektów rekonstrukcji, co objawiało się stopniową poprawą jakości generowanych spektrogramów w trakcie procesu treningu.

Jednakże, ocena skuteczności proponowanej metody GAN w rekonstrukcji nagrań ujawniła znaczące ograniczenia. Mimo obiecujących rezultatów widocznych w domenie częstotliwościowej, próby konwersji zrekonstruowanych spektrogramów z powrotem do domeny czasowej nie przyniosły zadowalających wyników. Wygenerowane sygnały audio charakteryzowały się wysokim poziomem szumów i zniekształceń, co uniemożliwiło ich subiektywną ocenę poprzez odsłuch.

Porównując osiągnięte rezultaty z początkowymi założeniami i oczekiwaniami, należy przyznać, że badanie nie spełniło wszystkich postawionych celów. Zakładana możliwość generowania wysokiej jakości rekonstrukcji nagrań, które byłyby percepcyjnie zbliżone do oryginałów, nie została osiągnięta. Model GAN, mimo swojej złożoności i zastosowania zaawansowanych technik optymalizacji, nie był w stanie w pełni odtworzyć subtelności i detali muzycznych niezbędnych do uzyskania satysfakcjonującej jakości dźwięku.

Niemniej jednak, przeprowadzone badania dostarczyły cennych informacji na temat wyzwań związanych z zastosowaniem uczenia maszynowego w dziedzinie rekonstrukcji audio. Zidentyfikowano kluczowe problemy, takie jak trudności w zachowaniu spójności fazowej sygnału czy ograniczenia związane z redukcją szumów w wygenerowanych nagraniach. Te obserwacje stanowią istotny wkład w zrozumienie kompleksowości zadania rekonstrukcji nagrań muzycznych i otwierają drogę do dalszych, bardziej ukierunkowanych badań w tej dziedzinie.

Podsumowując, mimo że proponowana metoda GAN nie osiągnęła wszystkich zakładanych celów w zakresie praktycznej rekonstrukcji nagrań, przeprowadzone badania przyczyniły się do pogłębienia wiedzy na temat zastosowania uczenia maszynowego w przetwarzaniu sygnałów audio. Zidentyfikowane wyzwania i ograniczenia stanowią cenny punkt wyjścia do dalszych prac nad udoskonaleniem technik rekonstrukcji nagrań dźwiękowych z wykorzystaniem sztucznej inteligencji.

8.2) Ograniczenia proponowanej metody

W trakcie realizacji badań nad zastosowaniem sieci GAN w rekonstrukcji nagrań dźwiękowych napotkano szereg istotnych wyzwań i ograniczeń, które wpłynęły na ostateczne wyniki pracy. Identyfikacja i analiza tych trudności stanowi kluczowy element w zrozumieniu aktualnych ograniczeń metody oraz wyznaczeniu kierunków dalszych badań.

Jednym z głównych wyzwań okazała się złożoność zadania rekonstrukcji pełnych nagrań muzycznych. W przeciwieństwie do prostszych zadań, takich jak usuwanie pojedynczych typów zakłóceń, pełna rekonstrukcja wymaga jednoczesnego adresowania wielu aspektów jakości dźwięku. Model musiał nie tylko usuwać szumy i trzaski, ale także odtwarzać utracone częstotliwości i zachowywać muzyczną spójność, co okazało się zadaniem przekraczającym możliwości opracowanej architektury.

Istotnym czynnikiem ograniczającym skuteczność rekonstrukcji była jakość i reprezentatywność danych treningowych. Mimo starań o stworzenie zróżnicowanego zbioru danych, symulowane zniekształcenia mogły nie w pełni odzwierciedlać złożoność rzeczywistych uszkodzeń występujących w historycznych nagraniach. Ograniczenie to mogło prowadzić do niedostatecznej generalizacji modelu na rzeczywiste przypadki.

Złożoność architektury sieci GAN, choć teoretycznie korzystna, w praktyce przyczyniła się do powstania szeregu problemów. Niestabilność procesu uczenia, charakterystyczna dla GAN, okazała się szczególnie problematyczna w kontekście danych audio. Balansowanie między uczeniem generatora i dyskryminatora było trudne, co często prowadziło do nieoptymalnych rezultatów lub zjawiska zapadania się modelu (mode collapse).

W domenie audio napotkano specyficzne problemy, które nie występują lub są mniej znaczące w innych obszarach zastosowań uczenia maszynowego. Kluczowym wyzwaniem okazało się zachowanie spójności fazowej w rekonstruowanych sygnałach. Nawet niewielkie błędy w odtworzeniu fazy prowadziły do znaczących zniekształceń percepcyjnych, co było szczególnie widoczne przy próbach konwersji spektrogramów z powrotem do domeny czasowej.

Poważnym ograniczeniem okazały się również kwestie sprzętowe i obliczeniowe. Wykorzystywana w badaniach karta graficzna Radeon 6950 XTX z 16 GB pamięci VRAM, mimo swoich wysokich parametrów, wielokrotnie okazywała się niewystarczająca do efektywnego treningu modelu na pełnym zbiorze danych. Problemy z brakiem pamięci wirtualnej wymuszały ograniczenie rozmiaru batchy lub stosowanie technik takich jak akumulacja gradientów, co z kolei wpływało na stabilność i efektywność procesu uczenia. Długi czas treningu, sięgający kilkudziesięciu godzin na pełny proces, znacząco ograniczał możliwości eksperymentowania z różnymi konfiguracjami modelu i hiperparametrami.

Dodatkowym wyzwaniem okazała się interpretacja wyników pośrednich. Mimo obserwowanych popraw w reprezentacjach częstotliwościowych (spektrogramach), przekładanie tych ulepszeń na percepcyjną jakość dźwięku okazało się nieoczywiste. Sugeruje to, że stosowane funkcje straty mogły nie w pełni odzwierciedlać aspekty istotne dla ludzkiej percepcji dźwięku.

Wreszcie, należy zwrócić uwagę na ograniczenia wynikające z samej natury podejścia opartego na uczeniu maszynowym. Model, ucząc się na podstawie dostarczonych przykładów, mógł mieć trudności z rekonstrukcją rzadkich lub unikalnych elementów muzycznych, które nie były dobrze reprezentowane w zbiorze treningowym.

8.3) Potencjalne kierunki dalszych badań

Przeprowadzone badania, mimo napotkanych ograniczeń, otwierają szereg interesujących ścieżek dla dalszych prac w dziedzinie rekonstrukcji nagrań dźwiękowych z wykorzystaniem metod uczenia maszynowego. Przyszłe badania mogłyby skupić się na kilku kluczowych obszarach.

W kontekście architektury sieci GAN, obiecującym kierunkiem wydaje się eksploracja bardziej zaawansowanych wariantów, takich jak Progressive GAN czy StyleGAN, które wykazały imponujące rezultaty w dziedzinie generowania obrazów. Adaptacja tych architektur do domeny audio mogłaby potencjalnie przezwyciężyć niektóre z napotkanych problemów, szczególnie w zakresie stabilności treningu i jakości generowanych wyników. Ponadto, warto rozważyć implementację technik takich jak self-attention czy transformer blocks w generatorze, co mogłoby poprawić zdolność modelu do uchwycenia długoterminowych zależności w sygnałach muzycznych.

Alternatywnym podejściem wartym eksploracji są modele dyfuzyjne, które w ostatnim czasie zyskały znaczącą popularność w zadaniach generatywnych. Modele te, takie jak DDPM (Denoising Diffusion Probabilistic Models), mogłyby okazać się szczególnie skuteczne w kontekście rekonstrukcji audio, ze względu na ich zdolność do stopniowego usuwania szumu z danych. Ich potencjał w generowaniu wysokiej jakości próbek dźwiękowych oraz stabilność treningu czynią je atrakcyjną alternatywą dla tradycyjnych GAN-ów.

Istotnym kierunkiem badań powinna być także głębsza integracja wiedzy dziedzinowej z zakresu przetwarzania sygnałów audio w procesie uczenia maszynowego. Można by rozważyć opracowanie specjalizowanych warstw sieciowych, które explicite modelowałyby zjawiska akustyczne, takie jak propagacja fal dźwiękowych czy rezonans. Implementacja zaawansowanych technik analizy częstotliwościowej, takich jak transformata falkowa czy analiza cepstralna, bezpośrednio w architekturze sieci neuronowej mogłaby znacząco poprawić jej zdolność do precyzyjnej rekonstrukcji sygnałów muzycznych.

Przyszłe badania powinny również rozszerzyć zakres eksperymentów o szerszy wachlarz gatunków muzycznych i typów nagrań. Szczególnie interesujące pozostaje badanie skuteczności modeli w rekonstrukcji nagrań wokalnych, muzyki elektronicznej czy zapisów koncertów na żywo.

8.4) Implikacje dla przyszłości rekonstrukcji nagrań muzycznych

Rozwój technik opartych na sztucznej inteligencji w dziedzinie rekonstrukcji nagrań muzycznych niesie ze sobą znaczące implikacje dla ochrony dziedzictwa kulturowego. Potencjał AI w tym kontekście jest ogromny — zaawansowane algorytmy mogą nie tylko przywrócić do życia historyczne nagrania, ale także uczynić je dostępnymi dla szerszej publiczności w niespotykanej dotąd jakości. Możliwość automatycznej poprawy jakości tysięcy godzin archiwalnych nagrań otwiera nowe perspektywy dla badaczy, muzykologów i miłośników muzyki, umożliwiając głębsze zrozumienie i docenienie muzycznego dziedzictwa ludzkości.

Jednocześnie, stosowanie AI w rekonstrukcji historycznych nagrań rodzi istotne pytania etyczne. Kluczowe jest zachowanie równowagi między poprawą jakości a zachowaniem autentyczności oryginału. Zbyt agresywna ingerencja algorytmów AI może prowadzić do zniekształcenia historycznego brzmienia, zacierając granicę między rekonstrukcją a reinterpretacją. Dlatego niezbędne jest wypracowanie standardów i wytycznych etycznych, które będą kierować wykorzystaniem AI w tym obszarze, zapewniając poszanowanie integralności artystycznej i historycznej rekonstruowanych dzieł.

Patrząc w przyszłość, można prognozować dynamiczny rozwój technologii AI w dziedzinie przetwarzania audio. Możemy spodziewać się pojawienia się coraz bardziej wyrafinowanych modeli, zdolnych do jeszcze dokładniejszej analizy i rekonstrukcji sygnałów dźwiękowych. Prawdopodobne jest również powstanie systemów hybrydowych, łączących klasyczne techniki przetwarzania sygnałów z zaawansowanymi algorytmami uczenia maszynowego, co może prowadzić do przełomów w jakości i efektywności rekonstrukcji.

Wpływ zaawansowanych technik rekonstrukcji na przemysł muzyczny i praktyki archiwizacyjne będzie najprawdopodobniej znaczący. Możemy oczekiwać rosnącego zainteresowania remasteringiem i ponownym wydawaniem historycznych nagrań w poprawionej jakości. To z kolei może wpłynąć na strategie wydawnicze i modele biznesowe w branży muzycznej. Dla archiwów i instytucji kulturalnych, nowe technologie rekonstrukcji mogą oznaczać rewolucję w sposobie przechowywania i udostępniania zbiorów audio, potencjalnie prowadząc do demokratyzacji dostępu do muzycznego dziedzictwa.

Podsumowując, mimo że obecne badania nad wykorzystaniem AI w rekonstrukcji nagrań muzycznych napotkały pewne ograniczenia, perspektywy na przyszłość są niezwykle obiecujące. Dalszy rozwój w tej dziedzinie ma potencjał nie tylko do znaczącego postępu technologicznego, ale także do głębokiej transformacji naszego podejścia do zachowania i interpretacji muzycznego dziedzictwa. Kluczowe będzie zrównoważone podejście, które zmaksymalizuje korzyści płynące z nowych technologii, jednocześnie zachowując szacunek dla integralności i autentyczności historycznych nagrań.

Lista symboli

Bibliografia

[1] L. Gitelman, *Always Already New: Media, History, and the Data of Culture*. 2006.

[2] O. Touloumi, „Sound in silence: design and listening cultures in the Woodberry Poetry Room", *The Journal of Architecture*, 2018.

[3] Ø. Eiksund *i in.*, *Music Technology in Education — Channeling and Challenging Perspectives*. 2020.

[4] R. Vyacheslavovich, „The First Jazz Gramophone Record: The Music of the Moment That Became Timeless", *Department of Theory and History of Music, Moscow State Insitute of Culture*, luty 2021.

[5] T. Chvanova, „Sound Recording as a Factor of Evolution of the Piano Performing Style in the 20th Century", *conservatory.ru*, lip. 2021.

[6] *Vinyl: A History of the Analogue Record*. 2013.

[7] A. J. Houtsma, T. Rossing, i W. Wagenaars, „Auditory demonstrations on compact disc", *Acoustical Society of America*, sie. 2005.

[8] V. Straebel, „From Reproduction to Performance: Media-Specific Music for Compact Disc", *Massachusetts Institute of Technology*, grudz. 2009.

[9] C. Jongjaihan i A. Kaewrawang, „Micromagnetic Simulation of L10-FePt-Based Transition Jitter of Heat-Assisted Magnetic Recording at Ultrahigh Areal Density", *MDPI*, wrz. 2022.

[10] J. Eisentraut, *The Accessibility of Music*. 2013.

[11] J. Lemercier, J. Richter, S. Welker, V. Valimaki, i T. Gerkmann, „Diffusion Models for Audio Restoration", *Signal Processing (SP), Department of Informatics, Universitat Hamburg, Germany Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*, lip. 2024.

[12] F. Bressan i R. Hess, „Non-Standard Track Configuration in Historical Audio Recordings: Technical and Philological Consequences for Preservation", *Fontes Artis Musicae*, 2020.

[13] M. Lagrange i F. Gontier, „Bandwidth Extension of Musical Audio Signals With No Side Information Using Dilated Convolutional Neural Networks", *IEEE*, kwi. 2020.

[14] X. Peng, „Hearing the Opera: "Teahouse Mimesis" and the Aesthetics of Noise in Early Jingju Recordings, 1890s—1910s ", *University of Hawai'i Press*, lip. 2017.

[15] G. Barlindhaug, „Artificial Intelligence and the Preservation of Historic Documents", *University of Tromsø*, 2022.

[16] P. Flanders, „Remanence loss in γ-Fe2O3tapes as related to reptation, viscosity, and print-through", *IEEE*, sty. 2003.

[17] J. Banks, „Brahms's Hungarian Dances and the Early 'Csárdás' Recordings", *Oxford University Press*, lip. 2021.

[18] A. Bowsher, *18 11-, 12-, and 13½-Bar Blues: Time and African American Country Blues Recordings (1925—1938)*. 2021.

[19] C. Ngoasheng, M. Ngoepe, i N. Marutha, „Sounds like a broken record: preservation and access of audio-visual records at the South African broadcasting corporation radio", *University of South Africa*\, cze. 2021.

[20] L. De Marchi i J. Ladeira, „Digitization of music and audio-visual industries in Brazil: new actors and the challenges to cultural diversity", *Les Cahiers d'Outre-Mer*, 2018.

[21] H. Stančić, A. Rajh, i M. Jamić, „Impact of ICT on archival practice from the 2000s onwards and the necessary changes of archival science curricula", *University of Zagreb*, lip. 2017.

[22] S. Canazza, E. Schubert, A. Chmiel, N. Pretto, i A. Rodà, „The Magnetic Urtext: Restoration as Music Interpretation", *Frontiers in Psychology*, kwi. 2022.

[23] S. Čubrilović, Z. Kuzmanović, i G. Kvaščev, „Audio Denoising using Encoder-Decoder Deep Neural Network in the Case of HF Radio", *The Department of Signals and Systems, School of Electrical Engineering, Belgrade, Serbia*, kwi. 2024.

[24] C. Zou, S. Rhee, H. Lin, D. Chen, i X. Yang, „Sounds of History: A Digital Twin Approach to Musical Heritage Preservation in Virtual Museums", *MDPI*, kwi. 2024.

[25] F. Bressan i R. Hess, „Non-Standard Track Configuration in Historical Audio Recordings: Technical and Philological Consequences for Preservation", *Fontes Artis Musicae*, 2020.

[26] N. Nurmusabih i A. Naufal, „Exploring the interpretation techniques in Bach's Cello Suite No. 1 Prelude: A theoretical and practical approach to baroque performance practice", *Indonesian Jorunal of Music Research, Development and Technology*, 2024.

[27] M. Kowal i M. Woźniak, „ Quantitative Method as a Tool in Musicological Interdisciplinary Research of Musical Rhetoric ", *Interdisciplinary Studies in Musicology*, cze. 2024.

[28] C. Martin, „Integrating Mobile Music with Percussion Performance Practice", *dblp*, 2013.

[29] A. Mattes, „What Else Can Grieg's Historical Recordings Tell Us?", *Studia Musicologica Norvegica*, lis. 2020.

[30] S. Bannister *i in.*, „Muddy, muddled, or muffled? Understanding the perception of audio quality in music by hearing aid users", *Sec. Auditory Cognitive Neuroscience, Frontiers in Psychology*, luty 2024.

[31] K. İnce, M. Madzhıdov, i H. Turhan, „ Zeybeks in Greek music culture in Türkiye: the example of Imroz-Gokceada ", *dergipark*, cze. 2024.

[32] T. Saeki, S. Takamichi, T. Nakamura, N. Tanji, i H. Saruwatari, „SelfRemaster: Self-Supervised Speech Restoration for Historical Audio Resources", *IEEE*, 2023.

[33] B. Morgan, „Revenue, access, and engagement via the in-house curated Spotify playlist in Australia", *Taylor & Francis Online*, grudz. 2018.

[34] A. Stefaniak, „Remixing Multimovement Works, Classical Music Concept Albums, and Twenty-First-Century Pianists' Interpretations of the Canon ", *Washington University in Saint Louis*, lip. 2023.

[35] G. Carfoot, J. Willsteed, i A. Arthurs, „Nostalgia, authenticity and the culture and practice of remastering music", *Queensland University of Technology*, cze. 2019.

[36] V. Danylov, „OPEN SOURCE AND PROPRIETARY SOFTWARE FOR AUDIO DEEPFAKES AND VOICE CLONING: GROWTH AREAS, PAIN POINTS, FUTURE INFLUENCE", *Baltija Publishing*, kwi. 2024.

[37] C. Steinmetz, T. Walther, i J. Reiss, „High-Fidelity Noise Reduction with Differentiable Signal Processing", *arXiv*, paź. 2023.

[38] J. Su, Y. Wang, A. Finkelstein, i Z. Jin, „Bandwidth Extension is All You Need", *IEEE*, maj 2021.

[39] F. Miotello, M. Pezzoli, L. Comanducci, F. Antonacci, i A. Sarti, „Deep Prior-Based Audio Inpainting Using Multi-Resolution Harmonic Convolutional Neural Networks", *IEEE*, paź. 2023.

[40] M. Gogate, K. Dashtipour, i A. Hussain, „Robust Real-time Audio-Visual Speech Enhancement based on DNN and GAN", *IEEE*, luty 2024.

[41] Y. He, K. Seng, i L. Ang, „Generative Adversarial Networks (GANs) for Audio-Visual Speech Recognition in Artificial Intelligence IoT ", *mdpi*, wrz. 2023.

[42] K. Mehta i S. Thakur, „Artistic Renaissance: Harnessing Artificial Intelligence for Cultural Restoration", *International Journal of Advanced Research in Science, Communication and Technology*, sty. 2022.

[43] P. Szczotka, „Projekt i implementacja wybranych algorytmów sztucznej inteligencji w procesie rekonstrukcji nagrań fonicznych", 2023.

[44] E. Moliner i V. Välimäki, „BEHM-GAN: Bandwidth Extension of Historical Music Using Generative Adversarial Networks", *IEEE*, cze. 2022.

[45] S. Lattner i J. Nistal, „Stochastic Restoration of Heavily Compressed Musical Audio Using Generative Adversarial Networks", *MDPI*, cze. 2021.

[46] A. Wilson i B. Fazenda, „Perception of Audio Quality in Productions of Popular Music", *Acoustics Research Centre, University of Salford, Greater Manchester, UK*, sty. 2016.

[47] A. Nogales, S. Donaher, i A. García-Tejedor, „A deep learning framework for audio restoration using Convolutional/ Deconvolutional Deep Autoencoders", *CEIEC*, mar. 2022.

[48] Z. Cheddad i A. Cheddad, „Active Restoration of Lost Audio Signals using Machine Learning and Latent Information", *Blekinge Institute of Technology, Universit´e Frères Mentouri*, sty. 2024.

[49] D. Jinhui, Z. Yue, Pengcheng X., i X. Xinzhou, „Super-Resolution for Music Signals Using Generative Adversarial Networks", 2021.

[50] O. Casey, R. Dave, N. Seliya, i E. Sowells Boone, „Challenges, Limitations, and Compatibility for Audio Restoration Processes", *University of Wisconsin-Eau Claire*, 2021.

[51] B. Lohani, C. Gautam, P. Kushwaha, i A. Gupta, „Deep Learning Approaches for Enhanced Audio Quality Through Noise Reduction", *IEEE*, maj 2024.

[52] G. Labrèche, C. Guzman, i S. Bammens, „Generative AI... in Space! Adversarial Networks to Denoise Images Onboard the OPS-SAT-1 Spacecraft", *IEEE*, mar. 2024.

[53] E. Moliner, F. Elvander, i V. Välimäki, „Blind Audio Bandwidth Extension: A Diffusion-Based Zero-Shot Approach", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, sty. 2024.

[54] M. Mrak, „AI Gets Creative", *Association for Computing Machinery*, paź. 2019.

[55] A. Gupta, „Restoring and reconstructing old and damaged compositions using AI to preserve the musical heritage", *medium.com*, lip. 2023, [Online]. Dostępne na: https://medium.com/illumination/restoring-and-reconstructing-old-and-damaged-compositions-using-ai-to-preserve-the-musical-heritage-b5eeb20b0039

[56] T. Bertin-Mahieux, G. Grindlay, R. Weiss, i D. Ellis, „Evaluating music sequence models through missing data", *IEEE*, maj 2011.

[57] J. Marus Coldenhoff i Z. Ren, „Music Super-resolution with Spectral Flatness Loss and HiFi-GAN Discriminators", *EPFL, Switzerland*, 2021.

[58] Q. Coleman, „MACHINE LEARNING APPROACHES TO HISTORIC MUSIC RESTORATION", 2021.

[59] O. Mohammed i B. Mahmood, „UNIVERSAL IMAGE AND AUDIO RESTORATION USING DEEP LEARNING", *Springer*, 2021.

[60] H. Zhao, „A GAN Speech Inpainting Model for Audio Editing Software", *Proc. INTERSPEECH 2023*, 2023.

[61] K. Mizuta, T. Koriyama, i H. Saruwatari, „Harmonic WaveGAN: GAN-Based Speech Waveform Generation Model with Harmonic Structure Discriminator", *Proc. Interspeech 2021*, 2021.

[62] P. Marry, A. Preethi, K. Bhuvan, A. Ravali, i C. Linesh, „A Dual-Step-U-Net for Crystal-Clear Restoration of Audio Recordings", *3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India*, 2023.

[63] C. Jin, W. Zhao, i H. Wang, „Research on Objective Evaluation of Recording Audio Restoration Based on Deep Learning Network", *Advances in Multimedia*, 2018.

[64] N. Torres-Reyes i S. Latifi, „Audio Enhancement and Synthesis using Generative Adversarial Networks: A Survey", *International Journal of Computer Applications*, 2019.

[65] P. Sharma, M. Kumar, i H. Sharma, „Generative adversarial networks (GANs): Introduction, Taxonomy, Variants, Limitations, and Applications", *Springer*, mar. 2024.

[66] S. Kim i V. Sathe, „Bandwidth Extension on Raw Audio via Generative Adversarial Networks", *arXiv*, 2019.

[67] H. Ashraf, Y.-S. Jeong, i C. H. Lee, „Improved CycleGAN for underwater ship engine audio translation", *The Journal of the Acoustical Society of Korea*, lip. 2020.

[68] R. Agrawal, K. Sharma, S. Gonge, R. Joshi, i D. Singh, „Towards the Applications of Generative Adversarial Networks Beyond Images", *International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, 2023.

[69] M. Baas i H. Kamper, „GAN You Hear Me? Reclaiming Unconditional Speech Synthesis from Diffusion Models", *IEEE Spoken Language Technology Workshop (SLT)*, 2022.

[70] S. Gupta, „MusicNet Dataset", kaggle.com. [Online]. Dostępne na: https://www.kaggle.com/datasets/imsparsh/musicnet-dataset

[71] „torchvggish", github.com. [Online]. Dostępne na: https://github.com/harritaylor/torchvggish