prof. dr hab. inż. Andrzej Materka
Politechnika Łódzka
Instytut Elektroniki

Łódź, 12 października 2023 r.

Recenzja rozprawy doktorskiej

opracowana na zlecenie Rady Dyscypliny Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne Politechniki Wrocławskiej

Tytuł rozprawy	Automatyczna klasyfikacja wybranych zniekształceń sygnałów muzycznych z wykorzystaniem konwolucyjno-rekurencyjnych sieci neuronowych bez porównania do sygnału referencyjnego
Autor rozprawy	mgr inż. Kamila Organiściak
Promotor	prof. dr hab. inż. Józef Borkowski
Dyscyplina	automatyka, elektronika, elektrotechnika i technologie kosmiczne

Recenzję opracowałem odpowiadając na pismo RDN AEETK/128/2023 z 31 lipca 2023 r. dostarczone mi 20 września 2023 r. za pośrednictwem Działu Kancelaryjno-Archiwalnego Politechniki Łódzkiej. Podstawą prawną recenzji jest ustawa Prawo o szkolnictwie wyższym i nauce z 20 lipca 2018 r. (Dz. U. z 2023 r. poz. 742).

Przedmiotem pracy jest poszukiwanie skutecznych metod automatycznej detekcji i klasyfikacji wybranych zniekształceń sygnałów muzycznych, bez dostępu do sygnału niezniekształconego. Sygnały muzyczne powstają w rezultacie ciągu operacji analogowo-cyfrowych, od nagrania przez zapis na różnego rodzaju nośnikach po transmisję i odtwarzanie za pomocą urządzeń elektroakustycznych. Każda z tych operacji może wprowadzać zmiany przebiegu sygnału, mogące skutkować jego zniekształceniami prowadzącymi do obniżenia jakości. Ocena jakości sygnałów jest niezbędna w wielu aspektach, m.in. dla identyfikacji źródła zniekształceń i przeprowadzenia korekcji samego sygnału (gdy to możliwe) lub zmodyfikowania operacji skutkującej nieakceptowanymi zmianami własności przebiegów. Potrzeba automatyzacji takiej oceny jest uzasadniona jej dużym kosztem i czasochłonnością oraz koniecznością rekrutacji i szkolenia ekspertów do badań odsłuchowych.

Odbiorcą sygnałów akustycznych jest złożony zmysł słuchu człowieka. Powszechnie stosowane algorytmy stratnej kompresji sygnałów wykorzystują własności tego zmysłu, np. redukują poziom składowych subiektywnie maskowanych w dźwiękach. Próbą obiektywizacji oceny jakości zapisów dźwięku jest porównanie badanego sygnału do jego wersji niezniekształconej (do sygnału referencyjnego), z uwzględnieniem modelu psychoakustycznego. W praktyce nierealistyczne jest oczekiwanie dostępu do takiej wersji sygnału. Autorka ocenianej pracy podjęła w związku z tym zadanie zbadania możliwości opracowania algorytmu detekcji i klasyfikacji wybranych zniekształceń próbek sygnału,

którym nie towarzyszą przebiegi oryginalne – nieprzetworzone, niezniekształcone. Jest to zadanie ambitne, które nie zawsze może mieć rozwiązanie, np. w odniesieniu do pewnych gatunków muzycznych. Nie ma jednak wątpliwości, że w wielu praktycznych przypadkach detekcja i klasyfikacja zniekształceń w ujęciu przedstawionym w dysertacji może przynieść istotne korzyści. Temat rozprawy jest w związku z tym ważny dla praktyki, jest również interesujący pod względem poznawczym. Jest to temat aktualny – podjęte zagadnienie naukowe było przedmiotem badań odnoszących się do sygnału mowy. Automatyczna ocena jakości sygnałów muzycznych z wykorzystaniem sztucznych sieci neuronowych, w sposób obiektywny, bez odniesienia do sygnału pozbawionego zniekształceń, ma charakter pionierski.

We wprowadzeniu Autorka zwięźle charakteryzuje istniejące metody oceny jakości sygnałów muzycznych. Na str. 13 przedstawia cel oraz tezę rozprawy i omawia zakres opisanych badań. Autorka twierdzi, że skuteczność klasyfikacji wybranych zniekształceń sygnału muzycznego [z wykorzystaniem sztucznych sieci neuronowych, bez sygnału odniesienia] można znacząco zwiększyć przez zastosowanie dwukierunkowych warstw rekurencyjnych, odpowiedniej architektury i pobudzenia wejść, a także "opracowania bazy zniekształceń do celów ewaluacji modelu" [sieci]. Pierwsze trzy części tego twierdzenia są klarowne. Wątpliwości budzi dobór bazy danych do celów ewaluacji, który miałby poprawiać skuteczność klasyfikacji. Być może w ujęciu Autorki ewaluacja modelu nie jest oceną modelu, która powinna być przeprowadzona za pomocą danych zebranych niezależnie od głównego zadania, czyli poprawnej klasyfikacji. Proszę o skomentowanie zamierzonego znaczenia tej części tezy w czasie obrony rozprawy.

Praca ma charakter teoretyczno-doświadczalny. Obejmuje wprowadzenie, krytyczny przegląd stanu wiedzy, sformułowanie hipotez dotyczących możliwości obiektywizacji oceny jakości sygnałów muzycznych, propozycję własnych rozwiązań, opracowanie stosownych algorytmów i programów komputerowych, zaplanowanie i przeprowadzenie eksperymentów, analizę i dyskusję wyników. Treść rozprawy jest podzielona na 11 rozdziałów, logicznie odpowiadających kolejnym etapom przeprowadzonych badań, zawiera też streszczenia, wykazy skrótów i oznaczeń oraz wykaz literatury.

Wprowadzenie do tematyki rozprawy i analiza stanu wiedzy zostały opracowane na podstawie literatury, głównie publikacji anglojęzycznych z lat dwutysięcznych. Całkowita liczba cytowanych źródeł jest znaczna – wynosi 163 (faktycznie 162: artykuł konferencyjny [161] został wymieniony w bibliografii dwukrotnie). Autorka ze swobodą korzysta z zawartych w nich informacji. Przegląd literatury jest krytyczny i szczegółowy, zarówno w zakresie podstaw teoretycznych jak i efektów automatycznej oceny zniekształceń. Przeprowadzona przez mgr Organiściak dyskusja zawartości źródeł świadczy o tym, że posiada ona wiedzę niezbędną do prowadzenia badań naukowych mieszczących się w zakresie dyscypliny automatyka, elektronika, elektrotechnika i technologie kosmiczne. Brakuje w rozprawie odwołań do literatury polskojęzycznej. Na przykład, umieszczenie na str. 50 odwołania do blisko związanych z tematem rozprawy monografii [R1, R2] zamiast do trudno dostępnego dzieła poświęconego zastosowaniom inteligencji obliczeniowej w ochronie zdrowia [135]

pomogłoby czytelnikowi zrozumieć istotę enigmatycznej opinii Autorki zawartej w stwierdzeniu "zastosowanie skali melowej podczas analizy sygnałów audio pozwala na dokładniejszą symulację rzeczywistej percepcji słuchu ludzkiego".

W celu udowodnienia słuszności tez Autorka opracowała bazę sygnałów muzycznych zniekształconych w zamierzony sposób i zaplanowała eksperymenty numeryczne obejmujące uczenie i testowanie sztucznych sieci neuronowych – splotowych i rekurencyjnych – pobudzonych parametrami uzyskanymi drogą przetwarzania przebiegów czasowych zapisanych w tej bazie oraz ich spektrogramów.

Rozwinięcie zagadnienia oceny jakości sygnałów muzycznych w rozdziale 1 zawiera informacje o zjawiskach psychoakustycznych mających wpływ na odbiór dźwięku przez człowieka, a także zmianach sygnału będących efektem podstawowych operacji cyfrowego przetwarzania i transmisji. Rozdział 2 zawiera charakterystykę subiektywnych testów odsłuchowych i dyskusję ich ograniczeń. Rozdział 3 jest wprowadzeniem do automatycznych testów jakości dźwięku, z podziałem na metody korzystające z sygnałów odniesienia i nieliczne metody nie wymagające tej dodatkowej informacji. Scharakteryzowano najbardziej popularne miary jakości definiowane w dziedzinie czasu i dziedzinie częstotliwości. Zwięzły podrozdział 3.4 jet przeglądem literatury na temat zastosowania sztucznych sieci neuronowych do przetwarzania i analizy sygnałów muzycznych. Zestawione tabelarycznie prace są ilustracją dużej skuteczności sieci neuronowych w automatyzacji takich zadań jak klasyfikacja emocji, systemy rekomendacji nagrań, generowanie muzyki, czy rekonstrukcja sygnału poddanego kompresji stratnej. Metody nie wymagające referencji zostały rozwinięte w odniesieniu do oceny jakości sygnału mowy. Celem badań Autorki była w związku z tym ocena skuteczności sztucznych sieci neuronowych w klasyfikowaniu zniekształceń sygnałów muzycznych, bez dostępu do sygnału odniesienia. W rozdziale 4 mgr Organiściak naszkicowała zwięźle swoją metodę klasyfikacji, ograniczając liczbę kategorii zniekształceń do czterech: związanych z kwantyzacją sygnału, charakterystyką wzmocnienia, obecnością dodatkowych dźwięków i brakiem oczekiwanych składowych. Z opisanych badań wyłączono również muzyczne sygnały wielokanałowe. Ten krótki rozdział zawiera również plan działań potrzebnych do implementacji metody i oceny jej własności. Ich opis oraz uzyskane rezultaty rozwinięto szczegółowo w kolejnych rozdziałach.

Rekomendacja ITU wymienia 11 rodzajów zniekształceń w odniesieniu do oceny jakości sygnałów akustycznych. Wprowadzone ograniczenie ich liczby jest uzasadnione pionierskim charakterem opisywanych badań. Z drugiej strony, nieuwzględnionym w rozprawie rodzajem zniekształceń, stosunkowo łatwo identyfikowanych przez człowieka są zniekształcenia nieliniowe. Proszę o komentarz w czasie obrony na temat przewidywanej skuteczności opracowanej metody w zastosowaniu do detekcji tego typu zmian kształtu sygnału.

Podstawą autorskiej bazy sygnałów przygotowanych na potrzeby opisywanego projektu doktorskiego był publicznie dostępny zbiór MUSDB18 sygnałów, przetwarzanych z szybkością 44100 próbek/s i zapisywanych bez kompresji w słowach szesnastobitowych, opracowany przez amerykańsko-francusko-niemiecko-brytyjskie konsorcjum [125]. Dostępne w bazie

MUSDB18 sygnały zostały uznane za niezniekształcone. Część z nich została zmodyfikowana w celu utworzenia pięciu grup przebiegów do uczenia, walidacji oraz testowania sztucznych sieci neuronowych. Niestety, lapidarny opis procesu modyfikacji dostępny na str. 46 nie precyzuje zakresu wprowadzonych zmian. Przedstawione w rozprawie informacje nie wystarczają do replikacji badań Doktorantki przez innych badaczy. Na przykład stwierdzenie, że "modyfikacja głośności polegała na zmianie zakresu dynamicznego wybranych fragmentów, przy czym każdy skok poziomu sygnału trwał co najmniej 20 ms" nie charakteryzuje wszystkich działań potrzebnych do utworzenia sygnałów kategorii 1. Jakiej wartości "skok" został uznany za progowy poziom zniekształceń? Jakie były maksymalne wartości takiego skoku? Czy wartości te były wybierane losowo? Czy chwile ich występowania też były losowane? Jaki był rozkład prawdopodobieństwa losowanych wielkości? Jaki zestaw zmian sygnału był przez Autorkę uznawany za wystarczający do przyporządkowania przebiegu do wybranej grupy sygnałów zniekształconych? Podobne pytania sformułowane w odniesieniu do pozostałych trzech kategorii przebiegów także pozostają bez odpowiedzi. Proszę o uzupełnienie potrzebnych informacji w trakcie obrony rozprawy.

Kolejne części rozdziału 5 są poświęcone encyklopedycznym informacjom na temat algorytmów wstępnego przetwarzania sygnałów oraz ekstrakcji cech sygnału – w dziedzinie czasu i częstotliwości. W rozdziale 6 zawarte są podstawowe informacje na temat zastosowanych architektur sztucznych sieci neuronowych – splotowych i rekurencyjnych dwukierunkowych. W opisanym projekcie sieci są pobudzane spektrogramami sygnałów, a w kolejnych jego etapach także parametrami obliczanymi na podstawie przebiegów w dziedzinie czasu i ich widma. Sieci pełnią funkcję klasyfikatorów (rozdz. 8, 9), a także służą ekstrakcji cech, o które powiększano sekwencje danych wejściowych w procesie ulepszania opracowanej metody (rozdz. 10). Wszystkie węzły sieci zawierały funkcje aktywacji "ReLU", z wyjątkiem warstw wyjściowych, w których zastosowano funkcję "softmax" dla przypadku pięciu rozpoznawanych klas. Funkcją kosztu minimalizowanego podczas uczenia była entropia krzyżowa.

Przeprowadzony dowód słuszności hipotezy naukowej sformułowanej przez Autorkę obejmuje etapy nadzorowanego uczenia sieci (z walidacją) oraz testowania. Metody te są odpowiednie do badania własności statystycznych systemów uczących się. Praca nie zawiera informacji na temat kryteriów doboru przykładów tworzących potrzebne trzy zbiory danych. Wiadomo jaka była liczność tych zbiorów (tab. 5.1.), ale nie wiadomo jak zadbano o niezależność przykładów. Ta kwestia wiąże się z pytaniami, które zadałem w akapicie na str. 3-4 recenzji.

Dokładność opracowanych klasyfikatorów została oceniona ilościowo za pomocą statystyk definiowanych dla klasyfikacji binarnej. Jest to pewne uproszczenie - problem klasyfikacji obejmuje 5 kategorii. Zamieszczenie macierzy błędów wszystkich testów dostarczyłoby więcej informacji dot. badanych klasyfikatorów. Nadmiernie zwięzły styl raportu z badań dodatkowo utrudnia ocenę otrzymanych wyników. Na przykład trudno dociec jak uzyskano "wynik ogólny (wspólny dla wszystkich kategorii)", str. 69, albo jak przeprowadzono test

pozwalający obliczyć wartości dane wzorem (7.2) dla "klasy negatywnej (nieprzynależącej do danej kategorii)", str. 70. Czy test był zrównoważony? Niektóre określenia są niepoprawne, jak np. "klasa negatywna", "klasa pozytywna". Anglojęzyczne terminy wybrane jako nazwy miar jakości klasyfikacji mają polskie odpowiedniki, np. *accuracy* to dokładność a *specificity* to swoistość. Nie uzasadniono tego wyboru, co jest niezbędne w odniesieniu do dzieła napisanego po polsku.

Częściowe wyniki pracy badawczej mgr Organiściak zostały opublikowane w recenzowanym artykule [118] na łamach indeksowanego czasopisma naukowego *Metrology and Measurement Systems,* umieszczonego w wykazie opublikowanym przez Ministra Edukacji i Nauki (1 grudnia 2021 r.). Doktorantka jest pierwszym autorem tego artykułu. Tym samym spełnione jest ustawowe wymaganie zapisane w art. 186 ust. 1 p. 3a dotyczące dorobku osoby ubiegającej się o nadanie stopnia doktora.

Struktura rozprawy w zasadzie odpowiada oczekiwanej kompozycji opracowań naukowych, choć nadmiernie zwięzły styl i skróty pojęciowe nie pozwalają czytelnikom uzyskać wszystkich informacji potrzebnych do replikacji przeprowadzonych badań i porównania wyników z rezultatami innych projektów. Zaznaczyłem to w treści recenzji. Poniżej zamieszczam moje komentarze dotyczące treści pracy oraz uwagi odnoszące się do niektórych słów i sformułowań niedokładnie odpowiadających sytuacji, w których zostały użyte.

1) Strona 8: Umieszczenie skrótów OBSC oraz ZCR w wykazie ułatwiłoby lekturę rozprawy.
2) Strony 9, 57: Zwrot „oczekiwane oznaczenie sygnału w bazie danych wejściowych" jest niezrozumiały. Oznaczenie jest oczekiwane, czyli go nie ma? Może jest to raczej etykieta sygnału w rzeczonym zbiorze? Jakie wartości może przyjmować?
3) Strona 9, 49 i inne: krótko-czasowa→krótkoczasowa
4) Strony 11, 26 i inne: niemożliwym jest zrekonstruowanie→niemożliwe jest zrekonstruowanie, istotnym jest fakt→istotny jest fakt
5) Strona 26 i inne: Określenie „manualne testy odsłuchowe" jest mylący, sugeruje wykonywanie testów z użyciem rąk. Przypuszczam, że intencją były testy nieautomatyczne. Według mnie przymiotnik „odsłuchowe" wystarczająco określa naturę testów.
6) Strona 42 i inne: Zdanie „Wykorzystując spektrogramy ... problem ... może być sformułowany jako ... ." jest przykładem błędnego użycia imiesłowu współczesnego. Problem nie jest podmiotem i nie wykonuje czynności wykorzystania spektrogramów równolegle z czynnością formułowania samego problemu. Podmiot obu części zdania powinien być ten sam albo należy użyć formy osobowej w takich przypadkach.
7) Strona 50: Zależność dana wzorem (5.2) jest nieliniową funkcją częstotliwości, ale co opisuje i do czego jest wykorzystana „funkcja melowa" posiadająca cechę nieliniowości opis nie wyjaśnia.
8) Strona 52: Na czym polega „nałożenie krzywej ważonej K"? Co reprezentuje ta krzywa (będąca tworem geometrycznym jako krzywa) i na co jest „nakładana"?

9) Strona 64: interpretacja fonemów może być bardziej skuteczna znając ...: Patrz uwaga do tekstu na str. 42.
10) Strona 64: Jak zdefiniowano „wartość odwrotną"?
11) Strona 68: Schemat przetwarzania informacji w warstwie wyjściowej klasyfikatora nie wydaje się być przykładem ilustrującym działanie funkcji „softmax". Jak w naszkicowanej strukturze sieciowej odbywałaby się normalizacja zapisana wzorem w środkowej kolumnie i ostatnim wierszu tab. 6.1.? Zaznaczenie wielkości z_i oraz y_i na schemacie byłoby pomocne...
12) Strona 70: Treść sekcji 7.3. nie zawiera dyskusji wymagań dotyczących sprzętu obliczeniowego niezbędnego do przeprowadzenia opisywanych badań, niezgodnie z tytułem tej sekcji („Wymagania sprzętowe"). Podano w niej informacje o wybranym przez Autorkę zestawie urządzeń, bez uzasadnienia wyboru. Czy projekt mógłby zostać zrealizowany bez dostępu do serwisu AWS? Jakie byłyby konsekwencje takiej decyzji? Czy niezbędne było wykorzystanie biblioteki *TensorFlow* czy też obliczenia mogły zostać wykonane za pomocą funkcji dostępnych w module *PyTorch*?
13) Strona 71: Skrót EBS w tab. 7.1. nie został wyjaśniony.
14) Zamieszczenie wykresów wartości funkcji strat w zbiorze uczącym i walidacyjnym w funkcji indeksu iteracji podczas uczenia pozwoliłoby czytelnikom porównać ich prace z badaniami Autorki.

Mam nadzieję, że zawarte w mojej recenzji komentarze i uwagi okażą się przydatne w kolejnych badaniach naukowych i publikacjach Autorki.

Magister inżynier Kamila Organiściak opracowała oryginalne rozwiązanie istotnego problemu naukowego w dziedzinie nauk inżynieryjno-technicznych. Kandydatka posiada wiedzę teoretyczną i praktyczną potrzebną do prowadzenia badań naukowych w dyscyplinie automatyka, elektronika, elektrotechnika i technologie kosmiczne. Stwierdzam w związku z tym, że oceniana rozprawa doktorska spełnia wymagania Ustawy z 20 lipca 2018 r. o stopniach naukowych i tytule naukowym... (Dz. U. z 2023 r. poz. 742).

Andrzej Materka

[R1] R. Makowski, Automatyczne rozpoznawanie mowy - wybrane zagadnienia, Oficyna Wydawnicza Politechniki Wrocławskiej, 2011
[R2] T. Zieliński, Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań. Wydawnictwa Komunikacji i Łączności, 2005