prof. dr hab. inż. Jan Żera
Instytut Radioelektroniki
i Technik Multimedialnych
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
ul. Nowowiejska 15/19
00-665 Warszawa
tel. 504-675179
e-mail: jan.zera@pw.edu.pl

RECENZJA
rozprawy doktorskiej mgr inż. Mariusza Klecia
pt. ,,Efektywna reprezentacja danych w systemach przetwarzania sygnałów dźwiękowych"

Podstawą wykonania recenzji jest pismo prof. dr hab. Marii Elżbiety Orłowskiej, Przewodniczącej Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja, Polsko-Japońskiej Akademii Technik Komputerowych z dnia 6 listopada 2024 roku.

Tematyka rozprawy doktorskiej mieści się w zakresie tematycznym Dyscypliny Informatyka Techniczna i Telekomunikacja. Praca została wykonana na Wydziale Informatyki PJATK. Promotorem rozprawy była dr hab. Alicja Wieczorkowska, a promotorem pomocniczym dr hab. inż. Krzysztof Szklanny.

Rozprawa doktorska mgr inż. Mariusza Klecia, jest cyklem następujących pięciu publikacji, w tym czterech publikacji współautorskich i jednej autorskiej:

A1. Kleć, M., Szklanny, K., Wieczorkowska, A. (2024). Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions. In B. Marcinkowski, A. Przybylek, A. Jarzębowicz, N. Iivari, E. Insfran, M. Lang, H. Linger, and C. Schneider (Eds.), *Harnessing Opportunities: Reshaping ISD in the post-COVID-19 and Generative AI Era* (ISD2024 Proceedings). Gdańsk, Poland: University of Gdańsk.

A2. Kleć, M., Korzinek, D. (2015). Pre-trained deep neural network using sparse autoencoders and scattering wavelet transform for musical genre recognition. Computer Science, 16 (2), 133–144.

A3. Kleć, M. (2018). Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation. In: Ceci, M., Japkowicz, N., Liu, J., Papadopoulos, G., Raś, Z. (Eds.) *Foundations of Intelligent Systems*. ISMIS 2018. Lecture Notes in Computer Science, vol 11177, Springer.

A4. Kleć, M., Wieczorkowska, A., Szklanny, K., Strus, W.: Beyond the Big Five personality traits for music recommendation systems. J. Audio Speech Music Proc. 2023, 4 (2023).

A5. Kleć, M., Wieczorkowska, A. (2021). Music Recommendation Systems: A Survey. In: Ras, Z.W., Wieczorkowska, A., Tsumoto, S. (eds) Recommender Systems for Medicine and Music. Studies in Computational Intelligence, vol 946, Springer.

Wspomniane publikacje są publikowane w czasopismach, lub monografiach pokonferencyjnych, które aktualnie są punktowane w ramach dorobku naukowego odpowiednio przez 140, 40, 20, 100 punktów (cztery pierwsze, A5 pomijam). Przy znacznej rozpiętości punktów od 20 do 140 należy uznać, że publikacje odpowiadają sumarycznej

punktacji ponad 300 punktów, co jest satysfakcjonujące w odniesieniu do publikacji pracy rangi rozprawy doktorskiej.

Ponadto, w dorobku doktoranta znajdują się jeszcze trzy publikacje konferencyjne na tematy ogólnie związane z tematyką rozprawy. Publikacje stanowiące osiągnięcie naukowe są opatrzone Autoreferatem o objętości 18 stron i 46 pozycjami bibliografii. Odniesienia bibliograficzne w publikacjach osiągnięcia naukowego liczą sumarycznie 265 pozycji (bez eliminacji powtórzeń).

Całość badań autora ma jeden wspólny cel – pokazanie, że systemy przetwarzania dźwięku, głównie oparte na stosowaniu wytrenowanych sieci neuronowych, można znacznie usprawnić w sensie skuteczności ich działania, przez włączenie w danych lub w formie dodatkowej informacji kontekstowej tych elementów, które zazwyczaj były pomijane przy realizacjach opartych na przetwarzaniu 'laboratoryjnie czystych' sygnałów. Podejście to zwraca uwagę na to, że przy realnym rozpoznawaniu obiektów dodatkowe zakłócenia istnieją, i jako takie powinny wchodzić w zakres systemu przetwarzania, lub powinny być uwzględnione w formie informacji dodatkowej.

Cel pracy jest sformułowany w postaci trzech tez rozprawy, z których pierwsza odnosi się do korpusów mowy i postuluje włączenie do zbiorów trenujących zjawisk związanych z pogłosem, szumami tła i przypadkowymi dźwiękami zakłócającymi standardowo uważanymi za powodujące wadliwość próbek trenujących. Teza druga odnosi się do dźwięków o pochodzeniu medycznym wykorzystywanych w modelach rozpoznawania wad serca. Teza ta postuluje wprowadzenie zmienności czasów obserwacji w postaci zmienności długości ramek czasowych analizy, a więc sugeruje elastyczność parametrów metod analizy. Wreszcie teza trzecia dotyczy uwzględniania informacji dodatkowej spoza materiału wejściowego systemu klasyfikacyjnego – w tym przypadku przy klasyfikacji w systemach rekomendacji muzycznej – sugerując wzięcie pod uwagę dodatkowej wiedzy o cechach osobowości użytkowników. Autorzy posługują się tym celu klasycznym modelem temperamentów obecnym w psychologii jako model *Wielkiej Piątki*. Należy podkreślić, że z pozoru przedstawione tezy odnoszą się do obszarów zasadniczo różnych i mogłyby stanowić tematykę różnych prac. Jednakże traktowane jako przykłady mogą tworzyć jedną całość wskazującą na zasadność sformułowanego ogólnego celu i postulowania konieczności uwzględniania rzeczywistych uwarunkowań sygnałów podlegających klasyfikacji metodami AI, w tym przypadku z użyciem sieci neuronowych.

Krótka charakterystyka treści zawartych w przedmiotowych artykułach jest następująca:

Artykuł A1—

Artykuł A1 stanowi propozycję tzw. korpusu nagrań mowy polskiej, co samo w sobie jest ważnym elementem, jako że absolutna większość prac z zakresu rozpoznawania mowy występujących w literaturze w naturalny sposób odnosi się do języka angielskiego. Autorzy zaproponowali włączenie w drodze symulacji do nagrań mowy polskiej zjawisk związanych z występowanie pogłosu, dźwięków tła, oraz zakłóceń w postaci tego, co nazwane jest nieplanowanymi zdarzeniami dźwiękowymi. Symulacja powstaje przez działania na osobno przechowywanych komponentach, czyli mowy, dźwięków tła, zakłóceń i pogłosu pomieszczenia w osobnych plikach dźwiękowych, dostępnych dodatkowo. Pozwala to na stworzenie naturalnych danych trenujących/ testujących na rzecz badań w drodze odpowiedniego dodawania i odejmowania warstw. Dla takich danych przeprowadzono (w architekturze sieci Conv-TasNet działającej w dziedzinie czasu) konstrukcję modeli wytrenowanych do separacji pojedynczego mówcy od dźwięków tła, do separacji dwóch

mówców od siebie w warunkach bez szumu i z szumem tła. Natomiast dowodzenie tezy pierwszej pracy przeprowadzono przez zastosowanie kombinacji obecności udziału czynnika hałasu tła, zakłóceń i pogłosu w zbiorach trenujących i testujących, samego hałasu tła, następnie z dodaniem pogłosu i dodaniem dźwięków zakłócających. Ogólnie wykazano, że brak czynnika zakłócającego w zbiorze trenującym obniżał skuteczność modelu przy obecności tego czynnika w zbiorze testującym, natomiast – co istotne – jego obecność w mniejszym znacznie stopniu skutkował w zmniejszenia poziomu rozpoznawania innych czynników w zbiorze testującym. W tym kontekście można uznać stwierdzenie podsumowujące tę część pracy odnoszące się do potwierdzenia pierwszej tezy pracy za prawdziwe.

Artykuł A2—

Należy do najstarszej pracy Autora (2015) i w moim przekonaniu najmniej wnosi do kwestii wykazania przedstawionych w rozprawie tez, jakkolwiek potwierdza tezę drugą wskazującą na znaczący wpływ szczegółowych parametrów analizy na skuteczność klasyfikacji. Praca odnosi się do często podejmowanej kwestii rozpoznawania gatunków muzycznych i jej esencja polega na zastosowaniu autoenkoderów w reprezentacji *scattering wavelet transform* (SWT). Pokazano, że połączenie obu technik pozwala na znaczący wzrost skuteczności rozpoznawania gatunków muzycznych, co skłoniło autorów do rozważania w modelowaniu sekwencji ramek czasowych analizy.

Artykuł A3—

Artykuł nr 3 jest pracą indywidualną doktoranta, w której przeprowadza szczegółową analizę kombinacji rozmiaru filtru splotowego CNN i długości ramki w analizie czasowo-spektralnej SWT. W największym skrócie praca ta pokazuje, że różne rodzaje sygnałów do zinterpretowania w zapisie stetoskopowych dźwięków serca, są w różny sposób skutecznie rozpoznawane w zależności od relacji między rozmiarem filtru CNN i czasem ramki SWT. A zatem praca ta odnosi się bezpośrednio do tezy drugiej stawianej w rozprawie, potwierdzając ją.

Artykuł A4—

Artykuł nr 4 stanowi obszerną pracę obejmującą połączenie klasycznych metod rekomendacji muzycznej według zasad *music information retrieval* (MIR) z oceną emocjonalności wg modelu *Wielkiej Piątki* cech osobowościowych. Ocenianiu profilu emocjonalnego podlegały zarówno osoby biorące udział w badaniach, jak i emocje niesione przez utwory. Dane te połączono z deskryptorami cech typowymi dla MIR. Praca ta jest zasadniczą badawczą odpowiedzią doktoranta na zagadnienie postawione w tezie trzeciej stanowiącej, że system rekomendacji muzycznej uzupełniony danymi osobowościowymi poprawia skuteczność rekomendacji. Wyniki tej pracy są wielowątkowe. Na uwagę zasługuje natomiast spostrzeżenie, że tylko niektóre cechy osobowościowe spośród badanych dwudziestu stanowi istotną część w ramach ocen utworów muzycznych. Pozwala to na znaczne zmniejszenie bazy danych i kwestionariuszy osób badanych.

Artykuł A5—

Artykuł nr 5, zgodnie z deklaracją doktoranta nie wchodzi w zakres dowodzenia tez, lecz jako artykuł przeglądowy stanowi uzupełnienie podając informację o najnowszych działaniach w zakresie systemów rekomendacji muzycznej. Publikacja ta świadczy o dobrym rozeznaniu literaturowym doktoranta w zakresie najnowszej literatury.

Chciałbym skierować dwie ogólne kwestie dyskusyjne do Doktoranta (które nie są bezpośrednią krytyką zawartości pracy):

W poruszanych zagadnieniach jednym ze standardów jest wyznaczanie Mel-Frequency Cepstrum Coefficients (MFCC). Natomiast jest to jedyne, jakie ja znam, praktyczne zastosowanie skali *meli* S.S. Stevensa, która to skala generalnie nie zyskała sobie innych znaczących zastosowań i jest wątpliwa patrząc z perspektywy praktykowanych skal muzycznych. W mojej opinii – choć nie prowadziłem jakichkolwiek badań dla uzyskania szczegółowego rozstrzygnięcia – jest to efekt przypadku, który polega na tym, że skala meli pod względem kształtu funkcji jest tak znacząco podobna do skali filtrów słuchowych w sensie szerokości filtrów ERB (czyli nie pasm krytycznych Zwickera, lecz filtrów R. Pattersona, B. C. J. Moora z Cambridge), że może filtry ERB reprezentować. W symulacjach numerycznych pewne różnice między skalą meli i ERB mogą okazać się nie istotne. Pytanie do doktoranta jest o jego spojrzenie na tę kwestię, przy obecności w literaturze prac porównujących MFCC i ERB.

Druga kwestia jest bardziej ogólna i dotyczy rozpoznawania gatunków muzycznych. Rozpoznawanie to jest dość powszechnym trendem, sterowanym komercjalizacją zastosowań. W pewnym sensie odróżnianie gatunków typu pop jest jednak relatywnie trywialne jeśli porównamy je do zadania – przykładowo – odróżnienia gatunku romantycznej muzyki symfonicznej od romantycznej muzyki operowej. Oczywiście w tym przypadku w zasadzie nie istnieje rynek zastosowań. Natomiast za najciekawsze uznałbym rozpoznawanie formy i stylu muzycznego tak, by można było uzyskać określone powiązania między kompozytorami, albo wręcz uzyskać możliwość wytypowania dla nierozpoznanych utworów określonych potencjalnych autorów z określonym prawdopodobieństwem. Tak kwestia dotyczy szczególnie muzyki dawnej, a była już skutecznie rozważana w pracach typu informatycznego w bardzo odległych czasach lat siedemdziesiątych XX w., w czasopismach takich, jak *Computer Music Journal*, czy *Computers and the Humanities* (obecnie już nie wydawane). W mojej opinii zagadnienia stylu i formy muzycznej przy obiektywizacji informatycznej są znacznie ciekawsze. Stymulacją do tego pytania jest artykuł A4, ponieważ, tak jak tam ważna jest informacja z obszaru poza dźwiękowego, to przy zastosowaniu AI w klasyfikacji i obiektywizacji relacji form muzycznych niezbędna jest wiedza teoretyczna z zakresu muzyki wykraczająca poza prosto definiowane wskaźniki czasowo-częstotliwościowe zjawisk dźwiękowych.

Ogólna ocena pracy.

Zgadzam się ze stwierdzeniem doktoranta wyrażonym w podsumowaniu autoreferatu, że w rozprawie udowodniono wszystkie postawione tezy. Również potwierdzam, że rozprawa zawiera znaczący przegląd najnowszej literatury odnoszącej się do systemów przetwarzania dźwięku. Wskazana przeze mnie wcześniej sumaryczna liczba 265 cytowanych artykułów we wszystkich przedstawionych pracach oczywiście musi zawierać konieczne powtórzenia cytowań artykułów w różnych pracach, jednak i tak stanowi o znaczącym rozeznaniu doktoranta w literaturze związanej z przedmiotem rozprawy.

Oceniając pracę należy określić sensowność i zasadność zaproponowanych tez. Uważam, że tezy pracy wnoszą znaczący element rozwoju w obszarze zagadnień, które obejmuje praca. A zatem świadczą o pozytywnej wartości przyjętego celu pracy. Nie wchodząc ponownie w szczegóły, ukierunkowują myślenie o zagadnieniach automatycznego rozpoznawania dźwięku w różnych aspektach, na znaczenie elementów, które powszechnie

uważane były za drugorzędne. Pogłos, hałas/szumy tła, lub zakłócenia, jeśli nie uwzględnione w materiale trenującym, powodują istotne obniżenie skuteczności systemu rozpoznawania. Osobnym nowatorskim podejściem są analizy osobowościowe przedstawione w pracy A4, które wskazują na znaczący wpływ efektów poza dźwiękowych na kwestie rozpoznawania i klasyfikacji rekomendacyjnej utworów.

Uważam, że obszar badań poruszony w publikacjach znacząco przewyższa standardy wymagane przy realizacji prac doktorskich. Gdyby forma doktoratu nie przyjmowała postaci zbioru publikacji, lecz stanowiła tradycyjną monografię, to z pewnością uznana by była za obszerną pod względem zakresu przeprowadzonych pomiarów i rozpatrywanych danych, rozmaitości przeprowadzonych analiz i obliczeń, oraz wielości kierunków badanych czynników.

Rozprawę doktorską mgr. inż. Mariusza Klecia pt. ,,Efektywna reprezentacja danych w systemach przetwarzania sygnałów dźwiękowych" zaliczam do kategorii jednoznacznie pozytywnej. W mojej opinii spełnia wszystkie kryteria zalecone do oceny przez Polsko-Japońską Akademię Technik Komputerowych. Dotyczy to układu pracy i sensowności jej celu, o czym pisałem już wcześniej. Zastosowane metody badawcze nie budzą zastrzeżeń, przy czym z racji na to, że rozprawa stanowi cykl publikacji ich szczegółowa ocena również podlegała wcześniej recenzentom przytaczanych prac. Rozprawa stanowi zbiór oryginalnego rozwiązanie kilku problemów naukowych. Z racji na wieloautorowość większości publikacji należy wskazać, że oświadczenia współautorów podają zasadniczy 60% udział doktoranta, przy czym informacja o wkładzie poszczególnych autorów dalej pokazuje, że jest to zasadniczy udział koncepcyjno-merytoryczny.

Stwierdzam, że rozprawa mgr inż. Mariusza Klecia pt. „Efektywna reprezentacja danych w systemach przetwarzania sygnałów dźwiękowych„ spełnia wymagania stawiane rozprawom doktorskim. Ponadto, biorąc pod uwagę jakość publikacji wnioskuję o wyróżnienie rozprawy doktorskiej.

Warszawa, 2.01.2025 r.