Efektywna reprezentacja danych w systemach przetwarzania sygnałów dźwiękowych

Rozprawa doktorska

Mgr inż. Mariusz Kleć

Polsko-Japońska Akademia Technik Komputerowych

Wydział Informatyki

Promotor:
Dr hab. Alicja Wieczorkowska

Promotor pomocniczy:
Dr hab. inż. Krzysztof Szklanny

Warszawa, 2024

Niniejszą pracę dedykuję Synowi Maksymilianowi

Oświadczenie

Ja, niżej podpisany Mariusz Kleć, autor rozprawy doktorskiej pt. ”Efektywna reprezentacja danych w systemach przetwarzania sygnałów dźwiękowych”, oświadczam, iż wyżej wskazaną rozprawę napisałem samodzielnie i żaden jej fragment lub całość nie był pisany przez osobę trzecią. Jednocześnie oświadczam, iż:

- praca nie była wcześniej podstawą nadania stopnia doktora innej osoby.
- załączona wersja elektroniczna jest tożsama z wydrukiem rozprawy.
- wszystkie elementy pracy, które zostały wykorzystane do jej realizacji, a nie będące mojego autorstwa, zostały odpowiednio oznaczone oraz zostało podane źródło ich pochodzenia.
- przedstawiona przeze mnie wyżej wskazana praca nie narusza przepisów ustawy z dnia 4 lutego 1994 r. o prawach autorskich i prawach pokrewnych (tj. z dnia 17 maja 2006 r. Dz.U. Nr 90, poz. 631 z późn. zm.).

Mam świadomość, że złożenie nieprawdziwego oświadczenia skutkować będzie niedopuszczeniem do dalszych czynności postępowania w sprawie nadania stopnia doktora lub cofnięciem decyzji o nadaniu mi stopnia doktora oraz wszczęciem postępowania dyscyplinarnego.

Data: 10.10.2024

Streszczenie

Skuteczność systemów przetwarzających sygnał dźwiękowy zależy m.in. od doboru architektury sieci neuronowej, zestawu danych trenujących oraz ich początkowej reprezentacji. Wytrenowane sieci neuronowe o odpowiednich architekturach (np. sieci splotowe lub autoenkodery) mogą stanowić dodatkowe detektory cech, uzupełniając początkową reprezentację danych i poprawiając w ten sposób ich zdolności dyskryminacyjne, jak również zdolności predykcyjne systemów wykorzystujących bezpośrednio dane dźwiękowe w postaci czasowej (ang. end-to-end learning). Poprawa skuteczności tych systemów wynika również m.in. z uzupełniania początkowych danych trenujących o dodatkowe dane kontekstowe, często odmiennej modalności. Dlatego też, odpowiednie modelowanie danych oraz ich reprezentacja są kluczowe w całym procesie przetwarzania sygnału dźwiękowego. Niniejsza rozprawa oparta jest na cyklu publikacji dotyczących metod usprawniania skuteczności systemów przetwarzających sygnał dźwiękowy, wykorzystujących wyżej wymienione podejścia.

W niniejszej pracy formułujemy tezę, iż włączenie zdarzeń dźwiękowych do korpusów mowy, wygenerowanych razem z pogłosem i dźwiękami tła, poprawia odporność modeli przetwarzających mowę na występowanie tych zakłóceń. Udowadniamy również, że wykorzystanie różnej długości ramek czasowych rozproszonej transformaty falkowej oraz splotowych sieci neuronowych przyczynia się do poprawy rozpoznawania wczesnych objawów chorób serca. Wreszcie, weryfikujemy tezę, że wykorzystanie modelu stanowiącego rozszerzenie standardowego modelu tzw. Wielkiej Piątki do reprezentacji osobowości użytkownika przyczynia się do redukcji błędu popełnionego przez systemy rekomendacji muzycznej, w porównaniu z modelem standardowym.

Często w nagraniach dźwiękowych zawierających mowę występują nieprzewidziane, często impulsowe zdarzenia dźwiękowe, które nie tylko zaburzają zrozumiałość, ale mogą również spowodować wykluczenie nagrania z korpusu z powodu niskiej jakości. W pracy pt. "Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions" poruszona została istotność problemu braku odpowiednio zdywersyfikowanego oraz odzwierciedlającego rzeczywiste warunki akustyczne zbioru trenującego do celów trenowania systemów służących poprawie zrozumiałości mowy oraz separacji mówców. W pracy tej proponujemy rozwiązanie generujące dowolną liczbę krótkich nagrań dźwiękowych, zawierających nagrania mowy polskiej na tle dźwięków otoczenia, z pogłosem oraz wspomnianymi nieprzewidzianymi zdarzeniami dźwiękowymi. Nasze eksperymenty z wykorzystaniem tych danych oraz treningiem głębokich sieci neuronowych potwierdzają ich użyteczność oraz potwierdzają stawianą tezę dotyczącą poprawy odporności modeli na występowanie tego typu dodatkowych zdarzeń dźwiękowych.

W pracy pt. "Music Recommendation Systems: a Survey" omówione zostały postępy w systemach rekomendacji muzycznej, w tym systemy uwzględniające głębokie sieci neuronowe. Praca ta przedstawia również wyzwania związane z personalizacją systemów rekomendacji muzycznej oraz zarysowuje przyszłe kierunki badań w tej dziedzinie. W innym artykule cyklu prac, pt. "Beyond the Big Five Personality Traits

for Music Recommendation Systems" zaproponowaliśmy wzbogacenie danych wejściowych dla systemów rekomendacji muzycznej o dodatkowe czynniki osobowościowe słuchaczy. Są one reprezentowane za pomocą rozszerzonego modelu tzw. Wielkiej Piątki (ang. Big Five), który oprócz pięciu podstawowych cech osobowościowych, mierzy trzy dodatkowe aspekty każdej z nich, prowadząc do uzyskania bardziej kompleksowej i szczegółowej reprezentacji. Za pomocą specjalnie stworzonej do tego celu aplikacji, zebrane zostały profile osobowościowe 279 uczestników. Dane te zostały wykorzystane w algorytmie hybrydowej rekomendacji muzycznej. Wyniki eksperymentów wykazały istotność zastosowania rozszerzonego modelu Wielkiej Piątki, przyczyniając się do redukcji błędu rekomendacji popełnianego przez system.

Istotnym aspektem skutecznej rekomendacji jest również skuteczna klasyfikacja utworów ze względu na ich gatunek lub nastrój. Skuteczność rozproszonej transformaty falkowej (ang. Scattering Wavelet Transform, SWT) w reprezentowaniu danych muzycznych do celów klasyfikacji została potwierdzona w pracy pt. "Pre-trained Deep Neural Network Using Sparse Autoencoders and Scattering Wavelet Transform for Musical Genre Recognition". W pracy tej zastosowano architekturę autoenkoderów celem wstępnego wytrenowania ostatecznej sieci neuronowej do klasyfikacji gatunku muzycznego. Eksperymenty podkreślają wysoką skuteczność SWT w reprezentacji danych muzycznych oraz umożliwiają wgląd w zachowanie funkcji kosztu w przypadku sieci z wstępnym wytrenowaniem za pomocą autoenkoderów. Wyniki uzyskane w tej pracy przyczyniły się do stworzenia kolejnej pracy pt. "Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation". Zajmujemy się w niej ekstrakcją dodatkowych cech sygnału bicia serca, poprzez zastosowanie SWT oraz splotowych sieci neuronowych (ang. Convolutional Neural Networks, CNN). Sieci tego typu, poprzez wykonywanie operacji splotu na kilku ramkach czasowych SWT jednocześnie umożliwiają wykrycie zależności występujących między nimi, co przyczyniło się do poprawy skuteczności rozpoznawania pierwszych objawów chorób serca, także z zaszumionych danych.

Podsumowując, w niniejszej rozprawie udowodniono postawione w niej tezy.

Abstract

The effectiveness of audio signal processing systems depends on several factors, including the choice of neural network architecture, the training data, and their initial representation. The trained neural networks of appropriate architectures (e.g., convolutional networks or autoencoders) can act as additional feature detectors, complementing the initial data representation and thus improving their discriminative capabilities. This can also improve the performance of the systems that directly use data in the time domain form (end-to-end learning). Additionally, enhancing the initial training data with additional contextual data of a different modality further improves the effectiveness of these systems. Therefore, proper data modelling and representation are crucial for audio signal processing. This dissertation is based on a series of publications focusing on methods for enhancing the effectiveness of audio signal processing systems using the abovementioned approaches.

In this dissertation, we hypothesise that including sound events in speech corpora, along with reverberation and background sounds, enhances the robustness of speech processing models to such disturbances. We also demonstrate that using various lengths of time frames of the Scattering Wavelet Transform (SWT) with Convolutional Neural Networks (CNN) helps improve the recognition of early symptoms of heart diseases. Finally, we validate the hypothesis that using the extended version of the standard Big Five model for representing user personality helps reduce the error made by music recommendation systems compared to the standard model.

The presence of unexpected sound events in audio recordings containing speech can disrupt intelligibility and cause the recording to be excluded from further processing due to its low quality. In the paper "Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions", we emphasised the significance of the lack of a suitably diversified training set that reflects real acoustic conditions for training systems aimed at improving speech intelligibility. The paper proposes a solution that generates short audio recordings containing Polish speech against background sounds, with reverberation and unexpected sound events. Our experiments with this data and training deep neural networks confirm their usefulness and support the hypothesis regarding improving the models' robustness to the occurrence of such additional sound events.

In the next paper, "Beyond the Big Five Personality Traits for Music Recommendation Systems", we proposed enriching the input data for music recommendation systems with the additional personality factors of listeners. These factors are represented by an extended Big Five model, which measures three additional aspects of each of the five main personality traits. Incorporating this extended model allowed us to reduce the recommendation error. The paper "Music Recommendation Systems: a Survey"discusses the advancements in music recommendation systems, including those involving deep neural networks. It also presents the challenges of personalising music recommendation systems and outlines future research directions.

An essential aspect of effective recommendation is the accurate classification of songs by genre or mood, which improves the quality of the music recommendation. The effectiveness of the SWT in representing music data was confirmed in the work titled "Pre-trained Deep Neural Network Using Sparse Autoencoders and Scattering Wavelet Transform for Musical Genre Recognition". The paper utilised the autoencoder architecture for pre-training the final neural network for music genre classification. The experiments highlighted the high efficiency of SWT in representing music data and provided insight into the behaviour of training the networks pre-trained with autoencoders. The results obtained in this work contributed to the subsequent work titled "Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation". This work addressed extracting additional features of the heartbeat signal using SWT and CNN. The convolution operations on several SWT time frames enabled the detection of dependencies between frames, improving the efficiency of recognising the first symptoms of heart diseases from noisy data.

To summarise, we verified all hypotheses of this dissertation.

Podziękowania

Chciałbym podziękować wszystkim którzy przyczynili się do realizacji niniejszej rozprawy doktorskiej. Przede wszystkim chciałbym złożyć podziękowania mojemu Promotorowi, dr hab. Alicji Wieczorkowskiej za wszelką pomoc w trakcie pisania rozprawy, cierpliwość i motywowanie do jej ukończenia. Bardzo dziękuję również mojemu Promotorowi Pomocniczemu dr hab. inż. Krzysztofowi Szklannemu za udostępnienie sprzętu, słowa życzliwości oraz cenne wskazówki. Podziękowania kieruję również do śp. prof. dr hab. Krzysztofa Maraska który służył wsparciem i umożliwił mi odbycie stażu w firmie SONY w Stuttgarcie, wyznaczając kierunek mojego dalszego rozwoju naukowego. Duże podziękowania kieruję również w stronę całej mojej Rodziny. W szczególności chciałem bardzo podziękować mojej Żonie Marlenie za dodawanie sił w chwilach zwątpienia i nieustannie okazywane wsparcie.

Spis treści

Rozdział 1

Wprowadzenie

Dobór architektury sieci, początkowej reprezentacji dźwięku, oraz zastosowanie odpowiedniego korpusu danych trenujących przyczynia się do nauki nowych cech audio przez modele neuronowe. Cechy te stanowią uzupełnienie początkowej reprezentacji oraz przyczyniają się do poprawy przetwarzania danych dźwiękowych w systemach klasyfikacji danych audio oraz w systemach wykorzystujących bezpośrednio dane dźwiękowe w postaci czasowej. Niniejsza rozprawa oparta jest na cyklu spójnych z tematem publikacji dotyczących metod poprawy skuteczności systemów przetwarzających sygnał dźwiękowy, w tym za pomocą sieci neuronowych, oraz dodatkowych danych trenujących.

1.1 Motywacja

Rozwój algorytmów głębokiego uczenia jest często napędzany przez duże korporacje, które dzięki posiadanym zasobom są w stanie wytrenować duże modele olbrzymią ilością danych w akceptowalnym czasie. Wytrenowana sieć może zostać wykorzystana w celu rozwiązania zupełnie innego problemu w procesie tzw. uczenia transferowego (ang. transfer learning). Przykładowo, sieci wytrenowane na danych obrazowych mogą zostać wykorzystane do ekstrakcji nowych cech z dwuwymiarowych reprezentacji sygnałów audio, celem poprawy skuteczności ich klasyfikacji. Zatem efektywność głębokich modeli neuronowych jest ściśle powiązana z jakością oraz wielkością danych trenujących. Dlatego też dostęp do wysokiej jakości danych trenujących jest kluczowy dla poprawy skuteczności działania głębokich sieci neuronowych, przyczyniających się do odkrywania nowych cech z danych dźwiękowych, a tym samym poprawy dyskryminacyjnych właściwości tych cech. W przypadku systemów przetwarzających mowę, dane te są często nagrywane w sposób spontaniczny, z reguły za pomocą urządzeń przenośnych, bez profesjonalnego zaplecza sprzętowego. Z tego powodu systemy poprawiające zrozumiałość mowy zyskują na znaczeniu. Należą do nich systemy redukujące szumy oraz systemy oddzielające mówców do osobnych kanałów. Przyczynia się to nie tylko do poprawy zrozumiałości mówców, ale również skuteczności systemów konwertujących mowę na tekst. Ponadto ma to kluczowe znaczenie w zastosowaniach telekomunikacyjnych, zarówno podczas rozmów telefonicznych jak i wideokonferencji.

Problem zaszumionych danych występuje również w nagraniach dźwięków fizjologicznych takich jak np. bicie serca, realizowanych za pomocą urządzeń przenośnych w nieprofesjonalnych warunkach. Choroby sercowo-naczyniowe są główną przyczyną śmiertelności na całym świecie, dlatego wczesna diagnoza chorób kardiologicznych ma istotne znaczenie dla zapobiegania poważnym powikłaniom i zmniejszenia

wskaźników śmiertelności. Dostęp do specjalistycznych narzędzi diagnostycznych w ubogich krajach jest ograniczony, jednak urządzenia mobilne są coraz bardziej powszechne na całym świecie. Wykorzystując mikrofony wbudowane w smartfony można przeprowadzić wstępną ocenę stanu zdrowia w ramach badań przesiewowych, ustanawiając początkowy i ważny krok w identyfikacji osób zagrożonych poważnymi powikłaniami sercowo-naczyniowymi. Często jednak urządzenia te są używane przez osoby bez technicznego przygotowania, dlatego też w nagraniach tego typu występują różnego rodzaju szumy i artefakty. Z tego powodu bardzo istotne jest rozwijanie systemów odpornych na występowania szumów w nagraniach.

Typem systemów które operują dużą ilością danych dźwiękowych są systemy rekomendacji muzycznej a platformy do strumieniowego słuchania muzyki stały się bardzo popularne w ostatnich latach, zapewniając użytkownikom dostęp do ogromnej liczby utworów muzycznych. Przeszukiwanie tak dużych zbiorów danych stanowi wyzwanie, któremu systemy rekomendacji muzycznej starają się sprostać. Odkrywanie nowych cech przez sieci neuronowe usprawnia przeszukiwanie dużych kolekcji danych muzycznych oraz ich klasyfikację, przyczyniając się do poprawy skuteczności tych systemów.

1.2 Tezy rozprawy

W niniejszej rozprawie postawiono następujące hipotezy badawcze:

1. Włączenie zdarzeń dźwiękowych do korpusów mowy, wygenerowanych razem z pogłosem i dźwiękami tła, poprawia odporność modeli przetwarzających mowę na występowanie tych zakłóceń.
2. Wykorzystanie różnej długości ramek czasowych rozproszonej transformaty falkowej oraz splotowych sieci neuronowych przyczynia się do poprawy rozpoznawania wczesnych objawów chorób serca.
3. Wykorzystanie modelu stanowiącego rozszerzenie standardowego modelu tzw. Wielkiej Piątki do reprezentacji osobowości użytkownika przyczynia się do redukcji błędu popełnionego przez systemy rekomendacji muzycznej, w porównaniu modelem standardowym.

Powyższe hipotezy badawcze zostaną udowodnione w dalszej części niniejszej rozprawy.

Rozdział 2

Przegląd literatury

W klasyfikacji sygnałów audio za pomocą sieci neuronowych, jednowymiarowy sygnał poddawany jest transformacji do postaci dwuwymiarowej. Najczęściej stosowane techniki służące do tej transformacji obejmują zastosowanie dyskretnej transformaty Fouriera (and. short-term Fourier Transform, STFT) lub transformat falkowych (ang. wavelet transform, WT). Wybór odpowiedniej początkowej reprezentacji sygnału wpływa na architekturę sieci neuronowej, proces treningu, wymagania co do mocy obliczeniowej, oraz ostatecznie na skuteczność klasyfikacji. W głębokich sieciach neuronowych nowe cechy są odkrywane warstwa po warstwie, wpływając na jakość cech w ostatniej warstwie. To właśnie ostatnia warstwa głębokich sieci neuronowych zawiera najbardziej wartościowe cechy wpływające na zdolności dyskryminacji nowych danych.

Rozdział tej zawiera przegląd najnowszej literatury w kontekście zastosowań głębokich sieci neuronowych w problemach przetwarzania dźwięku, ze szczególnym naciskiem na problem klasyfikacji. Przegląd uwzględnia różnorodne architektury sieci neuronowych, począwszy od sieci splotowych (ang. Convolutional Neural Networks, CNN) i sieci rekurencyjnych (and. Recurrent Neural Networks, RNN), poprzez autoenkodery (ang. autoencoders, AE) skończywszy na transformerach (ang. Transformers, TR). Niniejszy przegląd dotyczy również systemów służących m.in. do separacji mówców, wykorzystujących bezpośrednio dane dźwiękowe w postaci czasowej.

2.1 Splotowe sieci neuronowe w ekstrakcji cech audio

Metody zastosowane w sieciach splotowych do rozpoznawania obiektów w obrazie [14, 19, 34, 36, 46], są wykorzystywane również w przetwarzaniu dwuwymiarowych reprezentacji sygnału audio [24, 25, 26, 27]. W pracy [25] wykorzystano jedenaście różnych sieci splotowych, wytrenowanych na obrazach, w problemie klasyfikacji dźwięków otoczenia. Wytrenowane architektury uwzględniają takie modele jak VGG [34], SqueezeNet [17], ResNet [14] czy AlexNet [19]. Eksperymenty z użyciem baz ESC-10, ESC-50 [28] oraz UrbanSound8K [3] pokazują, iż podejście uczenia transferowego pozwala na uzyskanie wysokiej skuteczności rozpoznawania dźwięków otoczenia. Przykładowo, ResNet-152 osiąga 99.04% skuteczności dla bazy ESC-10 oraz 99.49% dla UrbanSound8K, natomiast DenseNet-161 [16] osiąga 97.57% skuteczności dla ESC-50.

W kolejnej pracy [24] autorzy proponują model CNN ze zintegrowanym mechanizmem uwagi czasowo-częstotliwościowej (ang. time-frequency attention mechanism), którego celem jest usprawnienie ekstrakcji

cech audio ze spektrogramu. Mechanizm uwagi wydobywa istotne ramki czasowe oraz pasma częstotliwościowe, minimalizując w ten sposób wpływ nieistotnych okienek spektrogramu oraz nieistotnych częstotliwości. Mechanizm ten znacząco poprawił zdolność sieci do wychwytywania kluczowych cech czasowo-częstotliwościowych, zwiększając możliwości dyskryminacyjne modelu przetestowanego na bazach UrbanSound8K oraz ESC-50. Podobne podejście zostało zastosowane w pracy [45] z tą różnicą, iż w tym przypadku autorzy zastosowali dodatkowo sieć rekurencyjną, która poprawiła wyniki uzyskane dla bazy ESC-50.

2.2 Sieci rekurencyjne w przetwarzaniu czasowym

Wcześniejsze prace podkreślają skuteczność sieci CNN w ekstrakcji cech z sygnału audio. Niemniej jednak, zależności czasowe występujące w następujących po sobie segmentach spektrogramu mogą nie być w całości wychwycone przez sieci CNN. W tych przypadkach pomocne okazują się sieci RNN. W pracy [32] zaprezentowano ich skuteczność w klasyfikacji zdarzeń dźwiękowych z placu budowy. Opisywany model otrzymuje na wejściu kombinację trzech reprezentacji spektralnych: współczynniki mel-cepstralne (ang. mel-frequency cepstral coefficients, MFCCs), spektrogram w skali melowej, oraz cechy chromatyczne. Autorzy za pomocą proponowanego podejścia uzyskali 97% skuteczności.

Interesujące podejście zostało opisane artykule [2], w którym autorzy opisują metodologię integracji różnych reprezentacji czasowo-częstotliwościowych do postaci trójwymiarowej. W skład tej reprezentacji wchodzą ciągła transformata falkowa, spektrogram w skali melowej, oraz spektrogram uzyskany z filtrów gammatone. W artykule wykorzystano CNN oraz RNN do przetwarzania wspomnianej trójwymiarowej reprezentacji, poprawiając skuteczność rozpoznawania wad wymowy.

2.3 Zastosowania transformerów

Transformery [40], zamiast modelować kolejne ramki czasowe jedna po drugiej (jak w przypadku sieci RNN), pobierają całą ich sekwencję na wejściu. Przegląd zawarty w pracy [18] opisuje zastosowanie transformerów w takich obszarach, jak przetwarzanie języka naturalnego, obrazów, danych multi-modalnych, oraz sygnałów audio. W pracy [44] zaproponowano koncepcję ”transformera spektrogramowego”, dedykowanego celom klasyfikacji sygnałów audio. Dzięki zastosowaniu transformera, opisywany model jest w stanie nauczyć się czasowych i częstotliwościowych cech ze spektrogramu, wykrywając w ten sposób skomplikowane zależności pomiędzy nimi.

Transformery wykazały również duży potencjał w zastosowaniach związanych z separacją mówców [8], jednak są wyjątkowo wymagające jeśli chodzi o wymaganą moc obliczeniową. W artykule [22] autorzy proponują kompaktową wersję transformera bez wpływu na skuteczność w separacji mówców.

2.4 Wydobywanie cech audio w sposób nienadzorowany

Powyższe metody stanowią przykłady treningu nadzorowanego. Uczenie bez nadzoru stanowi alternatywę dla powyższych metod, również umożliwiając ekstrakcję cech sygnału audio. Typowym przykładem architektury stosowanej do treningu bez nadzoru są autoenkodery. Uczą się one kompresować dane w taki sposób, aby zrekonstruować dane wejściowe na wyjściu sieci z ich zakodowanej i skompresowanej

reprezentacji w warstwie ukrytej. Przykład stanowi praca [9], wykorzystująca autoenkodery do wykrywania w sygnale mowy wysokopoziomowych cech, takich jak fonemy, nie ulegając jednocześnie wpływom zmiennych czynników mowy, takich jak wysokość dźwięku, czy szumy. W kolejnej pracy [30] autorzy wykorzystali wariacyjny autoenkoder (ang. variational autoencoder, VAE) do generowania syntetycznych przykładów trenujących celem zbalansowania zbioru danych treningowych do rozpoznawania dźwięków oddechu. Autorzy udowodnili w ten sposób skuteczność tej metody, w której zrównoważone dane trenujące są kluczowe dla efektywnego treningu modeli klasyfikacyjnych.

Zastosowanie autoenkoderów w celu poprawy klasyfikacji zaburzeń mowy zostało opisane w pracy [29]. Autorzy zastosowali hierarchiczny autoenkoder wariacyjny (ang. Factorized Hierarchical Variational Autoencoder, FHVAE) do ekstrakcji cech semantycznych z uwzględnieniem następstw czasowych występowania kolejnych ramek nagrań mowy zaburzonej.

2.5 Wykorzystanie surowych danych w głębokim uczeniu

Podejścia opisane do tej pory zakładały początkowe przekształcenie sygnału do postaci dwuwymiarowej, najczęściej reprezentującej moduł transformaty Fouriera. Głębokie sieci neuronowe umożliwiają jednak stosowanie surowej postaci czasowej sygnału, dokonując parametryzacji w sposób automatyczny. Podejście takie stawia jednak wyzwania. Jednym z nich jest dostęp do znaczących mocy obliczeniowych, szczególnie w przypadku treningu za pomocą dużych zbiorów danych. Drugim wyzwaniem jest konieczność przywrócenia postaci czasowej z powrotem na wyjściu modelu. Rekonstrukcja sygnału z transformaty Fouriera wymaga składowych fazowych, które są pomijane na wejściu. Faza sygnału musi podlegać zatem estymacji na wyjściu, co skutkuje zniekształceniami oraz niedoskonałą rekonstrukcją dźwięku. Rozwiązaniem jest przekształcenie jednowymiarowych segmentów sygnału audio na jego wielowymiarową postać, używając operacji jednowymiarowego splotu (ang. 1-D convolution). Sygnał z postaci wielowymiarowej może zostać odtworzony przy zastosowaniu transponowanej warstwy splotu jednowymiarowego (ang. transposed 1-D convolution) na wyjściu modelu, pomijając w ten sposób konieczność estymacji fazy. Podejście to zostało wprowadzone w pracy [23], opisującej model o nazwie Conv-TasNet do separacji mówców. Jednakże w przypadku zaszumionych nagrań, separacja może skutkować niezadowalającymi rezultatami. W artykule [15] autorzy proponują rozwiązanie tego problemu poprzez integrację podejść służących do poprawy zrozumiałości mowy oraz separacji mówców w jednym modelu. Zaproponowane rozwiązanie wzmacnia odporność modelu na zaszumione dane, a dodatkowo proponowana strategia modulacji gradientowej poprawia jakość rekonstrukcji czystego sygnału mowy na wyjściu.

W pracy [41] zaprezentowano zastosowanie tranformerów do surowych danych audio, bez korzystania z reprezentacji dwuwymiarowej. Autorzy przedstawiają wyższe wyniki klasyfikacji dźwięków z bazy Free Sound 50K dataset [12] z użyciem transformerów w porównaniu z klasyfikacją z użyciem sieci CNN.

2.6 Inżynieria cech dla systemów rekomendacji muzycznej

Celem rekomendacji muzycznej jest predykcja oceny określającej preferencję użytkownika względem określonego utworu. W przypadku, gdy predykcja zwraca wartości ciągłe, problem ten jest traktowany jako problem regresji, jednakże predykcja preferencji użytkownika może odbywać się również na skali dyskretnej (np. ”lubię”, ”nie lubię”); w takim przypadku będzie to problem klasyfikacji.

Kluczową rolę w systemach rekomendacji muzycznej odgrywa klasyfikacja utworów muzycznych, gdyż

znajomość kategorii utworu umożliwia rekomendację utworów dostosowanych do preferencji słuchacza. Jednakże zawartość dźwiękowa muzyki jest znacząco odmienna od mowy czy szumów otoczenia, dlatego ekstrahowane cechy powinny odzwierciedlać specyficzną naturę sygnałów muzycznych. W pracy [31] autorzy opisują metodę niejawnej ekstrakcji cech specyficznych dla gatunku muzycznego z surowych danych wejściowych. Wyekstrahowane cechy zostały ocenione w sposób jakościowy i ilościowy, bazując na skuteczności rozpoznawania gatunku muzycznego za pomocą prostej sieci neuronowej.

Muzyka jest tym szczególnym typem zawartości dźwiękowej, której znaczenie opiera się na czasowych następstwach zdarzeń muzycznych, takich jak rytm czy melodia. Do modelowania tych następstw mogą służyć sieci RNN. Praca [13] porusza problem klasyfikacji utworów muzycznych za pomocą sieci CNN oraz dwukierunkowych sieci RNN, osiągając 93,1 % dokładności predykcji dla bazy GTZAN [35].

Poprawa skuteczności systemów rekomendacji muzycznej może odbywać się również poprzez wykorzystanie danych kontekstowych o użytkowniku, szczególnie, że preferencje muzyczne uzależnione są m.in. właśnie od tych danych [37], np. danych o osobowości słuchacza. W pracy [11] autorzy analizują systemy rekomendacji muzycznej uwzględniające osobowość.

2.7 Analiza nagrań fizjologicznych

Klasyfikacja dźwięków fizjologicznych, takich jak oddech czy bicie serca, stanowi przykład systemów wspierających stawianie diagnoz dotyczących stanu zdrowia. W pracy [20] autorzy opisują metodę klasyfikacji dźwięków bicia serca pochodzących z fonokardiogramu. Metoda wykorzystuje połączenie dwóch reprezentacji falkowych oraz metodę łączenia wielu klasyfikatorów CNN do poprawy wyników rozpoznawania chorób serca. Autorzy oceniają wytrenowane modele używając baz PhysioNet/CinC 2016 [10] oraz PASCAL [6].

Kolejna praca [21] podkreśla potencjał SWT w reprezentowaniu sygnału bicia serca pochodzącego z elektrokardiogramu. Autorzy dokonują redukcji wymiarowości danych za pomocą analizy składowych głównych (ang. principal component analysis, PCA), by następnie przeprowadzić ostateczną klasyfikację za pomocą metod uczenia maszynowego, w tym sieci neuronowych.

Inne przykłady klasyfikacji dźwięków fizjologicznych obejmują próbę diagnozowania chorób układu oddechowego, w szczególności COVID-19 [5], oraz rozpoznawanie nieprawidłowych dźwięków oddechowych [5]. Oba podejścia wykorzystują klasyfikatory oparte na sieciach CNN oraz RNN, osiągając wysokie wyniki. Jednak badanie tego typu problemów pociąga za sobą konieczność pozyskania odpowiednich danych trenujących, co często bywa utrudnione z powodu niskiej ich dostępności.

2.8 Wykorzystanie samodzielnie opracowanych zbiorów danych

Rozwój metod klasyfikacji pociąga często za sobą potrzebę budowania własnych zbiorów danych. Dla przykładu, autorzy pracy [33] zajmują się klasyfikacją dźwięków występujących w budynkach mieszkalnych, wykorzystując przy tym wcześniej wytrenowane głębokie sieci neuronowe takie jak DenseNet [16], ResNet, Inception, and EfficientNet [39], osiągając wysoką skuteczność rozpoznawania 95% w przypadku modelu ResNet. W kolejnej pracy [7] autorzy koncentrują się na rozwoju systemu klasyfikacji do identyfikowania pojazdów z niestandardowo głośnymi układami wydechowymi. Stosując sieć AlexNet autorzy uzyskali 95% skuteczności korzystając z samodzielnie opracowanych danych. Artykuł [43] opisuje metodologię klasyfikacji mowy wypowiadanej z różnym akcentem języka angielskiego. Badania wskazują, iż

liniowo wyskalowany spektrogram skutkuje wyższą skuteczności rozpoznawania akcentu aniżeli spektrogram w skali melowej. Pomimo niewielkiego rozmiaru zbioru danych Accent Archive [42], autorzy uzyskali wysokie wyniki, tj. ponad 96% skuteczności rozpoznawania akcentu z użyciem sieci CNN. Badania opisane w [4] dotyczą rozróżniania dźwięków mowy sfałszowanej (and. fake voice recognition) od prawdziwej przy użyciu sieci CNN. Autorzy wykorzystali do tego celu samodzielnie skonstruowaną bazę, osiągając wysoką skuteczność rozpoznawania, tj. ponad 98 %.

Rozdział 3

Przegląd opublikowanych artykułów

Niniejszy rozdział omawia publikacje włączone do rozprawy, wraz z wykazaniem słuszności tez postawionych w rozprawie, oraz ich dane bibliograficzne.

Dane bibliograficzne publikacji włączonych do rozprawy

1. 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, & C. Schneider (Eds.), Harnessing Opportunities: Reshaping ISD in the post-COVID-19 and Generative AI Era (ISD2024 Proceedings). Gdańsk, Poland: University of Gdańsk. ISBN: 978-83-972632-0-8. https://doi.org/10.62036/ISD.2024.37.

2. 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,

3. 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, Cham. https://doi.org/10.1007/978-3-030-01851-1_3 http://dx.doi.org/10.7494/csci.2015.16.2.133

4. 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). https://doi.org/10.1186/s13636-022-00269-0

5. 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, Cham. https://doi.org/10.1007/978-3-030-66450-3_7

3.1 Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions

Systemy poprawiające zrozumiałość mowy wymagają odpowiednio przygotowanych danych trenujących, które oprócz zaszumionych przykładów powinny zawierać ich wysokiej jakości odpowiedniki. Systemy tego typu są często używane z danymi niskiej jakości, zawierające różnego rodzaju niestacjonarne szumy i zdarzenia dźwiękowe. Z tego względu zbiór trenujący powinien odzwierciedlać rzeczywiste i różnorodne warunki występujące w nagraniach. Należą do nich szum otoczenia, pogłos, oraz inne nieplanowane zdarzenia dźwiękowe. Dobrej jakości nagrania studyjne mogą zostać wykorzystane w symulacji nagrań zaszumionych, poprzez dodanie do nich szumów o różnym charakterze oraz pogłosu, odzwierciedlając w ten sposób warunki rzeczywiste, a jednocześnie dostarczając informacji o danych (ang. ground truth) niezbędnych do treningu i testów. Niemniej jednak, obecnie dostępne bazy nagrań mowy występują głównie w języku angielskim oraz uwzględniają relatywnie niską różnorodność dźwięków otoczenia. Niektóre z nich uwzględniają pogłos i dźwięki otoczenia, jednak wciąż nie odzwierciedlają bardziej specyficznych i rzeczywistych przypadków nagrań, w których mogą wystąpić nieprzewidziane zdarzenia dźwiękowe w tle, takie jak dźwięk budzika lub klakson samochodu.

W niniejszym artykule proponujemy metodę generowania symulowanych nagrań mowy polskiej, w szczególnie wymagającym środowisku akustycznym uwzględniającym występowanie pogłosu, dźwięków tła, oraz nieplanowanych zdarzeń dźwiękowych. Praca ta została nagrodzona na konferencji 32nd International Conference on Information Systems Development, ISD 2024. Opublikowanie tej metody stanowi istotne uzupełnienie luki w dostępie do darmowych i dużych zbiorów danych dla języka polskiego, z możliwościa zastosowania dla dowolnego języka. Za pomocą tej metody istnieje możliwość generowania ogromnej ilości nagrań. Innowacyjną cechą proponowanego rozwiązania jest możliwość łatwego dostosowania już wygenerowanych nagrań do różnych zadań klasyfikacji, np. klasyfikacji tła lub zdarzeń dźwiękowych. Jest to możliwe dzięki przechowywaniu każdego komponentu wchodzącego w skład nagrania (mowa, dźwięki tła, zdarzenia dźwiękowe oraz odpowiedź akustyczna pomieszczenia jako pogłos) w osobnych plikach dźwiękowych, dostępnych dodatkowo oprócz nagrania zaszumionego. Dodanie odwróconej fazy danego komponentu do ostatecznego nagrania powoduje jego eliminację z tego nagrania, dzięki zjawisku kasowania się fal dźwiękowych znajdujących się w przeciwfazie. Umożliwia to dowolne kształtowanie zakresu komponentów wchodzących w skład nagrania, bez potrzeby generowania korpusu ponownie.

W omawianym artykule zaprezentowana została użyteczność proponowanego rozwiązania w procesie trenowania kilku modeli działających na surowych danych dźwiękowych. W tym celu zastosowana została architektura sieci działająca w dziedzinie czasu, o nazwie Conv-TasNet [23]. Do celów eksperymentalnych wygenerowanych zostało 250 000 cztero sekundowych nagrań, co przekłada się na ponad 11 dni ciągłej mowy w różnorodnym środowisku akustycznym. Pierwszy z modeli wytrenowany został do separacji pojedynczego mówcy od dźwięków tła. Kolejny model separuje dwóch mówców od siebie, w nagraniu bez szumów. Ostatni model został wytrenowany pod kątem separacji dwóch mówców występujących na tle złożonych szumów. Ze względu na ograniczone zasoby obliczeniowe w eksperymentach zredukowaliśmy liczbę kanałów w koderze splotowym z 512 do 256, zmniejszając w ten sposób wymiarowość reprezentacji sygnału, a tym samym liczbę parametrów potrzebnych do wytrenowania, aby przyspieszyć obliczenia, szczególnie biorąc pod uwagę wielkość wygenerowanego korpusu.

Modele separacji mówców zostały porównane ze znacznie bardziej złożonym modelem, opartym na architekturze transformera o nazwie SepFormer [22]. Model ten został wcześniej wytrenowany na danych niezaszumionych.

Porównując wyniki Conv-TasNet z SepFormer uzyskane na proponowanej w niniejszej pracy bazie, pomimo krótkiego czasu treningu (15 epok) i ograniczonej liczby wag, Conv-TasNet zwraca pozytywny niezmienny w skali stosunek sygnału do szumu (SI-SNR) 1.7 dB, natomiast SepFormer zwraca negatywny wynik SI-SNR -0.48 dB. Negatywny wynik SI-SNR świadczy o tym, iż na wyjściu modelu znajdowało się więcej szumu niż użytecznej mowy. Potwierdza to niezdolność modelu SepFormer do radzenia sobie z zaszumionymi danymi w zadaniu separacji mówców, a jednocześnie podkreśla znaczenie występowania zdarzeń dźwiękowych oraz pogłosu w danych trenujących. Można to zaobserwować porównując wyniki uzyskane przez te same modele na bazie Libri2Mix, która zawiera dźwięki tła, lecz bez pogłosu oraz bez zdarzeń dźwiękowych. W tym przypadku model SepFormer, pomimo tego że był wytrenowany na czystych danych, był w stanie odseparować mówców od siebie ze skutecznością 6.84 dB SI-SNR, a zatem występowanie zdarzeń dźwiękowych oraz pogłosu w danych trenujących znacząco utrudnia zadanie separacji mówców, biorąc pod uwagę negatywny wcześniejszy wynik -0.48 dB SI-SNR, uzyskany na zbiorze danych zawierającym te dwa składniki dźwięków tła.

Kolejne eksperymenty uwzględniały ten sam zestaw danych, jednak z wyłączonym jednym mówcą, celem przeprowadzenia treningu separacji pojedynczej mowy od szumów tła, przyczyniając się do poprawy zrozumiałości mowy. Model ten został porównany z modelem o podobnej architekturze o nazwie CTNoar [38], wytrenowanym wcześniej metodą zwracania pojedynczego odseparowanego mówcy na jednym kanale i pozostałych mówców na drugim. Model był wcześniej wytrenowany na czystych danych zawierających od 2 do 3 mówców jednocześnie.

Modele zostały przetestowane z użyciem naszego zbioru testowego oraz Libri1Mix, zawierającego sygnały pojedynczych mówców na tle szumów otoczenia, bez pogłosu oraz zdarzeń dźwiękowych. W tym przypadku uzyskano wynik ponad 12 dB SI-SNR w naszym modelu, co stanowi znaczącą poprawę w stosunku do modelu separującego dwóch mówców. Świadczy to o zdolności architektury Conv-TasNet do skutecznej separacji tła od mowy. Testy z wykorzystaniem zbioru testowego zawierającego dodatkowo pogłos oraz zdarzenia dźwiękowe skutkowały nieznacznie gorszym wynikiem, tj. ponad 10 dB SI-SNR. Jest to jednak wciąż relatywnie wysoka wartość, biorąc pod uwagę to, że model CTNoar zupełnie nie sprawdza się w obu przypadkach, zwracając negatywne wyniki.

Uzyskane wyniki podkreślają potrzebę trenowania rozwiązań separacji mówców zaszumionymi danymi, aby mogły radzić sobie z nagraniami zawierającymi szumy. Przeprowadzone eksperymenty uwypukliły szczególnie znaczenie nieprzewidywalnych zdarzeń dźwiękowych oraz pogłosu w zbiorze trenującym dla tych systemów; w obecności takich zdarzeń model SepFormer staje się bezużyteczny.

Podobna sytuacja występuje w przypadku modelu poprawiającego zrozumiałość pojedynczego mówcy, gdzie model CTNoar również zwraca negatywne wyniki, niezależnie od tego, czy w warstwie szumów występują zdarzenie dźwiękowe, czy nie, podczas gdy model wytrenowany na wygenerowanym zbiorze zwraca relatywnie wysokie wyniki, tj. ponad 10 dB SI-SNR.

Opublikowane w tym artykule wstępne wyniki eksperymentów nie wyczerpują jednak wszystkich możliwości proponowanej metody generowania korpusów; prace te są obecnie kontynuowane, a rozszerzona wersja artukulu zostanie opublikowana (na zapeoszenie organizatorów konferencji ISD) jako rozdział w książce z serii Lecture Notes in Information Systems and Organization series wydawnictwa Springer. Z danych użytych do eksperymentów opisanych w artykule przedstawionym na ISD, wyodrębniono podzbiór 10000 przykładów trenujących oraz 1000 przykładów testowych, aby przyspieszyć obliczenia oraz dokonać analizy porównawczej w celu dalszej weryfikacji Hipotezy 1.

Za pomocą tego podzbioru wytrenowane zostały 3 identyczne modele Conv-TasNet separujące pojedynczego mówcę od dźwięków tła (CTN-SE1). W każdym przypadku był to jeden mówca oraz tło

dźwiękowe składające się różnych składników: sceny dźwiękowej (S), zdarzeń dźwiękowych (E) oraz pogłosu (R). Każdy z modeli trenowany był dokładnie tymi samymi przykładami trenującymi przez 20 epok, z identycznym stanem wag początkowych; jedyną różnicą był skład dźwięków tła w danych trenujących. Modyfikacja składników dźwięków tła była możliwa dzięki zastosowaniu metody modyfikacji nagrań z użyciem zjawiska znoszenia się fal dźwiękowych znajdujących się w przeciwfazie, opisanej w artykule.

Pierwszy z modeli wytrenowany został przykładami zawierającymi wszystkie składniki dźwięków tła (SER), kolejny wszystkimi za wyjątkiem zdarzeń dźwiękowych (SR), oraz ostatni wytrenowany został wyłącznie dźwiękami mówcy oraz sceny dźwiękowej, bez pogłosu ani zdarzeń dźwiękowych (S). Modele te zostały przetestowane z użyciem 1000 przykładów danych testowych (dalej nazywanych PolSE1), poddanych również takiej samej procedurze modyfikacji składników tła dźwiękowego jak w przypadku danych trenujących. Dla przykładu, zbiór testowy oznaczony jako PolSE1-SER uwzględnia pojedynczego mówcę oraz wszystkie trzy składniki tła dźwiękowego, S, E, oraz R. Analogicznie, PolSE1-SR uwzględnia w tle jedynie scenę dźwiękową oraz pogłos, bez dźwięków zdarzeń dźwiękowych.

Następnie, wytrenowane modele zostały przetestowane z użyciem każdej wersji danych testowych oraz dodatkowo zbioru Libri1Mix-Noisy. Celem eksperymentów jest sprawdzenie odporności modeli separacji tła na występowanie zdarzeń dźwiękowych w zbiorze testowym. Odporność w tym przypadku jest rozumiana jako wysokie wartości SI-SNR oraz PESQ w sytuacji występowania dodatkowych zdarzeń dźwiękowych w danych testowych.

Tabela 3.1: Wyniki eksperymentów przeprowadzonych w celu dalszej weryfikacji Hipotezy 1. Najwyższe wartości SI-SNR oraz PESQ dla danego zbioru testowego zostały wytłuszczone.

	PolSE1-SER		PolSE1-SR		PolSE1-S		Libri1Mix-Noisy		
	SI-SNR	PESQ	SI-SNR	PESQ	SI-SNR	PESQ	SI-SNR	PESQ	średnia
CTN-SE1-SER	**7.7**	**2.38**	13.26	3.01	16.74	3.32	9.73	2.51	**7.33**
CTN-SE1-SR	3.01	2.14	**14.05**	**3.14**	17.96	3.46	**10.12**	**2.55**	7.05
CTN-SE1-S	0.46	1.88	9.35	2.56	**20.07**	**3.48**	8.38	2.38	6.07

W wynikach przedstawionych w Tabeli 3.1 można zaobserwować wpływ występowania zdarzeń dźwiękowych na separację czystej mowy od tła w modelach trenowanych do poprawy zrozumiałości mowy. Model uwzględniający tego typu zdarzenia w zbiorze trenującym (CTN-SE1-SER) zwraca 7.7db SI-SNR oraz 2.38 PESQ dla danych testowych, które również zawierają tego typu zdarzenia dźwiękowe. Model wytrenowany bez udziału zdarzeń dźwiękowych w tle (CTN-SE1-SR) zwraca 3.01 db SI-SNR oraz 2.14 PESQ. Różnica wyników dla tych dwóch modeli wynosi 4.69 db SI-SNR.

Warto zauważyć, iż ten sam model (CTN-SE1-SER) zwraca relatywnie wysokie wyniki niezależnie od zawartości składników tła. Biorąc pod uwagę wszystkie zbiory testowe, model wytrenowany z udziałem zdarzeń dźwiękowych zwraca najwyższą uśrednioną wartość z obydwu metryk SI-SNR oraz PESQ. W ten sposób potwierdzamy Hipotezę 1, iż włączenie zdarzeń dźwiękowych do korpusów mowy, wygenerowanych razem z pogłosem i dźwiękami tła, poprawia odporność modeli przetwarzających mowę na występowanie tych zakłóceń.

Tabela 3.1 pokazuje również, iż najlepsze wyniki można uzyskać stosując model wytrenowany tymi samymi rodzajami składników tła, które występują w zbiorze testowym. Stąd powstał pomysł rozszerzenia rozwiązania poprawiającego zrozumiałość mowy o podejście adaptacyjne, w którym wyjście wytrenowanego dodatkowego klasyfikatora, wytrenowanego do rozpoznawania złożoności tła (tj. czy zawiera dodatkowe zdarzenia dźwiękowe, pogłos i scenę dźwiękową), mogłoby wskazywać odpowiedni model do

separacji tła. System ten planujemy zbudować jako rozszerzenie opublikowanego artykułu.

3.2 Pre-trained Deep Neural Network Using Sparse Autoencoders and Scattering Wavelet Transform for Musical Genre Recognition

Niniejszy artykuł omawia metodę trenowania autoenkoderów z użyciem reprezentacji SWT w celu wstępnego wytrenowania sieci do rozpoznawania gatunku muzycznego. W celu wytrenowania autoenkoderów pobranych zostało 10 000 utworów z serwisu udostępniającego muzykę na darmowej licencji. Pierwszy z autoenkoderów wytrenowany został z użyciem wspomnianych utworów sparametryzowanych za pomocą SWT, natomiast każdy kolejny autoenkoder był trenowany na podstawie cech wydobytych z poprzednich autoenkoderów. Wagi wytrenowanych autoenkoderów posłużyły do inicjalizacji wag sieci neuronowej, której celem było rozpoznawanie gatunków muzycznych z bazy GTZAN, składającej się z 1000 plików reprezentujących 10 gatunków.

W pracy przeanalizowana i porównana została zmienność funkcji kosztu w trakcie treningu sieci z wstępnym wytrenowaniem i bez wstępnego wytrenowania, z uwzględnieniem od 1 do 4 warstw ukrytych. Wyniki pokazują, iż wstępne wytrenowanie za pomocą autoenkoderów pozwala uniknąć zjawiska nadmiernego dopasowania wag do danych trenujących (ang. overfitting), szczególnie gdy wzrasta liczba warstw ukrytych. Zaobserwowano również, że minimalna wartość bezwzględna funkcji kosztu, w przypadku sieci z wytrenowaniem wstępnym, była wyższa niż w przypadku sieci bez wstępnego wytrenowania. Wstępne wytrenowanie zadziałało zatem jak regularyzacja sieci, która zazwyczaj poprawia generalizowalność modelu kosztem spadku jego dokładności. Im głębszy wstępnie wytrenowany model tym jego zdolność generalizacji powinna być większa. Najniższa możliwa wartości funkcji kosztu wystąpiła w modelu z jedną warstwą ukrytą, bez wczesnego wytrenowania. Jednakże najlepsze rezultaty, jeśli chodzi o dokładność rozpoznawania gatunków, osiągnięto w przypadku dwóch warstw ukrytych ze wstępnym wytrenowaniem. Wskazuje to na rozbieżność między zmiennością wartości funkcji kosztu a zmiennością dokładności klasyfikacji relatywnie niewielkiego zbioru GTZAN. Przeprowadzone eksperymenty dają jednak wgląd w zachowanie funkcji kosztu podczas treningu modelu, w którym wagi zostały wcześniej wytrenowane bez nadzoru za pomocą autoenkoderów. Za pomocą dwóch warstw ukrytych uzyskano wynik 90.2% dokładności, co w porównaniu z innymi pracami [13] jest wysokim rezultatem. Przeprowadzona analiza funkcji kosztu wskazuje ponadto na wysoką efektywność SWT w reprezentowaniu gatunków muzycznych, co również potwierdza artykuł [1].

Użycie SWT do wytrenowania autoenkoderów za pomocą korpusu wielotysięcznych utworów muzycznych, celem wstępnego wytrenowania sieci neuronowej, zostało zaproponowane w omawianej pracy po raz pierwszy. Dalsze badania powinny uwzględniać modelowanie następstwa czasowego ramek SWT za pomocą sieci splotowych, rekurencyjnych lub tranformerów. Jak wskazują dotychczasowe badania [13, 32, 45], najlepsze rezultaty można uzyskać stosując podejście hybrydowe.

Wnioski wyciągnięte z powyższych eksperymentów zainspirowały autorów do rozwinięcia tej tematyki w kolejnym artykule, uwzględniającym nie tylko modelowanie pojedynczych ramek, ale również modelowanie sekwencji kilku ramek jednocześnie za pomocą sieci CNN.

3.3 Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation

Badania opisane w tym artykule dotyczą modelowania następstw czasowych kolejnych ramek SWT za pomocą CNN, celem ekstrakcji dodatkowych cech i poprawy skuteczności rozpoznawania symptomów chorobowych w nagraniach dźwiękowych bicia serca.

W sieciach CNN dokonywana jest operacja splotu na danych wejściowych za pomocą filtrów w warstwach konwolucyjnych (splotowych). Filtry te mogą być różnej wielkości (szerokości i wysokości) i mogą dokonywać splotu na kilku ramkach czasowych SWT jednocześnie. Modelowanie współzależności szerokości filtrów i długości ramek SWT, może przyczynić się do dokładniejszego rozpoznawania różnych aspektów sygnału bicia serca. Dlatego też stawiamy hipotezę, iż wykorzystanie różnej długości ramek SWT w treningu głębokich sieci splotowych przyczynia się do poprawy skuteczności wczesnego rozpoznawania chorób serca.

W celu sprawdzenia tej hipotezy przeprowadzono eksperymenty z udziałem nagrań pochodzących z bazy PASCAL [6]. We wstępnej fazie eksperymentów nagrania bicia serca zostały zakodowane z uwzględnieniem 3 różnych długości ramek SWT: 0,74s, 0,37s, oraz 0,185s. Następnie przeprowadzono eksperymenty klasyfikujące sygnał za pomocą zaproponowanej architektury sieci CNN. Przetestowano 8 różnych wartości szerokości filtra splotowego, od 1 do 8, oraz wymienione wcześniej trzy długości ramek SWT.

Wyniki pokazują, że dobór szerokości filtru oraz długości ramki SWT wpływa na skuteczność rozpoznawania różnych czynników chorobowych serca. Przykładowo, ponad 99% skuteczności uzyskano w rozpoznawaniu szmerów sercowych przy szerokości filtra równej 6, gdzie każda ramka reprezentuje sygnał o krótkiej długości (0.185s).

Dodatkowo, zaproponowana metoda skutecznie rozpoznaje artefakty w nagraniu, wskazując z dużą dokładnością (96%), kiedy nagranie powinno zostać powtórzone. Ten aspekt jest szczególnie istotny dla zastosowań medycznych, gdyż przeprowadzanie diagnozy na błędnie wykonanym nagraniu może być niebezpieczne dla pacjenta. Wysoką precyzję w rozpoznawaniu artefaktów nagrania uzyskano stosując ramkę czasową o długości 0,74s oraz szerokości filtru równej jeden.

Porównując wyniki do innych metod stosujących reprezentacje spektralne, proponowana metoda osiąga najwyższe wyniki, stosując znormalizowaną precyzję, która bierze pod uwagę nierówną liczbę nagrań w każdej kategorii w zbiorze testowym. Ponadto, proponowana metoda wyjątkowo skutecznie odróżnia prawidłowe bicie serca od chorobowego.

W ten sposób Hipoteza 2, iż wykorzystanie różnej długości ramek czasowych SWT w treningu sieci CNN przyczynia się do poprawy skuteczności wczesnego rozpoznawania chorób serca, została potwierdzona.

Poprawa skuteczności tej metody może uwzględniać, oprócz CNN do ekstrakcji cech, dodatkowe włączenie w proces klasyfikacji sieci rekurencyjnych bądź transformerów.

3.4 Beyond the Big Five Personality Traits for Music Recommendation Systems

W niniejszej pracy opisano metodę rozszerzenia danych wejściowych do systemu rekomendacji muzycznej o reprezentację cech osobowości słuchacza. Zastosowany model tzw. Wielkiej Piątki (ang. Big Five), mierzący pięć głównych cech osobowościowych (neurotyczność, ekstrawersja, otwartość na doświadczenia,

ugodowość i sumienność) został rozszerzony o dodatkowe piętnaście aspektów osobowościowych (cech drugorzędnych), po trzy dla każdej z pięciu głównych cech. Eksperymenty zostały przeprowadzone we współpracy z dr. hab. Włodzimierzem Strusem z Instytutu Psychologii Uniwersytetu Kardynała Stefana Wyszyńskiego w Warszawie. Psychologiczne aspekty tej pracy mają zatem silne fundamenty naukowe. W celu pozyskania danych osobowościowych użytkowników, które mogą zostać wykorzystane w systemach rekomendacji muzycznej, została stworzona aplikacja z wbudowanym formularzem osobowości o nazwie BFI-2. Aplikacja ta pozwala słuchać muzyki oraz ją oceniać z użyciem pięciostopniowej skali Likerta. W badaniu wzięło udział 279 uczestników. Zebrane w ten sposób dane zostały opublikowane, a na ich podstawie przeprowadzone zostały opisane niżej eksperymenty. Baza zawiera cechy osobowościowe uczestników, wyrażone dwudziestowymiarowym wektorem cech osobowości (5 głównych i 15 pobocznych), oceny utworów, oraz same utwory sparametryzowane do postaci 29 wymiarowego wektora cech. Cechy plików muzycznych uwzględniają cechy spektralne, amplitudowe, a także cechy wysokiego poziomu (takie jak klarowność rytmu czy nieharmonijność) oraz emocje rozpoznane w muzyce (takie jak strach, smutek, złość, czy radość).

Naszym celem było sprawdzenie hipotezy, iż wykorzystanie rozszerzonego modelu Wielkiej Piątki do reprezentacji osobowości użytkownika przyczynia się do redukcji błędu popełnionego przez systemy rekomendacji muzycznej, w porównaniu ze standardowym modelem osobowościowym.

Aby zweryfikować tę hipotezę, zaimplementowano hybrydowy system rekomendacji muzycznej oraz przeprowadzono ocenę jakości generowanych rekomendacji z zastosowaniem wcześniej zebranych danych. Oceny dokonywano porównując wyniki eksperymentów wykorzystujących pięciowymiarowy wektor cech osobowości z wynikami uwzględniającymi rozszerzony model dwudziestowymiarowy. Eksperymenty polegały na przeprowadzeniu predykcji oceny utworów przez użytkownika. Funkcja predykcyjna uwzględniała pomiar podobieństwa zarówno użytkowników, jak i utworów. Eksperymenty opierały się na poszukiwaniu optymalnego zbioru cech osobowości, które w wyniku pomiaru podobieństwa między użytkownikami skutkują najmniejszym błędem predykcji. Poszukiwania opierały się na braniu pod uwagę każdej możliwej kombinacji cech osobowości, zaczynając od pojedynczych cech i stopniowo zwiększając wymiarowość wektora cech stosowanego w obliczeniach podobieństwa. Poszukiwania przerywano, gdy zwiększanie wymiarowości skutkowało pogorszeniem wyników rekomendacji. W ten sposób wyselekcjonowano tylko ten podzbiór cech osobowości, który skutkował najniższym błędem popełnianym przez algorytm. Wyniki eksperymentów pokazały, iż zastosowanie cech osobowościowych niższego rzędu zwraca mniejszy błąd predykcji niż te same eksperymenty z wykorzystaniem cech Wielkiej Piątki bez cech niższego rzędu, potwierdzając jednocześnie Hipotezę 3.

Eksperymenty ujawniły ponadto optymalny podzbiór 4 aspektów osobowościowych, skutkujący najniższym błędem predykcji. Są to: ciekawość, odpowiedzialność, wrażliwość, oraz zaufanie. Odkrycie, iż nie wszystkie cechy osobowości są niezbędne w celu osiągnięcia najlepszych rezultatów, pośrednio przyczynia się również do poprawy satysfakcji użytkownika z użytkowania systemu. Użytkownik nie musi odpowiadać na wszystkie 60 pytań zawartych w kwestionariuszu osobowościowym; wystarczy, by odpowiedział na 16 pytań mierzących wyłącznie wspomniane 4 cechy osobowości.

3.5 Music Recommendation Systems: a Survey

Pomimo tego, iż ten artykuł nie wspiera bezpośrednio żadnej hipotezy, jego dołączenie w tym miejscu rozszerza przegląd literatury dotyczący systemów rekomendacji muzycznej. Artykuł zawiera przegląd tego typu systemów w kontekście przetwarzania sygnałów muzycznych oraz ich personalizacji. Publikacja ta

skupia się na podejściach skoncentrowanych na użytkowniku oraz danych kontekstowych, takich jak emocje, osobowość, czy dane uzyskane z platform społecznościowych. Artykuł omawia postępy w systemach rekomendacji muzycznej, w tym systemy uwzględniające głębokie sieci neuronowe. Praca ta przedstawia również wyzwania związane z przetwarzaniem dużej ilości danych w platformach streamingowych, oraz zarysowuje przyszłe kierunki badań w tej dziedzinie.

Rozdział 4

Podsumowanie

W rozprawie udowodniono wszystkie postawione hipotezy. Rozprawa zawiera także wyczerpujący przegląd najnowszej literatury dotyczącej systemów przetwarzania danych dźwiękowych, ze szczególnym uwzględnieniem sieci neuronowych. Do udowodnienia tez pracy przyczyniły się m.in. zdolności sieci neuronowych do odkrywania nowych cech audio, które często rozszerzają możliwości początkowych reprezentacji, np. poprzez zwiększenie ich zdolności dyskryminacyjnych. Podejście to zostało opisane w artykułach "Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation" oraz "Pre-trained Deep Neural Network Using Sparse Auto-encoders and Scattering Wavelet Transform for Musical Genre Recognition". W pracach tych opisano zastosowanie autoenkoderów, wraz z reprezentacją SWT, do wstępnego wytrenowania sieci neuronowej celem klasyfikacji muzyki oraz zastosowanie sieci splotowych, wraz z SWT, do klasyfikacji nagrań bicia serca. Przeprowadzone eksperymenty wskazały najbardziej optymalne wartości szerokości filtra oraz długości ramek SWT ze względu na skuteczność rozpoznawania nie tylko poszczególnych czynników chorobowych serca, ale także ze względu na wysoką precyzję w rozpoznawaniu artefaktów nagrania oraz odróżniania nagrania zdrowego bicia serca od chorobowego. Tym samym potwierdzono hipotezę, iż wykorzystanie zmiennej długości ramek czasowych SWT oraz splotowych sieci neuronowych przyczynia się do poprawy rozpoznawania wczesnych objawów chorób serca.

Zaproponowane w artykule "Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions" podejście do generowania zaszumionych korpusów mowy, dzięki wykorzystaniu czystych odpowiedników mowy zaszumionej, przyczyniło się do wytrenowania modeli operujących w domenie czasowej dźwięku, które dzięki automatycznej parametryzacji danych wejściowych, są w stanie nauczyć się separować mówców od siebie oraz mówców od tła. Zaproponowana metoda przyczynia się do rozwoju modeli służących do separacji mówców oraz poprawiających zrozumiałość mowy, odpornych na występowanie w nagraniach nieprzewidzianych zdarzeń dźwiękowych oraz pogłosu. Ponadto, zaproponowana metoda do generowania symulowanych nagrań może służyć do treningu wielu modeli rozwiązujących inne zadania, bez potrzeby generowania osobnych korpusów dla każdego z nich. Przykładami tych zadań może być klasyfikacja tła sceny dźwiękowej, zdarzeń dźwiękowych, lub próba predykcji czasu pogłosu. Możliwość dostosowywania składników symulowanych nagrań wykorzystując zjawisko znoszenia się fal dźwiękowych została zaproponowana w kontekście korpusów trenujących po raz pierwszy. Eksperymenty opisane w tym artykule nie wyczerpują w pełni potencjału zaproponowanego rozwiązania, zaś wielkość generowanych korpusów i ich zróżnicowanie umożliwia trening dużych i złożonych modeli, w tym trans-

formerów.

Rezultatem badań opisanych w pracy "Beyond the Big Five personality traits for music recommendation systems" jest m.in. publicznie dostępna baza danych zawierająca cechy osobowościowe 279 uczestników, wraz z ich preferencjami muzycznymi, oraz cechy dźwiękowe plików muzycznych. Baza tego typu została opublikowana po raz pierwszy. Z punktu widzenia analizy wpływu osobowości na preferencje muzyczne użytkownika stanowi ona cenne źródło danych nie tylko dla badaczy nauk technicznych, ale również psychologicznych. Zebrane dane pozwoliły na przeprowadzenie eksperymentów polegających na wzbogaceniu danych wejściowych do systemów rekomendacji muzycznej o dodatkowe dane osobowościowe słuchacza. W pracy wykazano, iż wykorzystanie modelu stanowiącego rozszerzenie standardowego modelu tzw. Wielkiej Piątki do reprezentacji osobowości użytkownika przyczynia się do redukcji błędu popełnianego przez systemy rekomendacji muzycznej, w porównaniu z modelem standardowym.

Dołączona do niniejszej rozprawy praca pt. 'Music Recommendation Systems: a Survey' omawia postępy w systemach rekomendacji muzycznej, w tym systemy uwzględniające głębokie sieci neuronowe. Praca ta przedstawia również wyzwania związane z personalizacją systemów rekomendacji muzycznej oraz przetwarzaniem dużej ilości danych w tych systemach. Przede wszystkim jednak, praca ta zarysowuje przyszłe kierunki badań w tej dziedzinie.

Rozwój sztucznej inteligencji zawdzięczamy rozwojowi sieci neuronowych oraz ich możliwościom wykrywania cech z danych wejściowych. Rozwój ten jest możliwy dzięki dostępności dużej ilości danych, zaś upublicznienie prezentowanych w rozprawie zasobów umożliwia innym badaczom wniesienie swojego wkładu do tych badań. Mamy nadzieję, iż niniejsza rozprawa i płynące z niej wnioski przyczynią się do dalszego rozwoju m.in. systemów klasyfikacji dźwięku oraz przetwarzania sygnałów mowy.

Rozdział 5

Bibliografia

[1] Andén, J., and Mallat, S. Deep scattering spectrum. *IEEE Transactions on Signal Processing 62*, 16 (2014), 4114–4128.

[2] Arias-Vergara, T., Klumpp, P., Vasquez-Correa, J. C., Nöth, E., Orozco-Arroyave, J. R., and Schuster, M. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Analysis and Applications 24* (2021), 423–431.

[3] Arnault, A., Hanssens, B., and Riche, N. Urban sound classification: striving towards a fair comparison. *arXiv preprint arXiv:2010.11805* (2020).

[4] Ballesteros, D. M., Rodriguez-Ortega, Y., Renza, D., and Arce, G. Deep4snet: deep learning for fake speech classification. *Expert Systems with Applications 184* (2021), 115465.

[5] Bansal, V., Pahwa, G., and Kannan, N. Cough classification for covid-19 based on audio mfcc features using convolutional neural networks. In *2020 IEEE international conference on computing, power and communication technologies (GUCON)* (2020), IEEE, pp. 604–608.

[6] Bentley, P., Nordehn, G., Coimbra, M., Mannor, S., and Getz, R. The pascal classifying heart sounds challenge 2011 (chsc2011). *Recuperado (Ago 2015) de: http://www. peterjbentley. com/heartchallenge/index. html* (2011).

[7] Cheng, K. W., Chow, H. M., Li, S. Y., Tsang, T. W., Ng, H. L. B., Hui, C. H., Lee, Y. H., Cheng, K. W., Cheung, S. C., Lee, C. K., et al. Spectrogram-based classification on vehicles with modified loud exhausts via convolutional neural networks. *Applied Acoustics 205* (2023), 109254.

[8] Chetupalli, S. R., and Habets, E. A. Speech separation for an unknown number of speakers using transformers with encoder-decoder attractors. In *INTERSPEECH* (2022), pp. 5393–5397.

[9] Chorowski, J., Weiss, R. J., Bengio, S., and Van Den Oord, A. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing 27*, 12 (2019), 2041–2053.

[10] Clifford, G. D., Liu, C., Moody, B., Springer, D., Silva, I., Li, Q., and Mark, R. G. Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016. In *2016 Computing in cardiology conference (CinC)* (2016), IEEE, pp. 609–612.

[11] Dhelim, S., Aung, N., Bouras, M. A., Ning, H., and Cambria, E. A survey on personality-aware recommendation systems. *Artificial Intelligence Review* (2022), 1–46.

[12] Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 30* (2021), 829–852.

[13] Gan, J. Music feature classification based on recurrent neural networks with channel attention mechanism. *Mobile Information Systems 2021*, 1 (2021), 7629994.

[14] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[15] Hu, Y., Chen, C., Zou, H., Zhong, X., and Chng, E. S. Unifying speech enhancement and separation with gradient modulation for end-to-end noise-robust speech separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), IEEE, pp. 1–5.

[16] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4700–4708.

[17] Iandola, F. N. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360* (2016).

[18] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., and Pedrycz, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications* (2023), 122666.

[19] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems 25* (2012).

[20] Lee, J.-A., and Kwak, K.-C. Heart sound classification using wavelet analysis approaches and ensemble of deep learning models. *Applied Sciences 13*, 21 (2023), 11942.

[21] Liu, Z., Yao, G., Zhang, Q., Zhang, J., and Zeng, X. Wavelet scattering transform for ecg beat classification. *Computational and mathematical methods in medicine 2020*, 1 (2020), 3215681.

[22] Luo, J., Wang, J., Cheng, N., Xiao, E., Zhang, X., and Xiao, J. Tiny-sepformer: A tiny time-domain transformer network for speech separation. *arXiv preprint arXiv:2206.13689* (2022).

[23] Luo, Y., and Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing 27*, 8 (2019), 1256–1266.

[24] Mu, W., Yin, B., Huang, X., Xu, J., and Du, Z. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports 11*, 1 (2021), 21552.

[25] Mushtaq, Z., Su, S.-F., and Tran, Q.-V. Spectral images based environmental sound classification using cnn with meaningful data augmentation. *Applied Acoustics 172* (2021), 107581.

[26] Nanni, L., Maguolo, G., Brahnam, S., and Paci, M. An ensemble of convolutional neural networks for audio classification. *Applied Sciences 11*, 13 (2021), 5796.

[27] Pham, L. D., McLoughlin, I., Phan, H., and Palaniappan, R. A robust framework for acoustic scene classification. In *INTERSPEECH* (2019), pp. 3634–3638.

[28] Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (2015), pp. 1015–1018.

[29] Qi, J., et al. Speech disorder classification using extended factorized hierarchical variational auto-encoders. *arXiv preprint arXiv:2106.07337* (2021).

[30] Saldanha, J., Chakraborty, S., Patil, S., Kotecha, K., Kumar, S., and Nayyar, A. Data augmentation using variational autoencoders for improvement of respiratory disease classification. *Plos one 17*, 8 (2022), e0266467.

[31] Sawhney, A., Vasavada, V., and Wang, W. Latent feature extraction for musical genres from raw audio. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)* (2021), pp. 2–8.

[32] Scarpiniti, M., Comminiello, D., Uncini, A., and Lee, Y.-C. Deep recurrent neural networks for audio classification in construction sites. In *2020 28th European Signal Processing Conference (EUSIPCO)* (2021), IEEE, pp. 810–814.

[33] Shin, H.-k., Park, S. H., and Kim, K.-w. Inter-floor noise classification using convolutional neural network. *Plos one 15*, 12 (2020), e0243758.

[34] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[35] Sturm, B. L. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461* (2013).

[36] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2818–2826.

[37] Szyca, B., Wejda, B., Muchewicz, M., and Kostek, B. Exploring music listening patterns: an online survey. *International Journal of Electronics and Telecommunication 70*, 2 (2024), 367–372.

[38] Takahashi, N., Parthasaarathy, S., Goswami, N., and Mitsufuji, Y. Recursive speech separation for unknown number of speakers. *Interspeech 2019* (2019).

[39] Tan, M. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946* (2019).

[40] Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[41] Verma, P., and Berger, J. Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions. *arXiv preprint arXiv:2105.00335* (2021).

[42] Weinberger, S. H., and Kunath, S. A. The speech accent archive: towards a typology of english accents. In *Corpus-based studies in language use, language learning, and language documentation*. Brill, 2011, pp. 265–281.

[43] Zaman, K., Direkoğlu, C., et al. Classification of harmful noise signals for hearing aid applications using spectrogram images and convolutional neural networks. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (2020), IEEE, pp. 1–9.

[44] Zhang, Y., Li, B., Fang, H., and Meng, Q. Spectrogram transformers for audio classification. In *2022 IEEE International Conference on Imaging Systems and Techniques (IST)* (2022), IEEE, pp. 1–6.

[45] Zhang, Z., Xu, S., Zhang, S., Qiao, T., and Cao, S. Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing 453* (2021), 896–903.

[46] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8697–8710.

Rozdział 6

Artykuły dołączone do rozprawy

6.1 Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions

Dane bibliograficzne pracy:

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, & C. Schneider (Eds.), Harnessing Opportunities: Reshaping ISD in the post-COVID-19 and Generative AI Era (ISD2024 Proceedings). Gdańsk, Poland: University of Gdańsk. ISBN: 978-83-972632-0-8. https://doi.org/10.62036/ISD.2024.37.

Developing a Corpus for Polish Speech Enhancement by Reducing Noise, Reverberation, and Disruptions

Mariusz Kleć
Polish-Japanese Academy of Information Technology
Warsaw, Poland *mklec@pjwstk.edu.pl*

Krzysztof Szklanny
Polish-Japanese Academy of Information Technology
Warsaw, Poland *kszklanny@pjwstk.edu.pl*

Alicja Wieczorkowska
Polish-Japanese Academy of Information Technology
Warsaw, Poland *alicja@poljap.edu.pl*

Abstract

This paper presents a solution for generating corpora of simulated Polish speech recordings in complex acoustic environments. The proposed method introduces a layer of unpredictable sound events, in addition to the acoustic scene noise and reverberation, making the solution unique. We generated a corpus comprising over 277 hours of training examples and over 5.5 hours for testing purposes using publicly available data sources. Next, we trained several Conv-TasNet networks on the generated data to enhance single speech and separate two speakers from complex noise. The results of the experiments indicated the potential of the generated corpora for solving these tasks. Researchers can use publicly available code to create their corpora tailored to the Polish language and solve various speech-related tasks.
Keywords: speech denoising, speech separation, speech enhancement.

1. Introduction

Speech recordings can be a valuable source of information in various fields of science, such as linguistics, history, psychology, sociology, and medicine, to name a few. However, non-professional microphones, various acoustic environments, and casual settings can significantly affect the intelligibility of speech. The audio files obtained this way are often noisy, with reverberation and random sound events like car horns or dog barking. Additionally, when two or more people are involved in the conversation, their utterances sometimes occur concurrently (i.e. crosstalk occurs). These issues greatly influence the performance of Automatic Speech Recognition (ASR) services, which require one person to speak at a time and a signal of relatively high signal-to-noise ratio (SNR) to transcribe speech accurately. Therefore, developing speech enhancement and speech separation methods are key preprocessing steps for ASR, speech corpus analysis, and real-time communication.

The speech enhancement can be accomplished by speech denoising [34], increasing the resolution of the signal [21], or by conditional speech synthesis [1]. However, this task can be challenging due to the complex and dynamic nature of the acoustic environment. Various disturbances such as non-stationary noise, reverberation, and other acoustic phenomena and unpredictable sound events may complicate the denoising process further. Recent research has revealed that Deep Learning (DL) techniques are more effective in speech denoising than conventional methods, such as spectral subtraction [41], Wiener filtering [37], and non-negative matrix factorization [14]. The DL-based denoising techniques comprise models based on Wave-U-Net [7, 46], Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks

[12, 8], Generative Adversarial Networks (GAN) [29, 36], Transformers [17], and recently also models that do not require clean data for training [15].

Researchers have also used speech separation techniques to denoise speech signals, as mentioned in [9, 16, 20]. Speech separation is the separation of concurrent speakers in monophonic audio recordings. To this end, the researchers in most cases employ time-domain models, which are more accurate than models based on time-frequency representations. The latter require estimating phase information, which can lead to distortions and inaccuracies during signal reconstruction [38]. Among the successful time-domain models, the Conv-TasNet neural network [23] has proven to be very effective in speech separation, outperforming other time-frequency methods. Another effective model is the Hybrid Tasnet network [43], which integrates the time and frequency domains to improve separation performance. In another paper [22] the model leverages Recurrent Neural Network for utterance-level sequence modelling. Wavesplit [44] is another end-to-end speaker separation model that achieves very high efficiency in various speech separation tasks, including clean mixtures of 2 speakers from WSJ0-2Mix dataset [10], and in noisy and reverberated settings from WHAMR dataset [24]. The authors of [44] obtained 22.2 dB on WSJ0-2Mix and 13.2 dB of Signal-to-Distortion Ratio (SI-SDR) on WHAMR. Moreover, a recent study [38] investigated the use of transformer architecture called SepFormer for the speech separation task, yielding promising results of 22.3 dB of SI-SDR on the WSJ0-2Mix. Additionally, the MossFormer2 model [45] currently achieves the best speech separation results for the Libri2Mix [4] and WSJ0-2Mix [10] datasets, achieving 24.1 dB and 21.7 dB of SI-SDR, respectively.

1.1. Contribution

Real-life speech recordings are often made during spontaneous situations and in uncontrolled acoustic environments. As a result, they can contain a lot of noise, which seriously affects speech intelligibility and may prevent further use of such recordings. The speech enhancement methods are constantly being developed, but most of them use English data sources and try to remove non-stationary noise and reverberation [44, 45]. Our approach is unique in that we build the speech corpora explicitly using the Polish language and introduce unpredictable sound events as an additional layer of noise, in addition to the acoustic scene signals and reverberation. The script we have published allows generating any number of such simulated real-world noisy speech recordings based on publicly available data sources. The corresponding components of speech recordings, such as clean speech, events, acoustic scenes and reverberation, are saved in separate files that prepare the created corpora for training deep models of Polish speech enhancement in noisy, reverberated and disturbed acoustic environments. Additionally, adding a phase-inverted version of one of the corresponding components to a simulated speech cancels that component from the file. This feature makes the created corpora higly customizable and easily adjustable for various other problems like scene and event recognition or dereverberation.

Using the script, we created a corpus containing over 277 hours of training examples and over 5.5 hours for testing purposes. Using the data generated by the script, we trained three models. The first enhances single speech by separating it from complex noise. The second model is used for speech separation when two speakers occur concurrently against a background noise. The third model separates two speakers, which occurs without any noise. We evaluated our models using the prepared test set and the Libri2Mix [4]. We also compared the performance of our models with other pre-trained solutions.

2. Corpus for Polish Speech Enhancement

Deep learning models for speech enhancement require a large number of noisy recordings and their corresponding clean speech signals for training. However, obtaining clean speech signals

can be difficult and expensive, often requiring access to professional recording studios. To address this problem, we have developed a script [1] that generates noisy speech corpora from the corresponding clean speech signals, which makes it ready to be used for training deep models. In the script, we use publicly available high quality data sources to create the corpora. The speech data sources are in Polish, and each generated audio file comprises a mix of seven distinct layers, as shown in Figure 1.

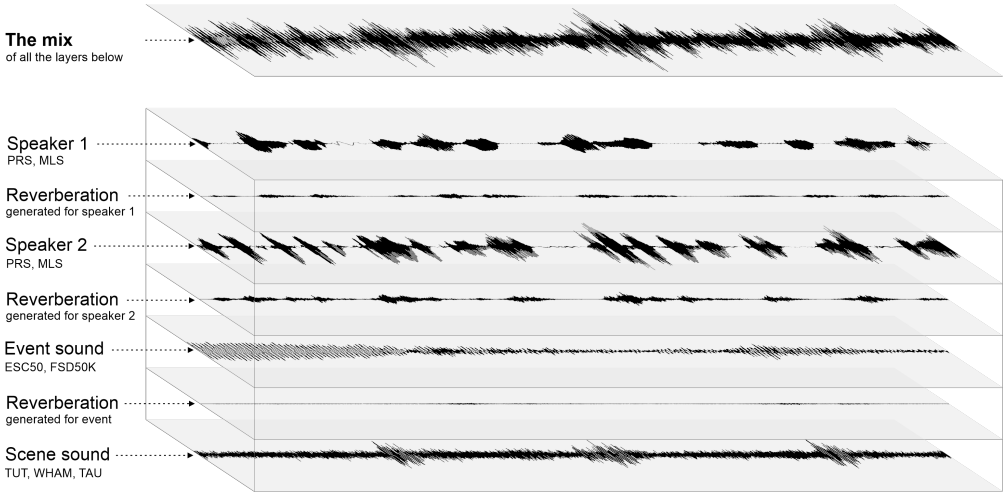


Fig. 1. The structure of a simulated speech recording generated by our script. The first four layers consist of two speech signals and the corresponding reverberation signals, based on room impulse responses. The speech signals are sourced from the Polish Read Speech corpus (PRS) [18] and the Polish section of The Multilingual LibriSpeech (MLS) [32]. Layers five and six contain additional sound events and their corresponding reverberation. The recordings of sound events are sourced from the ESC50 [31] and FSD50K [5] datasets. The final layer of the mix comprises the sound of the environmental scene taken from TUT Acoustic Scenes 2017 [26], TAU Urban Acoustic Scenes 2019 [25], and the WHAM [42] datasets.

Each layer of the mix is saved as a separate audio file in its designated folder. This allows for easy cancellation of a component from the mix by adding its phase-inverted version. It is possible to adjust the created corpora to a specific task using this principle. For instance, one speaker can be removed from the mix by adding its phase-inverted version. The same should be done with the reverberation for this speaker in such a case to remove this speaker completely. Therefore, it is easy to obtain the corpus with one speaker instead of the two without the need to generate the corpus again. Other possibilities are also feasible; exemplary ideas are presented in Table 1.

Our goal was to develop a solution that is accessible to everyone to stimulate the research in this field. To achieve this, each layer of the mix required to contain publicly available, high-quality data sources. In our quest for data sources to use in particular layers of the mixes, we provide an overview of the most popular data sources in the following section.

2.1. Review of Existing Data Sources

Many data sources are released when scientific challenges and workshops are organized, but most of them are in English and only occasionally provide sources of noisy and corresponding clean data. For example, the CHiME [2] challenge frequently releases new datasets. The CHiME-5 [3] provides conversational speech recordings in everyday home environments, but is

[1] https://github.com/mklec/PolSMSE
[2] https://www.chimechallenge.org/

Table 1. Possible corpus adjustments to make it suitable for training different speech enhancement or speaker separation problems. This can be achieved by cancelling a given component from the mix without generating the new corpus from scratch. Each column shows the layers left in a mix to solve a particular problem. For instance, one can remove one speaker while leaving another, or cancel out all reverberation. Other examples include cancelling the background scene while retaining only the sound events, with or without reverberation, etc.

Layers of the mix	Speaker separation							Speech enhancement				
Speaker 1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speaker 2	✓	✓	✓	✓	✓	✓	✓					
Scene sound	✓			✓		✓		✓	✓	✓		
Event sound	✓	✓		✓	✓			✓	✓		✓	
Reverb for speaker 1	✓	✓	✓					✓				✓
Reverb for speaker 2	✓	✓	✓									
Reverb for event	✓	✓						✓				

intended to advance ASR performance rather than improve speech enhancement. On the other hand, the Deep Noise Suppression challenge (DNS) [34] focuses on improving overall speech quality when recorded in a challenging background. The dataset from this challenge contains over 10000 hours of noise and over 1000 hours of clean speech signals for training the noise suppression models and a representative test set of real-world scenarios consisting of both synthetic and real recordings.

The Wall Street Journal (WSJ0) corpus provides a speech source for creating other corpora. WSJ0-2Mix [10] takes two speakers and mixes them at speech-to-speech ratios (SSR) between 0 and 5 dB. It provides 30 hours of training examples. However, WSJ0 is not open-sourced. Therefore, the Libri2Mix [4] dataset is an alternative to WSJ0-2Mix. It is based on LibriSpeech [28], and consists of mixes of two speakers combined with ambient noise sampled from the WHAM [42] dataset. It provides 212 hours of training examples. The WHAMR dataset [24] extends the WSJ0-2Mix by noise from the WHAM and reverberation and mixes with the loudest speaker at SNR between -6 and 3 dB. It provides 58 hours of noisy and corresponding clean speech examples for training.

Other corpora can also provide valuable speech sources; however, their recordings often contain distortions and reverberation and are made with poor-quality microphones. Therefore, their usage for speech enhancement should include a selection of the best quality candidates. One example of such a data source is VoxCeleb [13], containing over 100000 utterances from more than 6000 speakers. Another example is the multi-language Common Voice dataset [2], which also contains the Polish subset of 177 hours of spontaneous speech. Librivox project [3] is another example of a multi-language source of speech data. It contains recordings of volunteers reading over 10000 public-domain books in various languages, also in Polish. The Polish clean and high-quality recordings are available from the Polish Read Speech corpus (PRS) [18]. It provides 56 hours of recordings featuring phonetically rich words and sentences spoken by 317 speakers in an acoustically treated recording studio. The Multilingual LibriSpeech (MLS) [32] is a multilingual dataset of read books (audiobooks), and the Polish section contains 137 hours of clean speech from 16 speakers reading 25 books. Finally, there are over 140 corpora available in the Common Language Resources and Technology Infrastructure (CLARIN) from the Polish language [11, 30].

As for the environmental scene and noise data sources, Audioset [6] provides a collection of about 2 million human-labeled 10s sound clips extracted from YouTube videos, which belong to about 600 audio classes. The TUT Acoustic Scenes 2017 dataset [26] includes 52 hours of

[3] https://librivox.org/

audio recordings from 15 locations, such as homes, city centres, forest paths, grocery stores, metro stations, and more. The WHAM dataset [42] provides 80 hours of background noise from urban environments, such as restaurants, bars, cafes, and parks. The TAU Urban Acoustic Scenes dataset 2019, introduced in [25], features 40 hours of audio recordings from various urban acoustic scenes such as pedestrian streets, trams, and airports.

The sound events can also be downloaded from the ESC50 [31], which comprises 2000 recordings, each five seconds long, classified into 50 semantic classes such as clapping, vacuum cleaners, fireworks, and more. The FSD50K [5] contains over 50000 audio clips, amounting to more than 100 hours of audio, organized into 200 classes drawn from the AudioSet Ontology. Examples of these classes include various musical instruments, splashes, zippers, telephones, and many others.

2.2. The Corpus Creation

We considered factors such as license, accessibility, language, and data quality when selecting the sources for our solution. Ultimately, we chose two spoken Polish recordings sources, PRS and MLS, and obtained the sounds of real-world environments from the TUT, WHAM, and TAU datasets. Sound events were sourced from the ESC50 and FSD50K datasets. Finally, we created simulated recordings of noisy speech by combining signals from the aforementioned sources and applying reverberation, using the following formula.

$$y(t) = \sum_{i=1}^{C=2} (s_i(t) + r_i(t)) + b(t) + e(t) + r_e(t) \tag{1}$$

where $y(t)$ represents the simulated speech recording in the time domain with a maximum of C speakers, $C = 2$ in our case. The signal $s_i(t)$ represents the ith speaker in the mix, which is mixed with their corresponding room response $r_i(t)$, generated earlier as reverberation. The signal $b(t)$ denotes the scene's ambient sound, and $e(t)$ represents a non-speech sound event, along with its corresponding reverberation $r_e(t)$. In this context, the background noise refers to the sum of four components: $r_i(t)$, $b(t)$, $e(t)$, and $r_e(t)$.

First, we divided all files from different sources into three subsets: training, validation, and testing, according to the instructions provided by each data source. This ensures that no mix component in the training subset is used for testing or validation. Next, we excluded files with speech-related events, such as whispering or singing, from the event sources, to avoid conflicts with speech layers. Figure 2 shows the remaining event classes used for creating our corpus. The two speakers were mixed at a speech-to-speech ratio that is randomly selected from -5 to 5 decibels, rounded to the nearest whole decibel. The resulting mix includes a randomly selected 4-second fragment from the given source files, making the final mixes always different.

In order to recreate the actual recording's conditions as much as possible we took into account several characteristics of them. To ensure that the selected sound events are suitable for a particular scene, we created a matrix that maps scene classes to possible event classes. This step helps prevent the random selection of mismatched events, such as the sound of a cow in an airport, which would be absurd. The matrix provides guidelines for the event and specific scene classes when mixed with the script. Additionally, reverberation was generated in Matlab only when the scene class represents an indoor category. We manually defined a dictionary with possible reverb parameters range for these classes to ensure that the characteristics of generated room reflections are suited to the particular scene class. The reverb parameters were randomly selected each time, but only from such predefined range. This approach allows us to avoid generating unsuitable reverbs, such as long reverbs for the library class, which typically has a short decay time. We controlled parameters like decay time, reflection diffusion, strength of high-frequency damping, and early reflection time in this manner. The same values of these

parameters were applied to the speech and event sources, except with different early reflection times for events, as these two sound sources usually have different placements in the recording room, affecting the time when the microphone captures their first reflection.

We created a training subset of 250000 files of $y(t)$ using our script. The testing and validation subsets contain 5000 files each. The duration of each file is 4 seconds, and the mix layers are saved in separate files, at 8000 Hz and 16 bits. This setting is justified by using the same values in [23]. However, generating the files encoded in 16 kHz is also possible by the script. The training subset provides over 277 hours (1 million seconds) of continuous noisy speech along with the corresponding clean sources and other layers for training purposes. This corpus is called PolSMSE-2-Noisy and is intended to separate speakers into two channels when they speak simultaneously over a noisy background, i.e., to separate noisy signals (SN). Therefore it contains recordings for all layers shown in Fig. 1. Next, we used the phase-inversion phenomenon to cancel out one speaker and their reverberation, creating another corpus called PolSMSE-1-Noisy. This corpus is intended to enhance single speech (ES), containing only one speech signal over the noisy background. The noisy background refers to all the other layers of the mix, except the speech.

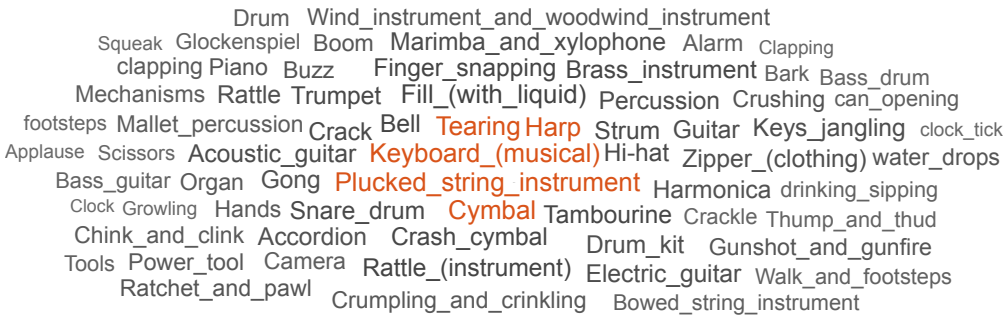


Fig. 2. The word cloud showing the classes of events used in the created corpus. For the sake of clarity, the least common events have been excluded. Four of the most frequent event classes have been highlighted in red. The most common event classes are related to musical instruments.

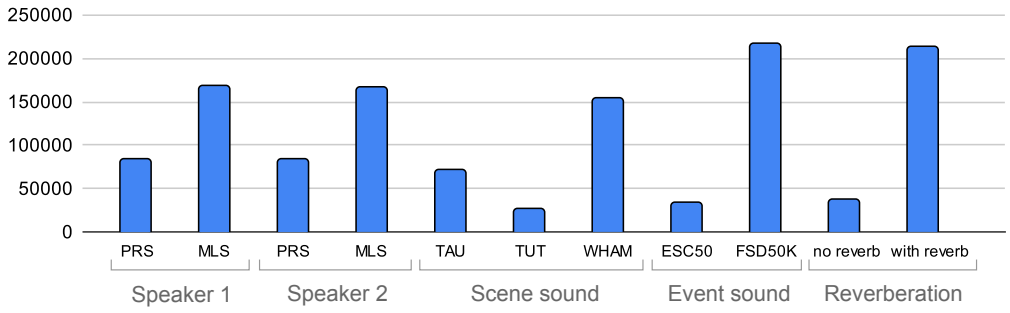


Fig. 3. The histogram illustrates each data source's contribution to creating layers of the PolSMSE corpus. The y-axis denotes the number of training examples that contain a given data source.

3. Experiments

Our research aimed to illustrate the feasibility of developing a corpus according to the ideas presented in Section 2.2 and utilizing it in practical experiments involving speech separation and enhancement through deep learning. To the best of our knowledge, no other speech corpora exhibit such a complex and noisy recording environment. The published script enables the

recreation of the testing subset, ensuring the comparison of the results and hopefully achieving state-of-the-art results with our testing subset in future work.

We utilized the Conv-TasNet architecture initially designed for speech separation [23]. Conv-TasNet uses a linear encoder to generate a representation optimized for learning masks for the speakers. Next, the mask is applied to the encoder output, which is then inverted back to the waveforms, representing the speech of separated speakers. The network finds the masks using a temporal convolutional network (TCN) consisting of stacked 1-D dilated convolutional blocks. We reduced the number of filters in the encoder and decoder from 512 to 256, which decreased the model's capacity, but significantly accelerated the calculations. Other hyperparameters of the network are: the length of the input filters equal to 20 samples, and 32 convolutional blocks in the TCN. The loss functions and training procedure followed [23], with the initial learning rate set to 1e-3 and Adam used as the optimizer. If the validation set's accuracy did not improve over two consecutive epochs, the learning rate was halved. The models were implemented in Matlab and trained from scratch for fifteen epochs on a single GPU GeForce RTX 3080.

We trained three models using PolSMSE to estimate $s_i(t)$ from $y(t)$ (see Equation 1). The first model was trained to enhance the speech (ES) using PolSMSE-1-Noisy, containing a single speaker in complex background noise. The second model was trained to separate two speakers conversing in a noisy environment and events (SN) using PolSMSE-2-Noisy, containing all the layers of speech and noise depicted in Figure 1. The third model was trained to separate two clean speeches (SC) using a mixture of two clean speakers without any noise, events, or reverberation. This version of the dataset will be referred to as PolSMSE-2-Clean. Figure 4 illustrates the objective of training the models based on the provided input data and desired output.

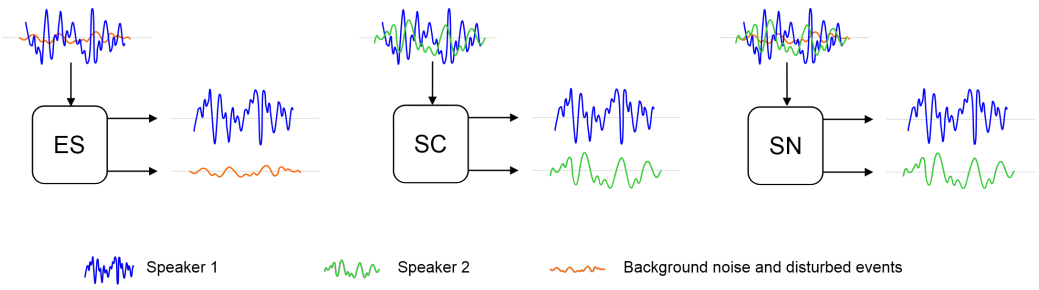


Fig. 4. The inputs and target outputs we utilized to train the experimental models. These models were created to address the following problems: improving the quality of single-speaker speech (ES), separating speech signals with noisy backgrounds (SN), and isolating clean speech signals without any interference (SC).

In evaluating the models' performance, we used the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [19]. Higher SI-SDR values indicate better separation quality. Besides, we used other pre-trained models and datasets to provide a more comprehensive evaluation. The first pre-trained model, SepFormer [38], was trained for speaker separation using Libri2Mix-Clean from SpeechBrain [33]. It achieved a 20.6 dB SI-SNR with this dataset. Additionally, the second pre-trained model, referred to as CTNoar [40], is the Conv-TasNet network pre-trained with 2-speaker mixtures from the WSJ0-2mix datasets, containing clean speech signals without noise. It achieved a 14.6 dB SI-SNR with this dataset.

4. Results and Discussion

The preliminary results summarized in Table 2 suggest that the ES model effectively enhances single speech by isolating it from complex noise, disturbing sound events, and reverberation,

Table 2. The dB values of SI-SDR representing the performance of three Conv-TasNet models trained using the PolSMSE and tested using both the PolSMSE and LibriMix testing subsets. These datasets come in two versions: clean and noisy. The clean version comprises a mixture of clean speech signals devoid of noise, events, or reverberation. The first model aims to separate two speakers in their respective channels when they are present amidst complex noise (SN). The second model is designed to separate two clean speech signals (SC). The third model was trained to isolate single speech from noise (ES). The table also includes other pre-trained networks: SepFormer, pre-trained with Libri2Mix-Clean from SpeechBrain [33], and another Conv-TasNet (CTNoar) [40], trained with clean WSJ0-2mix. Best results are shown in bold.

Speech separation				Single speech enhancement		
Testing subset	SN	SC	SepF	Testing subset	ES	CTNoar
PolSMSE-2-Noisy	**1.69**	-3.55	-0.48	PolSMSE-1-Noisy	**10.24**	-1.39
PolSMSE-2-Clean	5.28	6.64	**18.25**	Libri1Mix-Noisy	**11.97**	-0.64
Libri2Mix-Noisy	4.59	1.73	**6.84**	-	-	-
Libri2Mix-Clean	6.71	7.85	**20.56**	-	-	-

even with a short training time limited to fifteen epochs. The results show almost 12 dB of SI-SDR for Libri1Mix-Noisy and over 10 dB for PolSMSE-1-Noisy, underscoring the proposed solution's potential. It is important to note that these datasets contain speech mixed with noise. However, the Libri1Mix-Noisy contains English speech without reverberation or disturbing sound events, explaining slightly better results in this case.

Despite the short training time, the SN model, trained with PolSMSE-2-Noisy, outperforms SepFormer, achieving an SI-SDR of 1.69 dB compared to -0.48 dB. These results indicate that the created corpus offers valuable data for addressing the speaker separation problem, especially in challenging and complex noise environments. The difference in the results also highlights the limitations of models trained with clean speech signals, as SepFormer was trained with Libri2Mix-Clean. The reported SI-SDR result of 20.56 dB for SepFormer with Libri2Mix-Clean aligns with previous research results [39].

In our evaluation of single speech enhancement, we compared our model to the CTNoar, which had been originally trained on a mix of 2 and 3 speech signals to separate one signal and put the others in a separate channel. The hypothesis was that the CTNoar model could identify one speaker and separate the noise into a distinct channel when tested with a noisy dataset. However, the results showed negative SI-SDR values for PolSMSE-1-Noisy and Libri1Mix-Noisy, indicating that this model was ineffective at enhancing single speech in a noisy environment. Our ES model successfully addresses this issue, achieving 10.24 dB for PolSMSE-1-Noisy and 11.97 dB for Libri1Mix-Noisy.

The research detailed in [16] demonstrated that a single-channel time-domain denoising technique could reduce the word error rate (WER) by 30%. These findings motivate us to enhance our models in the future by integrating more data sources, generating additional training examples, increasing the model's capacity, and extending the model training time. Furthermore, the speech data sources should include transcriptions to help other researchers evaluate their models with ASR in terms of WER. Currently, only the MLS data source contains the transcription. Another potential source is the newly released Polish speech corpus discussed in [30], which offers transcriptions and conversational Polish speech, complementing the supervised speech recordings used in the current study.

Turning off specific training layers, as presented in Table 1, may also help address other problems, such as event recognition or dereverberation in complex and non-stationary background noise. Additionally, our future research will further explore this specific feature of the proposed solution, by investigating the effect of particular noise layers on the performance of speech separation, enhancement, and recognition tasks. The corpus can also be utilized to train

voice activity detection, which is crucial for effectively operating in unpredictable and noisy environments, for example, recognizing speech in cars or voice-controlled machines [27, 35]. Accurately distinguishing between noise and speech is essential in such scenarios.

5. Conclusion

This paper addresses the challenges of enhancing and separating speech from noisy, reverberant, and disrupted backgrounds, especially in the context of Polish speech recordings. The publicly available, customizable, and scalable corpora generated by the proposed data generation script can be valuable for training deep models and facilitating further research. We hope the proposed solution will contribute to developing new models that leverage the layers' interrelation in the data and potentially set a new state-of-the-art for separating Polish speech from complex noise with disruptions and reverberation.

References

1. E. A. AlBadawy, A. Gibiansky, Q. He, J. Wu, M.-C. Chang, and S. Lyu. Vocbench: A neural vocoder benchmark for speech synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 881–885. IEEE, 2022.
2. R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, 2020.
3. J. Barker, S. Watanabe, E. Vincent, and J. Trmal. The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines. In *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*, 2018.
4. J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent. Librimix: An open-source dataset for generalizable speech separation. 2020.
5. E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.
6. J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.
7. H. R. Guimarães, H. Nagano, and D. W. Silva. Monaural speech enhancement through deep wave-u-net. *Expert Systems with Applications*, 158:113582, 2020.
8. X. Hao, X. Su, R. Horaud, and X. Li. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6633–6637. IEEE, 2021.
9. T. Hasumi, T. Kobayashi, and T. Ogawa. Investigation of network architecture for single-channel end-to-end denoising. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 441–445. IEEE, 2021.
10. J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 31–35. IEEE, 2016.
11. E. Hinrichs and S. Krauwer. The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, May 2014.

12. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*, 2020.
13. J. Huh, A. Brown, J.-w. Jung, J. Son Chung, A. Nagrani, D. Garcia-Romero, and A. Zisserman. Voxsrc 2022: The fourth voxceleb speaker recognition challenge. *arXiv e-prints*, pages arXiv–2302, 2023.
14. H. Kagami, H. Kameoka, and M. Yukawa. Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2018.
15. M. M. Kashyap, A. Tambwekar, K. Manohara, and S. Natarajan. Speech denoising without clean training data: A noise2noise approach. *arXiv preprint arXiv:2104.03838*, 2021.
16. K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7009–7013. IEEE, 2020.
17. Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi. Speech enhancement using self-adaptation and multi-head self-attention. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185. IEEE, 2020.
18. D. Koržinek, K. Marasek, and Ł. Brocki. Polish read speech corpus for speech tools and services. In *CLARIN Annual Conference 2016 in Aix-en-Provence, France*, 2016.
19. J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.
20. D. Lee, S. Kim, and J.-W. Choi. Inter-channel conv-tasnet for multichannel speech enhancement. *arXiv preprint arXiv:2111.04312*, 2021.
21. T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson. Time-frequency networks for audio super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2018.
22. Y. Luo, Z. Chen, and T. Yoshioka. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50. IEEE, 2020.
23. Y. Luo and N. Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266, 2019.
24. M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux. Whamr!: Noisy and reverberant single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2020.
25. A. Mesaros, T. Heittola, and T. Virtanen. A multi-device dataset for urban acoustic scene classification. In *Scenes and Events 2018 Workshop (DCASE2018)*, page 9.
26. A. Mesaros, T. Heittola, and T. Virtanen. Acoustic scene classification: an overview of dcase 2017 challenge entries. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 411–415. IEEE, 2018.
27. S. Mihalache and D. Burileanu. Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection. *Sensors*, 22(3):1228, 2022.

28. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.
29. S. Pascual, A. Bonafonte, and J. Serrà. Segan: Speech enhancement generative adversarial network. *Interspeech 2017*, 2017.
30. P. Pęzik, S. Karasińska, A. Cichosz, Ł. Jałowiecki, K. Kaczyński, M. Krawentek, K. Walkusz, P. Wilk, M. Kleć, K. Szklanny, et al. Spokesbiz–an open corpus of conversational polish. *arXiv e-prints*, pages arXiv–2312, 2023.
31. K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.
32. V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert. Mls: A large-scale multilingual dataset for speech research. 2020.
33. M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.
34. C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. 2020.
35. D. Rho, J. Park, and J. Ko. Nas-vad: Neural architecture search for voice activity detection. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 3754–3758. International Speech Communication Association, 2022.
36. A. Satheesh and K. Muthu-Manivannan. Denoising speech signals with hifi-coulomb-gans. *Journal of Student Research*, 11(3), 2022.
37. P. Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632. IEEE, 1996.
38. C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2021.
39. C. Subakan, M. Ravanelli, S. Cornell, F. Grondin, and M. Bronzi. Exploring self-attention mechanisms for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
40. N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji. Recursive speech separation for unknown number of speakers. *Interspeech 2019*, 2019.
41. N. Upadhyay and A. Karmakar. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. *Procedia Computer Science*, 54:574–584, 2015.
42. G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*, 2019.
43. G.-P. Yang, C.-I. Tuan, H.-Y. Lee, and L.-s. Lee. Improved speech separation with time-and-frequency cross-domain joint embedding and clustering. *arXiv preprint arXiv:1904.07845*, 2019.
44. N. Zeghidour and D. Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021.
45. S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang, T. H. Nguyen, K. Zhou, J. Q. Yip, D. Ng, and B. Ma. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP 2024-2024 IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 10356–10360. IEEE, 2024.

46. S. Zhao, T. H. Nguyen, and B. Ma. Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6648–6652. IEEE, 2021.

6.2 Pre-trained Deep Neural Network Using Sparse Autoencoders and Scattering Wavelet Transform for Musical Genre Recognition

Dane bibliograficzne pracy:

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, http://dx.doi.org/10.7494/csci.2015.16.2.133

Computer Science • 16 (2) 2015 http://dx.doi.org/10.7494/csci.2015.16.2.133

Mariusz Kleć
Danijel Koržinek

PRE-TRAINED DEEP NEURAL NETWORK USING SPARSE AUTOENCODERS AND SCATTERING WAVELET TRANSFORM FOR MUSICAL GENRE RECOGNITION

Abstract *Research described in this paper tries to combine the approach of Deep Neural Networks (DNN) with the novel audio features extracted using the Scattering Wavelet Transform (SWT) for classifying musical genres. The SWT uses a sequence of Wavelet Transforms to compute the modulation spectrum coefficients of multiple orders, which has already shown to be promising for this task. The DNN in this work uses pre-trained layers using Sparse Autoencoders (SAE). Data obtained from the Creative Commons website jamendo.com is used to boost the well-known GTZAN database, which is a standard benchmark for this task. The final classifier is tested using a 10-fold cross validation to achieve results similar to other state-of-the-art approaches.*

Keywords Sparse Autoencoders, deep learning, genre recognition, Scattering Wavelet Transform

Citation Computer Science 16 (2) 2015: 133–144

1. Introduction

Genre recognition has been a staple of Music Information Retrieval (MIR) since the very beginning. Initial approaches relied mostly on Data Mining and Natural Language Processing, but audio analysis became popular when Machine Learning techniques improved to a substantial degree. MIR, as a concept, involves many diverse fields of study: classification, analysis, organization, recommendation, and various areas of research: signal processing, music theory, linguistics, sociology, psychology, and others. Many tasks that involve MIR will concentrate on a single problem by utilizing a particular method, but we are more often faced with projects that involve a variety of concepts spanning a couple of different domains.

If we take music recommendation as an example, it is clear that taking a single criterion into account most likely will not suffice to meet our goals, whether they be measured in terms of commercial success or user satisfaction. Problems like this have to be viewed from different angles and utilize several approaches to find a solution. The goal of finding the right music for the customer should consider not only the musical piece, but also the user and his needs. Additionally, both the music and the user have to be considered in context. The context of the music can be extracted from its acoustic features (genre, style, tempo, emotion, etc.) and meta-information (band, language, historic, social, etc.), while the user will have its own internal context (social, psychological, emotional), but also external context (environmental, situational). All of these can affect the quality of the system to a certain degree.

Even though genre recognition is not the best feature when it comes to MIR problems, it is still very popular among researchers. This comes as a consequence of its simplicity, both as a computational problem and a topic that is easy to understand by someone without a deep technical or music background. Everyone has heard of music genres, and it is very simple to construct the task as a classification problem with various types of inputs and a set of discrete classes as the output. In reality, genre recognition is a quite difficult and poorly defined problem. Not only is it difficult to assign a single class to any random musical piece, but even the classification taxonomy can not be defined without dissension. This has not stopped people from trying, and several standard benchmarks have been created to tackle this particular problem.

One of the most popular is the GTZAN [22] database, which is available for free and very easy to use. It consists of 1000 tracks (each 30 seconds in length) and organized into 10 classes (each consisting of 100 tracks). Initial experiments relied on simple musical descriptors (rhythm, pitch, timbre) as well as classic music analysis features like the Mel-Frequency Cepstral Coefficients (MFCC) and, less frequently, the Wavelet transform [8]. In [22], Tzanetakis utilized Gaussian Mixture Models on MFCCs, to achieve 61% baseline accuracy in the first ever GTZAN experiment. The authors in [19] reported 83% accuracy using Deep Neural Networks and spectral features. A breakthrough in feature quality was presented in the paper about the Scattering Wavelet Features (SWT) [1], where a simple SVM classifier achieved 89.3% accuracy. Later, the same features were utilized in a better Sparse Representation

Classifier [6], improving the result slightly with a reported accuracy of 91.2%. Other experiments on GTZAN utilized Wavelets [14] to achieve 78.5% accuracy, Deep-Belief Networks [9] for 84.3% accuracy, various representation based on the properties of the auditory cortex [17] for 92.4% accuracy and Compressive Sampling techniques [5] reporting 92.7% accuracy.

It is worth noting that the margin of error between these results is quite wide, and the difference at the high end becomes quite negligible due to the fairly small size of the corpus, compounded by the numerous reported inconsistencies within the database [21]. Ultimately, as mentioned in the previous paragraph, the genre taxonomy cannot be too objective, and individual sample track classification can often be fuzzy. As an example, many of the experiments mentioned above used voting to determine the final class; but, if the distribution of classes for individual frames gives, for example, 49% to one class and 51% to another, it may be difficult to say that either class is more relevant.

The goal of this paper is to combine the SWT described in [1] with the power of a Deep Neural Network (DNN) consisting of multiple layers of Sparse Autoencoders (SAE). To improve the Unsupervised Pre-training phase, a much larger database (acquired from the jamendo.com website) was prepared to match the GTZAN database. Jamendo is a music-sharing platform which publishes music on a Creative Commons license. A publicly-available API allowed the authors to download more than 80,000 musical tracks, nearly 10,000 of which were selected according to the GTZAN genres.

2. Background

This section includes background information of various components used in the experiments described in this paper.

2.1. Scattering Wavelet Transform

Most of the research behind MIR relies on Mel-Frequency Cepstral Coefficients (MFCCs), which are a Fourier-based feature set designed specifically for analyzing speech and music. MFCCs are calculated as the Fourier transform of the logarithm of the Fourier transform of the signal that was partitioned using standard windowing techniques (like in the STFT). The resulting features can be used to estimate a smoothed spectral envelope that is robust to small intra-class changes, but loses information [15].

Unlike the Fourier transform (which decomposes the signal into sinusoidal waves of infinite length), the Wavelet Transform (WT) encodes the exact location of the individual components. The Fourier transform encodes the same information as the phase component, but this is usually discarded in the standard MFCC feature set. This means that Fourier-based methods are very good at modeling harmonic signals, but are very weak at modeling sudden changes or short-term instabilities of the signal – something that the WT seems to deal with very well. The WT begins by defining

a family of dilated signals known as wavelets. A single mother wavelet $\psi(t)$ is expanded to a dictionary of wavelets $\psi_{u,s}$, translated to u and scaled by s, using the formula:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \tag{1}$$

These wavelets are then used to decompose the input signal by using a convolution operator (denoted by $< \cdot >$):

$$Wf(u,s) =< f, \psi_{u,s} >= \int f(t)\frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)dt \tag{2}$$

The Scattering Wavelet Transform (SWT) [15] works by computing a series of Wavelet decompositions iteratively (the output of one decomposition is decomposed again), producing a transformation which is both transformation invariant (like the MFCC) and experiences no information loss (proven by producing an inverse transform – something which cannot be done using MFCC without loss).

In [1], the SWT is used in the problem of phoneme classification and musical genre recognition. The paper also points out a similarity between the multilayer structure of the SWT and other deep structures, such as the Convolutional Neural Network [12]. This would hint at a certain level of redundancy of using DNNs with SWT features, but [6] demonstrates that certain improvements can still be achieved using better classifiers, and this paper intends to explore this.

2.2. Unsupervised feature learning

Training an Artificial Neural Network (ANN) with multiple layers (i.e., more than 2 or 3 hidden layers) using backpropagation does not fully utilize its theoretical capabilities. This is caused by the weakness of the gradient descent optimization method, where gradients that are computed by backpropagation rapidly diminish in magnitude as the depth of the network increases. As a result, the final layers don't receive meaningful training data [7]. This problem was well known and has been studied for decades. It was especially troubling that a Multi-Layer Perceptron often performed worse than its shallow counterparts (e.g., SVM) even though its expressiveness was theoretically more powerful.

A breakthrough happened in 2006 when G. E. Hinton introduced a fast-learning algorithm for training, which he named Deep Belief Networks [10]. This method uses a greedy layer-wise training to train one layer at a time in an unsupervised manner. This step is called pre-training, and its aim is to prepare the weights of the model in such a way that they better represent local feature states. Following this, the final fine-tuning of the weights using labeled data creates a model which performs far better than one that is trained on randomly-initialized weights alone.

This unsupervised pre-training approach started a new research trend called "deep learning." Deep learning takes advantage of unlabeled data to learn a good representation of the features space [2] – each layer representing another abstraction

of the features pre-trained from a previous layer. Layer-wise, bottom-up pre-training (one layer at a time) is possible by incorporating Restrictive Boltzman Machines (RBM) or Autoencoders (AE) [3]. Stacking RBMs or AEs (as features detectors) forms a "deep structure" which can be fine-tuned using gradient-based optimization methods with respect to labeled data (i.e., supervised training).

2.3. Sparse Autoencoders

An Autoencoder (AE) is an ANN with an odd number of hidden layers, where the number of units in the output layer is set to be equal to the number of units in the input. In other words, AEs try to reconstruct the input at the output passing data through hidden layers. To ensure that the mapping is non-trivial, various constraints can be used to force the network to learn useful representations of the data. When the number of units in the hidden layer is smaller than the input, the AE learns a compressed form of the data, similar to the Principle Component Analysis (PCA). Unlike PCA, however, the learned compression is non-linear and more robust. If we use more hidden than input units, the AE can still learn meaningful representations of the data [3], provided it uses proper constraints.

One of the constraints that can be applied to AE training is trying to reconstruct the input from its corrupted version. This is the basic idea behind Denoising Autoencoders [23]. Another type of AE (used in this paper) is the Sparse Autoencoder (SA). [18, 13]. The idea behind it is to enforce activations of hidden units to be close to zero for most of the time during training. This can be achieved by applying the measure of Kullback-Liebler Divergence (KL) to the cost function:

$$KL = \rho \log \frac{\rho}{\hat{\rho}} + (1-\rho) \log \left(\frac{1-\rho}{1-\hat{\rho}} \right) \tag{3}$$

$$J_{sparse}(W,b) = J(W,b) + \beta \cdot KL(\rho || \hat{\rho}) \tag{4}$$

KL measures the difference between the two distributions: $\hat{\rho}$, which represents the average activations of hidden units over the training set, and ρ, which represents the target distribution. $J_{sparse}(W,b)$ denotes the sparse cost function with respect to weights W and biases b. Because we want to keep hidden units inactive most of the time, the target distribution should be set close to zero. In our experiments (described below), the target distribution ρ was always set to 0.1. In other words, we wanted to enforce $\hat{\rho} = \rho$. In order to penalize an average activation of hidden units which deviates too much from its target value of ρ, a special penalty term β is introduced to control the weight of the sparsity term.

2.4. DNN implementation

A neural network with mini-batch stochastic gradient descent (SGD) was developed in Matlab. The core of the code was written according to the guidelines presented in CS294A Lecture notes [16]. Additionally, the part of the code responsible for gradient calculation is compatible with the minFunc function that uses the L-BFGS [20]

optimization algorithm. This algorithm uses a limited amount of computer memory, and was used in this paper for training the Autoencoders to improve training speed. The code, besides having an implemented square-error cost function, was extended to operate on cross entropy error [4] and to use momentum. The regularization term of $\|W\|^2$ was added to the cost error function, for the purpose of decreasing the magnitude of the weights and help to prevent overfitting. As a weight initialization for training AEs and NNs (in the case of experimenting without pre-training phase), we used a random uniform distribution U from the range described by formula 5, as it is recommended in [7]. The $n_{visible}$ and n_{hidden} denote the number of visible and hidden units in a given layer.

$$W_{init} = U\left[-\sqrt{\frac{6}{n_{visible}+n_{hidden}}}, \sqrt{\frac{6}{n_{visible}+n_{hidden}}}\right] \tag{5}$$

3. Data preparation

Two databases were used in the experiments. First is the well-known GTZAN dataset [22], consisting of 1000 musical files that are each 30 seconds long. They are categorized into 10 genres with 100 musical pieces per category (rock, blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae). The second data collection was obtained from the jamendo.com website, which offers music ready to download for free due to the Creative Commons license. A publicly-available API allowed us to download over 80,000 musical tracks, together with meta-data in an XML format. The meta-data contains, among other features, a genre association of each file. There are three attributes containing this information: "album genre", "track genre", and "tags". The "album genre" and "track genre" contain ID3 genre names, and "tags" can contain genres and other information (without restrictions) as annotated by users.

The goal was to create a much bigger database than GTZAN yet organized in the same manner. From the 80,000 files, only those that belonged to one of the 10 musical genres were taken into consideration. To avoid ambiguities, all of the files were passed through a couple of filters. Initially, files that had the same values in all attributes were immediately accepted. This assumption gave the highest probability that a particular file belonged to the given genre. For the genres that thusly resulted in less than 1000 musical files (this occurred with blues, country and reggae which are more specific than pop or rock), the filter was made less restrictive. First, only "track genre" and "album genre" had to be equal to choose a song (ignoring the tags); if there were still too few songs, only "track genre" was considered, ignoring the rest of the attributes. This generated a list of 9966 musical files organized into 10 musical genres with nearly 1000 track per genre.

Out of each file, a 30-second fragment starting at 30 seconds from the beginning of the file (to skip the potential problems which occur in the beginnings of some tracks) was extracted and down-sampled to 22,050 Hz (to match the GTZAN format).

The features were extracted from the files using the ScatNet toolbox. The SWT transform was computed to the depth of 2, as this was shown as the optimal setting

in [1]. The first layer contained 8 wavelets per octave of the Gabor kind, and the second had 2 wavelets per octave of the Morlet type. The window length was set to 740 ms. After the transformation, we obtained 81,052 training examples from GTZAN and 802,925 training examples from the JAMENDO database – each with 747 features. The resulting databases can be acquired by contacting the authors.

4. Experiments

One of the goals of the experiments was to determine if the JAMENDO database could be used as an additional source of data for pre-training the DNN. We assumed that this data was completely independent from that in GTZAN, which was used for fine-tuning. In the first step, each SAE was trained using the songs from JAMENDO. The SAEs were trained with the L-BFGS optimizer for 300 epochs. When the training of the first SAE finished, the new data representation was derived by feeding the original data through the SAE's hidden layer. This representation was used for pre-training the second SAE and so on, until the whole DNN was pre-trained. This process of NN pre-training is illustrated in Figure 1.

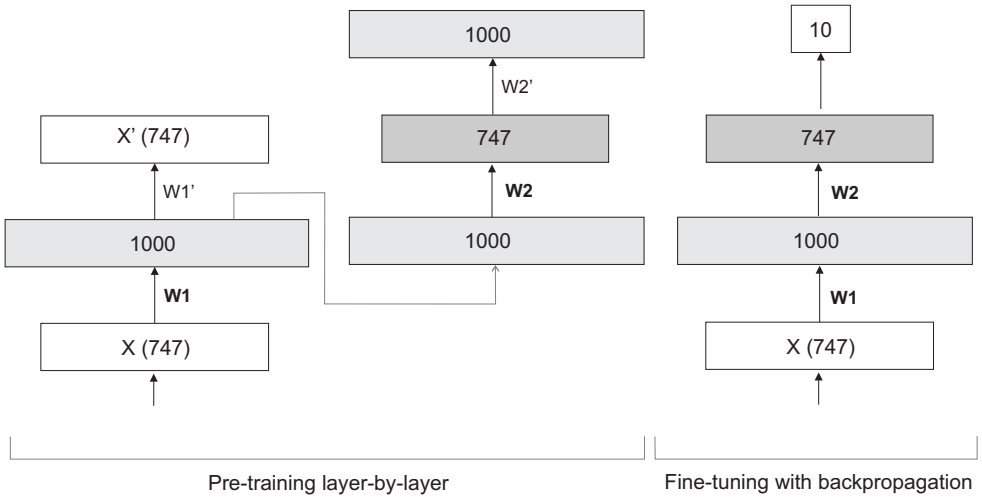


Figure 1. The process of pre-training the two hidden layers is illustrated. Two SAEs are trained. The weights from the encoder parts (W1 and W2) are used to initialize the final NN. Finally, the whole structure is fine-tuned using backpropagation, with the cross-entropy cost function.

To estimate the strength of the sparsity constraint β for the SAE, logistic regression was trained on the SAE representation derived from the GTZAN. The highest accuracy in this test determined the parameter β for the final SAE training. In each case, the target distribution of hidden activation ρ was set to 0.1.

Our experiments were based on pre-training and fine-tuning different topologies of neural networks. The GTZAN songs were randomly shuffled and divided into 10 folds for cross-validation (CV) tests. During CV, one fold was always reserved for validation and didn't take part in training. Its error rate was monitored during

training to determine the early stopping criterion. The training was terminated when the cost value on the validation set didn't decrease by more than $1e-4$.

Before training, the data was standardized to achieve zero mean and a standard deviation of 1. The mean and standard deviation were calculated once in each fold and then used to standardize a test and validation set. Maximum voting was used to predict the label (genre) of the whole track in the test set. Classification error rates were averaged over all 10 folds. The final DNN had a topology consisting of 1,000, 747, 625, and 1,000 units in individual hidden layers respectively (see Table 1). The input vector had 747 dimensions. A log-sigmoid transfer functions were used in the DNNs and SAEs.

Table 1

Results after performing 10 fold CV on different topologies of DNNs pre-trained using SAEs. The results were determined by early-stopping. The experiment with the asterisk used momentum and a larger batch size.

Topology	Error %
747/1000/747/625/1000/10	11.1
747/1000/747/625/1000/10*	10.8
747/1000/747/625/10	10.9
747/1000/747/10	9.8
747/1000/10	12.1

Some additional experiments were also performed. A single fold of data was trained through 200 epochs. We plotted the changes of cost values for different topologies of NNs. Figure 2 presents the NNs with no pre-training whilst Figure 3 shows NNs with pre-training using SAEs. The experiments were performed with the following settings in both cases: learning rate: 3e−2; batch size: 80, momentum: 0.5, regularization: 1e−4.

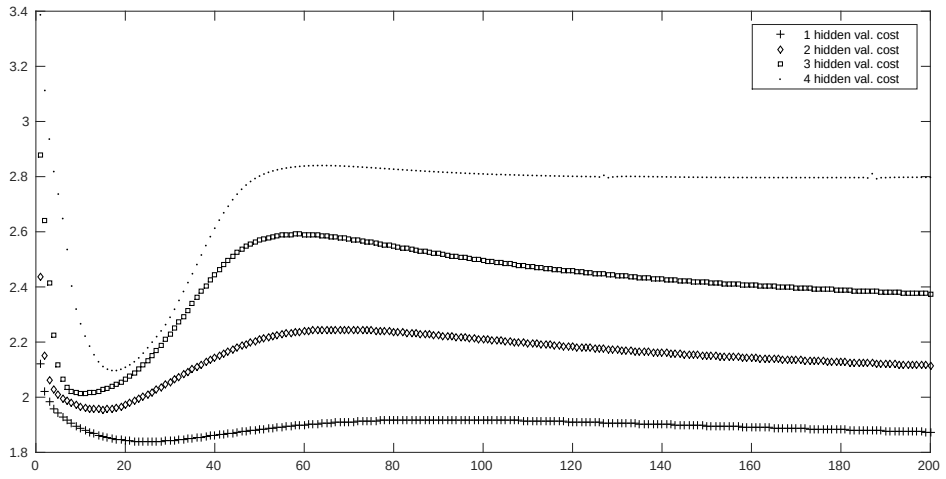


Figure 2. Validation set cost value of the network without pre-training on one fold of data.

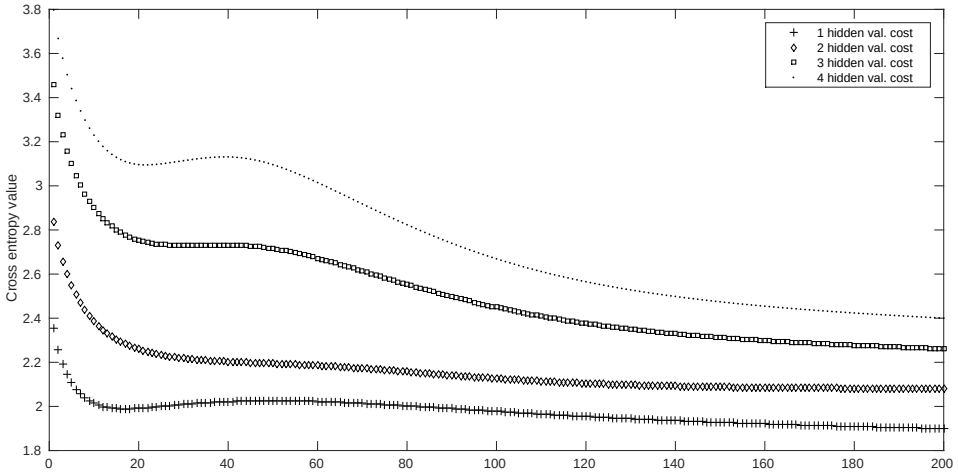


Figure 3. Validation set cost value of the network with pre-training on one fold of data.

Experiments were performed on a computer with 8 CPU threads (Intel i7 3820 3.6 Ghz) and also on an NVidia GeForce GTX TITAN Black GPU. The experiments used about 3 GBs of memory. Moving the calculation on the GPU improved the calculation slightly, but further code optimization is required for more significant improvements.

5. Conclusions and discussion

In our previous paper, we showed that adding more layers to the MLP does not improve the accuracy of genre recognition, and may even diminish it if the number of parameters becomes too high [11]. Using a hidden layer that was pre-trained with an SAE did improve the accuracy, however.

The purpose of the experiments in this paper was to examine whether pre-training using SAE improved the genre recognition in more than one hidden layer. This is the basic principle for building a DNN that has been proven to work for many tasks, including genre recognition [9]. The difference in our work is the utilization of SWT, which already outperforms many of the other approaches, including the DBN mentioned earlier.

The fine-tuning by using gradient descent didn't always improve the final network error rate. The best result was obtained with two hidden layers (9.8% error rate), but we weren't able to reproduce this improvement for other topologies with higher number of layers.

The graphs in Figures 2 and 3 demonstrate this problem very well. The network that didn't use pre-training achieved its minimum cost very early in the training (around 20–30 epochs) and didn't improved that score after that. It seems to over-fit quickly and converges to a worse value than achieved in 20–30 epochs. The shape of the cost for the network with pre-trained weights, however, has a much different

shape. Not only is the over-fitting much less pronounced, but the network seems to generally improve much better than its randomly-initialized counterpart.

One of the reasons for our results may be the early stopping strategy that we employed in our experiments. The problem with having such a small corpus is that small differences in training can cause large jumps in the test error rate, and the error rate is poorly correlated with the network loss. Some initial experiments showed that training the network for much longer than the early stopping suggested could improve the error rate significantly, but we are not sure about the objectivity of such a result.

Nevertheless, even if the end result could be improved in individual layers, it seems that when comparing the results between layers, adding more layers to the DNN simply doesn't improve either the network cost or the error rate in any of the training, validation, or test sets. It is not clear whether this is the consequence of using the SWT features or an issue with the training methodology.

More tests are planned for this problem, especially with respect to the early stopping issue mentioned above. Different methods of pre-training also need to be tested; for example, de-noising AE and RBMs. Finally, attempts at studying the feature-space in a spatio-temporal manner could enable completely different approaches to this problem. The current system models the problem in the feature-space of individual frames describing the spectral content of the sound at a certain point in time, completely disregarding the temporal aspects of the signal (i.e the change of frequencies in time). It is likely that analysing several samples at once will allow the system to recognize certain temporal patterns in the signal. Furthermore, such methods as Convolutional Neural Networks have shown very promising in analysing 2-D signals and would be worth investigating here as well.

Acknowledgements

We would like to thank prof. Krzysztof Marasek, Thomas Kemp, and Christian Scheible for their support. This work was funded by a grant agreement no. ST/MN/MUL/2013 at the Polish-Japanese Academy of Information Technology.

References

[1] Andén J., Mallat S.: *Deep Scattering Spectrum. CoRR*, vol. abs/1304.6763, 2013, `http://arxiv.org/abs/1304.6763`.

[2] Bengio Y.: Learning Deep Architectures for AI. *Foundations Trends Machine Learning*, vol. 2(1), pp. 1–127, `http://dx.doi.org/10.1561/2200000006`.

[3] Bengio Y., Lamblin P., Popovici D., Larochelle H., et al.: Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, vol. 19, p. 153, 2007.

[4] Bishop C. M.: *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.

[5] Chang K. K., Jang J. S. R., Iliopoulos C. S.: Music Genre Classification via Compressive Sampling. In: *ISMIR*, pp. 387–392, 2010.

[6] Chen X., Ramadge P. J.: Music genre classification using multiscale scattering and sparse representations. In: *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pp. 1–6, IEEE, 2013.

[7] Glorot X., Bengio Y.: Understanding the difficulty of training deep feedforward neural networks. In: *International conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[8] Grimaldi M., Cunningham P., Kokaram A.: A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. In: *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 102–108, ACM, 2003.

[9] Hamel P., Eck D.: Learning Features from Music Audio with Deep Belief Networks. In: *ISMIR*, pp. 339–344, Utrecht, The Netherlands, 2010.

[10] Hinton G., Osindero S., Teh Y. W.: A fast learning algorithm for deep belief nets. *Neural Computation*, vol. 18(7), pp. 1527–1554, 2006.

[11] Kleć M., Koržinek D.: Unsupervised Feature Pre-training of the Scattering Wavelet Transform for Musical Genre Recognition. *Procedia Technology*, vol. 18, pp. 133–139, 2014.

[12] LeCun Y., Bengio Y.: The Handbook of Brain Theory and Neural Networks. chap. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258, MIT Press, Cambridge, MA, USA, 1998, `http://dl.acm.org/citation.cfm?id=303568.303704`.

[13] Lee H., Ekanadham C., Ng A.Y.: Sparse deep belief net model for visual area V2. In: *Advances in neural information processing systems*, pp. 873–880, MIT Press, 2008.

[14] Li T., Ogihara M., Li Q.: A comparative study on content-based music genre classification. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 282–289, ACM, 2003.

[15] Mallat S.: Group invariant scattering. *Communications on Pure and Applied Mathematics*, vol. 65(10), pp. 1331–1398, 2012.

[16] Ng A.: Sparse autoencoder. *CS294A Lecture Notes*, vol. 72, pp. 1–19, 2011.

[17] Panagakis Y., Kotropoulos C., Arce G. R.: Music Genre Classification Using Locality Preserving Non-Negative Tensor Factorization and Sparse Representations. In: *ISMIR*, pp. 249–254, 2009.

[18] Poultney C., Chopra S., Cun Y.L., et al.: Efficient learning of sparse representations with an energy-based model. In: *Advances in neural information processing systems*, pp. 1137–1144, 2006.

[19] Sigtia S., Dixon S.: Improved music feature learning with deep neural networks. In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6959–6963, IEEE, 2014.

[20] Skajaa A.: Limited memory BFGS for nonsmooth optimization. Master's thesis, Courant Institute of Mathematical Science, New York University, 2010.

[21] Sturm B.L.: The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

[22] Tzanetakis G., Cook P.: Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, vol. 10(5), pp. 293–302, 2002.

[23] Vincent P., Larochelle H., Lajoie I., Bengio Y., Manzagol P. A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

Affiliations

Mariusz Kleć
Polish-Japanese Academy of Information Technology, Warsaw, Poland, mklec@pjwstk.edu.pl

Danijel Koržinek
Polish-Japanese Academy of Information Technology, Warsaw, Poland,
danijel@pjwstk.edu.pl

Received: 19.01.2015
Revised: 09.03.2015
Accepted: 17.03.2015

6.3 Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation

Dane bibliograficzne pracy:

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, Cham. https://doi.org/10.1007/978-3-030-01851-1_3

Early Detection of Heart Symptoms with Convolutional Neural Network and Scattering Wavelet Transformation

Mariusz Kleć

Polish-Japanese Academy of Information Technology,
Multimedia Department, Warsaw, Poland.
mklec@pjwstk.edu.pl

Abstract. The paper utilizes Convolutional Neural Network (CNN) for preliminary screening of cardiac pathologies by classifying the signal of heartbeat, recorded by digital stethoscope and mobile devices. The Scattering Wavelet Transformation (SWT) was used for the heartbeat representation. The experiments revealed the optimum concatenation size of SWT windows to obtain the state-of-the-art in the majority of metrics, coming from the PASCAL Classifying Heart Sounds Challenge.

Keywords: Heartbeat classification, Convolutional Neural Network, Scattering Wavelet Transformation

1 Introduction

The World Health Organization (WHO) states that mortality from heart disease is a plague of the 21st century. Annually, 17.5 million people die due to cardiovascular diseases.[1] In Poland, the Central Statistical Office (GUS)[2] reports that cardiovascular diseases are also the most common cause of mortality which states to be 46% of all deaths in Poland. Therefore, early diagnosis becomes a huge challenge for the medical community in the field of implementation of preventive care, also in Third World countries where the access to medical care and medical devices is limited.

Every effort to delay or prevent the morbid events is worth considering. Correct assessment of the heart function, recorded by electrocardiogram (ECG), attracts more attention from the community of computer scientists [1, 14, 15, 17]. Convolutional Neural Networks (CNN) have found their application for ECG data classification. The Arrhythmia Detection (AD) is the most popular field of research among the cardiology fields. The comprehensive survey of ECG-based heartbeat classification for AD is carried out in [12]. They report that a lot of

[1] http://new.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/statystyka-zgonow-i-umieralnosci-z-powodu-chorob-ukladu-krazenia,22,1.html

research rely on publicly available MIT-BIH[3] database (which is recommended by ANSI/AAMI [7] for validation of medical equipment) and AHA.[4] However, the main disadvantage of using popular benchmarks is that they do not represent very big volume of data, what in the domain of deep learning could be utilized efficiently and could presumably boost performance results. This is exactly what Andrew Ng's scientific group did. They obtained state-of-the-art in both recall and precision in cardiology performance using CNN with 34-layers and the dataset with more than 500 times of data than the previously studied corpora [16]. In another paper [14], the authors used CNN and a large volume of raw ECG time-series data to obtain the feature representation for identifying patients with paroxysmal atrial fibrillation (PAF) (life threatening cardiac arrhythmia). They experimentally verified that the learned representation can effectively replace the user's hand-crafted features, as CNN learned the key ones, unique to the PAF. They have also conducted the comparison with several conventional machine learning classifiers and indicated that combining the learned features with other classifiers significantly improves the performance of the patient screening systems. Their findings were verified by many researchers who claim that the problem of ECG classification heavily depends on the appropriate features that represent the data [16, 15, 17].

The Wavelet Transform (WT) is very often used by researchers for representing ECG signals [8, 11, 10]. The WT allows information extraction from both frequency and time domains, different from what is usually achieved by the traditional Fourier transform, which permits the analysis of only the frequency domain [6]. Within the types of WT, the Discrete Wavelet Transform (DWT) is the most popular [10]. Apart from the DWT, Continuous Wavelet Transform (CWT) has also been used to extract features from the ECG signals [2], since it overcomes some of the DWT drawbacks, such as the coarseness of the representation and instability.

This paper extends these methods by using Scattering Wavelet Transform (SWT) [3] as it was not used so far for the problem of heartbeat classification, especially with the data coming from the PASCAL Classifying Heart Sounds Challenge[4]. In [20], the authors classified this data directly from its Fourier transformation with CNN, omitting the segmentation phase to detect the fundamental physical characteristics of the heartbeat (WSCNN). Similar approach is presented in [5] where the authors described the framework based on the autocorrelation feature and diffusion maps, that are further provided to the SVM classifier (SVM-DM). Another paper describes a scaled spectrogram and partial least squares regression (SS-PLSR) for classifing the hearbeat signal [21]. In [22], they used tensor decomposed features (SS-TD) for the same purpose. All the results from these papers are aggregated in Table 2. The method described in this paper is abbreviated to CSWT for convenience.

The paper is organized as follows. Section 1 presents the recent works and introduction to the field. In Section 2 the data and pre-procesing are described.

[3] http://ecg.mit.edu/

[4] http://www.ahadata.com/

The Section 3 contains the description of performed experiments. Finally, the results and conclusions are presented in Sections 4 and 5 respectively.

2 Data Description and Pre-processing

Two datasets were provided for the PASCAL Classifying Heart Sounds Challenge [4]. Dataset A comprises data recorded by iPhone app called iStethoscope. Dataset B comprises data collected from a clinical trial in hospitals using the digital stethoscope DigiScope. The data was gathered in real-world situations and frequently contained the background noise such as speech, traffic, brushing the microphone against cloth or skin etc. The audio files were of different lengths, between 1 and 30 seconds. Datasets were divided into training and testing sets with 4 categories for Dataset A (Normal, Murmur, Extra Heart Sound and Artifact) and 3 categories for Dataset B (Normal, Murmur and Extrasystole). Table 1 contains statistics of both datasets used in the experiments. The meaning of the categories is described in [19] in more detail.

The audio files were re-sampled to the equal value of 22050Hz in both Datasets. The high-pass filter was applied below 200Hz as the heartbeat information exists in the low frequencies.

Table 1. The number of files in Dataset A and B with respect to categories

Dataset A, 44100Hz, 16bit			Dataset B, 4000Hz, 16bit		
Category	train	test	Category	train	test
Normal	31	14	Normal	200	136
Murmur	34	14	Murmur	66	39
Extrasound	19	8	Extrasystole	46	20
Artifact	40	16	-	-	-

In this paper, the SWT was computed to a depth of 2 with 8 wavelets per octave of the Gabor kind in the first layer and 2 wavelets per octave of the Morlet type in the second layer. The author decided to use the same settings as used for Automatic Genre Recognition [3], assuming that music and heartbeat share the similar subset of structures (like sudden changes and short-term instabilities) which can be well captured by SWT [3]. The longest SWT window length T was 0.74s. The two additional window lengths were half size of the previous ones: $T = 0.37s$ and $T = 0.185s$. Eventually, the data was standardized to have zero mean and unit variance.

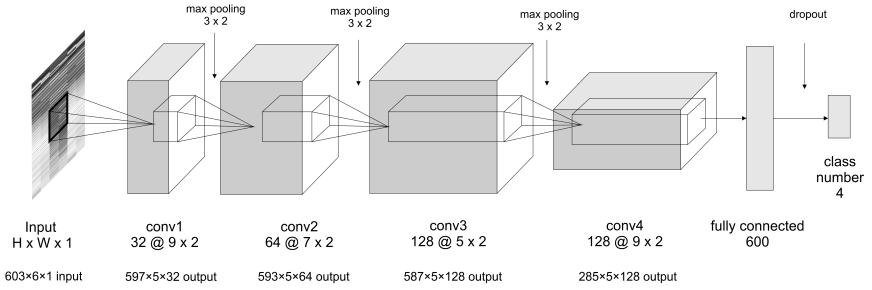
3 Experiments

The experiments were on training CNNs to classify the real heartbeat recordings described in Section 2. The use of the three SWT window lengths was evaluated:

$T = 0.74s$, $T = 0.37s$ and $T = 0.185s$. Training and testing examples were constructed by the concatenation of several SWT windows into frames. The overall frame length was not longer than the shortest testing file (1.75s for Dataset A and 1s for Dataset B). The frames half overlapped one another within the overall length of audio file.

The CNN contained 4 convolutional layers with max-pooling layers in between. The *stride* value was set to $1x1$, apart from the last max-pooling layer, where the *stride* was set to $2x1$ in order to reduce the output dimensions from the previous layer. The detailed architecture of CNN is presented in Figure 1.

Fig. 1. The picture presents the architecture of CNN used for all experiments. The input size differed depending on the size of the input frames. The numbers below the name of layers indicate: the number of filters @ height x width of the filters. Furthermore, the instance of input frame is presented (603x6x1) together with the size of output from each subsequent convolutional layer. The instance of input frame represents the concatenation of six SWT windows with 0.185s length each.



The validation accuracy has been monitored in subsequent epoch during training. The training was terminated when the validation accuracy has not been larger or equal to the previously highest accuracy for 10 epochs. The evaluation metrics were derived from the state of the network when the accuracy was the highest. However, the author noticed a slight deviation in results when running the same experiment several times. To provide a fair comparison, the final values were averaged after running the same experiment ten times. Additionally, the author observed that the highest validation accuracy in Dataset B always occurred after the first epoch. The model becomes over-fitted every time training is being continued beyond the first epoch.

For the loss function optimization, Adaptive Moment Estimation (ADAM) was chosen [9] with the following options: $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\alpha = 0.0005$. The initial learning rate was reduced by the factor of 0.9 every 5 epoch. The decision of choosing ADAM was dictated by its documented superior performance over Stochastic Gradient Descent [9]. The majority voting was used to derive

the final category of audio file from the sequence of frame labels, outputted from the model.

The evaluation follows the same work-flow as used for the PASCAL Classifying Heart Sounds Challenge [4]. The results are compared to those presented in Table 2. The evaluation takes into account Precision per category, the Youden Index, the F-score (only for the Dataset A) the Discriminant Power (only for the Dataset B) and Normalized Precision. The more detailed definitions of the evaluation criteria are described in [22].

4 Results

The Table 2 highlights the best values for each metric, gathered from the literature. There is no strong conclusion which method overcomes the other ones in all possible aspects. However, the Normalized Precision (NP) can be treated as a good comparative determinant as it aggregates the precision of recognizing all heart categories together, taking into account the imbalanced number of files in each category. The highest value of NP in the Dataset A (0.80) was possible to obtain by representing the data with the concatenation of six SWT frames, 0.185s long each. In turn, in the Dataset B, the best value of NP (0.75) belongs to the concatenation of two SWT frames, 0.37s long each. In both cases the NP exceeds all the other results reported so far (see Table 2).

Table 2. The table presents the results on the Datasets A and B, gathered from the literature: SS-PLSR [21], SVM-DM [5], SS-TD [22], WSCNN [20]. The method proposed in this paper is abbreviated to CSWT. The values in column CSWT also contain the standard deviation from the results being averaged after running the same experiment ten times.

Results on Dataset A					
Evaluation criteria	SS-PLSR	SVM-DM	SS-TD	WSCNN	**CSWT** T=0.185s, 6F
Precision of Normal (PN)	0.60	0.62	**0.67**	0.61	0.57 $\pm$0.03
Precision of Murmur (PM)	0.91	0.91	**1.00**	0.91	**1.00** $\pm$0.00
Precision of Extra Heart Sound (PE)	0.44	**1.00**	0.43	0.50	0.57 $\pm$0.19
Precision of Artifact (PA)	**0.94**	0.64	0.80	**0.94**	**0.94** $\pm$0.05
Artifact Sensitivity (ASe)	**1.00**	**1.00**	**1.00**	**1.00**	0.80 $\pm$0.20
Artifact Specificity (ASp)	0.64	0.58	0.64	**0.67**	**0.66** $\pm$0.03
Youden Index of Artifact (YIx)	0.64	0.58	0.64	**0.67**	**0.66** $\pm$0.03
F-score (FS)	0.30	0.31	0.30	**0.67**	0.30 $\pm$0.02
Total Precision (TP)	2.89	**3.17**	2.90	2.96	3.07 $\pm$0.19
Normalized Precision (NP)	0.76	0.76	0.76	0.77	**0.80** $\pm$0.03
Results on Dataset B					
Evaluation criteria	SS-PLSR	SVM-DM	SS-TD	WSCNN	**CSWT** T=0.37s, 2F
Precision of Normal (PN)	0.76	0.77	**0.83**	0.81	0.78 $\pm$0.01
Precision of Murmur (PM)	0.65	0.76	0.70	0.67	**0.96** $\pm$0.05
Precision of Extrasystole (PE)	0.33	**0.50**	0.15	0.14	0.15 $\pm$0.33
Heart Problem Sensitivity (HPSe)	0.34	0.34	**0.49**	0.51	0.34 $\pm$0.04
Heart Problem Specificity (HPSp)	0.90	0.95	0.84	0.80	**0.99** $\pm$0.01
Youden Index of heart problem (YIxp)	0.24	0.29	**0.33**	0.31	**0.33** $\pm$0.03
Discriminant Power (DP)	0.36	0.54	0.39	0.34	**1.39** $\pm$0.91
Total Precision	1.75	**2.03**	1.68	1.62	1.89 $\pm$0.35
Normalized Precision (NP)	0.69	0.74	0.74	0.71	**0.75** $\pm$0.04

Analyzing each metric separately, we can conclude that the proposed method (CSWT) performs very well in recognizing Murmur. The method returns the high values of PM in both Datasets (1.00 on the Dataset A and 0.96 on the Dataset B), resulting the state-of-the-art in this metric, especially on the Dataset B (see Table 2). The DP evaluates how well the algorithm distinguishes between normal and problematic heartbeats (Murmur and Extrasystole categories combined). The high value of PM boosted the DP to the value of 1.39 and the HPSp to the value of 0.99, resulting the state-of-the-art also in these metrics on Database B (see Table 2).

Table 3. The first table presents the evaluation results on the Dataset A: Precision of Normal (PN), Precision of Murmur (PM), Precision of Extrasound (PE), Precision of Artifact (PA), Artifact Sensitivity (ASe), Artifact Specificity (ASp), The Youden Index of Artifact (YIx), F-Score of Heartproblem (FS), Total Precision (TP) and Normalized Precision (NP). The second table presents the results on the Dataset B: The Sensitivity of heart problems (HPSe), The Specificity of heart problems (HPSp), The Youden Index of Heartproblem (YIxp), Discriminant Power (DP). The columns refer to the SWT window size T and the number of concatenated frames F.

Results on the Dataset A

	$T = 0.74s$		$T = 0.37s$			$T = 0.185s$				
	$1F$	$2F$	$1F$	$2F$	$4F$	$1F$	$2F$	$4F$	$6F$	$8F$
PN	0.49	0.52	0.57	0.58	0.55	0.53	0.57	0.56	**0.57**	0.54
PM	0.82	0.86	0.98	0.96	0.90	0.82	0.95	0.97	**1.00**	0.90
PE	0.43	0.35	0.47	0.46	0.50	0.43	0.49	**0.64**	0.57	0.54
PA	**0.96**	0.95	0.93	0.90	0.94	0.92	0.92	0.89	0.94	0.94
ASe	0.8	0.8	0.80	**0.96**	0.8	0.75	0.84	0.8	0.8	0.8
ASp	0.57	0.57	0.65	0.64	0.62	0.57	0.64	0.64	**0.66**	0.62
YIx	0.57	0.57	0.65	0.64	0.62	0.56	0.64	0.64	**0.66**	0.62
FS	0.28	0.27	0.30	0.30	0.30	0.27	0.30	**0.31**	0.30	0.29
TP	2.70	2.69	2.95	2.91	2.89	2.7	2.92	3.05	**3.07**	2.92
NP	0.71	0.72	0.77	0.76	0.76	0.71	0.77	0.78	**0.8**	0.76

Results on the Dataset B

	T=0.74s	T=0.37s		T=0.185s		
	1f	1f	2f	1f	2f	4f
PN	**0.79**	0.77	0.78	0.75	0.76	0.76
PM	0.85	0.88	**0.96**	0.90	0.92	0.86
PEs	**0.21**	0.14	0.15	0.07	0	0.12
HPSe	**0.39**	0.33	0.34	0.24	0.27	0.29
HPSp	0.96	0.96	**0.99**	0.98	**0.99**	0.98
YIxp	**0.34**	0.28	0.33	0.22	0.25	0.26
DP	0.66	0.63	**1.39**	0.77	0.84	0.8
TP	1.84	1.80	**1.89**	1.72	1.79	1.68
NP	0.74	0.73	**0.75**	0.71	0.72	0.71

The model's ability to detect Artifacts from the heartbeat signal is important to inform the user to repeat the recording and avoid failure. Interestingly, the state-of-the-art in the PA (0.96) is possible to obtain with the use of only one and wide singular SWT window, 0.74s long (see Table 3). It might be caused by the high ability of SWT to capture more short-term instabilities (that exist in the Artifacts) with a wider window [13]. The high ability of avoiding failures

is highlighted by the high value of YIx on the Databaset A (see Table 3). This finding can guide the future solutions to treat the Artifact recognition differently than other categories.

However, there is still room for improvement of Precision of Normal, Precision of Extra Heart Sound and the Precision of Extrasystole. The highest values still belong to the other methods (see Table 2)

5 Conclusions

In this paper Convolutional Neural Network is utilized for preliminary screening of cardiac pathologies by classifying the heartbeat signals recorded by the digital stethoscope and the mobile phone. The Scattering Wavelet Transformation (SWT) is used to represent the data coming from the PASCAL Classifying Heart Sounds Challenge. The experiments reveal the optimum concatenation size of SWT windows to obtain the state-of-the-art in the Normalized Precision, Precision of Murmur, Precision of Artifact, Heart Problem Specificity and Discriminant Power. However, the Datasets used for these experiments are quite limited in terms of the number of training and testing examples. The PhysioNet/Computing in Cardiology Challenge 2016 (CinC)[5] addresses this problem by assembling the largest public heart sound database, aggregated from eight sources obtained by seven independent research groups around the world. The author plans to run CSWT method on that data and aims to fully validate the findings by running additional experiments and incorporate comparative statistical inference.

As it is pointed out in [18], data captured with off-the-person based devices (like mobile and wearable devices or electronic stethoscopes) can be highly correlated to those captured with traditional on-the-person based equipment (ECG systems). The author believes that the off-the-person approach is worth to further research as it can extend preventive medicine practices by allowing the heartbeat monitoring without interference on daily routine. It could help people to avoid serious problems and hopefully significantly improve the health statistics.

References

1. Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adam, M., Gertych, A., San Tan, R.: A deep convolutional neural network model to classify heartbeats. Computers in biology and medicine 89, 389–396 (2017)
2. Addison, P.S.: Wavelet transforms and the ecg: a review. Physiological measurement 26(5), R155 (2005)
3. Andén, J., Mallat, S.: Deep scattering spectrum. Signal Processing, IEEE Transactions on 62(16), 4114–4128 (2014)

[5] `https://www.physionet.org/challenge/`

4. Bentley, P., Nordehn, G., Coimbra, M., Mannor, S.: The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results. http://www.peterjbentley.com/heartchallenge/index.html
5. Deng, S.W., Han, J.Q.: Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps. Future Generation Computer Systems 60, 13–21 (2016)
6. Dokur, Z., Ölmez, T.: Ecg beat classification by a novel hybrid neural network. Computer methods and programs in biomedicine 66(2-3), 167–181 (2001)
7. EC57, A.A.: Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms. Association for the Advancement of Medical Instrumentation, Arlington, VA (1998)
8. Güler, İ., Übeylı, E.D.: Ecg beat classifier designed by combined neural network model. Pattern recognition 38(2), 199–208 (2005)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Kutlu, Y., Kuntalp, D.: Feature extraction for ecg heartbeats using higher order statistics of wpd coefficients. Computer methods and programs in biomedicine 105(3), 257–267 (2012)
11. Lin, C.H., Du, Y.C., Chen, T.: Adaptive wavelet network for multiple cardiac arrhythmias recognition. Expert Systems with Applications 34(4), 2601–2611 (2008)
12. Luz, E.J.d.S., Schwartz, W.R., Cámara-Chávez, G., Menotti, D.: Ecg-based heartbeat classification for arrhythmia detection: A survey. Computer methods and programs in biomedicine 127, 144–164 (2016)
13. Mallat, S.: Group invariant scattering. Communications on Pure and Applied Mathematics 65(10), 1331–1398 (2012)
14. Pourbabaee, B., Roshtkhari, M.J., Khorasani, K.: Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2017)
15. Pyakillya, B., Kazachenko, N., Mikhailovsky, N.: Deep learning for ecg classification. In: Journal of Physics: Conference Series. vol. 913, p. 012004. IOP Publishing (2017)
16. Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y.: Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836 (2017)
17. Rubin, J., Abreu, R., Ganguli, A., Nelaturi, S., Matei, I., Sricharan, K.: Recognizing abnormal heart sounds using deep learning. arXiv preprint arXiv:1707.04642 (2017)
18. da Silva, H.P., Carreiras, C., Lourenço, A., Fred, A., das Neves, R.C., Ferreira, R.: Off-the-person electrocardiography: performance assessment and clinical correlation. Health and Technology 4(4), 309–318 (2015)
19. Yuenyong, S., Nishihara, A., Kongprawechnon, W., Tungpimolrut, K.: A framework for automatic heart sound analysis without segmentation. Biomedical engineering online 10(1), 13 (2011)
20. Zhang, W., Han, J.: Towards heart sound classification without segmentation using convolutional neural network. Computing 44, 1 (2017)
21. Zhang, W., Han, J., Deng, S.: Heart sound classification based on scaled spectrogram and partial least squares regression. Biomedical Signal Processing and Control 32, 20–28 (2017)
22. Zhang, W., Han, J., Deng, S.: Heart sound classification based on scaled spectrogram and tensor decomposition. Expert Systems with Applications 84, 220–231 (2017)

6.4 Beyond the Big Five Personality Traits for Music Recommendation Systems

Dane bibliograficzne pracy:

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). https://doi.org/10.1186/s13636-022-00269-0

RESEARCH

Beyond the Big Five Personality Traits for Music Recommendation Systems

Mariusz Kleć[1*], Alicja Wieczorkowska[1], Krzysztof Szklanny[1] and Włodzimierz Strus[2]

[*]Correspondence:
mklec@pjwstk.edu.pl
[1]Multimedia Dept.,
Polish-Japanese Academy of Information Technology, Warsaw, Poland
Full list of author information is available at the end of the article

Abstract

The aim of this paper is to investigate the influence of personality traits, characterized by the BFI (Big Five Inventory) and its significant revision called BFI-2, on music recommendation error. The BFI-2 describes the lower-order facets of the Big Five personality traits. We performed experiments with 279 participants, using an application (called Music Master) we developed for music listening and ranking, and for collecting personality profiles of the users. Additionally, 29-dimensional vectors of audio features were extracted to describe the music files. The data obtained from our experiments were used to test several hypotheses about the influence of personality traits and the audio features on music recommendation error. The performed analyses take into account three types of ratings that refer to the cognitive-emotional, motivational, and social components of the attitude towards the song. The experiments showed that every combination of Big-Five personality traits produces worse results than using lower-order personality facets. Additionally, we found a small subset of personality facets that yielded the lowest recommendation error. This finding can condense the personality questionnaire to only the most essential questions. The collected data set is publicly available and ready to be used by other researchers.

Keywords: music recommendation systems; personality traits; the Big Five Inventory-2; collaborative filtering

Introduction

The volume of music data uploaded to the Internet has increased radically. The expanding number of music collections, mobile access to audio files and streaming services pose challenges to finding appropriate songs. Today, thanks to the popularity of streaming services such as Spotify [1], Last.fm [2], Tidal [3], Pandora [4] or Qobuz [5], music discovery and recommendation systems have become much more popular than they were several years ago. Most of these services are hybrid systems (HS) that combine collaborative filtering (CF) and content-based (CB) approaches.

CF analyzes the community's ratings to conclude one's musical preference. The underlying assumption is that if a person A highly rates the same music as person B, then the system is more likely to recommend to user A songs unheard by A from the music pool of user B than that from any randomly chosen user [1, 2]. Although this approach is widely adopted and computationally fast, it has limitations. First,

[1] http://www.spotify.com
[2] https://www.last.fm/
[3] http://www.tidal.com
[4] https://www.pandora.com/
[5] http://www.qobuz.com

CF assumes that musical taste is fixed and does not change over time, which is not always true [3]. Another limitation is the tendency to recommend popular music over those pieces that have few ratings. Individual and unique preferences have no chance of being discovered by this algorithm. Therefore, the most critical obstacle is the Cold-Start (CS) problem [4].

The CS problem occurs when the system has not yet gathered sufficient information about the user or item to infer precise recommendation. One of the strategies for tackling this problem is to resort to the user's contextual data (e.g. social network), in order to enrich rating profiles. The enhanced information about the user can be further used for clustering "similar" users and personalize the recommendation [5–7]. The user personality is a special case of such contextual data. The assumption is that people with similar personalities have similar interests and behavioral patterns [8], so they will also rate the music in a similar way. Personality can be derived implicitly from social networks [9] or explicitly from users [10]. The latter is obtained by asking the user to answer a list of personality questions. However, the personality questionnaires are well established in the psychology field, but not for recommendation systems. Additionally, they may be very long (some of them contain even 240 items [11]). Therefore, in this paper we also want to address this problem and select only the most relevant personality traits for making recommendations. This approach allows reducing the number of personality questions, and presumably increasing the satisfaction from using the system.

The CB approach can also alleviate the CS problem. It focuses on the content of items, which can be the meta-data or audio features. In this case, a single song's rating from the user is enough to calculate the similarity of that song's features to the others and to make the recommendation. However, it leads to the recommendations that are "too similar", without a chance to surprise the user (low serendipity). Hybridizing these two approaches (i.e. CF and CB) can give satisfactory results. The hybrid approach is used today by large companies like Spotify or Pandora. The significant contribution to this field comes from adopting Deep Learning (DL) [12–14], which allows automatic feature extraction from audio signals [15], or learning latent factors from user-item rating data [16, 17].

However, the factors that influence musical taste vary among individuals. Therefore, music information retrieval systems need to go beyond these approaches to deliver better recommendations. The type of music that one wants to listen to depends not only on listening history but also on one's current disposition, activity, as well as health condition, education, gender, and musical training [18–21].

Factors Underpinning Musical Preferences

A positive correlation between a specific situation (context) and the preference for the music exists [18, 21]. It is possible to track the listener's context (e.g. time [22], weather [23], location [24]) and derive the musical taste in that context implicitly [25–27]. In the works [3, 28, 29], the authors utilized the surrounding environment (e.g. noise, time, light, and weather) to suggest music.

Other essential factors that influence musical preferences are emotions [20]. While listening to music, people want to relieve stress, change or match their current emotions with those expressed by the music. Descriptions exist on how to communicate

emotions via musical structure and how our emotions are influenced by listening to music [30]. Tracking the listener's emotions can help to improve the quality of a recommendation [31]. It is usually achieved implicitly by tracking the context, such as keywords from an extensive collection of documents written by users [32], or extracting the users' texts from social networks [33, 34]. Another approach is to derive emotions from the user's face using the inbuilt camera of a mobile phone [31, 35] or from the signals obtained via wearable physiological sensors [36]. Consequently, research on Context-Aware Music Recommendation Systems (CA-MRS) has gained importance in recent years [37].

However, musical preferences depend not only on the way people regulate their emotions, and current situation, but also on their personality [38]. For example, people who are neurotic (i.e., have low emotional stability) are more likely to use music to foster emotions [20]. Conversely, people who are conscientious and low in creativity (low open-mindedness) are more likely to use music for emotional change and emotional regulation [39]. The systems that incorporate the user's personality into the recommendation process are called Personality-Aware Music Recommendation Systems (PA-MRS) and are a branch of the CA-MRS [40].

In 2003, Rentfrow and Gosling [41] empirically found the relationships between personalities and musical preferences. Namely, reflective, complex music (e.g. blues, jazz or folk) and intense and rebellious music (e.g. rock, alternative or heavy metal) are positively related to Openness to experience. On the other hand, upbeat and conventional music (e.g. country or pop) negatively correlates with Openness, but it positively correlates with Extraversion, Agreeableness and Conscientiousness. Finally, energetic and rhythmic music (e.g. hip-hop, dance or electronic) is positively correlated with Extraversion and Agreeableness. Classical music positively correlates with Neuroticism [42]. In 2011, Rentfrow et al. in [43] provided an improved description of musical preferences. Their findings demonstrate a latent five-factor structure underlying music preferences (further called MUSIC factors): Mellow (comprising smooth and relaxing styles), Urban (defined largely by rhythmic and percussive music), Sophisticated (includes classical, operatic, world music, and jazz), Intense (defined by loud, forceful, and energetic music) and Campestral (comprising a variety of various styles of direct and rootsy music, often found in country and singer-songwriter genres).

In [44] Bansal and co-authors confirmed that the music genre relates to the Big Five personality traits. They analyzed a global music-download database consisting of millions of entries with music metadata describing people downloading songs onto Nokia mobile phones. They showed that many genres in people's music collections are positively associated with Openness and (unexpectedly) Agreeableness, suggesting that individuals with high Openness and Agreeableness have broader musical tastes than those with high levels of other personality traits. The outcomes also aligned with literature showing that individuals who prefer jazz and folk score highly in Openness [45]. Such persons also tend to avoid genres like pop [46]. Since the level of Openness is related with the level of IQ [47], the findings above also find confirmation in the work of [48]. The authors indicate that people with higher IQ tend to prefer reflective and complex (e.g. jazz, classical, folk, blues) to upbeat and conventional music (e.g. pop). It is because the complex and reflective music

is more likely to suit those who seek intellectually stimulating experiences. These people use music in rational or intellectual rather than emotional ways, implying higher levels of cognitive processing.

In [42] the authors indicate strong positive correlations between Neuroticism and classical music preference. Interestingly, they did not find Conscientiousness, Extraversion, or Neuroticism to be predictors of genre exclusivity. However, in [49] the authors analyzed a large dataset consisting of music listening histories and personality scores of 1415 Last.fm users. Their results corroborate the prior work, but also show a negative correlation between Conscientiousness and folk music. They also report positive correlations between Extraversion and such genres as R&B or rap, between Agreeableness and country or folk, and also between Neuroticism and alternative music.

However, musical genre is a conventional term and often the border between different musical genres is quite blurry. The authors in [50] investigated how different musical taxonomies (e.g. mood, activity, genre) influence the user experience and satisfaction of using music streaming services. Their findings are correlated with the Big-Five personality traits. They also describe the link between the musical expertise of the listener and the number of categories within the given taxonomy. Their outcomes show that musically sophisticated users (e.g. experts) enjoy using the system more when exposed to a broader set of categories. This is also confirmed in [51], where experts enjoyed the music more when having a more diverse choice of categories.

Still, there is a need to describe the link between personality and music in a more quantitative way. Such an approach is presented in [52]. The authors correlated such audio features as dynamics, mode, register, and tempo with the Big Five. They have also shown that slow tempo is rated higher by listeners high in Conscientiousness, major mode is preferred by persons low in Conscientiousness but high in Extraversion, and piano dynamics are rated higher by listeners high in Openness. In general, audio features are expressed in a quantitative way and can be used together with personality traits in PA-MRS. Interesting approach is described in [53], where the authors are trying to predict the personality trait (Extraversion or Introversion) on the basis of the audio features of the excerpt by employing several classification algorithms.

The authors of [54] showed that the recommendation accuracy could be improved by integrating personality traits. They also demonstrated that the accuracy depends on the recommendation domain: higher accuracy can be achieved in the movie domain than in the music domain. In another paper [55], the authors analyze the influence of personality traits and emotional states (among others) on ratings. They found that the users with a high degree of Agreeableness rate at least 0.5 stars higher compared to the users with low Agreeableness (on a rating scale from 1 to 5) [56]. In [57] the authors compared the contribution of personality features and physiological signals (recorded by a wearable device) to the accuracy of their recommendation system. They found that the physiological features contributed less than the personality features.

It is also worth mentioning that users with different personalities show different preferences, regarding not only the recommendation accuracy, but also such proper-

ties of recommendation as diversity, popularity, and serendipity [58, 59]. The personalization of diversity is described in [60] and used in [61]. The authors demonstrated increased user satisfaction and recommendation diversity when they personalized the system according to the user's personality.

Personality Acquisition

Developing the most efficient acquisition for Music Recommendation Systems (MRS) is a challenge. The review of personality assessment questionnaires can be found in [40]. The most popular one is the Big Five Inventory (BFI) questionnaire, used for Big Five personality acquisition [62, 63]. The Ten Item Personality Inventory (TIPI) is another common option [64]. Generally, the questionnaires vary in the number of questions that the user is to answer. The TIPI is a very short questionnaire containing only 10 items. However, most questionnaires contain more than 50 items (some even 100, 200 and more). Longer questionnaires provide higher reliability, but, at the same time, require more effort from the user. Therefore, researchers try to acquire personality factors implicitly, e.g. using machine learning techniques with features extracted from social media streams [9]. The implicit acquisition does not require any action from the user, but its performance is much worse than explicit methods. For example, in [65] the authors were able to predict personality parameters from Twitter within 11%-18% of their actual value, by looking at the content of the user's tweets. Thus, the obtained results were very low, which was also confirmed in [66].

Contribution

We hypothesize that selecting only the most relevant personality traits for doing recommendations allows for reducing the recommendation error and limiting the number of questions the user needs to answer. To verify this hypothesis, we aimed at selecting the most relevant personality traits. In our study, we decided to use an explicit method for personality acquisition. Since using long questionnaires may be fatiguing for users, we wanted to find a trade-off between the reliability of the user personality representation and the length of the questionnaire. We used the revised version of the BFI (i.e. BFI-2) [67], as it contains 60 items (questions) and allows to go beyond the Big Five personality traits, by also measuring the lower-order level (i.e., facets) of the Big Five. We developed an application (called Music Master) for gathering users' personality information, listening to music, and rating it. Based on the data collected from the listening sessions, a memory-based hybrid music recommendation system has been developed and evaluated in an offline manner. The memory-based approach allows us to measure (among others) the similarities between users, in terms of various subsets of their personality traits, and clearly interpret the recommendation process. The system takes into account the similarities between users; the similarities are measured using various subsets of personality traits. Based on the results, we selected only those traits (and their corresponding questions from the BFI-2 questionnaire) that contributed most to the system's performance. To the best of our knowledge, the BFI-2 has not been used before in any recommendation system. Additionally, we have published the collected data with ratings, features, and personality traits, to make them available for further investigations by other researchers.

Personality

Personality describes how individuals differ in their permanent emotional, interpersonal, experiential, attitudinal and motivational styles [68]. Over the past quarter-century, personality psychology has been dominated by theories of traits. There are several established and at the same time competing models of personality trait structure, such as the so-called Giant Three model by Eysenck [69], six-factor HEXACO model [70], or Two-Factor Model of higher-order personality factors [71, 72]. However, the Five-Factor Model, which is also known as the Big Five [11, 62, 73], is the prevailing conceptualisation of personality structure and its basic dimensions. According to the Big Five model, most of the significant individual differences in people's patterns of thinking, feeling, and behaving are embraced by five personality domains: Extraversion, Agreeableness, Conscientiousness, Neuroticism (or Negative Emotionality) and Openness to experience (alternatively labeled Intellect or Open-Mindedness) [11, 62, 67]. These domains are basic personality dimensions, and each of them is a quantitative variable with a positive and negative pole (e.g. the negative pole of Extraversion is introversion, and the negative pole of Neuroticism is emotional stability).

Most papers focus on the Big Five model [40], possibly because of the ease of its interpretation and because the results can be expressed quantitatively [74]. A discussion on the usability of this model in recommendation systems can be found in [75]. However, in our study, the revised version of the BFI (i.e. BFI-2) [67] was used. This psychometric model contains scales for the 5 primary domains and 15 subscales, nested within the primary ones (in total 20 personality dimensions, further referred to as traits). Brief characteristics of primary personality domains, and a list of their lower-order subscales (further referred to as facets) are given below:

- **Extraversion**: characterizes the activity (energy) level, the number of social interactions and social self-confidence, as well as positive emotionality.
  - Sociability, Assertiveness, Energy Level;
- **Agreeableness**: general disposition toward other people: positive, trustful, polite, empathic and altruistic vs. negative, antagonistic, and egocentric.
  - Compassion, Respectfulness, Trust;
- **Conscientiousness**: revealed in relation to work, rules and obligations and characterizes the level of orderliness, dutifulness, as well as perseverance and diligence.
  - Organization, Productiveness, Responsibility;
- **Neuroticism**: contains negative emotionality, over-sensitivity, volatility and irritability, as well as vulnerability, lack of resistance to stress, and low self-esteem.
  - Anxiety, Depression, Emotional Volatility;
- **Openness**: positive (cognitive) attitude towards novelty and both intellectual stimuli (abstract ideas), as well as aesthetic (or artistic) experiences; vivid imagination and complex thinking.
  - Aesthetic Sensitivity, Intellectual Curiosity, Creative. Imagination

Music Master Application

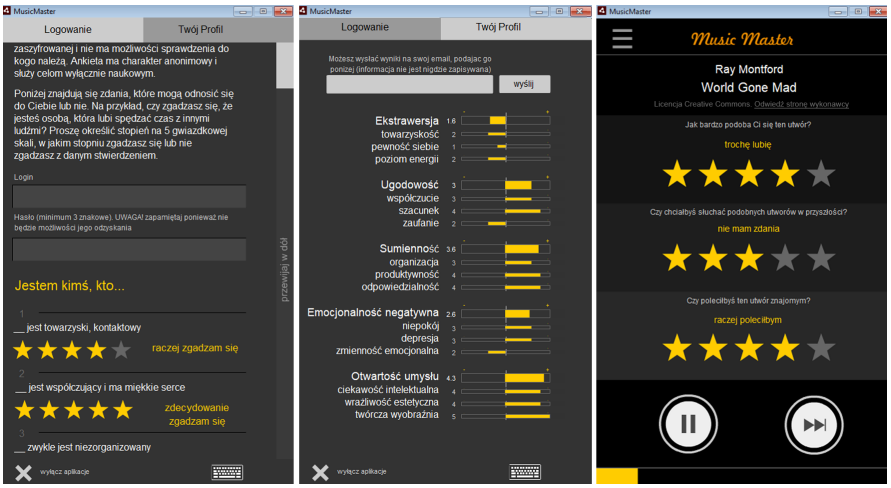


Figure 1: The main screens of the Music Master application used in the experiments. The left screen shows the registration process, with 60 questions from the BFI-2 questionnaire. The middle screen displays the intensity of negative and positive poles of the personality dimensions. The right screen presents a music player where the participants can rate the song in terms of three different aspects (cognitive, motivational and social), using a 5-point Likert scale. The application is for Polish users, so the user interface is in Polish.

We developed an application to gather the listener's personality profiles and musical ratings. The application communicates with the server using the TCP/IP protocol. The client part is called Music Master (MM). Its User Interface (UI) is divided into three main views: personality registering, personality visualization, and music player (see Figure 1). First, the user needs to create an account by assigning a username and password and then rates the phrases about oneself. The phrases came from BFI-2, e.g. *"I am someone who is outgoing"* or *"I am someone who is compassionate"* [67]. Next, the personality profile is calculated and presented visually. When saving, the data is encrypted to ensure anonymity. Setting up a new account allows the user to start listening to music. The application is prepared to propose one song at a time or to generate a set of songs as a playlist. However, only the first option was used for gathering the data described in this paper. The client part has been written in Action-Script 3.0 in the Adobe Animate CC software. It allowed easy deployment on various platforms, such as for PC or mobile applications with Android or iOS operating systems. The server part has been written in JAVA. Its role is to communicate with the client and stream audio files. It saves the music meta-data, audio features, the user's profiles, ratings, and user actions. The recommendation engine has been written in Matlab. The 29-dimensional feature vector represents each song. The description of the features is presented below.

Audio features

There were 29 features calculated from each of the songs: 11 amplitude-based features, 6 spectrum-based features, 4 high-level features, and 8 emotion-based fea-

tures. They were calculated using 50 ms frame length with Hamming windowing and half-frame overlapping by means of the MIRtoolbox in Matlab [76, 77]. The values of the audio features were averaged across all the frames within the length of the audio file. Some features are based on the statistics of occurring sudden bursts of signal energy that usually corresponds to such events as notes, chords and rhythm beats. Additional information about each feature can be found in [76–78].

Amplitude-based features

- Attack time: the mean, standard deviation, slope, and entropy of the duration of events' attack phase, detected in the amplitude of the signal (**AttackTimeMean, AttackTimeStd, AttackTimeSlope, AttackTimeEntropy**).
- Attack slope: the mean, standard deviation, slope, and entropy of the average slope of events' attack phase, detected in the amplitude of the signal (**AttackSlopeMean, AttackSlopeStd, AttackSlopeSlope, AttackSlopeEntropy**).
- Zero crossing rate (**Zerocross**) is a simple indicator of the noisiness of the signal. It counts the average number of times that the signal changes sign in the frame.
- **RMS** measures the global energy of the signal. It is defined as the root mean square of the energy of the amplitude.
- **Lowenergy** is the percentage of frames that show less than average energy [79].

Spectrum-based features

- Centroid, spread, skewness, kurtosis, flatness, entropy are statistical descriptions of spectral distribution and are described by statistical moments.
  - **Centroid** indicates the center of mass of the spectrum. It has a connection with the impression of the brightness of a sound. A higher value of centroid corresponds to a brighter sound (i.e. with more energy of the signal being concentrated within higher frequencies).
  - **Spread** is the indicator of how a spectrum is spread in the frequency domain. Noises have a high spectral spread, whereas sounds with isolated peaks in the spectrum have a low spectral spread. Noisy signals are more challenging to interpret. Spectral spread is used as an indication of the dominance of a tone because the spread is low in this case; pitched sounds have low spectral spread. For complex sounds, the spread increases as the tones diverge and decreases as the tones converge.
  - **Skewness** measures the symmetry of the distribution. A distribution can be positively skewed in the case when it has a long tail to the right, while a negatively skewed distribution has a longer tail to the left. Symmetrical distribution has a skewness of zero. For harmonic signals, the spectral skewness indicates the relative strength of higher and lower harmonics.
  - **Kurtosis** measures the flatness or non-Gaussianity of the spectrum around its centroid. It is used to indicate the "peakiness" of a spectrum. For example, if the white noise is occurring within the signal, then the kurtosis decreases.

  - **Flatness** can be used to distinguish between a harmonic (flatness close to zero) and a noisy signal (flatness close to one for white noise).
  - **Entropy** is low for a spectrum with many distinct spectral peaks and high for a flat spectrum. Spectral entropy is a measure of signal irregularity.

Higher-level features

- **EventDensity** estimates the average frequency of events per second.
- **PulseClarity** estimates the rhythmic clarity, indicating the strength of the beats [80].
- **Inharmonicity** estimates the number of partials that are not multiples of the fundamental frequency. It takes into account the amount of energy outside the ideal harmonic series.
- **Brightness**. Although spectral centroid can be used as brightness predictor, we decided to use an improvement to it, namely to calculate centroid only for signal energy above a particular frequency; we chose 1500 Hz [81, 82]. This feature might be used to quantify the sensation of sharpness, related to the high frequency content of a sound.

Emotion-based features

- **Activity**, **Valence**, **Tension**, **Happy**, **Sad**, **Tender**, **Anger**, **Fear**: emotions evoked in music can be described using two paradigms: in terms of five basic emotions (i.e. happy, sad, tender, anger, and fear) and in terms of three dimensions: activity (or energetic arousal), valence (a pleasure-displeasure continuum) and tension (or tense arousal). The output of the predictive model of emotions, found on the basis of parameters from musical signal [77, 83] gives the localisation of emotional content within the five basic classes and within the three dimensions.

The Experiment Setup

In the presented work, 279 participants were invited to take part in the experiment. They were mainly students from the Faculty of Information Technology and the Faculty of New Media Arts of the Polish-Japanese Academy of Information Technology. The listening sessions were organized only for volunteers in classrooms with a small number of students. Each participant was asked to set up an account with their personality profile in the Music Master (MM) application. It was preceded by a short presentation about the data encryption in the code because it was necessary to convince the participants that the research was entirely anonymous. Over-ear semi-open headphones AKG K-240 were used in the experiments. The participants were informed that they can listen to as many songs as they want for at least 10 minutes and they should not perform any other tasks on the computer. The songs were on Creative Commons license, randomly chosen from the pool of 745 songs downloaded from the magnatune.com website. The details about the pool of songs used in our experiments are given in Table 1. The participants were informed that they could skip the song after the minimum 20 seconds of continued listening (with the option to pause or skip to any desired point) and when the song received

Table 1: The number of songs per genre used for the experiment

Genre	Number of songs	Genre	Number of songs
classical	123	world	142
jazz	63	hard rock	113
alternative rock	145	electronic rock	42
electronica	117		

ratings. The three types of ratings we gathered denote answers, using a five-point Likert scale, to the following three questions:

- **Q1**: How much do you like this song?
  - (1) "I definitely don't like it", (2) "I rather do not like it", (3) "I have no opinion", (4) "I rather like it", and (5) "I definitely like it".
- **Q2**: Would you like to listen to similar songs in the future?
  - (1) "I definitely would not want to", (2) "I rather would not want to", (3) "I have no opinion", (4) "I rather would want to", and (5) "I definitely would want to".
- **Q3**: Would you recommend this song to your friend?
  - (1) "I definitely would not recommend", (2) "I rather would not recommend", (3) "I have no opinion", (4) "I would rather recommend", and (5) "I would definitely recommend."

The majority of music recommendation systems ask users about *"how much do you like this song?"* (Q1 rating type) and try to predict the same for unknown songs. This question refers to the cognitive-emotional component of the attitude towards a particular song (i.e. simply to the actual opinion and belief concerning the reaction to music). The question Q2, *"would you like to listen to similar songs in the future?"*, refers to the motivational component of the attitude, reflecting possible engagement in future contacts with the song. It is worth noting that the prediction of future engagement with similar songs is something the recommendation systems try to do. Finally, the question Q3 *"would you recommend this song to your friend?"* refers to the social component of the attitude, reflecting a willingness to share a given song with the user's friends. To summarize, we can say that while Q1 refers to just intrapsychic elements of the song preference, Q2 and Q3 are markers of its more extrinsic and behavioral aspects.

Collected Data

In total, 5278 data items have been recorded. Each item represents the ratings of a particular song by one user, according to the three questions (Q1, Q2 and Q3). The answers to these three questions are further referred to as three rating types. The collected data set contains the values of 20 personality traits, the ratings for Q1, Q2, and Q3, and audio features extracted from musical files. The data set is publicly available.

Afterwards, three user-item matrices (each containing a different rating type) with 279 rows (users) and 745 columns (songs) were created. The sparsity of the matrices is equal to 0.9764. The global averages of the ratings are 2.85 for Q1, 2.58 for Q2, and 2.28 for Q3. We also studied the relationships between personality traits and ratings. We used Pearson's correlation coefficient to measure the strength and direction of each relationship (see Figure 2). The correlation between Q1 and Q2

equals 0.871, between Q1 and Q3 0.764 and between Q2 and Q3 0.806. Figure 3 presents the distribution of each rating type across the Likert scale.

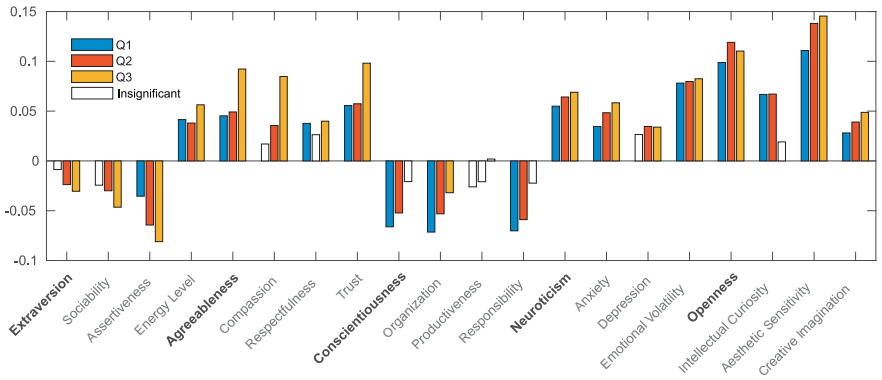


Figure 2: Pearson's correlation coefficients, describing correlations between three rating types and personality domains. The Big-Five main traits are marked in bold. Statistically insignificant correlations ($p > 0.05$) are marked in white.

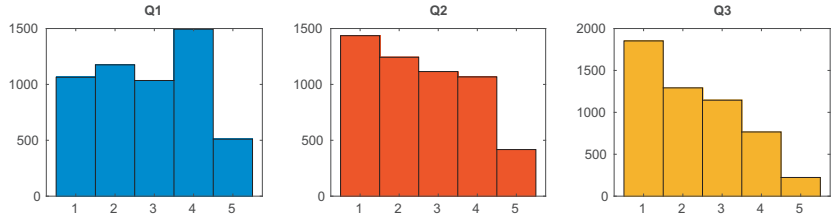


Figure 3: The histograms of each rating type in the collected data set.

Proposed Methodology

The role of MRS is to predict the user's rating value for an unknown song. The prediction is perfect when it is equal to the rating value that the user would give. More formally, having the group of users U and the set of songs S, the system's task is to learn a function f, which predicts the recommendation value $r \in R$ for a song s to user u: $f(u, s) : U \times S \rightarrow R$.

Model based approaches, especially those incorporating DL techniques, can learn the recommendation function f to predict ratings with high accuracy [84]. This requires a sufficient amount of data to prevent the models from over-fitting during the training. However, the size of our data-set is not sufficient for DL models. Moreover, we wanted to obtain high interpretability of the learning model, and to analyze the results and interactions between variables in the prediction from a psychological point of view. This would be cumbersome or impossible in the case of DL. Therefore, we decided to implement an easy to interpret memory-based Collaborative Filtering (CF) algorithm. It utilizes the k most similar users (user-based) or similar items (item-based) for predicting rating for a given item, i.e. song

[1, 2]. Cosine similarity is one of the most common measures used to calculate the similarity of two vectors of ratings [2], and we decided to use this measure.

For rating similarity calculations, the item and user ratings were first normalized using $rnorm_{u,i} = \mu + b_i + b_u$, to remove user and item bias. The $rnorm_{u,i}$ represents the normalized rating for user u and item i, μ denotes the global rating average, b_i and b_u are item and user bias, respectively. The biases are calculated as the difference between the global average and the average item or user ratings.

In order to determine the set of k most similar users (user-based) or items (item-based), we first calculate a similarity matrix for each approach, using cosine similarity, and k most similar users/items are found in the corresponding similarity matrix. In an item-based approach, the rating prediction for a song s and a user u is determined according to the following formula:

$$predItemBased(u, s) = \frac{\sum_{n \in K} sim(n, s) * (r_{u,n})}{\sum_{n \in K} sim(n, s)} \tag{1}$$

where $sim(n, s)$ denotes the similarity between the song s and its n'th most similar neighbor. The $r_{u,n}$ is the rating for the n'th item given by the user u.

In a user-based approach, the rating prediction can be defined according to the following formula:

$$predUserBased(u, s) = \frac{\sum_{n \in K} sim(n, u) * (r_{s,n})}{\sum_{n \in K} sim(n, u)} \tag{2}$$

where $sim(n, u)$ denotes the similarity between the user u and its n'th most similar neighbor. The $r_{s,n}$ is the rating for the item s given by the n'th user.

Next, the user and item based approach were combined in the following formula of the hybrid rating prediction:

$$predHybrid(u, s) = \frac{\sum_{n \in K} sim(n, s) * (r_{u,n}) + \sum_{n \in K} sim(n, u) * (r_{s,n})}{\sum_{n \in K} sim(n, s) + \sum_{n \in K} sim(n, u)} \tag{3}$$

Besides the similarity of ratings, we also used the similarity of audio features (instead of $sim(n, s)$) and personality domains (instead of $sim(n, u)$) in our experiments. These data were normalized to have zero mean and standard deviation equal to one (z-score normalization), and cosine similarity was applied.

We evaluated the experiments by calculating Root Means Square Error (RMSE), using the 10-fold cross validation approach (10-CV) to evaluate the predictions of ratings. Therefore, the "recommendation quality" in the further text refers to the quality measured by RMSE obtained via the 10-CV procedure, and the lower the RMSE, the higher the recommendation quality. We will report the results for all three rating types: Q1, Q2, and Q3.

Experiments

In our experiments we studied two main hypotheses:

1 The recommendation quality differs when employing various personality domains (user-based approach) or audio features (item-based approach).

2 There is a difference in recommendation quality when using solely Big-Five personality traits, or their low-level facets (using a hybrid approach).

In order to examine these hypotheses, baseline recommendation quality values (in terms of RMSE) were calculated first for various settings. First of all, a global average value of ratings was calculated as a baseline prediction. Next, we calculated baseline RMSE values for simple user-based and item-based CF. Subsequently, the similarity of ratings ($sim(n,u)$ and $sim(n,s)$) were replaced with the similarity of all personality traits and the similarity of all the audio features. Finally, we calculated baselines for the hybridized approaches. The results of the baseline RMSE values are presented in Table 2.

Table 2: RMSE (10-CV) calculated in baseline experiments, for each rating type. The k denotes the number of neighbors used for prediction, chosen experimentally.

Experiment name	**RMSE for:**	**Q1**	**Q2**	**Q3**
Global rating average		1.296	1.292	1.203
Item-based with ratings similarity ($k = 50$)		1.192	1.173	1.052
User-based with ratings similarity ($k = 10$)		1.326	1.313	1.199
Item-based with all features similarity ($k = 10$)		1.163	1.144	1.029
User-based with all personalities similarity ($k = 10$)		1.365	1.361	1.262
Hybrid with ratings similarity ($k_s = 50, k_u = 10$)		1.180	1.163	1.043
Hybrid with all personalities and features similarity ($k_s = 10, k_u = 10$)		1.150	1.139	1.028

In order to study the influence of individual personality traits on the quality of music recommendations, we used a user-based CF. The influence of individual audio features was examined using an item-based approach. In order to measure the similarity between individual personality domains and individual audio features, which are 1-dimensional vectors (scalars), $1 - d$ was applied as a similarity measure (instead of cosine similarity), where d denotes Euclidean distance, The results are presented in Figures 4 and 5.

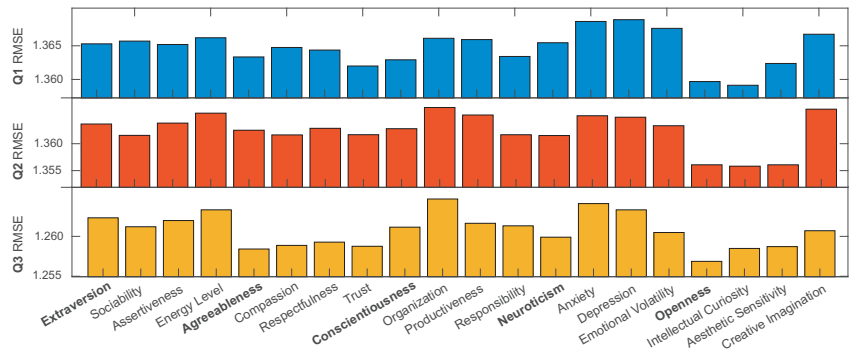


Figure 4: The comparative analysis of recommendation quality in a user-based approach, with singular personality domains used in similarity calculations.

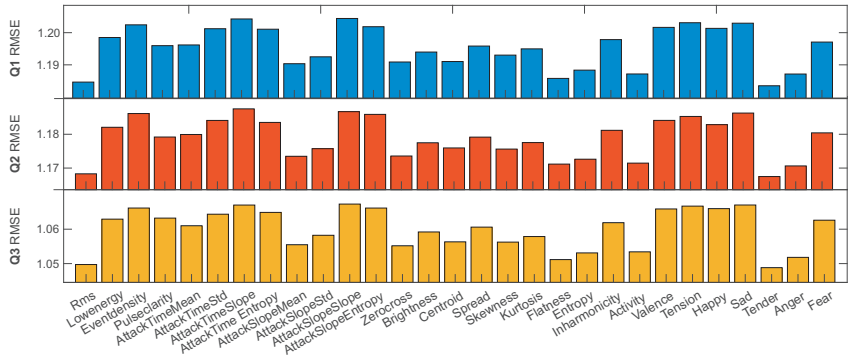


Figure 5: The comparative analysis of recommendation quality in an item-based approach, with singular audio features used in similarity calculations.

For studying the differences in recommendation quality between Big-Five and their low-level personality facets, we used the hybrid model for rating prediction (see Eq. 3). First, we calculated the similarity values for simplified models, namely for each pair consisting of one personality trait and one audio feature, and these values were applied to calculate predictions, for each rating type (see Figure 6). Next, from all performed experiments, two minimum RMSE values were chosen for each rating type: 1) belonging to one of the Big-Five traits and 2) belonging to one of the personality facets. Together with their corresponding audio feature, these results were saved for further experiments. The pair (personality dimension and audio feature) that gave the lowest RMSE for each rating type, will be further called the "best pair". Therefore, we obtained 6 best pairs, i.e. two pairs for each of the three ratings Q1, Q2, and Q3.

In the next steps, we gradually improved the results. We started with the two pairs, for which minimal values of Q1 are obtained (see Figure 6), i.e. curiosity, tender, and Openness, tender. Next, for each of the two previously selected best pairs, the next best pair was added and selected in the same manner as the first one. Namely, we added one personality trait (domain or facet) and one audio feature, together with the previous pair yields minimal RMSE for Q1. The difference was that the first selection used Euclidean distances (as we had one-dimensional vectors, for which cosine distance would not work), and in the next steps, cosine similarities were applied (as in this case we had multi-dimensional vectors). Every selection was performed in two ways: selecting only among Big Five domains and only low level facets. This process was repeated step by step until the RMSE error started to grow. In each step, we reported the minimum RMSE results. The same procedure was also performed for Q2 and Q3. The results are presented in Figure 7.

Results and Discussion

Every comparative analysis of results in the description below will concern the values of Q1, unless indicated otherwise.

Aesthetic Sensitivity (Openness's facet) has the highest and positive correlation with all rating types (see Figure 2). Interestingly, Assertiveness (Extraverion's facet)

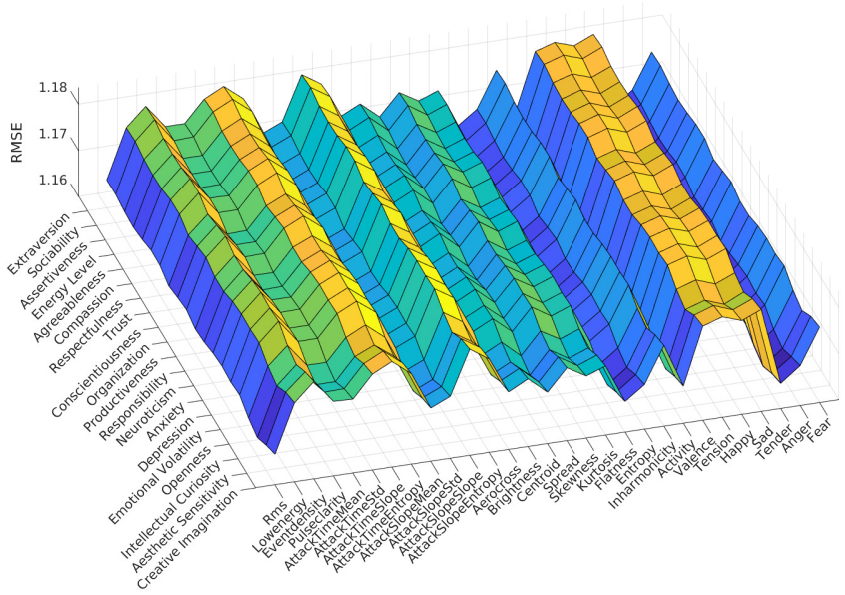


Figure 6: The prediction error for Q1 ratings using a hybrid model based on the similarity of a single personality trait and a single audio feature. The minimum point (best pair) with respect to the 15 personality facets is obtained for curiosity and tender (RMSE = 1.1628). The best pair with respect to the Big Five personality domains is obtained for Openness and tender (RMSE = 1.1639).

negatively correlates with all ratings. We believe that this can be explained by the genres used in our experiment (classical, world, jazz, hard rock, alternative rock, electronic rock, and electronica). Rentfrow et al. [85] show that people with high Openness usually prefer more complex music, like blues, jazz, folk, and rock, than Extravert people, who usually appreciate upbeat music like hip-hop, funk and electronic [53]. Our experiments corroborate these findings.

As shown in Figure 2, persons of high Openness usually give higher ratings, in contrast to the persons of high Extraversion, who usually give lower ratings. Additionally, the genres used in the experiments seem to be preferred by persons high in the trait of Openness. However, as described in [44], people of high Openness have broader musical tastes (and enjoy more genres) than Extraverted people. Therefore, they may rate the music higher because they generally like to listen to it, not only because of the preferred genres. To summarize, even though we found statistically significant correlations between ratings and personality domains, these correlations are relatively weak. The performed meta-analysis described in [86] confirms weak connections between personality and five-dimensional MUSIC factors for music preferences. The authors in [52] also confirm that associations between personality and acoustic features exist, though this association is relatively weak. Nevertheless, it is worth noting that some lower-order facets show higher correlation with ratings than their main personality domains (see Figure 2).

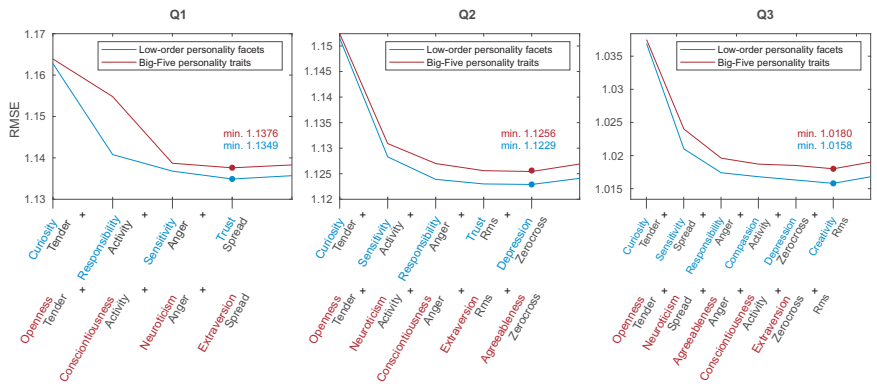


Figure 7: The recommendation quality when gradually adding consecutive the best pairs to the previous ones. They were selected in two ways: only the best results belonging to the low-level personality facets (blue colors) and only the best results belonging to the Big-five personality domains (red color). This gradual improvement method revealed the moment (the subset of personality domains and audio features), marked by dots on each graph, after which the errors started to grow.

Looking at the baseline results presented in Table 2 we can conclude that predicting Q2 and Q3 gives lower RMSE than predicting Q1 in all performed experiments. This means that our models make more accurate predictions for Q2 (*how much the user wants to listen to similar songs in the future*) than for Q1 (*how much the user likes the song*). Furthermore, the models perform even better when predicting Q3 (*how much the user would like to share the song with friends*). We believe these differences can be explained by the different distribution of these rating types (see Figure 3). In the case of Q2 and Q3, we can see that participants tended to give lower ratings more often than for Q1. Therefore, a system that predicts lower ratings for Q2 and Q3 will achieve lower RMSE. Q1 refers to the opinion or belief concerning a particular song that the listener has just heard, and therefore it could be treated as a somewhat superficial aspect of the attitude. In contrast, Q2 and Q3 reflect socio-motivational and therefore more behavioral aspects of the attitude towards the music, requiring more engagement. Therefore, Q2 and Q3 can be seen as concerning more profound psychological characteristics, more strongly related to stable personality dispositions (traits).

When comparing all the rating-based CF results, we can see that the user-based approach performs much worse than the item-based one (1.326 vs 1.192). It is not surprising as the user-item matrix had only 275 users but 745 songs. Thus, the model had a more limited number of user neighbors for making the prediction, compared to songs. Regarding the item-based approach, replacing rating with audio feature-based similarities improves the results (1.192 vs 1.163). The reason is that the similarity of the audio features expresses the actual nearness of the songs (taking into account their audio content) better than the similarity of rating vectors. On the other hand, we have not observed the improvement when replacing ratings

with the personality traits' similarities in the case of the user-based approach (1.326 vs 1.365). This result may suggest that people with similar personalities might not share similar musical tastes with the same strength as people with similar song ratings. However, in [10] the authors have shown that combining the personality similarity with a rating-based CF can bring improvement in rating prediction, compared to predictions based on rating data only. Therefore, we think we could not get a lower error because we used personality similarity alone, without the similarity of rating data. To confirm whether this combination will improve the results, as stated in [10], there is a need to combine personality and ratings similarity in the future work.

It is worth noting that only the similarity of Intellectual Curiosity gives a lower error than the similarity of all personality traits together (1.359 vs 1.365). It confirms the findings of Braunhofer et al. [87] who have shown that exploiting even a single personality trait may lead to a considerable improvement in recommendation accuracy. Still, even if the improvement was observed for a single personality trait (Curiosity), the error (1.359) is still higher than user-based CF with rating similarity (1.326). Therefore, additional experiments with more data that combine the similarity of ratings and personalities are needed in the future.

When analyzing the recommendation quality of the hybrid model, and using the similarity of a single personality trait and a single audio feature, we can see that Intellectual Curiosity and Tender (emotion-based audio feature of music) result in the lowest error, see Figure 6. Furthermore, this hybrid model slightly outperforms the item-based CF that considers all the personality and audio features dimensions (RMSE=1.1630 for CF vs 1.1628 for this particular hybrid model).

Prediction using the similarity of personality facets yields a lower error for all Qs than prediction based on the similarity calculated for any combination of the main Big-Five personality domains (see Figure 7). Error reduction is relatively small, but always exists. However, the main gain results from the reduction of the set of personality facets (together with the appropriate set of audio features) applied in the similarity calculations. We found that Intellectual Curiosity, Responsibility, Aesthetic Sensitivity and Trust yielded the lowest recommendation error for Q1 (see Figure 7). In this case, the RMSE error was reduced from 1.1628 to 1.1349. The characteristic of the above set of personality facets is as follows: people of high Intellectual Curiosity desire to acquire general knowledge about the world, such as on how systems work, about mathematical relationships, what objects are composed of, etc. Responsible people are being accountable or blamed for something. Therefore, they feel a moral obligation to behave correctly, so other people usually perceive them as reliable. Aesthetic Sensitivity describes the ability to detect and appreciate beauty wherever it exists.

We used miremotion library [83] to calculate all the audio features, including the description of music-evoked emotions, based on the analysis of the audio signal of the recordings. These emotions have been described using two representations: 1) a discrete model with five basic emotions: happy, sad, tender, anger, and fear,

2) a three-dimensional model, where these five basic emotions can also be placed: with the following dimensions: activity (energetic arousal), valence (a pleasure-displeasure continuum), and tension (or tense arousal). From Figure 7, we can see that the similarity of activity of the tender and anger emotions evoked in music contributed most to the reduction of the recommendation error. This conclusion can also be drawn from an item-based approach, with single audio features used in similarity calculations (see Figure 5). It is also worth noting that, among other features, the indicator of how a spectrum is spread in the frequency (spread) contributed to reaching the minimum RMSE for Q1 and Q3. In addition, the global energy of the signal (rms) and its noisiness (zero-crossing rate) also contributed to reaching the minimum RMSE for Q2 and Q3.

Since the prediction highly depends on the similarity measure, further experiments may incorporate dimensionality reduction techniques (such as Singular Value Decomposition (SVD) or Principal Component Analysis (PCA)), together with clustering algorithms (such as k-means of Self Organizing Maps (SOM)), in order to find similar users or items [88]. SOM produce clusters in an unsupervised manner from multi-dimensional data. Since the prediction could also depend on Q2 and Q3 values, the SOM can be used to group similar users or items, based on the three rating types, and also on other available observations together (personality and audio features). The clusters obtained in this way can bring improvements in rating prediction [89]. Additionally, further analysis is required to investigate how motivational (Q2) and social (Q3) components of the attitude towards the song influence the cognitive-emotional (Q1) component, which may depend on the personality. This can be inferred from the dataset by employing appropriate statistical analysis. Another interesting hypothesis to check, similar to the described in [52], is the existence of the difference between features rated low and high, which may depend on the level of personality traits. The authors of this paper leave this (and also others) hypotheses to investigate by other researchers.

The recommendation quality does not depend on the prediction accuracy only. The prediction is needed for the recommender systems to build the list of songs for which the highest prediction of Q1 is obtained, indicating that the user would probably like to listen to these songs. Therefore, the songs are added to the recommendation list in the order of increasing RMSE for Q1 prediction. However, the user may actually prefer listening to other songs at the moment. In our opinion, it seems reasonable to select the song for which the predictions of Q1 and Q2 were both high. This means that the system could recommend songs similar to those the user would like to listen to in the future (Q2) and reject those for which Q2 was low. It presumably would increase user satisfaction with the recommended items. As far as Q3 rating is concerned, the system could favor the songs that received high Q3 ratings from the user's friends. However, the link between "being the nearest neighbor" (used in the recommendation algorithm) and "being a friend" is unclear. Another idea is to formulate a confidence measure that tells the system how trustworthy a particular prediction is. This measure would need to incorporate additional knowledge about the interactions between the three rating types, the number of ratings

in neighbors, and possibly other factors. The authors leave these issues to be investigated in the future, and keep it as open research questions, to be discussed by other researchers.

One of the limitations of our experiments is that we used the random selection of music from the magnatune.com website. It offers both Eastern and Western music to download. As stated in [90], in terms of BFI, only the preferences for Western music are universal across 53 countries, but we do not know whether it is true for Eastern music as well. Additionally, the range of music genres was limited, and a more elaborate genre taxonomy would allow us to compare the results with other researchers [49, 50, 53], in terms of the preferences to genres by personality traits. In the future study, there may also be a control question that verifies answers related to personality types. Additionally, there is still an open question, how do our findings correlate with real world scenarios and how the preferred use of music (relaxing or jogging) influences the way participants rate the music in the experimental controlled environment with the use of headphones. However, the most important conclusion from the experiments performed in this paper is that utilizing BFI-2 (instead of BFI) is worth considering with every rating type.

Conclusions

This article describes the effect of utilizing BFI-2 personality domains in the music recommendation systems on the recommendation error. The BFI-2 allowed performing the analysis with more granularity due to the availability of low-order facets of Big-Five personality domains. We collected the personality profiles and three music rating types (related to cognitive, motivational and social components of the attitude towards the music) from 279 users of the newly developed Music Master application. In addition, 29-dimensional vectors of audio features were incorporated into the analysis. To the best of our knowledge, a dataset with BFI-2 personality profiles, three rating types, and audio features has never been published before.

The experiments with our hybrid recommendation model showed interesting interactions between personality domains and audio features. It turned out that only the several low-order personality facets were enough to obtain the lowest recommendation error. The Intellectual Curiosity, Responsibility and Aesthetic Sensitivity decreased the error significantly for predicting all three rating types. It is essential to note, when using memory-based methods, any combination of Big-Five personality traits produced a higher error than lower-order personality facets. However, there is still an open question whether the results scale to the real world scenarios or to model based methods.

The experiment also revealed the subset of audio features that contributed most to obtaining the lowest error. These features refer to the activity of tender and anger emotions (i.e. two basic emotions, tender and anger, as represented along the activity axis in 3-dimensional space) evoked in music. These features were calculated based on the analysis of the audio contents of the recordings. More details about the predictive models of emotions can be found in [83].

We performed our experiments on a small dataset (5278 ratings from 279 users) and a relatively simple recommendation model based on user or item similarity. Unfortunately, our initial trials with training Singular Value Decomposition (SVD)

caused over-fitting with the dataset due to its relatively small size. Therefore, a more extensive setup and even live system, working in real time, are required to prove that the reported subset of personality domains scales well with different recommendation algorithms. Nevertheless, the proposed simple hybrid model allowed a detailed analysis, based on the similarity of users and the similarity of songs.

An additional conclusion is that, instead of implementing the complete BFI-2 questionnaire, it is more practical and more effective to implement only a small subset of its questions. We observed that the best trade off between the performance and the number of questions is to have the following three personality traits: Intellectual Curiosity, Aesthetic Sensitivity and Responsibility, and the following three audio features: tender, anger, and activity (see Figure 7). When adding additional ones, the error improvement is negligible. Therefore, instead of 60 questions (4 questions per personality facet), only the 12 of them would result in a better recommendation performance and higher user satisfaction than a full questionnaire.

The authors hope that other researchers will find the data set practical and stimulating to design other experiments, and prove other hypotheses that relate to three aspects of ratings (Q1, Q2, and Q3), recommendation models, and personalities.

Acknowledgements
The authors want to thank all participants who agree to take part in the experiments described in this paper.

Funding
Not applicable

Abbreviations
Not applicable

Availability of data and materials
The dataset analysed during the current study are available in the figshare repository, `https://dx.doi.org/10.6084/m9.figshare.19678962`

Ethics approval and consent to participate
Not applicable

Competing interests
The authors declare that they have no competing interests.

Consent for publication
All authors have approved the paper for being published.

Authors' contributions
MK implemented the Music Master application, gathered and analysed the data. AW and KS interpreted the results, substantively revised them and improved the language and descriptions. WS provided the idea of incorporating BFI-2 into the research and the results validation from the psychological point of view. All authors read and approved the final manuscript.

Authors' information
M.K, MSc, endlessly fascinated by technology and music, he has found curiosity driving his research in the direction where music and technology meet together. He conducts research on music processing by deep neural networks, while working to complete his doctoral dissertation. Mariusz is also a full-stack web developer. What excites him more than anything else is the prospect of combining all of his skills and interests together: web development, machine learning, music, and sound processing.
A.W. PhD, DSc is a computer scientist, specialising in multimedia. She is presently an Associate Professor and the Head of Multimedia Laboratory at the Polish-Japanese Academy of Information Technology (PJATK), Warsaw, Poland. Additionally, she is also an associate member of the Graduate Faculty at the University of North Carolina at Charlotte. She has always been interested in music, and graduated from the F. Chopin State School of Music (Second Level) in Gdansk. Her scientific interests include multimedia, music and audio information retrieval, human-computer interaction, as well as automated identification of emotions from various signals, data mining, and computer graphics. She has co-authored over 100 scientific works.
K.S (PhD, DSc) is a computer scientist. He received his PhD in 2009 in the area of human-computer interaction. His PhD thesis was on a multimodal speech synthesis system, the first non-commercial system of this kind in Poland. In 2020, he earned his DSc. His areas of expertise include voice quality issues, as well as classification and the digital processing of speech signals using electroglottography. Krzysztof Szklanny is the author and co-author of

many papers, and he has also participated in several national and international research projects. Dr. Szklanny is also a professional photographer.
W.S., PhD, is a personality psychologist specialising in research on personality structure. In particular, he is interested in basic dimensions and higher-order factors of personality, as well as in personality disorders and optimal functioning. Together with Jan Cieciuch and Tomasz Rowiński, he developed the Circumplex of Personality Metatraits - a synthetising model of personality.

Author details
[1]Multimedia Dept., Polish-Japanese Academy of Information Technology, Warsaw, Poland. [2]Institute of Psychology, Cardinal Stefan Wyszyński University, Warsaw, Poland.

References

1. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: The Adaptive Web, pp. 291–324 (2007)
2. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 230–237 (1999)
3. Tao, Y., Zhang, Y., Bian, K.: Attentive context-aware music recommendation. In: 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), pp. 54–61 (2019). IEEE
4. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Recommender Systems Handbook, pp. 1–35 (2011)
5. Herce-Zelaya, J., Porcel, C., Bernabé-Moreno, J., Tejeda-Lorente, A., Herrera-Viedma, E.: New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests. Information Sciences **536**, 156–170 (2020)
6. Ojagh, S., Malek, M.R., Saeedi, S.: A social–aware recommender system based on user's personal smart devices. ISPRS International Journal of Geo-Information **9**(9), 519 (2020)
7. Camacho, L.A.G., Alves-Souza, S.N.: Social network data to alleviate cold-start in recommender system: A systematic review. Information Processing & Management **54**(4), 529–544 (2018)
8. Gizaw, T.Z., Dong Jun, H., Oad, A.: Solving cold-start problem by combining personality traits and demographic attributes in a user based recommender system. International journal of advanced research in computer science and software engineering **7**(5) (2017)
9. Tiwari, V., Ashpilaya, A., Vedita, P., Daripa, U., Paltani, P.P.: Exploring demographics and personality traits in recommendation system to address cold start problem. In: ICT Systems and Sustainability, pp. 361–369 (2020)
10. Hu, R., Pu, P.: Enhancing collaborative filtering systems with personality information. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 197–204 (2011)
11. McCrae, R.R., Costa, P.T.: Personality in Adulthood: A Five-factor Theory Perspective, (2003)
12. Schedl, M.: Deep learning in music recommendation systems. Frontiers in Applied Mathematics and Statistics **5**, 44 (2019)
13. Fessahaye, F., Perez, L., Zhan, T., Zhang, R., Fossier, C., Markarian, R., Chiu, C., Zhan, J., Gewali, L., Oh, P.: T-recsys: A novel music recommendation system using deep learning. In: 2019 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–6 (2019). IEEE
14. Khoali, M., Tali, A., Laaziz, Y.: Advanced recommendation systems through deep learning. In: Proceedings of the 3rd International Conference on Networking, Information Systems & Security, pp. 1–8 (2020)
15. Irene, R.T., Borrelli, C., Zanoni, M., Buccoli, M., Sarti, A.: Automatic playlist generation using convolutional neural networks and recurrent neural networks. In: 2019 27th European Signal Processing Conference (EUSIPCO), pp. 1–5 (2019). IEEE
16. Aljunid, M.F., Dh, M.: An efficient deep learning approach for collaborative filtering recommender system. Procedia Computer Science **171**, 829–836 (2020)
17. Chang, S.-H., Abdul, A., Chen, J., Liao, H.-Y.: A personalized music recommendation system using convolutional neural networks approach. In: 2018 IEEE International Conference on Applied System Invention (ICASI), pp. 47–49 (2018). IEEE
18. Knees, P., Schedl, M., Ferwerda, B., Laplante, A.: User awareness in music recommender systems. Personalized human-computer interaction, 223–252 (2019)
19. Bauer, C., Novotny, A.: A consolidated view of context for intelligent systems. Journal of Ambient Intelligence and Smart Environments **9**(4), 377–393 (2017)
20. Juslin, P.N., Laukka, P.: Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. Journal of new music research **33**(3), 217–238 (2004)
21. North, A.C., Hargreaves, D.J.: Situational influences on reported musical preference. Psychomusicology: A Journal of Research in Music Cognition **15**(1-2), 30 (1996)
22. Bai, K., Kawagoe, K.: Background music recommendation system based on user's heart rate and elapsed time. In: Proceedings of the 2018 10th International Conference on Computer and Automation Engineering, pp. 49–52 (2018)
23. Lavanya, S., Saranya, G., Navin, K.: Weather based playlist generation in mobile devices using hash map. In: 2017 International Conference on IoT and Application (ICIOT), pp. 1–7 (2017). IEEE
24. Álvarez, P., Zarazaga-Soria, F., Baldassarri, S.: Mobile music recommendations for runners based on location and emotions: The dj-running system. Pervasive and Mobile Computing, 101242 (2020)
25. Su, J.-H., Yeh, H.-H., Philip, S.Y., Tseng, V.S.: Music recommendation using content and context information mining. IEEE Intelligent Systems **25**(1), 16–26 (2010)
26. Chen, J., Ying, P., Zou, M.: Improving music recommendation by incorporating social influence. Multimedia Tools and Applications **78**(3), 2667–2687 (2019)
27. Wu, D.: Music personalized recommendation system based on hybrid filtration. In: 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), pp. 430–433 (2019). IEEE

28. Wang, R., Ma, X., Jiang, C., Ye, Y., Zhang, Y.: Heterogeneous information network-based music recommendation system in mobile networks. Computer Communications **150**, 429–437 (2020)
29. Jin, Y., Htun, N.N., Tintarev, N., Verbert, K.: Contextplay: Evaluating user control for context-aware music recommendation. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp. 294–302 (2019)
30. Juslin, P.N., Sloboda, J.: Handbook of Music and Emotion: Theory, Research, Applications, (2011)
31. Gilda, S., Zafar, H., Soni, C., Waghurdekar, K.: Smart music player integrating facial emotion recognition and music mood recommendation. In: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 154–158 (2017). IEEE
32. Hyung, Z., Park, J.-S., Lee, K.: Utilizing context-relevant keywords extracted from a large collection of user-generated documents for music discovery. Information Processing & Management **53**(5), 1185–1200 (2017)
33. Polignano, M., Basile, P., de Gemmis, M., Semeraro, G.: Social tags and emotions as main features for the next song to play in automatic playlist continuation. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 235–239 (2019)
34. Lopes, P.S., Lasmar, E.L., Rosa, R.L., Rodríguez, D.Z.: The use of the convolutional neural network as an emotion classifier in a music recommendation system. In: Proceedings of the XIV Brazilian Symposium on Information Systems, pp. 1–8 (2018)
35. Iyer, A.V., Pasad, V., Sankhe, S.R., Prajapati, K.: Emotion based mood enhancing music recommendation. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1573–1577 (2017). IEEE
36. Ayata, D., Yaslan, Y., Kamasak, M.E.: Emotion based music recommendation system using wearable physiological sensors. IEEE transactions on consumer electronics **64**(2), 196–203 (2018)
37. Kulkarni, S., Rodd, S.F.: Context aware recommendation systems: A review of the state of the art techniques. Computer Science Review **37**, 100255 (2020)
38. Xu, L., Wen, X., Shi, J., Li, S., Xiao, Y., Wan, Q., Qian, X.: Effects of individual factors on perceived emotion and felt emotion of music: Based on machine learning methods. Psychology of Music (2020)
39. Juslin, P.N., Sloboda, J.A.: Music and Emotion: Theory and Research., (2001)
40. Dhelim, S., Aung, N., Bouras, M.A., Ning, H., Cambria, E.: A survey on personality-aware recommendation systems. Artificial Intelligence Review **55**(3), 2409–2454 (2022)
41. Rentfrow, P.J., Gosling, S.D.: The do re mi's of everyday life: the structure and personality correlates of music preferences. Journal of personality and social psychology **84**(6), 1236 (2003)
42. Dunn, P.G., de Ruyter, B., Bouwhuis, D.G.: Toward a better understanding of the relation between music preference, listening behavior, and personality. Psychology of Music **40**(4), 411–428 (2012)
43. Rentfrow, P.J., Goldberg, L.R., Levitin, D.J.: The structure of musical preferences: a five-factor model. Journal of personality and social psychology **100**(6), 1139 (2011)
44. Bansal, J., Flannery, M.B., Woolhouse, M.H.: Influence of personality on music-genre exclusivity. Psychology of Music (2020)
45. Zweigenhaft, R.L.: A do re mi encore: A closer look at the personality correlates of music preferences. Journal of individual differences **29**(1), 45–55 (2008)
46. Bansal, J., Woolhouse, M.: Predictive power of personality on music-genre exclusivity. In: ISMIR, pp. 652–658 (2015)
47. Kaufman, S.B.: Opening up openness to experience: A four-factor model and relations to creative achievement in the arts and sciences. The Journal of Creative Behavior **47**(4), 233–255 (2013)
48. Chamorro-Premuzic, T., Furnham, A.: Personality and music: Can traits explain how people use music in everyday life? British journal of psychology **98**(2), 175–185 (2007)
49. Ferwerda, B., Tkalcic, M., Schedl, M.: Personality traits and music genres: What do people prefer to listen to? In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 285–288 (2017)
50. Ferwerda, B., Yang, E., Schedl, M., Tkalcic, M.: Personality and taxonomy preferences, and the influence of category choice on the user experience for music streaming services. Multimedia tools and applications **78**(14), 20157–20190 (2019)
51. Ferwerda, B., Tkalčič, M.: Exploring online music listening behaviors of musically sophisticated users. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 33–37 (2019)
52. Flannery, M.B., Woolhouse, M.H.: Musical preference: Role of personality and music-related acoustic features. Music & Science **4**, 20592043211014014 (2021)
53. Dorochowicz, A., Kurowski, A., Kostek, B.: Employing subjective tests and deep learning for discovering the relationship between personality types and preferred music genres. Electronics **9**(12), 2016 (2020)
54. Fernández-Tobías, I., Braunhofer, M., Elahi, M., Ricci, F., Cantador, I.: Alleviating the new user problem in collaborative filtering by exploiting personality information. User Modeling and User-Adapted Interaction **26**(2), 221–255 (2016)
55. Atas, M., Felfernig, A., Polat-Erdeniz, S., Popescu, A., Tran, T.N.T., Uta, M.: Towards psychology-aware preference construction in recommender systems: Overview and research issues. Journal of Intelligent Information Systems, 1–23 (2021)
56. Karumur, R.P., Nguyen, T.T., Konstan, J.A.: Exploring the value of personality in predicting rating behaviors: a study of category preferences on movielens. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp. 139–142 (2016)
57. Liu, R., Hu, X.: A multimodal music recommendation system with listeners' personality and physiological signals. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 357–360 (2020)
58. Nguyen, T.T., Maxwell Harper, F., Terveen, L., Konstan, J.A.: User personality and user satisfaction with recommender systems. Information Systems Frontiers **20**(6), 1173–1189 (2018)
59. Wu, W., Chen, L., He, L.: Using personality to adjust diversity in recommender systems. In: Proceedings of the

24th ACM Conference on Hypertext and Social Media, pp. 225–229 (2013)
60. Onori, M., Micarelli, A., Sansonetti, G.: A comparative analysis of personality-based music recommender systems. In: Empire@ RecSys, pp. 55–59 (2016)
61. Lu, F., Tintarev, N.: A diversity adjusting strategy with personality for music recommendation. In: IntRS@ RecSys, pp. 7–14 (2018)
62. John, O.P., Naumann, L.P., Soto, C.J.: Paradigm shift to the integrative big five trait taxonomy. Handbook of personality: Theory and research **3**(2), 114–158 (2008)
63. Raad, B.d.E., Perugini, M.E.: Big Five Factor Assessment: Introduction., (2002)
64. Ehrhart, M.G., Ehrhart, K.H., Roesch, S.C., Chung-Herrera, B.G., Nadler, K., Bradshaw, K.: Testing the latent factor structure and construct validity of the ten-item personality inventory. Personality and individual Differences **47**(8), 900–905 (2009)
65. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 149–156 (2011). IEEE
66. Dunn, G., Wiersema, J., Ham, J., Aroyo, L.: Evaluating interface variants on personality acquisition for recommender systems. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 259–270 (2009). Springer
67. Soto, C.J., John, O.P.: The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. Journal of personality and social psychology **113**(1), 117 (2017)
68. John, O.P., Srivastava, S., *et al.*: The big five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research **2**(1999), 102–138 (1999)
69. Eysenck, H.J.: Dimensions of personality: 16, 5 or 3?—criteria for a taxonomic paradigm. Personality and individual differences **12**(8), 773–790 (1991)
70. Ashton, M.C., Lee, K.: Empirical, theoretical, and practical advantages of the hexaco model of personality structure. Personality and social psychology review **11**(2), 150–166 (2007)
71. Cieciuch, J., Strus, W., Zeigler-Hill, V., Shackelford, T.K.: Two-factor model of personality (2018)
72. Strus, W., Cieciuch, J.: Are the questionnaire and the psycho-lexical big twos the same? towards an integration of personality structure within the circumplex of personality metatraits. International Journal of Personality Psychology **5**, 18–35 (2019)
73. Goldberg, L.R.: An alternative" description of personality": the big-five factor structure. Journal of personality and social psychology **59**(6), 1216 (1990)
74. Nunes, M.A.S.N.: Recommender systems based on personality traits. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc (2008)
75. Tkalcic, M., Chen, L.: Personality and recommender systems. In: Recommender Systems Handbook, pp. 715–739 (2015)
76. Lartillot, O., Toiviainen, P.: A matlab toolbox for musical feature extraction from audio. In: International Conference on Digital Audio Effects, vol. 237, p. 244 (2007). Bordeaux
77. Lartillot, O.: Mirtoolbox 1.7. 2 user's manual. Oslo: University of Oslo.[Google Scholar] (2019)
78. Alías, F., Socoró, J.C., Sevillano, X.: A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Applied Sciences **6**(5), 143 (2016)
79. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on speech and audio processing **10**(5), 293–302 (2002)
80. Lartillot, O., Eerola, T., Toiviainen, P., Fornari, J.: Multi-feature modeling of pulse clarity: Design, validation and optimization. In: ISMIR, pp. 521–526 (2008). Citeseer
81. Pearce, A., Brookes, T., Mason, R.: Modelling the microphone-related timbral brightness of recorded signals. Applied Sciences **11**(14), 6461 (2021)
82. Juslin, P.N.: Cue utilization in communication of emotion in music performance: Relating performance to perception. Journal of Experimental Psychology: Human perception and performance **26**(6), 1797 (2000)
83. Eerola, T., Lartillot, O., Toiviainen, P.: Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In: Ismir, pp. 621–626 (2009)
84. Zhang, S., Yao, L., Sun, A., Tay, Y.: Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR) **52**(1), 1–38 (2019)
85. Rawlings, D., Ciancarelli, V.: Music preference and the five-factor model of the neo personality inventory. Psychology of music **25**(2), 120–132 (1997)
86. Schäfer, T., Mehlhorn, C.: Can personality traits predict musical style preferences? a meta-analysis. Personality and Individual Differences **116**, 265–273 (2017)
87. Braunhofer, M., Elahi, M., Ricci, F.: User personality and the new user problem in a context-aware point of interest recommender system. In: Information and Communication Technologies in Tourism 2015, pp. 537–549 (2015)
88. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A.: Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal **16**(3), 261–273 (2015)
89. Nilashi, M., bin Ibrahim, O., Ithnin, N.: Hybrid recommendation approaches for multi-criteria collaborative filtering. Expert Systems with Applications **41**(8), 3879–3900 (2014)
90. Greenberg, D.M., Wride, S.J., Snowden, D.A., Spathis, D., Potter, J., Rentfrow, P.J.: Universals and variations in musical preferences: A study of preferential reactions to Western music in 53 countries. Journal of Personality and Social Psychology **122**(2), 286 (2022)

6.5 Music Recommendation Systems: a Survey

Dane bibliograficzne pracy:

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, Cham. https://doi.org/10.1007/978-3-030-66450-3_7

Music Recommendation Systems: a Survey

Mariusz Kleć and Alicja Wieczorkowska

Abstract This introductory chapter presents an overview of music recommendation systems, supported by a comprehensive list of references. We pay special attention to user-centric systems which personalize their output by tracking the context of the user, including the user's emotions and personality. Besides, we emphasize the importance of social media, the usability of user interfaces, and the application of deep learning techniques in the recent developments in music recommendation systems. We also outline inspirations for the future research in the field.

1 Introduction

The underlying purpose of Recommendation Systems (RS) is to help users to find appropriate items (such as books, news, products, web sites, courses, music) from big repositories of data. Manual searching would be very ineffective. The music domain deserves special attention, as listening to short samples of songs to find an appropriate piece of music from thousands of titles is tiresome, frustrating and time-consuming. A great example is Spotify application, which has over 200 million active users monthly [61]. It gives access to more than 40 million tracks with the click of a button. That would give over 220 years of constant listening. Therefore, the process of searching musical content in a big data repository is a crucial concern for many researchers [39, 44].

A comprehensive study of currently existing Music RS (MRS) can be found in [58]. However, the heterogeneity and variety of approaches possible to implement in MRS continuously pose new challenges [69]. In this survey, we try to outline the diversity of these approaches and provide a survey and literature review through the recent advances in MRS and their applications. Since the music preferences can be inferred from a mood, time, activity, personality, place and other environmental fac-

Polish-Japanese Academy of Informtion Technology, Koszykowa 86, 02-008 Warsaw, Poland,
e-mail: mklec@pjwstk.edu.pl, alicja@poljap.edu.pl

tors [79, 9, 71], we decided to pay special attention to context-aware and user-centric MRS.

We performed the literature review by searching the ACM Digital Library, IEEE-explore and ScienceDirect databases. We also searched Google Scholar, as it provides results from a wide variety of other databases. We mainly choose Conference Proceedings and Journals published after 2016. The results of our bibliography search are shown in Table 1.

Table 1 Literature review on Music Recommendation Systems with respect to keywords and search databases; 87% of these papers have been published after 2016.

Keywords	ACM[1]	IEEE[2]	SD[3]	Other[4]	Literature
Review and current challenges	16%	16%	16%	50%	[69, 11, 68, 58, 41, 13]
Sequence modelling for playlists	38%	44%	0%	16%	[69, 11, 68, 13, 44, 45, 75, 76, 71, 36, 83, 20, 27, 29, 84, 74, 60, 35]
Deep learning	29%	47%	5%	17%	[68, 76, 71, 19, 57, 6, 27, 29, 74, 12, 2, 34, 21, 28, 48]
Cold start	40%	10%	10%	40%	[69, 57, 65, 25, 16, 77, 36, 83, 12, 2]
User interface	57%	14%	14%	14%	[16, 53, 30, 5, 37, 8]
Social context	50%	33%	0%	16%	[59, 78, 60, 10, 79, 14]
Personality context	53%	6%	0%	40%	[69, 18, 15, 73, 80, 56, 59, 51, 46, 50, 25, 16, 49]
Emotions context	20%	40%	10%	30%	[69, 18, 6, 78, 60, 79, 21, 28, 48, 5]
Physiological, activity, weather, time, text analysis, and other contexts	29%	41%	11%	17%	[69, 58, 44, 46, 50, 77, 82, 26, 43, 7, 42, 6, 35, 36, 79, 21, 28]

[1] ACM Digital Library
[2] IEEExplore
[3] ScienceDirect
[4] Google Scholar

As one can see, our search included various aspects of research on MRS. We searched for papers on modeling for playlists

2 Music Recommendation Systems

Systems for automatic recommendation of music to users have become significant in the recent years, as we are facing the expansion of gigantic sets of music files. Therefore, music recommender systems have been a topic of interest in the music information retrieval domain, starting approximately from the beginning of the 21th century, see [70] for a survey of early examples.

Music Recommendation Systems perform rating predictions for songs that suit the current music preference. They are closely related to content filtering and searching. An annual event called Music Information Retrieval (MIR) Evaluation eXchange (MIREX[1]) stimulates the development of MIR algorithms. These algorithms combine musicology, psychology, signal processing and machine learning. These areas are also reflected in the current development of MRS. Most of them rely on user-based collaborative filtering (CF), and content-based (CB) approaches.

The main advantage of CF is no need to store features and any meta-data about recommended items. This approach is domain-independent and guarantees effective computation and easy maintenance [64]. However, it suffers some drawbacks. Namely, CF-based systems suffer from popularity bias. They tend to recommend popular music, and the effect of "surprise" is scarce [65]. Moreover, the small number of ratings produces poor predictions. The recommendation might not be even possible for a song that appears as a new item in the system and has not received ratings from others yet. This effect is known as Cold Start (CS) problem [64]. Many researchers try to overcome the CS problems by applying a variety of techniques [57, 36, 77]. However, the most popular approach to partially overcome these limitations is incorporating the CB approach. It compares the song features against a user's profile to calculate recommendations. These systems do not need ratings from others to start to operate. Therefore, the cold-start problem does not exist in this case. The most successful MRS systems in the market (Mufin[2], Pandora[3], Spotify[4], Tidal[5], Qobuz[6]) use hybrid approach, combining CF and CB approaches together. This strategy allows them to take advantage of the strengths of both approaches and eliminate their weaknesses which would occur when used separately.

The most common output from MRS is a playlist, defined as a finite set of songs, often in the form of an ordered list. People use different types of playlists for different activities such as driving, reading, working or working out [3, 82]. The main aim for which users create playlists is not only to satisfy their music expectations about the music itself but also about the order of songs [11]. The modelling of sequences of songs is an essential part of research [45, 27, 20, 29, 84].

The Automatic Playlist Continuation (APC) is a recent trend of research [75, 76, 13]. The goal of APC is to automatically generate a sequence of songs for an existing playlist so that they reflect the target characteristics of the original one. The user can enjoy a continuous and endless music session [60, 83]. The 2018 ACM RecSys Challenge [13] evaluates and advances the current state-of-the-art in APC using a large scale data-set released by Spotify [76, 74].

---

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

[2] https://www.mufin.com

[3] https://www.pandora.com/

[4] http://www.spotify.com

[5] http://tidal.com

[6] http://www.qobuz.com

Recently, the Hidden Markov model (HMM) also turned out to perform very well for sequence prediction [45]. The evaluation on a large-scale real-world data-set in a Kaggle[7] competition shows that the mixture model significantly outperforms traditional methods [45].

The paper [71] proposes a novel MRS, called MusicRecLSTM, which models the changes of music taste over time. They leverage modified LSTM network to learn the embeddings of both the user and the music (an embedding space is a low-dimensional space, in which similar items are close), based on the sequential data and the temporal context. The system also learns whether a user would tend to listen to a newly released piece of music at a specific time. Evaluations over three real-world datasets show that the system can provide high precision, compared to several state-of-the-art recommender systems.

Deep Learning (DL) is increasingly adopted in MRS. Researchers use DL for automatic feature learning from audio signals, or extracting metadata [27], finding latent factors from user-item rating data for CF models [2, 12], and for learning sequential patterns of songs from musical playlists [27, 74]. The latent factors are integrated into content-based and hybrid systems [68, 19, 34]. In [19], the authors propose the algorithm, Tunes Recommendation System (T-RECSYS), that uses CB and CF approaches as input to a deep classification model to produce an accurate recommendation with real-time prediction. They evaluate their approach with the data obtained from the ACM Recsys Challenge.

Convolutional Neural Networks (CNN) have gained so much popularity after successful usage in image classification tasks that the authors of [12] applied a deep CNN to develop the emotion aware music recommendation system. To get correlation between the music and user data two approaches are combined: deep CNN, and weighted feature extraction. For the purpose of classification, latent features are extracted from audio by using deep CNN. Weighted feature extraction approach is used to get correlation between the music and the user data to finally generate user rating. User rating is generated by inverse document frequency and the term frequency weighted feature extraction approach. After that, emotion aware MRS recommends song based on user rating. The Million Song Dataset[8] is used for training in this paper. Based on the results, the system shows better performance than the content similarity based one.

Still, MRS are far from being perfect, due to a magnitude of factors that influence musical tastes and needs. To satisfy the users' musical need, the researchers need to take into account intrinsic, extrinsic, and contextual aspects of the listener. Therefore, the next section focuses on the review of user-centric MRS, taking into account various types of contexts for personalizing the recommendations.

[7] https://www.kaggle.com/

[8] http://millionsongdataset.com/

3 Personalization of Music Recommendation Systems

The perception of music does not solely depend on the individual's listening history, but also on the user's context, such as emotional state [3], current activity [43], health (personal context) [7], location [3], time (physical context) [7], weather [42], or social activity (social context) [78, 10]. Therefore, recommendation strategies should go deeper into the very essence of the listener's needs, preferences, and intentions of listening to music [9]. Furthermore, the musical taste might depend on gender, personality, education, and the life experience of the listener [54]. Therefore, the subjective character of music enforces researchers to go beyond hybridizing CF and CB approaches with DL techniques [50]. The authors of [41] provide a comprehensive study of existing Context-Aware Music Recommendation Systems (CA-MRS). The article [69] highlights several promising newest trends in research, such as music recommendation inspired by the listener's personality and emotions [46, 59].

3.1 Emotions

Emotions play a crucial role in the motivation for listening to music and thus in MRS. Music helps people to improve their emotional regulation, achieve their self-awareness, and express social relatedness [67]. In general, listeners use music to change emotions, or to sustain them [33]. Music can stimulate emotions in people; also, emotions can be identified in musical pieces [32, 62, 23, 22]. Tracking listener's emotions is one of the ways to improve the quality of recommendation [21]. It is usually achieved implicitly, by tracking the context (like keywords) from an extensive collection of documents written by users [26], or extracting the users' texts from social networks [60, 48]. Another approach is to derive emotions from the user's face using the mobile camera [21, 28], or from the signals obtained via wearable physiological sensors [6]. The authors of [21] identify the user's mood with an accuracy of 90.23%. They also classify the emotions perceived in music with 97.69% accuracy. Their system suggests songs to the user by mapping the user's emotions to the emotion type of the song, taking into consideration the preferences of the user.

3.2 Personality

The way people regulate their emotions depends not only on the current situation but also on their personality [81]. Users with different personalities tend to prefer different music [63]. Knowledge about the influence of personality traits on musical taste has been discussed in [63, 38, 51, 66].

Personality represents people's differences in their enduring emotional, interpersonal, experiential, attitudinal, and motivational styles [31]. Most of the papers in this field focus on the Big-Five Factor Model [31, 55], which makes the results possible to interpret, and compare. This personality model defines personality as five factors: Openness to Experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. The discussion of the usability of this model in recommendation systems can be found in [73]. However, the main challenge is to identify the most efficient way of personality acquisition in MRS. In [16] the authors compared three personality acquisition variants and demonstrated that the explicit personality acquisition interface was preferred by most of the experiment participants in terms of satisfaction and ease of use. The most common and relatively short personality questionnaire is the Ten Item Personality Inventory (TIPI) [17]. The TIPI was also used in [46] to explore users' personality traits together with physiological signals recorded by a wearable, namely a wristband. Experiments with four regression algorithms show that personality features contributed significantly to the improvement of RS accuracy, while physiological features contributed less.

Knowledge on the influence of personality traits on musical taste was also exploited to improve the accuracy in the user similarity calculations and was used to address the cold-start problem [18, 25, 24]. In [25] the authors enhanced CF with personality information. They show that this approach can significantly improve the performance of the traditional rating-based CF in terms of the evaluation metrics MAE (mean absolute error) and ROC (receiver operating characteristic curve, sensitivity vs. fall-out graph).

Since people's attitude towards new or diverse experiences varies considerably [72], the degree of diversity can also be personalized [80]. In [49] the authors proposed a system which incorporates the strategy of adjusting diversity of music recommendation, according to the user's personality. The authors of [56] implemented four MRSs and evaluated them on a sample of real users and real-world datasets. The experimental results show that MRSs which rely purely on users' personality information perform comparable or even better than state-of-the-art MRSs in terms of the diversity of recommendation.

3.3 Social Context

Nowadays, social media contain rich information about the user's frequent actions and may constitute a vast data mine for gathering information about similar interests between users. Users reveal much information about their lives during their interactions with others using mobile devices. This is an opportunity for researchers to gather contextual information instantly [60, 78, 10, 59].

In [14], the authors explore the effects of social influence on developing MRS. The experiments verify that this approach can outperform current state-of-the-art music recommendation methods substantially. Personality can also be derived from social

media and utilized for music recommendation [59]. The article [79] utilizes social media context such as posts, comments, and interactions for music recommendation to relax the user's mind, considering the current mood (happy, sad, calm, and angry) extracted from the social media profile of the user. The songs are classified to moods based on the lyrics and audio of the songs.

3.4 New Interfaces

Recent work has unveiled the importance of recommendation explanation and transparency [53, 52, 40], especially regarding user experience (UX) [4, 30, 5, 37, 8]. One of the most exciting trends is the rise of smart assistants from some of the world's largest technology companies, such as Apple's Siri[9], Google Assistant[10] and Amazon Alexa [47]. The companies have integrated voice recognition technology into their products, as a way of interacting with users. Amazon's Alexa is a crucial feature for the company's music[11] and their smart speakers, the Amazon Echo [47]. Amazon Echo is, somewhat controversially, equipped with the functionality enabling this device to listen to contextual information in the user's environment, such as conversations and background music. However, this information is used to infer mood and music preferences which enable the device to recommend songs to the user. Additionally, voice recognition allows users to use a more intuitive way of music search. For example, the user may say that she or he wants to listen to happy funk from the 1980s. The user might not know the best tunes from that era, but the system will recommend her or him the song that suits the voice query and surroundings. With the context of this survey, the authors see an expressive potential of incorporating voice assistants, together with music recommendation, to help people to get therapeutic music in medical care places and users' homes.

3.5 Automatic Music Generation

Throughout history, the vast majority of music we consume was written by another human being. However, recently computers start to compose music that is difficult to distinguish from those composed by humans. The automatically generated music has started to be recommended by Spotify. Just like AiMi[12] generates electronic music for personal purposes and AIVA[13] which generates soundtrack for a podcast

[9] https://www.apple.com/siri/

[10] https://assistant.google.com/

[11] https://www.amazon.com/music/

[12] https://aimi.fm/

[13] https://www.aiva.ai/

of short movies, it is not difficult to imagine a system that generates therapeutic musical compositions with the helpful impact on our health, and adjust the music for ongoing therapy. The example of such systems are ENDEL[14], which provides personalized sound environments to help the user to focus, and Tinnitracks system that was designed to treat tinnitus with music [1].

4 Summary

Music recommendation systems are becoming more and more popular. The environmental factors that influence the perception of music are pervasive and varied. Including all of them in the music recommendation system is still a big challenge and difficult to evaluate. Nevertheless, the increasing amount of personal data left by users on social media, together with the recent development of deep learning techniques, contribute to more accurate inference of the user's musical preferences implicitly. The knowledge on this topic is still being expanded. Environmental factors that induce musical preferences and needs are wider utilized these days. Still, we see a significant potential for this kind of data. We hope that this survey will contribute to the future development of user-centric and context-aware music recommendation systems.

References

1. Tinnitracks: Treat Tinnitus With Your Favorite Music! (06 Jul 2020). URL https://www.tinnitracks.com/en
2. Aljunid, M.F., Dh, M.: An efficient deep learning approach for collaborative filtering recommender system. Procedia Computer Science **171**, 829–836 (2020)
3. Álvarez, P., Zarazaga-Soria, F., Baldassarri, S.: Mobile music recommendations for runners based on location and emotions: The dj-running system. Pervasive and Mobile Computing p. 101242 (2020)
4. Alves, T., Natálio, J., Henriques-Calado, J., Gama, S.: Incorporating personality in user interface design: A review. Personality and Individual Differences **155**, 109709 (2020)
5. Andjelkovic, I., Parra, D., O'Donovan, J.: Moodplay: Interactive music recommendation based on artists' mood similarity. International Journal of Human-Computer Studies **121**, 142–159 (2019)
6. Ayata, D., Yaslan, Y., Kamasak, M.E.: Emotion based music recommendation system using wearable physiological sensors. IEEE transactions on consumer electronics **64**(2), 196–203 (2018)
7. Bai, K., Kawagoe, K.: Background music recommendation system based on user's heart rate and elapsed time. In: Proceedings of the 2018 10th International Conference on Computer and Automation Engineering, pp. 49–52 (2018)

[14] https://www.endel.io/

8. Baig, M.H., Varghese, J.R., Wang, Z.: Musicmapp: A deep learning based solution for music exploration and visual interaction. In: Proceedings of the 26th ACM international conference on Multimedia, pp. 1253–1255 (2018)
9. Bauer, C., Novotny, A.: A consolidated view of context for intelligent systems. Journal of Ambient Intelligence and Smart Environments **9**(4), 377–393 (2017)
10. Bauer, C., Schedl, M.: Cross-country user connections in an online social network for music. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2019)
11. Bonnin, G., Jannach, D.: Automated generation of music playlists: Survey and experiments. ACM Computing Surveys (CSUR) **47**(2), 1–35 (2014)
12. Chang, S.H., Abdul, A., Chen, J., Liao, H.Y.: A personalized music recommendation system using convolutional neural networks approach. In: 2018 IEEE International Conference on Applied System Invention (ICASI), pp. 47–49. IEEE (2018)
13. Chen, C.W., Lamere, P., Schedl, M., Zamani, H.: Recsys challenge 2018: automatic music playlist continuation. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 527–528 (2018)
14. Chen, J., Ying, P., Zou, M.: Improving music recommendation by incorporating social influence. Multimedia Tools and Applications **78**(3), 2667–2687 (2019)
15. Cheng, R., Tang, B.: A music recommendation system based on acoustic features and user personalities. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 203–213. Springer (2016)
16. Dunn, G., Wiersema, J., Ham, J., Aroyo, L.: Evaluating interface variants on personality acquisition for recommender systems. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 259–270. Springer (2009)
17. Ehrhart, M.G., Ehrhart, K.H., Roesch, S.C., Chung-Herrera, B.G., Nadler, K., Bradshaw, K.: Testing the latent factor structure and construct validity of the ten-item personality inventory. Personality and individual Differences **47**(8), 900–905 (2009)
18. Ferwerda, B., Schedl, M.: Enhancing music recommender systems with personality information and emotional states: A proposal. In: Umap workshops (2014)
19. Fessahaye, F., Perez, L., Zhan, T., Zhang, R., Fossier, C., Markarian, R., Chiu, C., Zhan, J., Gewali, L., Oh, P.: T-recsys: A novel music recommendation system using deep learning. In: 2019 IEEE International Conference on Consumer Electronics (ICCE), pp. 1–6. IEEE (2019)
20. Furini, M., Martini, J., Montangero, M.: Automated generation of user-tailored and time-sensitive music playlists. In: 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 1–6. IEEE (2019)
21. Gilda, S., Zafar, H., Soni, C., Waghurdekar, K.: Smart music player integrating facial emotion recognition and music mood recommendation. In: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 154–158. IEEE (2017)
22. Grekow, J.: From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces, *Studies in Computational Intelligence*, vol. 747. Springer (2018)
23. Grekow, J.: Musical performance analysis in terms of emotions it evokes. Journal of Intelligent Information Systems **51**(2), 415–437 (2018)
24. Hu, R., Pu, P.: A study on user perception of personality-based recommender systems. In: International conference on user modeling, adaptation, and personalization, pp. 291–302. Springer (2010)
25. Hu, R., Pu, P.: Enhancing collaborative filtering systems with personality information. In: Proceedings of the fifth ACM conference on Recommender systems, pp. 197–204 (2011)
26. Hyung, Z., Park, J.S., Lee, K.: Utilizing context-relevant keywords extracted from a large collection of user-generated documents for music discovery. Information Processing & Management **53**(5), 1185–1200 (2017)
27. Irene, R.T., Borrelli, C., Zanoni, M., Buccoli, M., Sarti, A.: Automatic playlist generation using convolutional neural networks and recurrent neural networks. In: 2019 27th European Signal Processing Conference (EUSIPCO), pp. 1–5. IEEE (2019)

28. Iyer, A.V., Pasad, V., Sankhe, S.R., Prajapati, K.: Emotion based mood enhancing music recommendation. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1573–1577. IEEE (2017)
29. Jiang, M., Yang, Z., Zhao, C.: What to play next? a rnn-based music recommendation system. In: 2017 51st Asilomar Conference on Signals, Systems, and Computers, pp. 356–358. IEEE (2017)
30. Jin, Y., Tintarev, N., Verbert, K.: Effects of individual traits on diversity-aware music recommender user interfaces. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, pp. 291–299 (2018)
31. John, O.P., Srivastava, S.: The big five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research **2**(1999), 102–138 (1999)
32. Juslin, P.N., Laukka, P.: Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. Journal of new music research **33**(3), 217–238 (2004)
33. Juslin, P.N., Sloboda, J.: Handbook of music and emotion: Theory, research, applications. Oxford University Press (2011)
34. Khoali, M., Tali, A., Laaziz, Y.: Advanced recommendation systems through deep learning. In: Proceedings of the 3rd International Conference on Networking, Information Systems & Security, pp. 1–8 (2020)
35. Kim, H.G., Kim, G.Y., Kim, J.Y.: Music recommendation system using human activity recognition from accelerometer data. IEEE Transactions on Consumer Electronics **65**(3), 349–358 (2019)
36. Kim, J., Won, M., Liem, C.C., Hanjalic, A.: Towards seed-free music playlist generation: Enhancing collaborative filtering with playlist title information. In: Proceedings of the ACM Recommender Systems Challenge 2018, pp. 1–6 (2018)
37. Kittimathaveenan, K., Pongskul, C., Mahatanarat, S.: Music recommendation based on color. In: 2020 6th International Conference on Engineering, Applied Sciences and Technology (ICEAST), pp. 1–4. IEEE (2020)
38. Kleć, M.: The influence of listener personality on music choices. Computer Science **18** (2017)
39. Kostek, B.: Wspomaganie procesu wyszukiwania nagrań w repozytoriach muzycznych. Przegląd Telekomunikacyjny+ Wiadomości Telekomunikacyjne (6), 200–205 (2011)
40. Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., Getoor, L.: Personalized explanations for hybrid recommender systems. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 379–390 (2019)
41. Kulkarni, S., Rodd, S.F.: Context aware recommendation systems: A review of the state of the art techniques. Computer Science Review **37**, 100255 (2020)
42. Lavanya, S., Saranya, G., Navin, K.: Weather based playlist generation in mobile devices using hash map. In: 2017 International Conference on IoT and Application (ICIOT), pp. 1–7. IEEE (2017)
43. Lee, W.P., Chen, C.T., Huang, J.Y., Liang, J.Y.: A smartphone-based activity-aware system for music streaming recommendation. Knowledge-Based Systems **131**, 70–82 (2017)
44. Leung, C.K., Kajal, A., Won, Y., Choi, J.M.: Big data analytics for personalized recommendation systems. In: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pp. 1060–1065. IEEE (2019)
45. Li, T., Choi, M., Fu, K., Lin, L.: Music sequence prediction with mixture hidden markov models. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 6128–6132. IEEE (2019)
46. Liu, R., Hu, X.: A multimodal music recommendation system with listeners' personality and physiological signals. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 357–360 (2020)
47. Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., Martinez, A.: Talk to me: Exploring user interactions with the amazon alexa. Journal of Librarianship and Information Science **51**(4), 984–997 (2019)

48. Lopes, P.S., Lasmar, E.L., Rosa, R.L., Rodríguez, D.Z.: The use of the convolutional neural network as an emotion classifier in a music recommendation system. In: Proceedings of the XIV Brazilian Symposium on Information Systems, pp. 1–8 (2018)
49. Lu, F., Tintarev, N.: A diversity adjusting strategy with personality for music recommendation. In: IntRS@ RecSys, pp. 7–14 (2018)
50. Lytvyn, V., Vysotska, V., Shatskykh, V., Kohut, I., Petruchenko, O., Dzyubyk, L., Bobrivetc, V., Panasyuk, V., Sachenko, S., Komar, M.: Design of a recommendation system based on collaborative filtering and machine learning considering personal needs of the user. Eastern European Journal of Advanced Technologies (4 (2)), 6–28 (2019)
51. Melchiorre, A.B., Schedl, M.: Personality correlates of music audio preferences for modelling music listeners. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 313–317 (2020)
52. Millecamp, M., Htun, N.N., Conati, C., Verbert, K.: To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 397–407 (2019)
53. Millecamp, M., Htun, N.N., Conati, C., Verbert, K.: What's in a user? towards personalising transparency for music recommender interfaces. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 173–182 (2020)
54. North, A.C., Hargreaves, D.J.: Situational influences on reported musical preference. Psychomusicology: A Journal of Research in Music Cognition **15**(1-2), 30 (1996)
55. Nunes, M.A.S.N.: Recommender systems based on personality traits. Ph.D. thesis (2008)
56. Onori, M., Micarelli, A., Sansonetti, G.: A comparative analysis of personality-based music recommender systems. In: Empire@ RecSys, pp. 55–59 (2016)
57. Oramas, S., Nieto, O., Sordo, M., Serra, X.: A deep multimodal approach for cold-start music recommendation. In: Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, pp. 32–37 (2017)
58. Patel, A., Wadhvani, R.: A comparative study of music recommendation systems. In: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), pp. 1–4. IEEE (2018)
59. Paudel, A., Bajracharya, B.R., Ghimire, M., Bhattarai, N., Baral, D.S.: Using personality traits information from social media for music recommendation. In: 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), pp. 116–121. IEEE (2018)
60. Polignano, M., Basile, P., de Gemmis, M., Semeraro, G.: Social tags and emotions as main features for the next song to play in automatic playlist continuation. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 235–239 (2019)
61. Prey, R.: Nothing personal: algorithmic individuation on music streaming platforms. Media, Culture & Society **40**(7), 1086–1100 (2018)
62. Raś, Z.W., Wieczorkowska, A.A. (eds.): Advances in Music Information Retrieval, *Studies in Computational Intelligence*, vol. 274. Springer (2010)
63. Rentfrow, P.J., Gosling, S.D.: The do re mi's of everyday life: the structure and personality correlates of music preferences. Journal of personality and social psychology **84**(6), 1236 (2003)
64. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM **40**(3), 56–58 (1997)
65. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: The adaptive web, pp. 291–324. Springer (2007)
66. Schäfer, T., Mehlhorn, C.: Can personality traits predict musical style preferences? a meta-analysis. Personality and Individual Differences **116**, 265–273 (2017)
67. Schäfer, T., Sedlmeier, P., Städtler, C., Huron, D.: The psychological functions of music listening. Frontiers in psychology **4**, 511 (2013)
68. Schedl, M.: Deep learning in music recommendation systems. Frontiers in Applied Mathematics and Statistics **5**, 44 (2019)
69. Schedl, M., Zamani, H., Chen, C.W., Deldjoo, Y., Elahi, M.: Current challenges and visions in music recommender systems research. International Journal of Multimedia Information Retrieval **7**(2), 95–116 (2018)

70. Song, Y., Dixon, S., Pearce, M.: A survey of music recommendation systems and future perspectives. In: 9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012), pp. 395–410. Queen Mary University of London (2012)
71. Tao, Y., Zhang, Y., Bian, K.: Attentive context-aware music recommendation. In: 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), pp. 54–61. IEEE (2019)
72. Tintarev, N., Dennis, M., Masthoff, J.: Adapting recommendation diversity to openness to experience: A study of human behaviour. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 190–202. Springer (2013)
73. Tkalcic, M., Chen, L.: Personality and recommender systems. In: Recommender systems handbook, pp. 715–739. Springer (2015)
74. Vagliano, I., Galke, L., Mai, F., Scherp, A.: Using adversarial autoencoders for multi-modal automatic playlist continuation. In: Proceedings of the ACM Recommender Systems Challenge 2018, pp. 1–6 (2018)
75. Vall, A., Dorfer, M., Schedl, M., Widmer, G.: A hybrid approach to music playlist continuation based on playlist-song membership. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 1374–1382 (2018)
76. Volkovs, M., Rai, H., Cheng, Z., Wu, G., Lu, Y., Sanner, S.: Two-stage model for automatic playlist continuation at scale. In: Proceedings of the ACM Recommender Systems Challenge 2018, pp. 1–6 (2018)
77. Vystrčilová, M., Peška, L.: Lyrics or audio for music recommendation? In: Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics, pp. 190–194 (2020)
78. Wishwanath, C.H., Ahangama, S.: A personalized music recommendation system based on user moods. In: 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), vol. 250, pp. 1–1. IEEE (2019)
79. Wishwanath, C.H., Weerasinghe, S.N., Illandara, K.H., Kadigamuwa, A., Ahangama, S.: A personalized and context aware music recommendation system. In: International Conference on Human-Computer Interaction, pp. 616–627. Springer (2020)
80. Wu, W., Chen, L., He, L.: Using personality to adjust diversity in recommender systems. In: Proceedings of the 24th ACM conference on hypertext and social media, pp. 225–229 (2013)
81. Xu, L., Wen, X., Shi, J., Li, S., Xiao, Y., Wan, Q., Qian, X.: Effects of individual factors on perceived emotion and felt emotion of music: Based on machine learning methods. Psychology of Music p. 0305735620928422 (2020)
82. Yakura, H., Nakano, T., Goto, M.: Focusmusicrecommender: a system for recommending music to listen to while working. In: 23rd International Conference on Intelligent User Interfaces, pp. 7–17 (2018)
83. Yang, H., Jeong, Y., Choi, M., Lee, J.: Mmcf: Multimodal collaborative filtering for automatic playlist continuation. In: Proceedings of the ACM Recommender Systems Challenge 2018, pp. 1–6 (2018)
84. Zhang, K., Zhang, Z., Bian, K., Xu, J., Gao, J.: A personalized next-song recommendation system using community detection and markov model. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), pp. 118–123. IEEE (2017)