AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Wydział Fizyki i Informatyki Stosowanej

Praca inżynierska

Paweł Pęksa

kierunek studiów: informatyka stosowana

Wykorzystanie algorytmów uczenia maszynowego w celu rozpoznawania nastroju muzyki

Opiekun: dr hab. Marcin Wolter

Kraków, styczeń 2016

Oświadczam, świadomy odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie i nie korzystałem ze źródeł innych niż wymienione w pracy.

.....
(czytelny podpis)

Merytoryczna ocena pracy przez opiekuna:

Ocena: ..... Podpis: .....

Data: .....

Merytoryczna ocena pracy przez recenzenta:

Ocena: ..... Podpis: .....

Data: .....

Składam serdeczne podziękowania
Panu dr hab. Marcinowi Wolterowi
za udzieloną pomoc i poświęcony czas.

Spis treści

1 Wprowadzenie

1.1 Przedmiot pracy

Muzyka towarzyszyła człowiekowi od czasów prehistorycznych. Z czasem stała się jedną z form sztuki. Niewątpliwie, gdy słyszymy jakaś melodię, nie sprawia nam większego kłopotu, aby określić emocje z nią związane. Należy jednak mieć na uwadze źródło emocji. Rozróżnić możemy emocje wyrażane przez muzykę oraz przez nią indukowane. Tematem niniejszej pracy jest rozpoznawanie nastroju muzyki przy użyciu uczenia maszynowego. Przez określenie „nastrój muzyki" mamy tutaj na myśli emocje, które ta muzyka reprezentuje. Konkretnym narzędziem wybranym w celu klasyfikacji muzyki jest sztuczna sieć neuronowa. Idea klasyfikacji utworów muzycznych pod tym kątem jest względnie nowa, lecz można zauważyć wzrastające zainteresowanie tym tematem[1]. Do tej pory powstał szereg różnych prac podejmujących to zadanie[2][3][4]. Niektóre korzystają nie tylko z samego sygnału audio, ale także np. z tekstu utworu[5]. W tej pracy jednak pod uwagę brany jest jedynie sygnał audio z którego wyekstrahowano odpowiednie cechy, które mogłyby pozwolić sieci neuronowej rozpoznawać emocje reprezentowane przez dany utwór muzyczny.

W kolejnych rozdziałach zostały opisane niezbędne podstawy teoretyczne, których przyswojenie pozwala zrozumieć w jaki sposób postawione zadanie jest realizowane. W rozdziale 2 znajduje się krótki opis sieci neuronowych, w rozdziale 3 czytelnik dowie się o tym jakie konkretnie atrybuty muzyki były rozważane, natomiast w rozdziale 4 opisany jest sposób w jaki matematyka pomaga nam w opisaniu emocji. Opis stworzonego systemu można znaleźć w rozdziale 5. Wyniki oraz wnioski wynikające z podjętej próby podejścia do opisywanego zagadnienia zostały przedstawione w rozdziale kolejno 6 oraz 7.

1.2 Problematyka pracy

Analiza muzyki pod kątem emocji jest zadaniem, które nie wiąże się tylko z przetwarzaniem sygnałów oraz uczeniem maszynowym, ale także z psychologią muzyki oraz jej teorią. Jest to bardzo wymagający problem, gdyż nastrój utworu muzycznego może być wysoce subiektywnym odczuciem. Wpływ na ocenę mogą mieć także wspomnienia danej osoby, nastrój w danej chwili, indywidualne preferencje czy poziom wykształcenia muzycznego. Wszystkie wspomniane jednak kwestie nie są na tyle znaczące, aby podjęcie pracy nad tym tematem było niemożliwe czy całkowicie nieskuteczne. Oczywiście nie da stworzyć się systemu, który będzie działał niezawodnie, bo tak jak zostało to wspomniane, zbyt wiele indywidualnych czynników ma wpływ na percepcje człowieka. Dotychczasowe badania udowadniają, że jest możliwe sprostanie temu zadaniu w zadowalającym stopniu[1] i taka próba zostaje podjęta w niniejszej pracy.

2 Sztuczne sieci neuronowe

Sztuczną siecią neuronową, dalej zwaną po prostu siecią neuronową, określamy model matematyczny służący jako system przetwarzania informacji. Źródłem inspiracji dla tegoż modelu była biologia, a mianowicie sieci neuronowe istniejące w ludzkim mózgu. Dużą zaletą sieci neuronowych jest fakt, iż nie działają jak tradycyjne algorytmy, lecz mają one zdolność do uczenia się. Jest to kolejna analogia związana z pracą ludzkiego mózgu, co powodowało wzrastające zainteresowanie tym tematem. Pomimo tego, że nie udało się z ich pomocą odtworzyć pracy tego niewątpliwie niesamowitego narządu, znajdują one zastosowanie w wielu dziedzinach nauki. Najbardziej podstawowym rodzajem sieci neuronowej jest sieć jednokierunkowa tj. czyli taka, w której nie występują sprzężenia zwrotne i takie też sieci są wykorzystane w tej pracy oraz opisane w tym rozdziale.

2.1 Budowa sieci neuronowej

Szczegółowy opis zagadnień związanych z budową sieci neuronowej można znaleźć w pracy [6].

2.1.1 Budowa neuronu

Sieć neuronowa zbudowana jest z neuronów, które są odpowiednikami komórek nerwowych. Synapsy łączące poszczególne komórki modelowane są przez wagi liczbowe, których wielkości z kolei można interpretować jako wpływ jednej komórki na drugą. Matematyczny model neuronu użyty w sieciach neuronowych jest przedstawiony na rysunku 1. Składa się on z $n + 1$ wejść oraz jednego wyjścia. Dodatkowym wejściem neuronu jest tzw. bias, który przyjmuje stałą wartość, ale jest także modyfikowany w procesie uczenia (podobnie jak pozostałe wagi). Najczęściej przyjmuje się, że przed rozpoczęciem uczenia jego wartość równa jest 1. Zależność pomiędzy sygnałami wejściowymi x_i, wagami w_i, a tzw. sygnałem sumarycznego pobudzenia φ, w najprostszym przypadku może być określana przez wzór:

$$\varphi = w_0 + \sum_{i=1}^{n} w_i x_i. \tag{1}$$

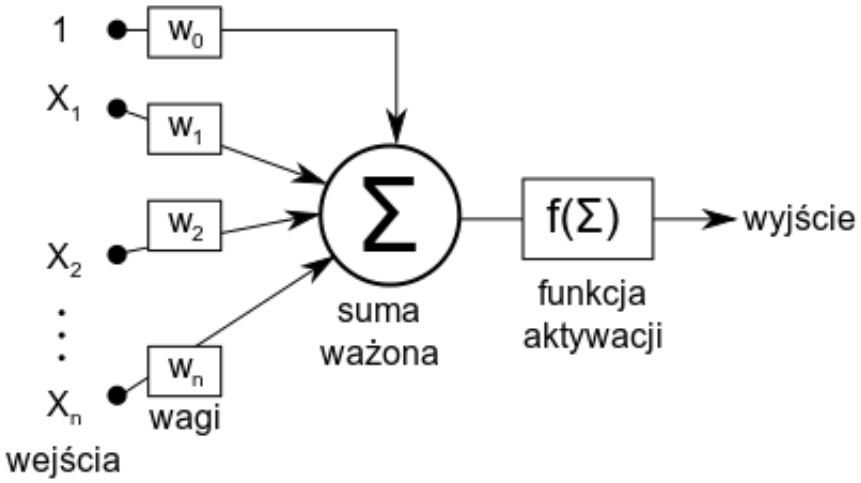
Sposób obliczenia sygnału wyjściowego neuronu określany jest przez funkcję aktywacji:

$$y = f(\varphi). \tag{2}$$

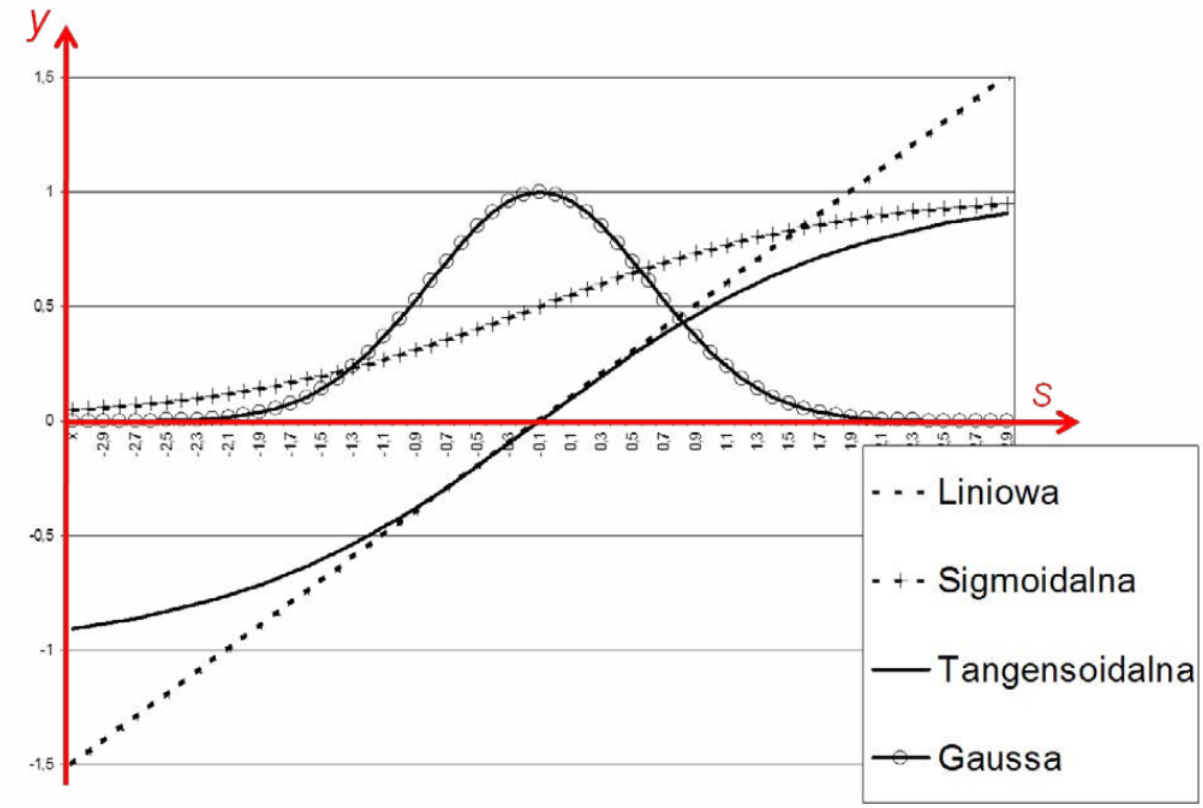
Powszechnie używa się czterech funkcji aktywacji: funkcji liniowej, funkcji sigmoidalnej, funkcji tangens hiperboliczny oraz funkcji Gaussa[7]. Wzory poszczególnych funkcji znajdują się w tabeli 1, natomiast na rysunku 2 przedstawione zostały ich wykresy.

funkcja	wzór
liniowa	$y = x$
sigmoidalna	$y = \frac{1}{1+e^{-x}}$
tangensoidalna	$y = \frac{1-e^{-2x}}{1+e^{-2x}}$
Gaussa	e^{-x^2}

Tabela 1: Funkcje aktywacji wraz z ich wzorami



Rysunek 1: Matematyczny model neuronu[1]



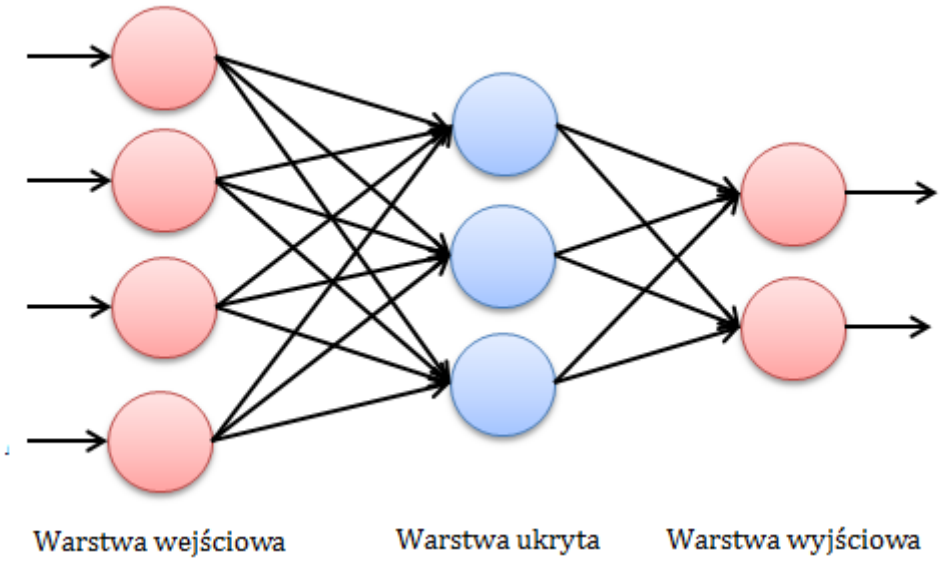
Rysunek 2: Najczęściej używane funkcje aktywacji[7]

2.1.2 Topologia sieci

Wszystkie neurony zgrupowane są w warstwy z których możemy wyróżnić:

- warstwę wejściową
- jedną lub więcej warstw ukrytych
- warstwę wyjściową

W każdej z warstw znajduje się dowolna liczba neuronów, który posiada połączenie do wszystkich neuronów znajdujących się w warstwie kolejnej. Przykład sieci neuronowej składającej się z dokładnie trzech warstw zawierającej różną ilość neuronów w każdej z nich został przedstawiony na rysunku 3.



Rysunek 3: Przykładowa jednokierunkowa sieć neuronowa[2]

2.2 Uczenie sieci neuronowej

Wyróżniamy dwa podstawowe sposoby uczenia sieci:

- uczenie z nauczycielem (uczenie nadzorowane) - gdy dysponujemy danym treningowymi
- uczenie bez nauczyciela (uczenie nienadzorowane) - gdy nie dysponujemy danymi treningowymi, a sieć neuronowa klasyfikuje dane znajdując podobieństwa pomiędzy przypadkami

[1] http://pl.wikipedia.org/wiki/Plik:Neuron_McCullocha-Pittsa.png

[2] https://2ml4pa.bn1303.livefilestore.com/y2p6no6Dn0weHW3FG9tceTUS9lohx5ldcxvFZRhKdbeFQi2kntad_77gKeKIC-INcsFRCvGI-_DY9lMdZzaX8jkSHDvqlcT3qRnftpAt7esi4s/1.PNG?psid=1

W niniejszej pracy zastosowane zostało uczenie sieci z nauczycielem. Dzięki możliwości uczenia sieci neuronowej nie jest konieczne projektowanie algorytmu, który przetwarza dla nas informacje w oczekiwany sposób. Sieć neuronowa korzystając z odpowiedniego algorytmu uczenia sama modeluje ten algorytm poprzez modyfikację wag. Należy też wspomnieć, że początkowe wagi sieci neuronowej zwykle inicjalizowane są wartościami losowymi.

2.2.1 Reguła delta

Proces uczenia polega na modyfikowaniu współczynników wagowych sieci neuronowej. Opisane w tym podrozdziale zostanie uczenie z nauczycielem, ponieważ takie właśnie zostało wykorzystane w tej pracy. W przypadku uczenia z nauczycielem potrzebujmy zbioru uczącego składającego się z wektora danych, które podajemy na wejście sieci i oczekiwanego rezultatu dla tego przypadku, co możemy oznaczyć jako pary (y_i, z_i), gdzie z_i jest oczekiwaną odpowiedzią dla sygnału wejściowego x_i. Zadaniem sieci jest modelowanie funkcji:

$$h(x) = z. \tag{3}$$

Uczenie jest procesem iteracyjnym, gdzie w każdej iteracji modyfikujemy wagi sieci. Liczba iteracji N równa jest liczbie par (x_i, z_i). W każdym kroku j procesu uczenia możemy zdefiniować wielkość błędu neuronu wyjściowego jako:

$$\delta_i^j = |z_i^j - y_i^j|, \tag{4}$$

gdzie y_i jest odpowiedzią sieci neuronowej dla sygnału x_i. Proces uczenia jest realizowany poprzez minimalizację funkcji:

$$Q = \frac{1}{2}\sum_{j=1}^{N}(\delta_i)^2 \tag{5}$$

będącej miarą dopasowania funkcji metodą najmniejszych kwadratów. Korzystając z metody gradientowej możemy zdefiniować poprawkę Δw dla wagi w neuronu i jako:

$$\Delta w_i = -\eta\frac{\partial Q}{\partial w_i}, \tag{6}$$

oraz, idąc dalej, zdefiniować wzór korygujący wagi w w kolejnych krokach:

$$w_i^{j+1} = w_i^j + \Delta w_i \tag{7}$$

przy czym η jest dodatkowym współczynnikiem liczbowym, który decyduje o szybkości uczenia. W ten sposób poprawiamy wagi sieci j razy. Problemem z którym borykano się do połowy lat 80-tych jest fakt, iż tym sposobem niemożliwe jest uczenie sieci, która składa się z więcej niż jednej warstwy, ponieważ nieznana jest oczekiwana odpowiedź neuronów warstwy innej niż wyjściowej. Wzór 6 jest jednak podstawą większości algorytmów automatycznego uczenia[6] i jest on określany w literaturze regułą delta[8].

2.2.2 Algorytm wstecznej propagacji błędów

W celu przeprowadzenia uczenia dla sieci wielowarstwowej spotykamy się z potrzebą określenia błędu δ dla neuronów, które należą do warstw ukrytych sieci neuronowej. Umożliwia nam to algorytm wstecznej propagacji błędów. Błąd dla takiego neuronu obliczamy korzystając ze wszystkich błędów neuronów do których wysłał on swój sygnał. Uwzględniane są także wagi połączeń. Mowa jest o wstecznej propagacji, ponieważ odbywa się ona przeciwnie do przepływów sygnałów w sieci. Błąd dla neuronów znajdujących się w warstwie innej niż wejściowa możemy określić jako:

$$\delta_i = f'(\varphi) \sum_{i}^{N} w_k \delta_k, \tag{8}$$

przy czym w_k oraz δ_k są kolejno wagami oraz błędami neuronów do których analizowany neuron wysyłał swój sygnał.

3 Ekstrakcja cech dźwiękowych

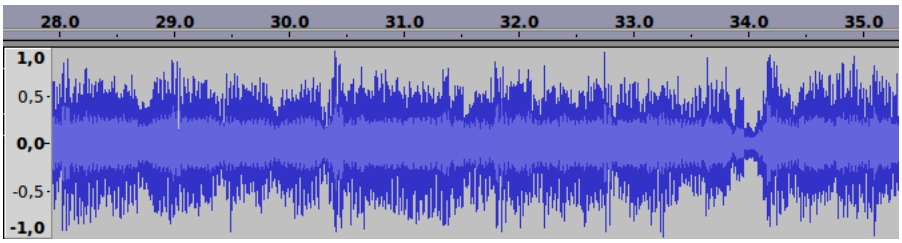
Cechami dźwiękowymi nazywamy zmienne wyekstrahowane z sygnału audio, które opisują ten sygnał i pozwalają uzyskać dodatkowe informacje na jego temat[9]. Opisane w tym rozdziale zostały cechy sygnału audio, które są wejściami dla użytego klasyfikatora tj. sieci neuronowej. Poszczególne cechy bazowały zarówno na czasowej jak i spektralnej reprezentacji sygnału.

3.1 Cyfrowa reprezentacja sygnału audio

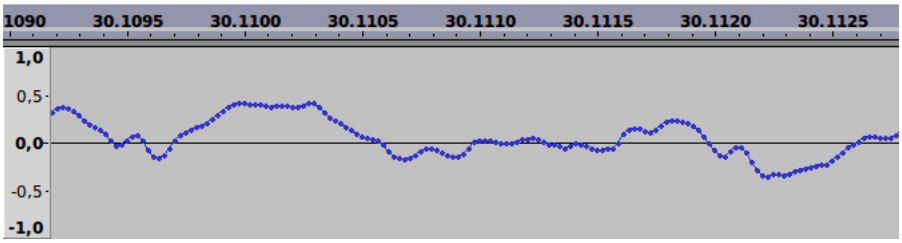
Dźwięk jest sygnałem analogowym. W celu przechowywania go na cyfrowych nośnikach spotkano się z potrzebą jego digitalizacji tj. reprezentacji w postaci cyfrowej. Najczęściej stosowaną w tym celu jest metoda PCM[3] w której to sygnał analogowy jest próbkowany w równych odstępach czasu i zapisywany cyfrowo. Powszechnie stosowana częstotliwość próbkowania wynosi 44 100 Hz ze względu na zakres częstotliwości słyszanych przez człowieka, który wynosi około 20 000 Hz, a zgodnie z prawem Nyquista sygnał powinien być próbkowany z dwa razy wyższą częstotliwością niż maksymalna częstotliwość sygnału w celu uzyskania dokładnej reprezentacji bez zniekształceń[10]. Przykładowy, reprezentowany cyfrowo, sygnał audio przedstawia rysunek 4 oraz 5. Na rysunku 5 przedstawiony został ten sam sygnał, który widzimy na rysunku 4, lecz w bardzo dużym przybliżeniu. Oba rysunki zostały wygenerowane za pomocą programu Audacity[11] z pliku audio w formacie mp3[4].

[3]PCM - Pulse Code Modulation

[4]Format cyfrowego zapisu i kompresji plików dźwiękowych



Rysunek 4: Przykładowy sygnał audio



Rysunek 5: Przykładowy sygnał audio - przybliżenie

3.2 Spektralna reprezentacja sygnału audio

3.2.1 Transformacja Fouriera

Każdy sygnał, który jest reprezentowany jako zmieniająca się w czasie amplituda posiada też odpowiadające spektrum częstotliwościowe tzw. widmo. Dotyczy to także sygnału audio. Dzięki przedstawieniu sygnału dźwiękowego w ten sposób możliwe jest uzyskanie dodatkowych informacji na jego temat[12]. Spektrum przedstawia skład częstotliwościowy dźwięku. Przy obliczaniu spektrum z pomocą przychodzi transformacja Fouriera. Oznaczając $x(t)$ jako sygnał, a $X(f)$ jako wynik transformacji, możemy zapisać:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt, \tag{9}$$

przy czym f to częstotliwość, natomiast t oznacza czas. W celu otrzymania spektrum amplitudowego, z którego szeroko korzystano w niniejszej pracy, należy z otrzymanego wyniku obliczyć wartość bezwzględną:

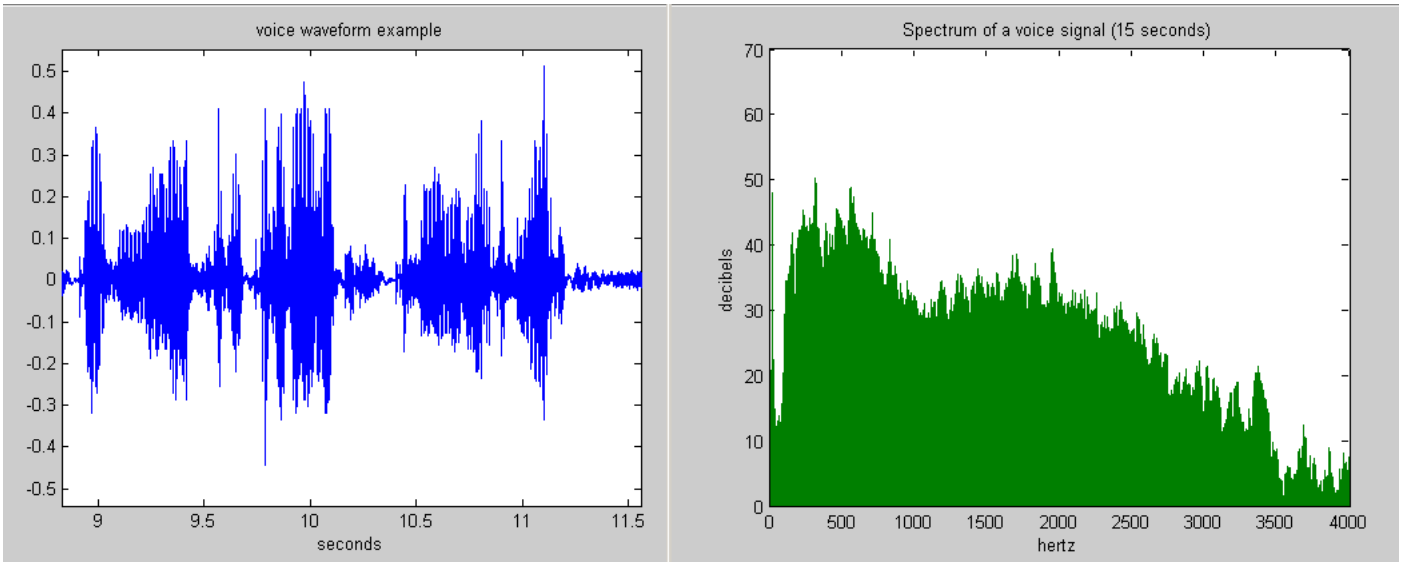
$$|X(f)|. \tag{10}$$

Trzeba jednak pamiętać, że w przypadku sygnału audio zapisanego w pamięci komputera mamy do czynienia z sygnałem dyskretnym, więc należy użyć DFT[5] - transformaty Fouriera dla sygnałów dyskretnych wyrażającej się wzorem:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi \frac{k}{N} n}, \tag{11}$$

przy czym x_n to kolejne wartości próbkowanego sygnału, X_k to wartości transformaty. Na rysunku 6 przedstawiony jest przykładowy sygnał wraz z odpowiadającym mu spektrum amplitudowym.

[5]Discrete Fourier Transform



Rysunek 6: Sygnał audio wraz z odpowiadającym mu spektrum amplitudowym [6]

3.3 Wstępna obróbka sygnału

Przed analizą sygnału dźwiękowego, jakim są utwory muzyczne, w celu poprawy jakości danych przeprowadza się obróbkę wstępną, co pozwala na bardziej efektywną jego analizę. Przed ekstrakcją cech dźwiękowych wykorzystana została funkcja okna czasowego oraz algorytm wyrównywania poziomu głośności dźwięku opisane w kolejnych dwóch podrozdziałach.

3.3.1 Okno czasowe

Ważną rolę przy korzystaniu z transformaty Fouriera odgrywa okresowość. W przypadku, gdy mamy do czynienia z danymi, które występują niecałkowitą ilość razy, na końcach analizowanych danych występują nieciągłości, które powodują, że otrzymane spektrum jest zniekształcone. Rozwiązaniem tego problemu jest okno czasowe. Jest to funkcja, którą mnożymy przy sygnał w celu zmniejszenia zniekształceń[12].

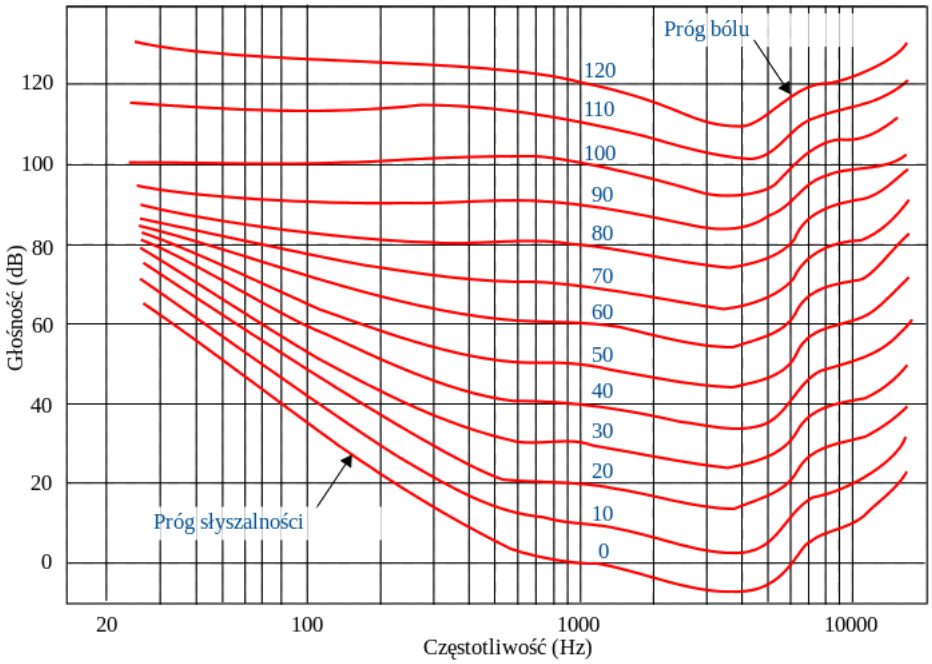
3.3.2 Algorytm wyrównywania poziomu głośności dźwięku

Ludzkie ucho nie słyszy dźwięków o wszystkich częstotliwościach jako dźwięków o tym samym poziomie głośności. Problem ten można rozwiązać częściowo przy użyciu algorytmu wyrównywania poziomu głośności, który filtruje dźwięk korzystając z krzywych izofonicznych. Pojecie izofony lub też krzywej izofonicznej określa w fonach[7] słyszalne natężenie dźwięku w zależności od częstotliwości[13]. Ze względu na subiektywność postrzegania głośności nie istnieją ściśle określone krzywe izofoniczne. Istnieje jednak model zaproponowany przez Fletchera oraz Munsona[14], który jest przedstawiony na rysunku 7.

[6] https://commons.wikimedia.org/wiki/File:Voice_waveform_and_spectrum.png

[7] Fon - jednostka poziomu głośności

[9] https://commons.wikimedia.org/wiki/File:FletcherMunson_ELC.svg



Rysunek 7: Izofony „normalnego" ucha według Fletchera i Munsona[9], wartości fonów dla poszczególnych krzywych określone są liczbami w kolorze niebieskim

3.4 Cechy dźwięku bazujące na czasowej reprezentacji dźwięku

W niniejszym rozdziale zostały krótko opisane cechy dźwięku, które zostały rozważone w pracy. W kolejnych podrozdziałach oznaczono f_i jako kolejne częstotliwości widma oraz a_i jako odpowiadające im amplitudy.

3.4.1 Wskaźnik zmiany znaku (*Zero Crossing Rate*)

Wskaźnik zmiany znaku (*Zero Crossing Rate*) jest jedną z najprostszych cech dźwiękowych obliczanych z wykorzystaniem reprezentacji dźwięku jako zmiana amplitudy w czasie. Wyraża on liczbę zmiany znaków w fali dźwiękowej w jednostce czasu. Możemy określić go wzorem:

$$ZCR = \frac{1}{2}\sum_{n=1}^{N} |sgn(x[n]) - sgn(x[n-1])|. \tag{12}$$

Jest to deskryptor często stosowany w pozyskiwaniu informacji z muzyki, ale także stosowany w rozpoznawaniu mowy. Zawdzięcza to łatwości jego obliczania, a także faktowi, że przechowuje informację o szumach występujących w dźwięku[9].

3.4.2 Wskaźnik zmian (*Onset rate*)

Wskaźnik zmian (*Onset rate*) jest podstawowym wskaźnikiem rytmu utworu muzycznego mającym duży wpływ na postrzeganie emocji reprezentowanych przez muzykę, ponieważ mówi o zmienności dźwięku. Jest on określany jako liczba ekstremów obwiedni dźwięku[10]. Zakłada

[10]Krzywa opisująca zmianę amplitudy sygnału[15]

się, że różnica czasowa pomiędzy dwoma zmianami branym pod uwagę w zliczaniu musi wynosić przynajmniej 60 ms[9].

3.5 Cechy dźwięku bazujące na spektralnej reprezentacji dźwięku

3.5.1 Złożoność spektralna (*Spectral complexity*)

Złożoność spektralna (*spectral complexity*) jest liczbą ekstremów w widmie amplitudowym sygnału dźwiękowego. Opisuje ona złożoność tego widma. Utwory z większą średnią złożonością spektralną charakteryzują się większą energicznością[9].

3.5.2 Kształt spektralny (*Spectral shape*)

Kształt spektralny jest kształtem widma amplitudowego danego sygnału dźwiękowego. W jego skład wchodzą m.in. środek masy widma sygnału (*spectral centroid*), współczynnik skośności widma sygnału (*spectral skewness*), kurtoza widma sygnału (*spectral kurtosis*), tzw. *spectral roll-off* oraz rozrzut spektralny (*spectral spread*). Wszystkie te cechy mają wpływ na odbiór utworu muzycznego przez słuchacza pod kątem reprezentowanych przez utwór emocji[9].

Moment centralny

W celu określenia kolejnych wzorów mówiących o kształcie spektralnym użyteczne jest zdefiniowanie momentu centralnego. Dla zmiennej dyskretnej moment centralny jest przedstawia się wzorem:

$$\mu = \sum_{i=1}^{N} \frac{(f_i - \overline{f})^r}{N}, \tag{13}$$

gdzie f_i są kolejnymi częstotliwościami występującymi w widmie, $\overline{f}$ średnią częstotliwością, natomiast N to liczba wszystkich częstotliwości. Moment centralny rzędu drugiego obliczamy korzystając ze wzoru 13:

$$\sigma^2 = \sum_{i=1}^{N} \frac{(f_i - \overline{f})^2}{N} \tag{14}$$

i nazywany wariancją. Pierwiastek kwadratowy z wariancji σ określany jest mianem odchylenia standardowego.

Środek masy widma

Środek masy widma wyrażamy wzorem:

$$SC = \frac{\sum f_i a_i}{a_i}, \tag{15}$$

gdzie f_i jest częstotliwością, natomiast a_i amplitudą dla poszczególnych częstotliwości[9].

Współczynnik skośności widma

Współczynnik skośności mówi o asymetryczności widma. W przypadku, gdy jest on mniejszy od zera, więcej danych znajduje się po lewej stronie widma, w przypadku, gdy jest on większy od zera, więcej danych znajduje się po prawej stronie widma. Określa się go wzorem:

$$\gamma = \frac{\mu_3}{\sigma^3} \tag{16}$$

Kurtoza widma

Kurtoza widma jest miarą jego spłaszczenia, wyraża się wzorem:

$$K = \frac{\mu_4}{\sigma^4} \tag{17}$$

***Roll-off* widma**

Cechą dźwięku określana mianem *Roll-off*'u widma jest częstotliwość, która dzieli widmo sygnału na dwie części według ustalonego progu T, który zazwyczaj wynosi 0.95[9]. Omawiana cecha jest zdefiniowana wzorem[16]:

$$\sum_{i=1}^{R_t} f_i = T \sum_{i=1}^{N} f_i. \tag{18}$$

Rozrzut spektralny

Rozrzut spektralny jest miarą mówiącą o szerokości widma sygnału, który wyraża się wzorem:

$$spectralSpread = \frac{\sum_i (f_i - SC)^2 a_i}{\sum_i a_i} \tag{19}$$

Utwory muzyczne o większym rozrzucie spektralnym charakteryzują się większą energicznością.

3.5.3 Płaskość spektralna *(Spectral flatness)*

Płaskość spektralna jest stosunkiem średniej arytmetycznej do średniej geometrycznej widma amplitudowego wyrażonym w decybelach:

$$spectralFlatness = 10 log_{10} \frac{G}{A} \tag{20}$$

przy czym G jest średnią geometryczną:

$$G = \sqrt[n]{\prod_{i=1}^{N} a_i} \tag{21}$$

oraz A jest średnią arytmetyczną:

$$A = \frac{\sum_{i=1}^{N} a_i}{N}. \tag{22}$$

Cecha ta określa jak bardzo spłaszczony jest wykres widma amplitudowego. Wraz ze wzrostem tego wskaźnika dźwięk bardziej przypomina szum. Wartość bliska 1 świadczy o występowaniu białego szumu[11][9].

[11] Szum akustyczny o prawie płaskim widmie

3.5.4 Dysonans (*Dissonance*)

Dysonans jest deskryptorem dźwięku obliczanym na podstawie odstępów pomiędzy ekstremami widma amplitudowego. W przypadku utworów muzycznych cechujących się mniejszym rozdźwiękiem obserwuje się większą równomierność tychże odstępów. Matematycznie dysonans można określić wzorem:

$$dissonance = \frac{1}{H}\sum_{h=1}^{H} a(h) - SE(h), \tag{23}$$

gdzie H jest liczbą ekstremów, $a(h)$ amplitudą dla danego ekstremum oraz $SE(h)$ amplitudą obwiedni spektrum dla częstotliwości $f(h)$[9].

3.5.5 Skala

Skala muzyczna składa się z dźwięków o różnych częstotliwościach ułożonych według ustalonego schematu. Możemy wyróżnić dwie podstawowe skale: molową oraz durową. Powszechnie uznaje się, że utwory muzyczne bazujące na skali durowej mają radosne brzmienie, natomiast na skali molowej smutne brzmienie[17]. W celu wyekstrahowania skali muzycznej utworu należy najpierw obliczyć jego HPCP[12], który obliczany jest na podstawie ekstremów widma amplitudowego według wzoru:

$$HPCP(n) = \sum_{i=1}^{N} w(n, f_i)a_i^2, \tag{24}$$

gdzie a_i oraz f_i są kolejno amplitudą oraz częstotliwością ekstremum, N jest liczbą wszystkich ekstremów, n kolejną wartością wektora HPCP, natomiast w funkcją wagową określającą w jaki sposób poszczególne ekstremum wpływa na wartość n wartości wektora HPCP. W celu określenia skali muzycznej obliczana jest korelacja pomiędzy wektorem HPCP, a odpowiednimi profilami dla obu skali[18].

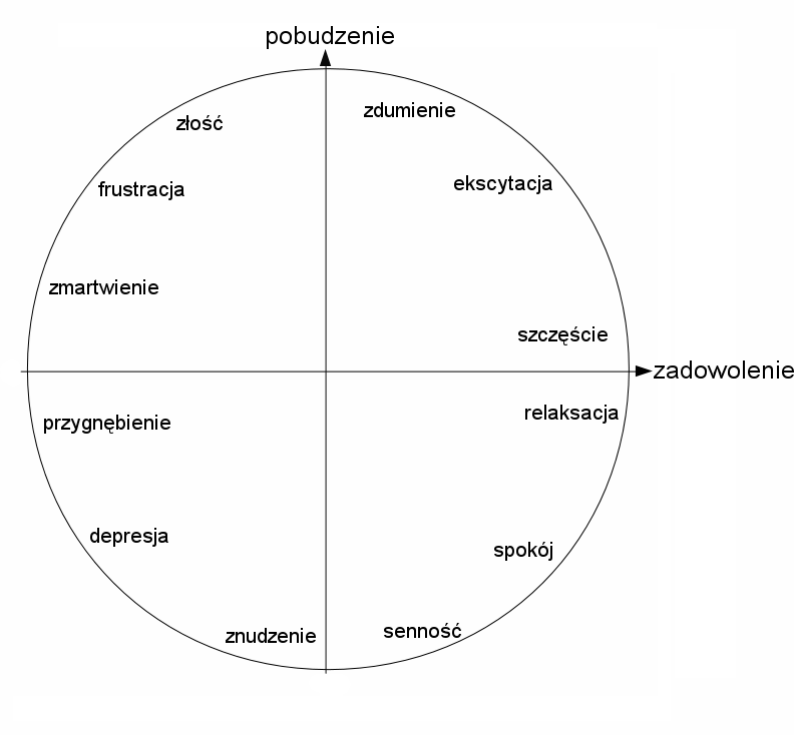
4 Matematyczny model emocji

W celu realizacji postawionego zadania należy zdefiniować emocje matematycznie. Jednym z podejść stosowanych w pracach o podobnej tematyce jest wykorzystanie etykiet mówiących o emocjach. Każdy z utworów etykietowano z wykorzystaniem przymiotników takich jak np. „smutny", „wesoły", „relaksujący". Innym rozwiązaniem jest określanie emocji z wykorzystaniem dwuwymiarowej przestrzeni opracowanej przez Russell'a[19] i podobnie postąpiono w niniejszej pracy. W tym przypadku, emocje określona są przy pomocy dwóch parametrów: pobudzenia(ang. *arousal*) oraz zadowolenia(ang. *valence*), co przedstawia rysunek 8. Model ten zakłada, że wszystkie emocje wynikają z dwóch niezależnych systemów neurofizjologicznych[13] w kontekście do poprzednich teorii zakładających istnienie niezależnych systemów dla każdej

[12]Harmonic Pitch Class Profile

[13]system neurofizjologiczny - system układu nerwowego

podstawowej emocji. Wyniki najnowszych badań są jednak bardziej konsystentne z nowszym podejściem[19].



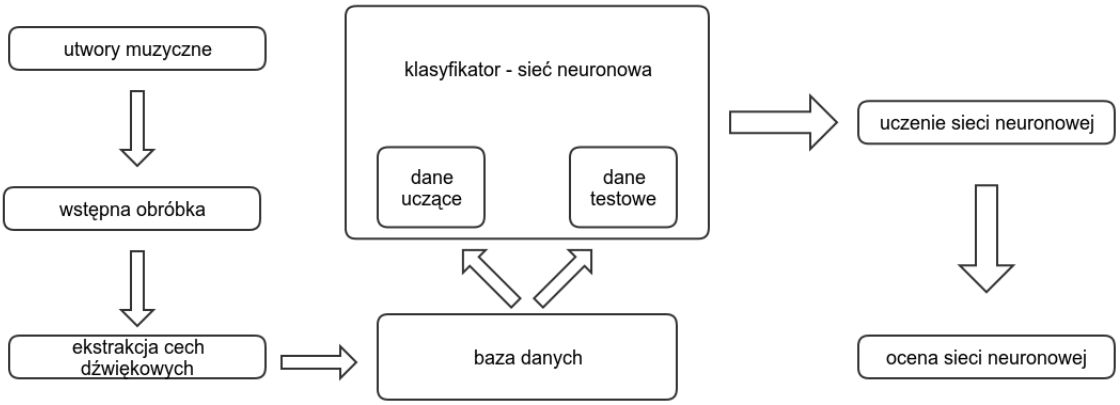
Rysunek 8: Model emocji według Russela

5 Opis stworzonego systemu

W niniejszym rozdziale opisana jest realizacja systemu stworzonego przez autora pracy. W kolejnych podrozdziałach znajdują się opisy poszczególnych jego elementów oraz użytych narzędzi.

5.1 Schemat systemu

Na rysunku 9 przedstawiony został schemat stworzonego systemu, który posłużył do realizacji klasyfikacji utworów muzycznych. Pokazuje on ogólny zamysł zaproponowany przez autora w celu rozwiązania problemu rozpoznawania nastroju muzyki. Pierwszym krokiem, który należy podjąć jest wstępna obróbka utworów muzycznych, a następnie wyekstrahowanie ich cech dźwięków. Na podstawie takie bazy danych należy podzielić dane na zbiór uczący oraz testowy, które będą służyć kolejno do uczenia sieci neuronowej, a następnie do jej oceny przez klasyfikator, którym jest sieć neuronowa.



Rysunek 9: Schemat systemu

5.2 Zastosowane narzędzia

W celu zbudowania systemu rozpoznającego nastrój muzyki należało wybrać język programowania oraz biblioteki umożliwiające analizę utworów muzycznych, ekstrakcję potrzebnych cech, a także implementujące sztuczne sieci neuronowe. Ponadto, niezbędny był wystarczająco duży zbiór utworów muzycznych, który mógłby posłużyć jako baza danych systemu.

5.2.1 Język programowania oraz biblioteki programistyczne

Język programowania jest językiem w którym człowiek może komunikować się z komputerem i przez to „kazać" mu wykonać określone zadanie. Niewątpliwie dobrze jest posiadać swobodę w porozumiewaniu się tym językiem i ten argument spowodował, że wybór padł na język programowania Python, który jest językiem wysokopoziomowym ogólnego przeznaczenia, który zdobywa coraz większą popularność. Jest to spowodowane jego prostą oraz dużymi możliwościami, co w połączeniu z dostępnymi bibliotekami naukowymi (matplotlib, numpy) powoduje, że staje się on dużą konkurencją dla popularnego Matlaba. W parze z językiem programowania idą także biblioteki programistyczne. W niniejszej pracy największą rolę odegrały dwie z nich tj. biblioteka Essentia[20] oraz NeuroLab. Pierwsza z nich została napisana w języku C++, ale oferuje także tzw. bindingi, które pozwalala na jej użycie także w języku Python. Posiada ona szeroką gamę algorytmów wykorzystywanych w analizie dźwięków i w tym celu została użyta w tej pracy,a także wielu innych związanych z MIR[14]. Druga natomiast, jest prostą i potężną biblioteką implementującą sieci neuronowe. Obie znacznie ułatwiły pracę nad problemem podjętym w niniejszej pracy.

5.2.2 System operacyjny

Systemem operacyjnym użytym przy realizacji założonego zadania był system Linux, dokładnie dystrybucja Xubuntu. Przenośność języka programowania jakim jest Python pozwala

[14]MIR - Music Information Retrieval

jednak na używanie aplikacji także na ich systemach pod warunkiem zainstalowania koniecznych do jego działania bibliotek jakimi są wspomniane biblioteki: Essentia oraz NeuroLab.

5.2.3 Baza utworów muzycznych

Koniecznością było pozyskanie odpowiedniej bazy danych, aby móc z jej wykorzystaniem nauczyć oraz przetestować sieć neuronową. Pierwszym pomysłem był odsłuch i ocena utworów przez samego autora i ewentualną osobę towarzyszącą. Biorąc jednak pod uwagę wysoką subiektywność percepcji muzyki prawdopodobnie nie byłby to miarodajny osąd. Z pomocą przyszedł zbiór danych[21] zawierający 744 utwory różnych gatunków muzycznych opublikowany właśnie dla prac naukowych o tematyce zbliżonej do niniejszej pracy. Wszystkie fragmenty zbioru pozyskane zostały z otwartego archiwum muzyki[15]. Każdy z utworów jest 45-cio sekundowym fragmentem uzyskanym z pełnego utworu oraz ocenionym zarówno pod względem pobudzenia jak i zadowolenia przez ponad 300 osób. Skala ocen należy do przedziału $[1, 9]$ dla obu parametrów, a każdy uczestnik przeprowadzonego badania oceniał utwór w czasie ciągłym[16], co dawało średnią częstotliwość oceniania 2 Hz biorąc pod uwagę możliwości techniczne przeglądarek internetowych oraz komputerów używanych przez badanych. Wszystko to sprawia, że ten zbiór danych stanowił dobrą podstawę dla systemu tworzonego przez autora.

5.2.4 Algorytm uczenia maszynowego

Temat pracy brzmi „Wykorzystanie uczenia maszynowego do rozpoznawania nastroju muzyki". Algorytmem, który jednak został wybrany do realizacji tego celu, zostały sztuczne sieci neuronowe, które pozwalają na pozostawienie problemu implementacji rozwiązania problemu samej sieci, co jest znacznym ułatwieniem.

5.3 Opis aplikacji

Pierwszy etapem działania programu jest wczytanie utworów muzycznych, które są zapisane w formacie mp3. Po tym następuje ich wstępna obróbka, która obejmuje zastosowanie okna czasowego dla przetwarzanego sygnału oraz algorytm wyrównywania poziomu głośności. Po wstępnej obróbce, obliczony zostaje wskaźnik zmiany znaku oraz wskaźnik zmian po której wykonywana jest transformacja Fouriera w celu otrzymania widma sygnału, które jest niezbędne w celu obliczenia kolejnych cech dźwiękowych, którymi są kolejno złożoność spektralna, dysonans, skala, płaskość spektralna oraz kształt spektralny w którego skład wchodzą środek masy widma, współczynnik skośności widma, kurtoza widma, *roll off* widma oraz rozrzut widma. Po ekstrakcji cech następuje uczenie oraz testowanie sieci neuronowej. W tym celu baza utworów została podzielona na dwa zestawy, pierwszy z nich składa się 674 utworów słu-

[15] Free Music Archive - `http://freemusicarchive.org/`

[16] `https://www.youtube.com/watch?v=G-GhONd_Wag`

żących do uczenia sieci neuronowej, natomiast drugi z 70 utworów, który służy jako zestaw do oceny sieci neuronowej. Na wejście sieci, zbudowanej z wykorzystaniem biblioteki NeuroLab, przekazywane są wyżej wymienione cechy. Jednokierunkowa sieć neuronowa (ang. *feed-forward*) zwraca dwie wielkości: zadowolenie oraz pobudzenie, które opisują nastrój reprezentowany przez utwór. Wszystkie oceny badanych z wykorzystywanej bazy danych zostały przeskalowane do przedziału $[-1, 1]$, ponieważ została użyta tangensoidalna funkcja aktywacji. Użycie liniowej funkcji aktywacji znacznie ogranicza możliwości sieci, co jest spowodowane tym, że funkcja modelowana przez złożenie funkcji liniowych wciąż pozostanie liniowa, natomiast funkcję liniową można modelować pojedynczym liniowym neuronem z czego wynika wniosek, że wielowarstwowa topologia zwiększa możliwości tylko w przypadku użycia nieliniowych funkcji aktywacji.

5.4 Uczenie sieci neuronowej

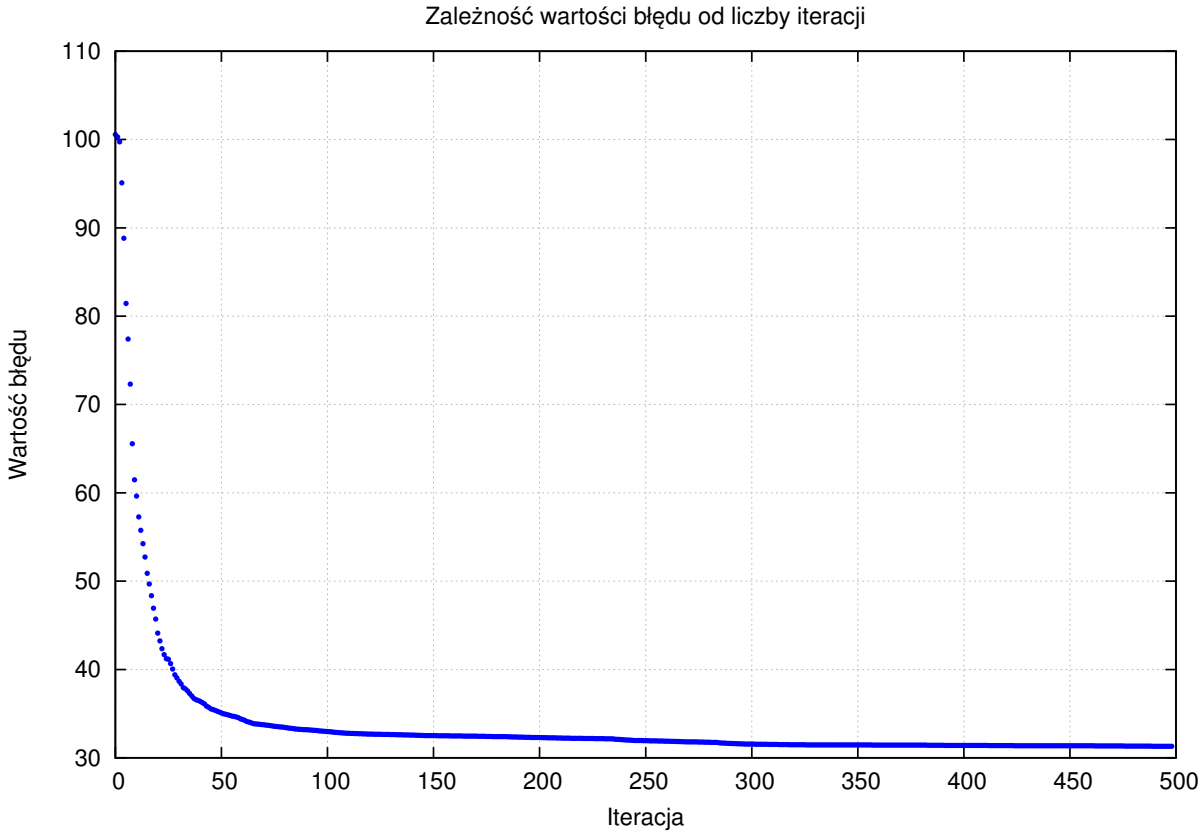
Jednym z podstawowych zadań przy treningu sieci neuronowej jest dobór optymalnej topologii sieci, a w szczególności liczby neuronów w warstwie ukrytej. Istnieją ogólne reguły jak należy w tym wypadku postępować. Z całą pewnością liczba neuronów w warstwie ukrytej nie powinna być zbyt duża ani zbyt mała. Zbyt duża, to znaczy znacznie przekraczająca ilość neuronów wejściowych, ponieważ może to spowodować przeuczenie sieci neuronowej. Przeuczenie, lub też przetrenowanie, sieci polega na tym, że będzie ona nadmiernie dopasowana do danych uczących przez co nie będzie odporna na szumy. Można powiedzieć, że sieć nauczy się „na pamięć" wyników dla danego zestawu danych, co nie jest oczekiwanym rezultatem. Zbyt mała liczba neuronów oznacza liczbę neuronów znacznie mniejszą od ilość neuronów wejściowych. Nie pozwoli to sieci odpowiednio modelować poszukiwanej zależności pomiędzy jej wejściami , a wyjściami. W przypadku uczenia znaczenie ma także czynnik losowy, ponieważ wagi sieci inicjalizowane są wartościami losowymi. Biorąc pod uwagę wspomniane fakty, w celu nauczenia sieci neuronowej wykonany został skrypt korzystając z języka skryptowego powłoki systemowej UNIX, który pozwalał na automatyczne wielokrotne uruchamianie programu. Korzystając z dostępnych flag sprawdzane były także wyniki uzyskiwane przez sieć dla różnej liczby neuronów w warstwie ukrytej. Użytym kryterium oceny sieci była wartość funkcji minimalizowanej przez algorytm uczący, która jest przedstawiona we wzorze 5.

Na wykresie 10 przedstawiona została wartość minimalizowanej funkcji w zależności od liczby iteracji. Łatwo można zauważyć, że już od około setnej iteracji sieć neuronowa praktycznie przestaje się uczyć, następują już wtedy tylko nieznaczne zmiany. Ostatecznie, najmniejsza uzyskana wartość błędu wynosi 4.299 dla zbioru testowego oraz 31.303 dla zbioru uczącego. Najlepszy wynik został otrzymany dla sieci o liczbie neuronów w warstwie ukrytej równej 7. Łatwo jest jest też doprowadzić do sytuacji przeuczenia sieci. W przypadku uruchomienia programu z 40 neuronami w warstwie ukrytej otrzymujemy błędy 13.426; 10.575 kolejno dla zestawu uczącego oraz testowego. Jak łatwo zauważyć, pomimo tego, że mogłoby się wydawać, że sieć dużo lepiej została wytrenowana, to jedna wyniki dla zestawu testowego są gorsze niż

liczba neuronów sieci	40	7
błąd dla utworów uczących / liczba utworów uczących	$0,012$	0.050
błąd dla utworów testowych / liczba utworów testowych	0.15	0.061

Tabela 2: Błąd dla 70 utworów dwóch różnych sieci

w przypadku tej pozornie gorzej wytrenowanej. W tabeli 2 przedstawione zostały wartości błędu dla zestawów uczących oraz testowych dla sieci z 40 neuronami oraz z 7 neuronami podzielone przez liczbę utworów uczących oraz testowych. Widać tutaj sporą różnicę pomiędzy błędem dla zestawu testowego oraz uczącego w przypadku sieci z 40 neuronami, co wskazuje zdecydowanie na jej przeuczenie.



Rysunek 10: Zależność wartości błędu od liczby iteracji

6 Uzyskane wyniki

6.1 Parametry zadowolenia oraz pobudzenia

Zależność pomiędzy odpowiedziami oczekiwanymi przez sieć, a jej rzeczywistymi odpowiedziami, w przypadku dobrze nauczonej i działającej sieci powinna być zbliżona do liniowej. Biorąc to pod uwagę, przedstawiono wykresy, których jedna oś wskazuje oczekiwaną wartość

parametr	zadowolenie	pobudzenie
współczynnik korelacji	0.66	0.55

Tabela 3: Współczynniki korelacji liniowych parametrów zadowolenia oraz pobudzenia

parametru, a druga rzeczywistą wartość parametru w celu zaobserwowania czy występuję liniowość, co też uczyniono, a efekty można zaobserwować na rysunkach 11 oraz 12. Jak można zauważyć, istnieje pewna korelacja dla obu parametrów. W celu określenia konkretnych wartości liczbowych wskazujących na korelację możemy posłużyć się współczynnikiem korelacji liniowej zmiennych x oraz y, który wyraża się wzorem:

$$\rho_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}, \tag{25}$$

gdzie σ_x, σ_y są odchyleniami standardowymi odpowiednio x oraz y, natomiast $cov(x,y)$ jest kowariancją i określamy ją wzorem:

$$cov(x,y) = \overline{xy} - \overline{x} * \overline{y}, \tag{26}$$

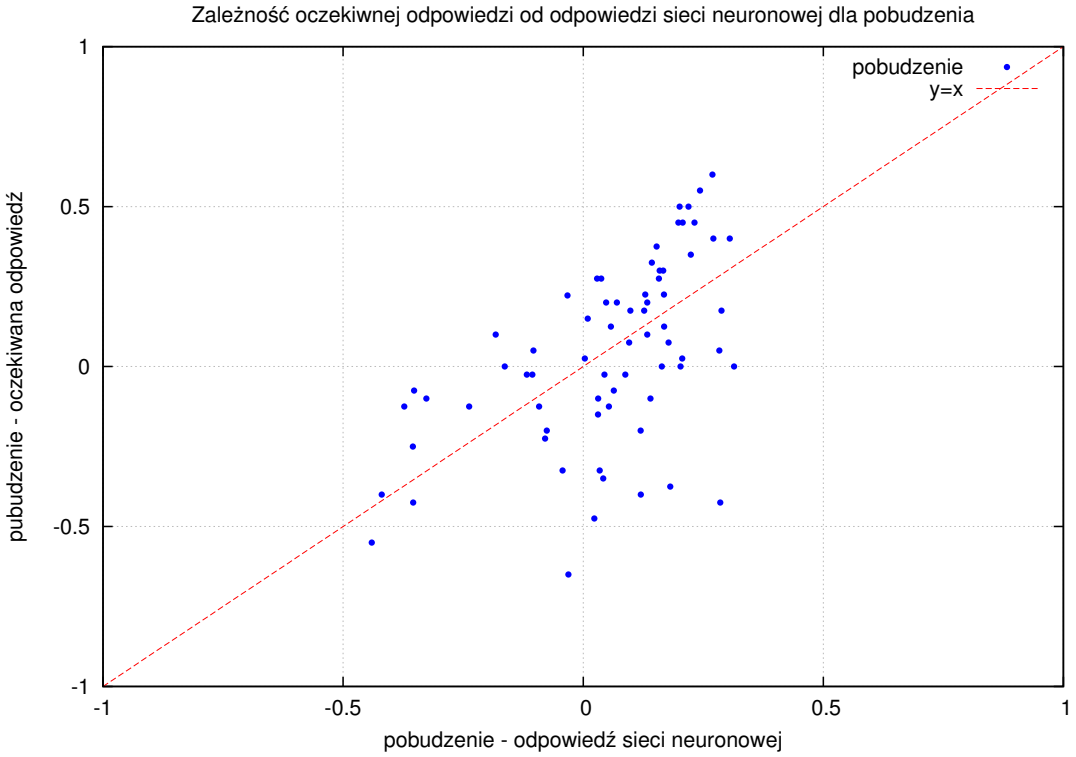
gdzie $\overline{xy}$, $\overline{x}$, $\overline{y}$ są kolejno wartościami średnimi xy, x oraz y. Obliczone współczynniki korelacji przedstawiono w tabeli 3. Zaobserwowana korelacja nie jest zbyt duża, niemniej jednak widzimy wyraźny związek pomiędzy odpowiedzią sieci neuronów, a wartościami pobudzenia oraz zadowolenia wyznaczonymi przez badane osoby. Biorąc pod uwagę subiektywny charakter oceny nastroju przez człowieka, jak również fakt, że sieć neuronowa używała tylko 11 cech uzyskanych z danego utworu muzyki, jest to wynik świadczący tym, że przedstawiony system do pewnego stopnia reprodukuje subiektywne oceny ludzi.

6.2 Zależność nastroju od cech dźwiękowych

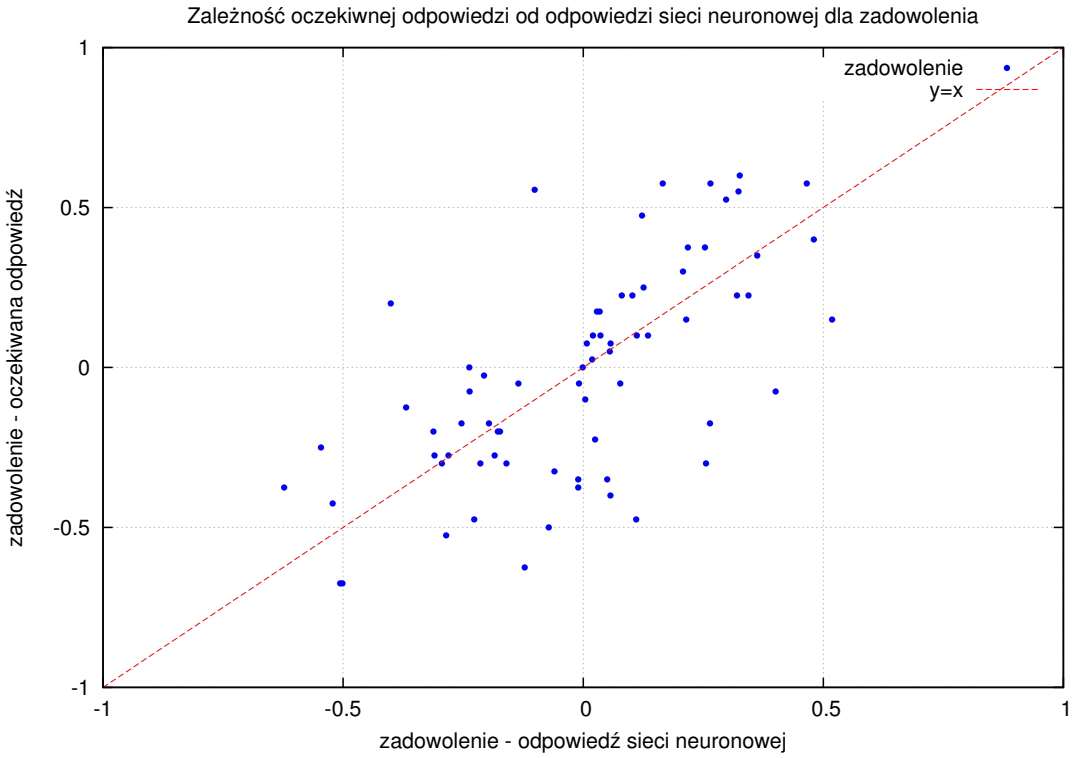
Pod uwagę zostało wziętych 11 cech dźwiękowych, które mają różny wpływ na subiektywny nastrój reprezentowany przez utwór muzyczny. W tym rozdziale przeanalizowana jest zależność wartości parametrów zadowolenia oraz pobudzenia od wartości cech dźwiękowych.

Wskaźnik zmiany znaku
Rysunek 13 przedstawia zależność wskaźnika zmiany znaku od subiektywnej oceny zadowolenia oraz pobudzenia. Możemy zaobserwować tendencję, że wraz ze wzrostem wartości tego wskaźnika wzrasta również wartość obu parametrów - zadowolenia jak i pobudzenia. Nie jest to jednak znaczący wzrost.

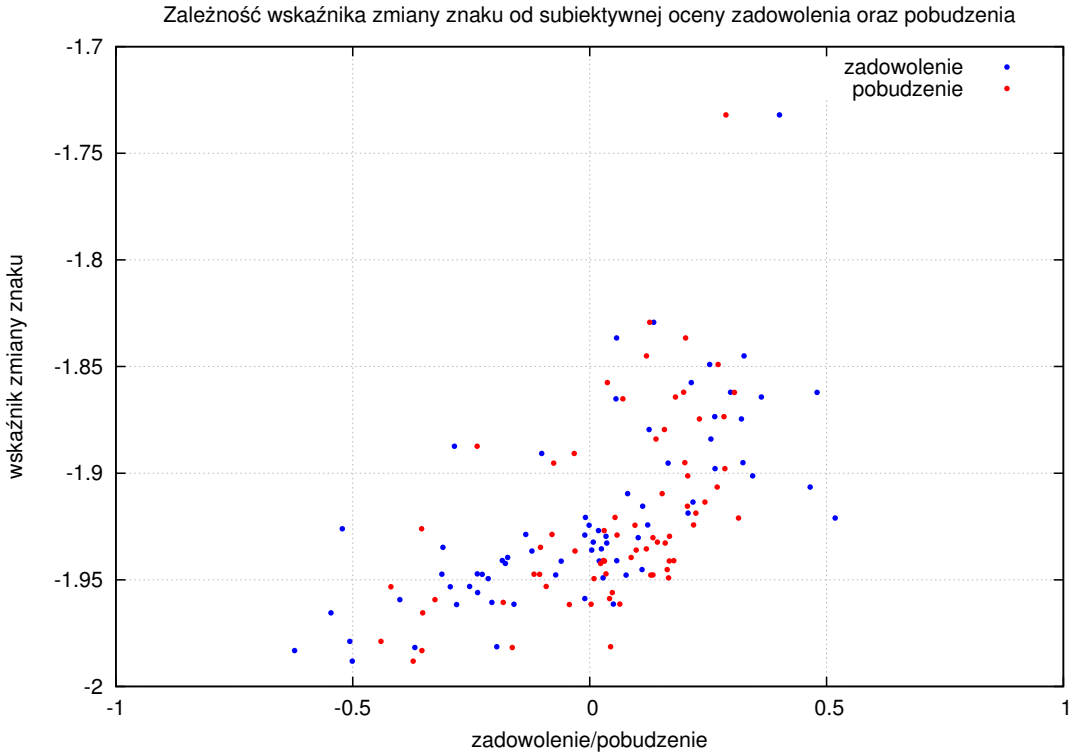
Wskaźnik zmian
Podobnie jak w przypadku wskaźnika zmiany znaku, oba parametry - zadowolenie oraz pobudzenie wzrastają nieznacznie wraz ze wzrostem wskaźnika zmian, ale wzrost ten jest nieznaczny. Wykres przedstawiony jest na rysunku 14.



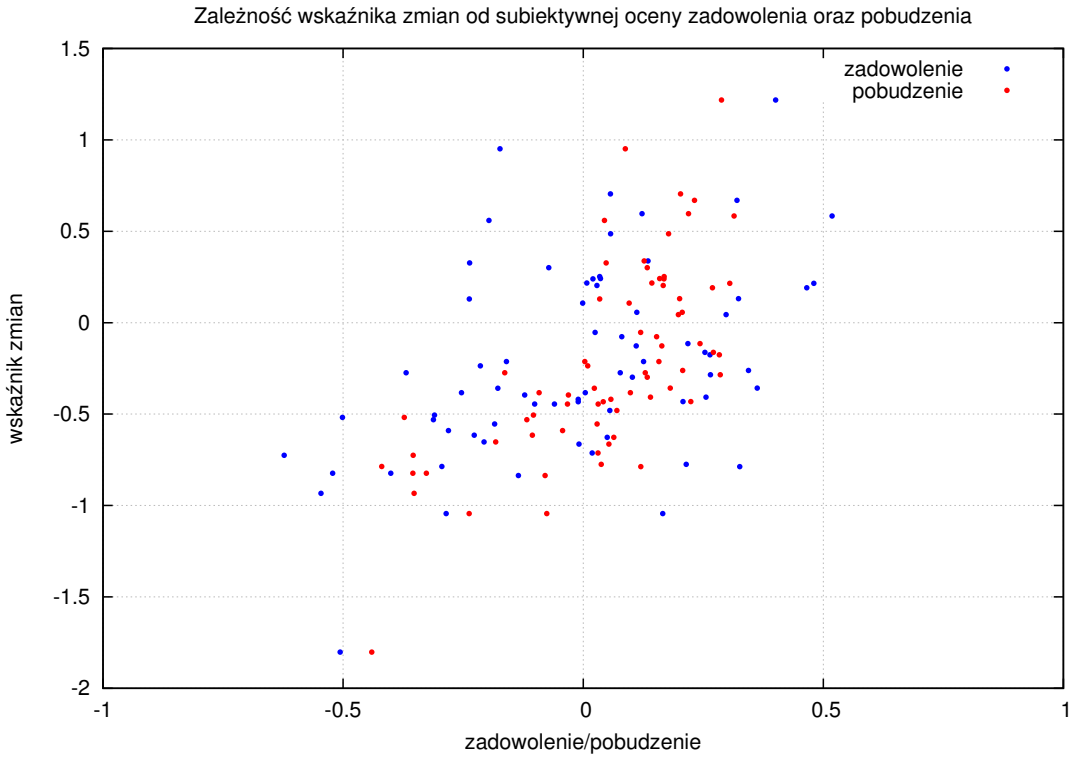
Rysunek 11: Zależność oczekiwanej odpowiedzi od odpowiedzi sieci neuronowej dla parametru pobudzenia



Rysunek 12: Zależność oczekiwanej odpowiedzi od odpowiedzi sieci neuronowej dla parametru zadowolenia



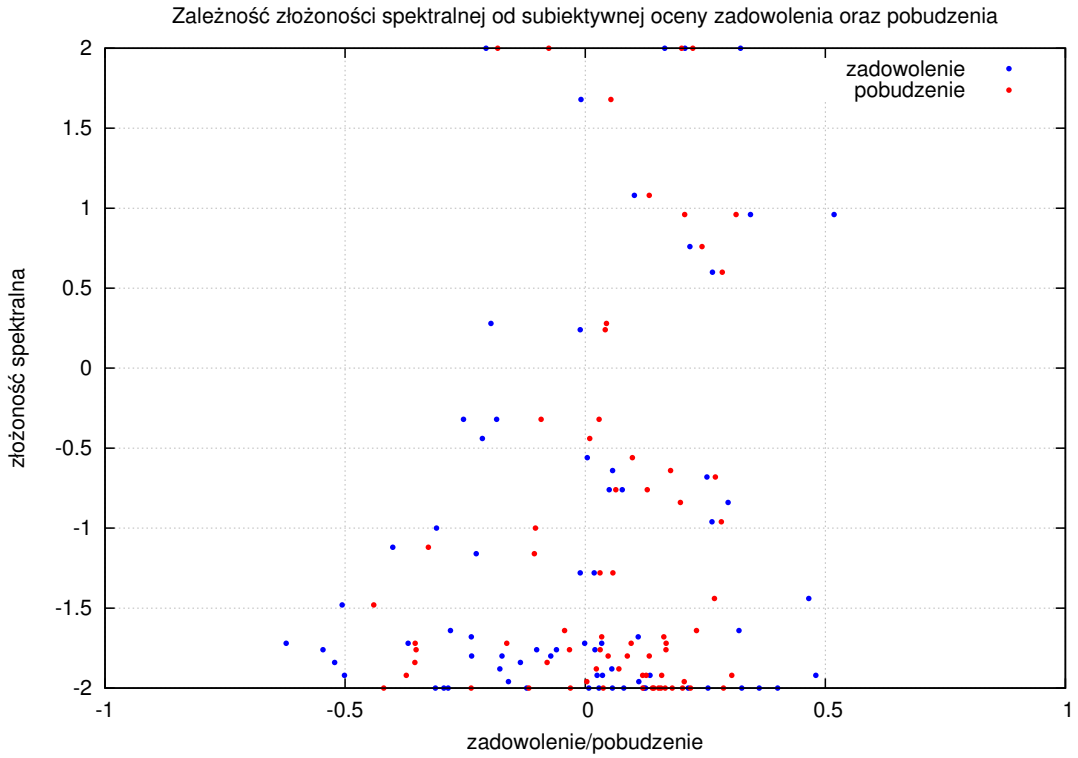
Rysunek 13: Zależność wskaźnika zmiany znaku od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)



Rysunek 14: Zależność wskaźnika zmian od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)

Złożoność spektralna

Zależność parametrów wskazujących na emocje od złożoności spektralnej została przedstawiona na rysunku 15. Obserwując wykres trudno mówić o jakiejkolwiek korelacji. Zdecydowana większość wartości wydaje być się losowo rozrzucona.



Rysunek 15: Zależność złożoności spektralnej od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)

Środek masy widma

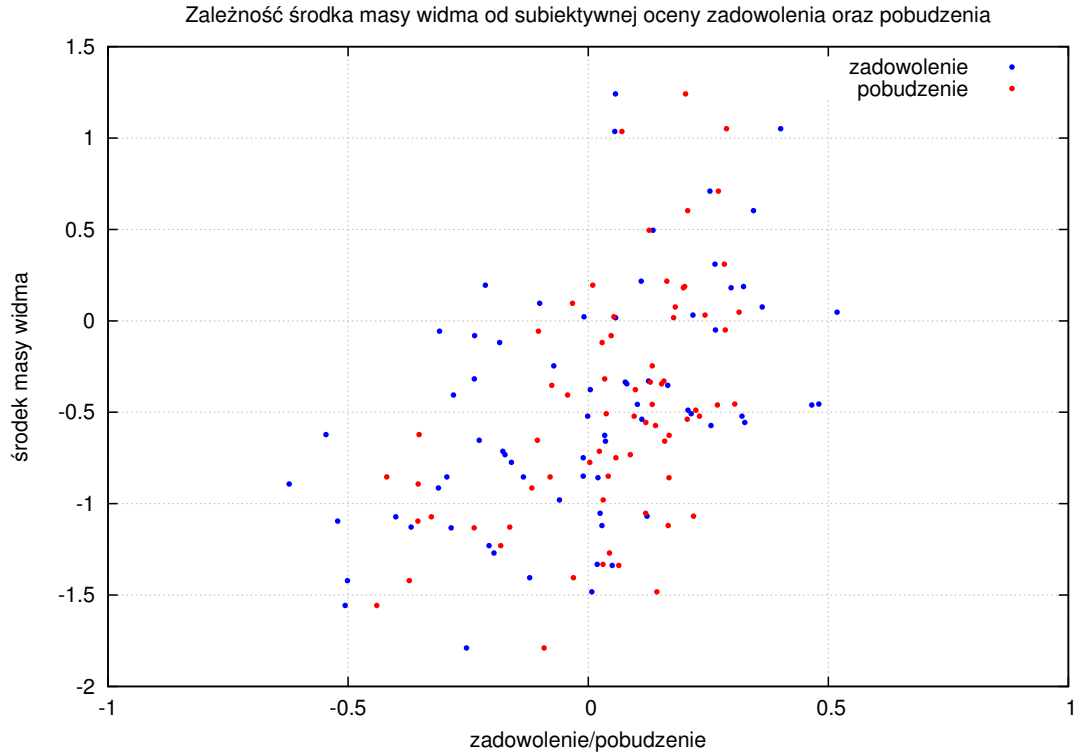
W przypadku środka masy widma możemy zauważyć nieznaczny związek pomiędzy zmianą wartości środka widma, a parametrami zadowolenia oraz pobudzenia. Zostało to przedstawione na rysunku 16.

Współczynnik skośności widma

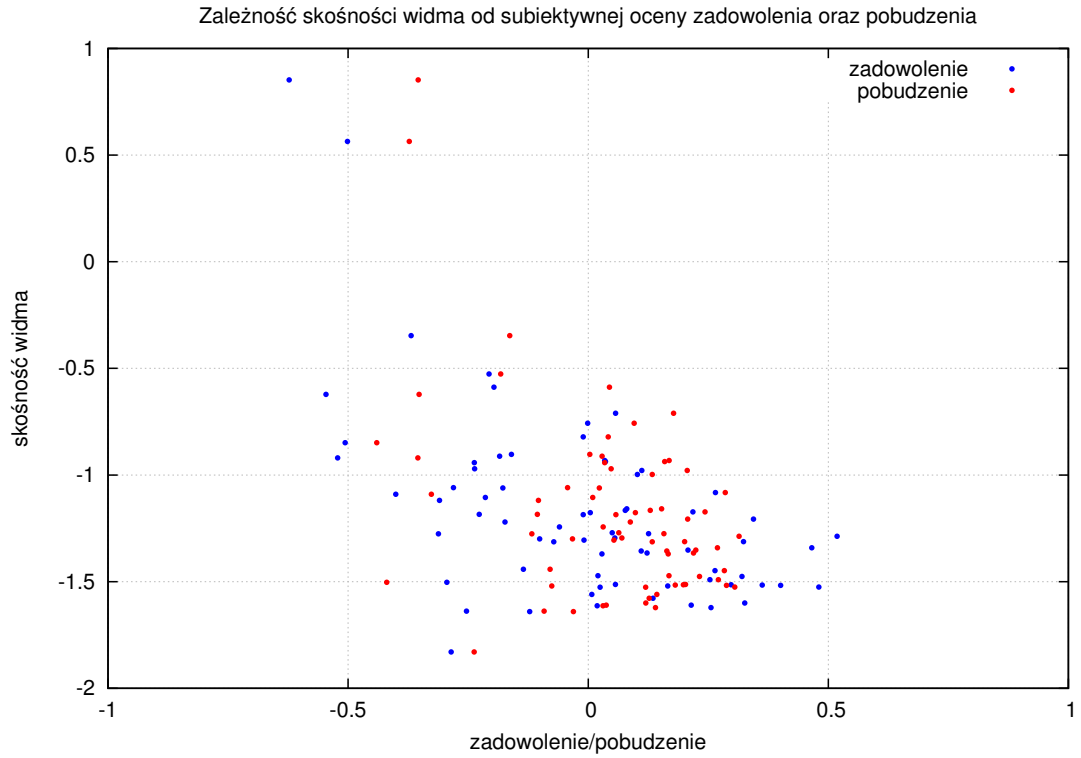
Jeśli chodzi o współczynnik skośności widma możemy zaobserwować, że przy wyższych jego wartościach parametry bardzo nieznacznie mogą być zwiększone, co jednak także nie jest wyraźne i przedstawia się na wykresie 17.

Kurtoza widma

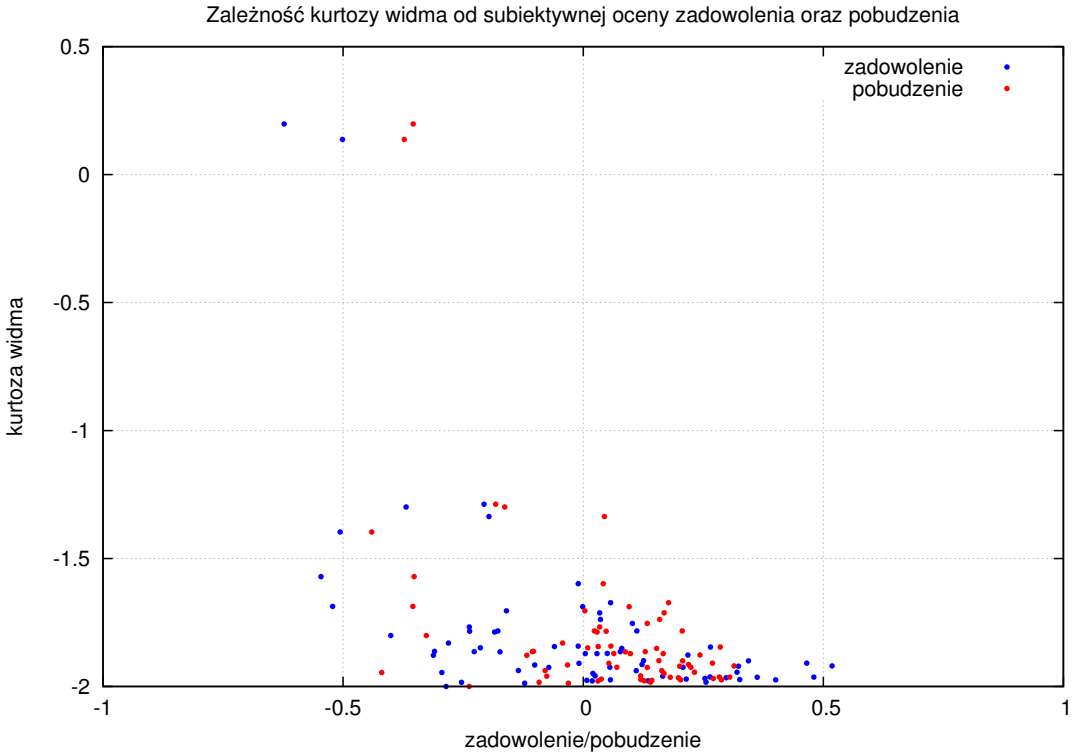
Kurtoza widma według wykresu 18 nie wydaje się mieć większego wpływu na wartości parametrów zadowolenia oraz pobudzenia.



Rysunek 16: Zależność środa masy widma od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)



Rysunek 17: Zależność skośności widma od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)



Rysunek 18: Zależność kurtozy widma od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)

Rozrzut spektralny

Rozrzut spektralny przedstawiony na wykresie 19 wydaje się być względnie znaczącym czynnikiem w ocenie parametrów nastroju muzyki w stworzonym systemie. Oba z nich wzrastają w przypadku wzrosty wartości tej cechy dźwięku.

Roll-off widma

Zależność *Roll off*u widma od parametrów pobudzenia oraz zadowolenia nie wskazuje na większą korelację (patrz Rys. 20).

Płaskość spektralna

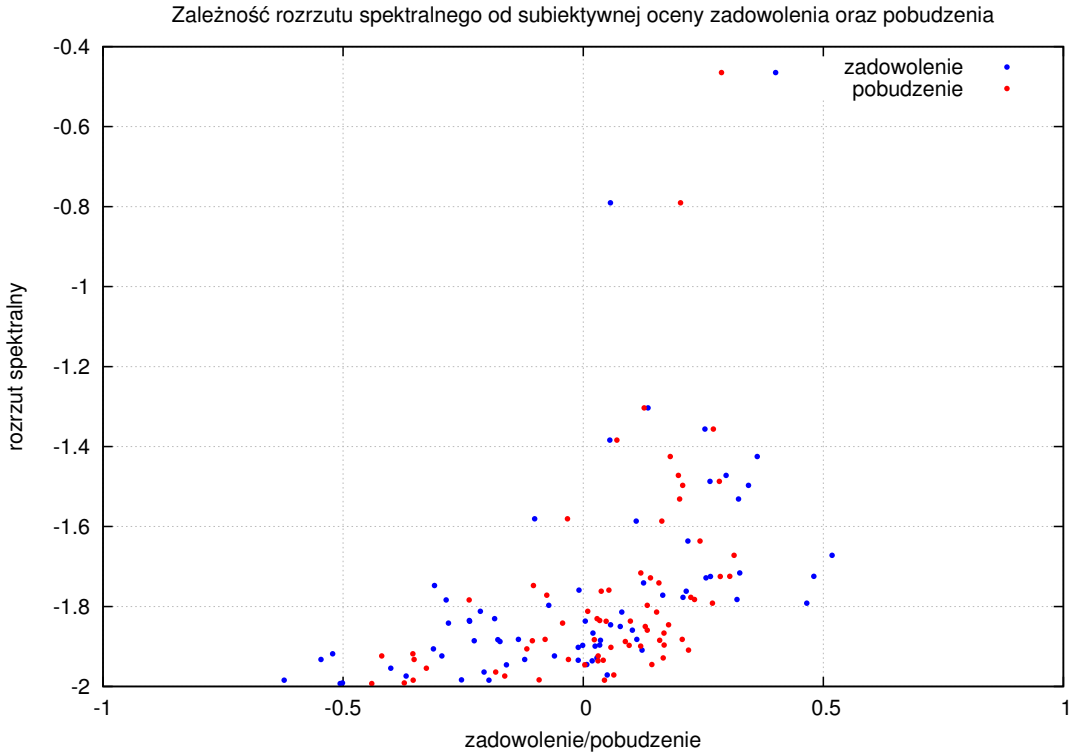
Na wykresie 21 przedstawiony został wpływ parametru płaskości spektralnej na parametry pobudzenia oraz zadowolenia. Również i w tym przypadku nie widać większych korelacji.

Dysonans

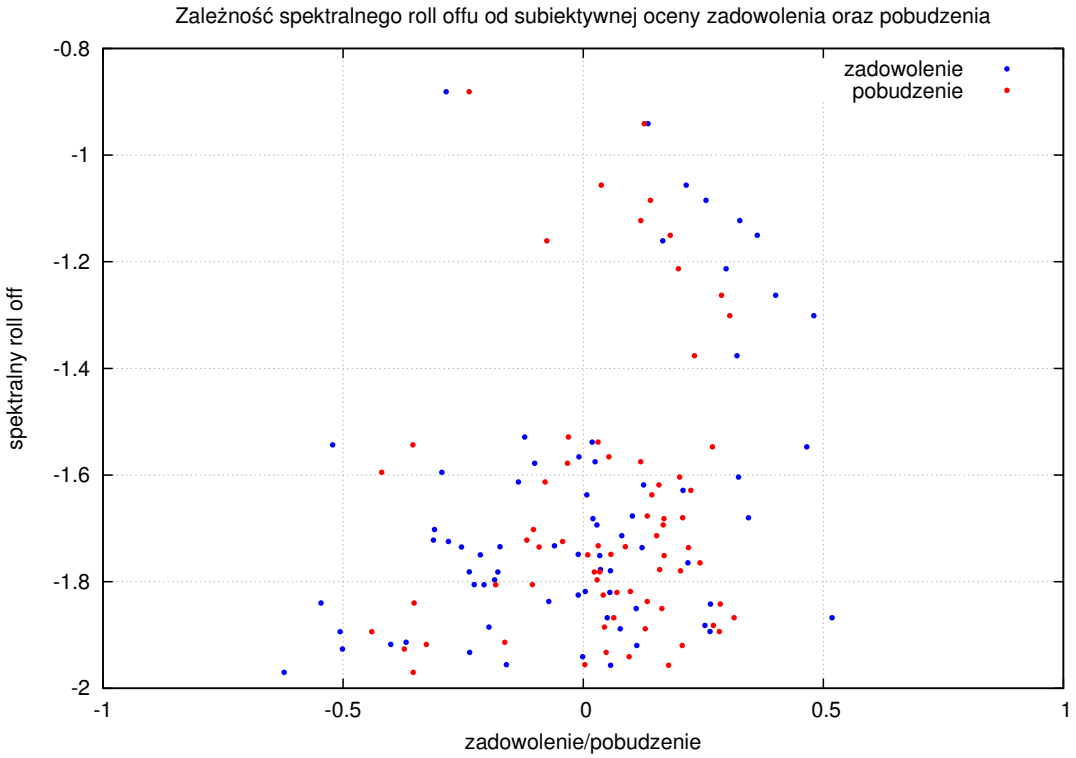
Podobnie jak w przypadku płaskości spektralnej nie istnieje większa korelacja pomiędzy parametrami nastroju muzyki, a dysonansem, co wizualizuje wykres22.

Skala

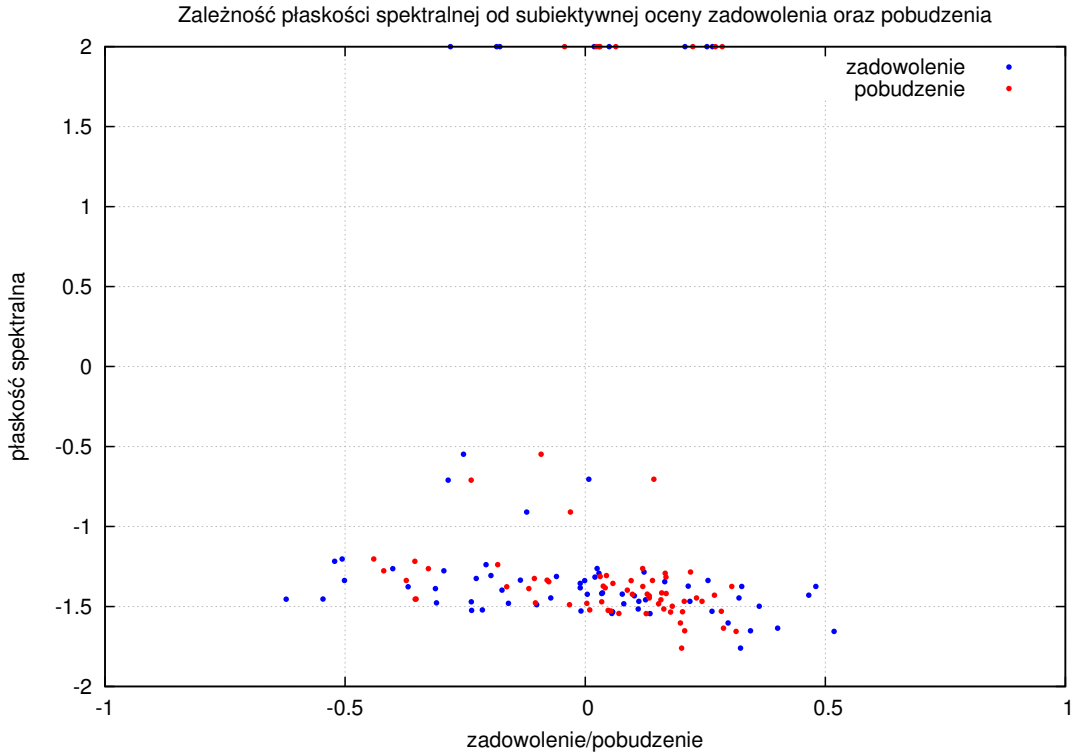
W tabeli 4 zostały przedstawione średnie wartości pobudzenia oraz zadowolenia dla skali muzycznych majorowej oraz minorowej, które jednak zbliżone są bardzo do siebie w przypadku obu



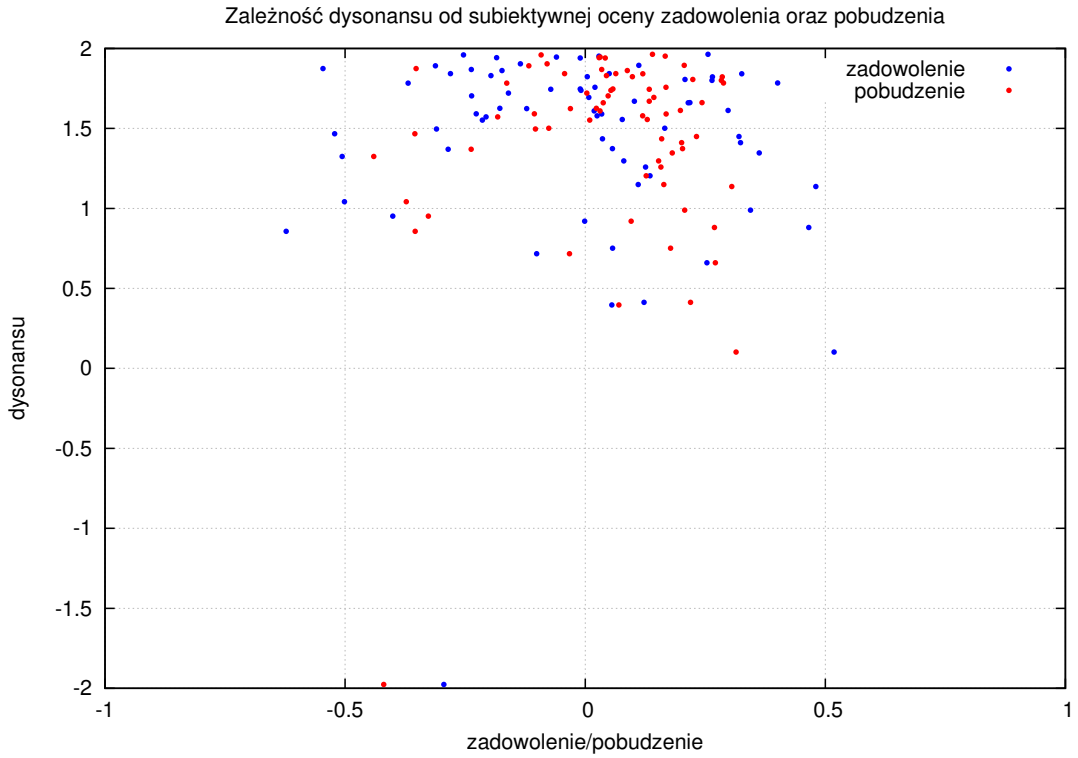
Rysunek 19: Zależność rozrzutu spektralnego od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)



Rysunek 20: Zależność spektralnego *roll off*'u od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)



Rysunek 21: Zależność płaskości spektralnej od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)



Rysunek 22: Zależność dysonansu od subiektywnej oceny zadowolenia (punkty niebieskie) oraz pobudzenia (punkty czerwone)

skala	minorowa (molowa)	majorowa (durowa)
wartość średnia dla zadowolenia	0.02 ± 0.25	-0.08 ± 0.30
wartość średnia dla pobudzenia	0.06 ± 0.19	0.03 ± 0.19

Tabela 4: Wartości średnie dla skali muzycznej wraz z odchyleniem standardowym

parametr	zadowolenie	pobudzenie
współczynnik korelacji	0.54	0.58

Tabela 5: Współczynniki korelacji liniowych parametrów zadowolenia oraz pobudzenia

skal. Sugeruje to, że ta cecha dźwiękowa nie była dobrym wskaźnikiem jeśli chodzi o parametry pobudzenia oraz zadowolenia.

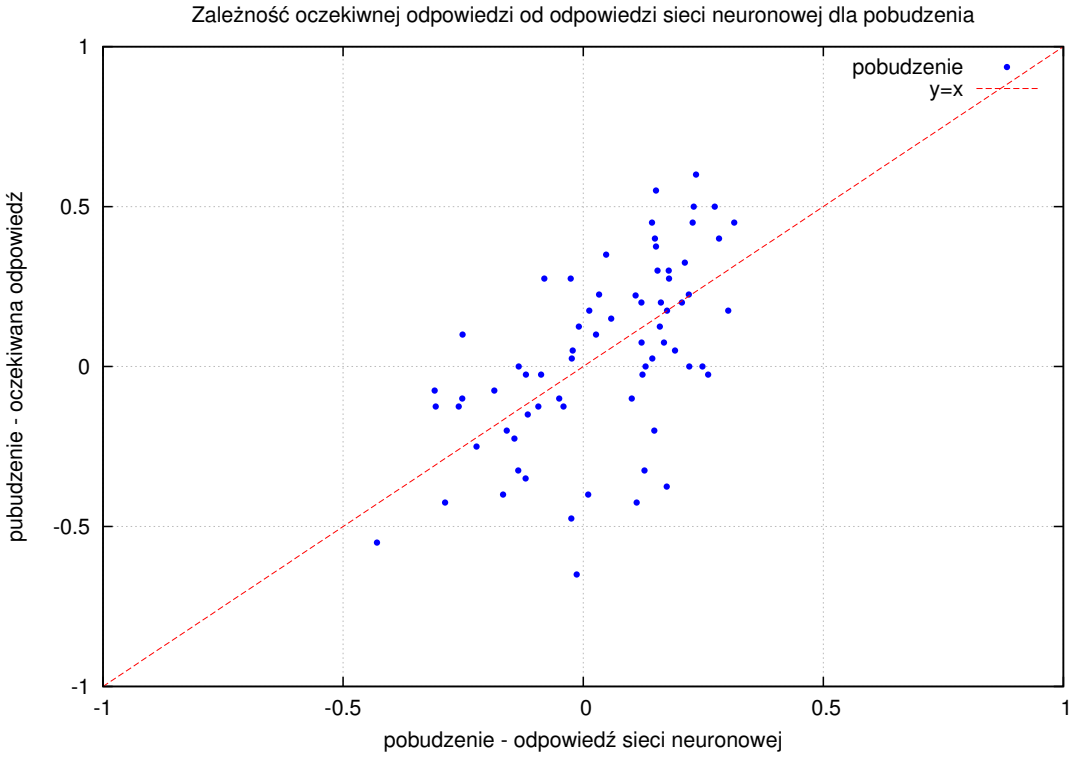
Podsumowanie

Obserwując wykresy dla poszczególnych cech dźwiękowych nie jest trudno dojść do wniosku, że nie wszystkie z nich mają wpływ na parametry nastroju muzyki, a te które mają, nie są silnymi zmiennymi. Tłumaczy to niezbyt wysoką korelację przedstawioną w tabeli 3. Analiza utworów została przeprowadzona powtórnie, ale biorąc pod uwagę tylko te cechy, które korelowały z wartościami pobudzenia oraz zadowolenia czyli: wskaźnik zmiany znaku, wskaźnik zmian, środek masy widma, skośność spektralną oraz rozrzut spektralny. Otrzymane rezultaty przedstawia tabela 5 oraz wykresy 23 i 24. Jak widać, usunięcie kilku zmiennych i pozostawienie tylko znaczących cech nie wpłynęło w wyraźny sposób na wyniki reprezentowane przez współczynniki korelacji. Świadczy to o tym, że na końcowy wynik efektywny wpływ ma tylko 5 pozostawionych cech.

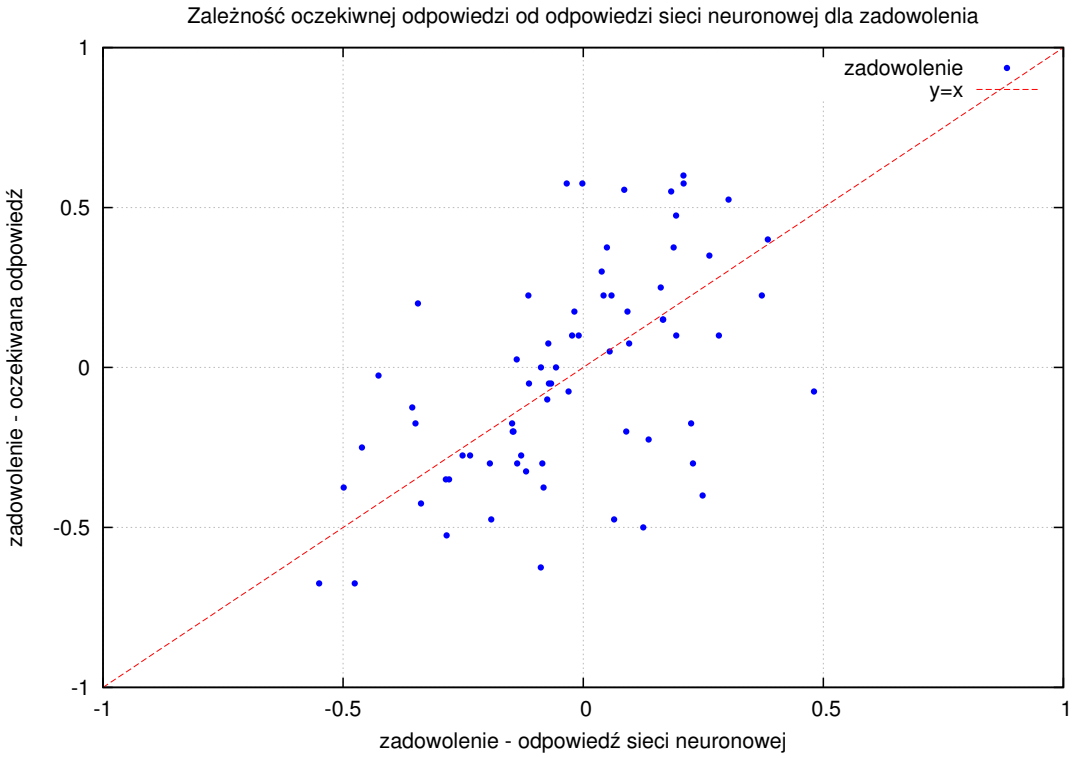
7 Wnioski

7.1 Ocena działania systemu

System do pewnego stopnia spełnił postawione przed nim zadania. Otrzymane wyniki pokazują, że stworzona sieć neuronowa nie modeluje w znaczącym stopniu zależności pomiędzy cechami dźwiękowymi utworów muzycznych, a parametrami określającymi nastrój tych utworów. Współczynniki przedstawione w tabeli 3 wskazują na słabą korelację pomiędzy odpowiedziami sieci, a oczekiwanymi wartościami. Można jednak zaobserwować pewną korelację pomiędzy uzyskaną przez system oceną nastroju, a oceną badanych. Wskazuje to, że możliwe jest zbudowanie efektywnego systemu klasyfikującego. Po lekturze rozdziału 6.2 zauważamy, że jedynie 5 z 11 analizowanych cech daje podstawy dla sieci pozwalające wyznaczyć wartości pobudzenia oraz zadowolenia, co nie wątpliwie jest jedną z przyczyn niskich współczynników korelacji. Należy jednak zauważyć, że wszystkie utwory były analizowane całościowo tzn. sygnał nie był dzielony na mniejsze fragmenty, co prawdopodobnie jest przyczyną niezadowalających wyników. Nastrój



Rysunek 23: Zależność oczekiwanej odpowiedzi od odpowiedzi sieci neuronowej dla parametru pobudzenia



Rysunek 24: Zależność oczekiwanej odpowiedzi od odpowiedzi sieci neuronowej dla parametru zadowolenia

muzyki zmienia się w czasie, co także powinno być wzięte pod uwagę w przypadku realizacji zadania rozpoznawania emocji reprezentowanych przez muzykę, co zostało zrobione w innych pracach podejmujących podobną tematykę[2][3] i dało wyraźnie lepsze rezultaty. Należy jednak mieć na uwadze jednak także bazę danych na której przeprowadzono badania, gdyż duże znaczenie może mieć także jej zróżnicowanie pod kątem gatunków muzycznych.

7.2 Propozycja usprawnienia

Biorąc pod uwagę wspominane prawdopodobne przyczyny niskiej efektywności systemu, pierwszą, najważniejszą propozycją usprawnienia jest dzielenie sygnału audio na mniejsze fragmenty, ocena ich oraz uśrednienie wyników, co mogłoby dać lepsze wyniki. Należałoby wtedy także przekonać się czy cechy dźwięku wykorzystane w programie wciąż nie mają wpływu na wartości parametrów zadowolenia oraz pobudzenia, a także rozważyć użycie innych cech, co nie stanowiłoby wielkiego wyzwania dzięki bibliotece programistycznej Essentia, gdyż oferuje ona kilkadziesiąt możliwych do wykorzystania algorytmów. Oprócz cech wynikających z samego sygnału audio, możliwe jest także analiza tekstu utworów, który również ma znaczący wpływ na nastrój utworu. Dodatkową sugestią w kwestii rozpoznawania nastroju mogą być także okładki płyt z których pochodzą utwory muzyczne, co również mogłoby usprawnić ocenę. Innym kierunkiem rozwoju aplikacji prezentowanej w pracy jest także ocena utworów muzycznych pod kątem ich gatunków.

Literatura

[1] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[2] Naresh N Vempala and Frank A Russo. Predicting emotion from music audio features using neural networks. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*. Lecture Notes in Computer Science London, UK, 2012.

[3] Eduardo Coutinho and Angelo Cangelosi. A neural network model for the prediction of musical emotions. *Advances in cognitive systems*, pages 331–368, 2010.

[4] Mudiana Binti Mokhsin, Nurlaila Binti Rosli, Wan Adilah Wan Adnan, and Norehan Abdul Manaf. Automatic music emotion classification using artificial neural network based on vocal and instrumental sound timbres. *New Trends in Software Methodologies, Tools and Techniques: Proceedings of the Thirteenth SoMeT_14*, 265:3, 2014.

[5] Ricardo Malheiro, Renato Panda, Paulo Gomes, and R Paiva. Music emotion recognition from lyrics: A comparative study. 6th International Workshop on Machine Learning and Music (MML13). Held in Conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPPKDD13), 2013.

[6] Ryszard Tadeusiewicz. *Sieci Neuronowe*. Akademicka Oficyna Wydawnicza, Warszawa, 1993.

[7] Maciej Szaleniec Ryszard Tadeusiewicz. *Leksykon Sieci Neuronowych*. Wydawnictwo Fundacji "Projekt Nauka", Wrocław, 2015.

[8] David Kriesel. A brief introduction to neural networks. *Retrieved August*, 15:2011, 2007.

[9] Cyril Laurier. *Automatic Classification of Musical Mood by Content Based Analysis*. Universitat Pompeu Fabra, 2011.

[10] Sophocles J Orfanidis. Introduction to signal processing. `http://www.ece.rutgers.edu/~orfanidi/intro2sp/orfanidis-i2sp.pdf`, 1995. Dostęp:29.12.2015.

[11] Audacity - free open source digital audio editor and recording computer software application. `http://audacity.pl/`.

[12] National Instruments White Papers. Understanding ffts and windowing. `http://www.ni.com/white-paper/4844/en/`. Dostęp:29.12.2015.

[13] Słownik Języka Polskiego PWN. Hasło izofona. http://encyklopedia.pwn.pl/haslo/;3915950. Dostęp:29.12.2015.

[14] Full revision of international standards for equal-loudness level contours (ISO 226). http://www.aist.go.jp/aist_e/latest_research/2003/20031114/20031114.html. Dostęp:29.12.2015.

[15] Janusz Słupik. Dźwięk cyfrowy. http://www.aist.go.jp/aist_e/latest_research/2003/20031114/20031114.html. Dostęp:02.01.2016.

[16] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[17] Janusz Jusiak. Podstawowe pojęcia muzyczne. http://bacon.umcs.lublin.pl/~jjusiak/dokumenty/konwersatoria/PodstawowePoj%C4%99ciaWMuzyce.pdf. Dostęp:01.01.2016.

[18] Emilia Gómez Gutiérrez et al. Tonal description of music audio signals. 2006.

[19] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(03):715–734, 2005.

[20] D. Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, P. Herrera, O. Mayor, Gerard Roma, J. Salamon, J. Zapata, and Xavier Serra. Essentia: an audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)*, pages 493–498, Curitiba, Brazil, 04/11/2013 2013.

[21] Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, pages 1–6, New York, NY, USA, 2013. ACM.

Dodatki

A Obsługa programu

A.1 Konfiguracja

Stworzony program posiada możliwości konfiguracyjne. W celu edycji ustawień należy edytować plik Configurations.py, który zawiera szereg stałych, których zdefiniowanie jest konieczne

do poprawnego działania aplikacji. Większość z nich są to ścieżki do plików, którymi są m.in. ścieżki dostępu do utworów muzycznych, oceny utworów muzycznych czy też ścieżka do pliku przechowującego logi uruchomienia oraz format samego logowania wiadomości. Można znaleźć tam także szczegóły konfiguracyjne dotyczące samej sieci neuronowej tj. liczba neuronów w warstwie ukrytej oraz liczba wyjść w zależności od potrzeb.

A.2 Uruchamianie

Wykonany program pozwala na uruchamianie go w różnych trybach w zależności od potrzeb. Podczas działania programu do konsoli wypisywane są logi, które informują użytkownika o tym co dzieje się w danym momencie lub co poszło nie tak w razie, gdy taka sytuacja nastąpi. Podstawowy przepływ jeśli chodzi o działanie aplikacji jest zgodny ze schematem systemu przedstawionym na rysunku9. Korzystając jednak z odpowiednich flag możemy go modyfikować. Jednym z dłużej trwających procesów podczas wykonywania jest ekstrakcja cech dźwiękowych z utworów ze względu na ich dużą ilość. Z tego względu możliwe jest podanie przy uruchomieniu flagi -a, która pozwala określić czy chcemy analizować wszystkie piosenki czy wczytać je z pliku o ile taki istnieje. Jego ścieżka znajduje się w pliku konfiguracyjnym. Po każdym uruchomieniu programu przeanalizowane cechy dźwiękowe są zapisywane do pliku, więc program należy przynajmniej raz uruchomić z flagą -a. Użyteczną flagą jest także flaga -t, która pozwala na wczytanie sieci neuronowej, ale także tyle wtedy, gdy istnieje plik wskazany w konfiguracji. Podobnie jak w przypadku flagi -a, należy przynajmniej raz uruchomić program z flagą -t, aby stworzyć chociaż jedną sieć neuronową. Inne dostępne flagi, których możemy użyć prezentują się następująco:

- -h, –help wyświetlenie pomocy
- -n N, określenie ilości neuronów N w warstwie ukrytej dla nowej sieci neuronowej (należy używać tylko w połączeniu z flagą -t)
- -e, ocena sieci neuronowej dla zbioru testowego
- -p, rysowanie wykresów dla parametrów zadowolenia oraz pobudzenia (należy używać tylko w połączeniu z flagą -e)
- -s FILE, ocena utworu muzycznego znajdującego się pod ścieżką FILE pod względem parametrów zadowolenia oraz pobudzenia (należy używać bez innych flag)