Politechnika Wrocławska

DZIEDZINA: NAUKI INŻYNIERYJNO-TECHNICZNE
DYSCYPLINA: Automatyka, elektronika, elektrotechnika i technologie kosmiczne

ROZPRAWA DOKTORSKA

Automatyczna klasyfikacja wybranych zniekształceń sygnałów muzycznych z wykorzystaniem konwolucyjno-rekurencyjnych sieci neuronowych bez porównania do sygnału referencyjnego

mgr inż. Kamila Organiściak

Promotor:
prof. dr hab. inż. Józef Borkowski

Słowa kluczowe: klasyfikacja zniekształceń, konwolucyjno-rekurencyjne sieci neuronowe, muzyka, audio

WROCŁAW 2023

Spis treści

Streszczenie *6*

Abstract *7*

Wykaz wybranych skrótów anglojęzycznych *8*

Wykaz wybranych oznaczeń *9*

Wprowadzenie *10*

Teza, cel i zakres pracy *13*

1. Ogólna charakterystyka oceny jakości sygnału audio *14*

1.1. Główne trudności oceny jakości sygnału audio **14**

1.2. Wpływ psychoakustyki na percepcję dźwięku **15**

1.2.1. Obszar słyszalności 15
1.2.2. Pasma krytyczne 16
1.2.3. Maskowanie 17
1.2.4. Głośność dźwięku 18
1.2.5. Krzywe jednakowego słyszenia 19

1.3. Wpływ cyfrowego przetwarzania i transmisji na jakość dźwięku **19**

1.3.1. Cyfrowe kodowanie dźwięku 20
1.3.2. Przetwarzanie typu downmix i upmix 22
1.3.3. Zmiany zakresu dynamiki sygnału 23
1.3.4. Normalizacja poziomu sygnału 23
1.3.5. Znakowanie wodne sygnałów fonicznych 24
1.3.6. Transmisja sygnału audio 25

2. Manualne metody oceny jakości audio *26*

2.1. Wprowadzenie do manualnych testów odsłuchowych **26**

2.2. Grupa słuchaczy **26**

2.3. Procedura testowa **26**

2.4. Metoda testowa **27**

2.5. Ocena jakości sygnału audio **28**

2.6. Najczęstsze błędy podczas testów odsłuchowych **30**

2.7. Podsumowanie **30**

3. Automatyczne metody oceny jakości audio *31*

3.1. Wprowadzenie do automatycznych testów jakości dźwięku **31**

3.2. Metody referencyjne **31**

3.2.1. Miara SNR 32
3.2.2. Miara *Segmental* SNR 32
3.2.3. Metoda PEAQ 33
3.2.4. Metoda PEMO-Q 34
3.2.5. Metoda ViSQOLAudio 35

3.3. Metody bezreferencyjne **36**

3.4. Zastosowanie sztucznych sieci neuronowych do przetwarzania muzyki **36**

3.5. Podsumowanie **38**

4. Autorska prototypowa metoda automatycznej klasyfikacji sygnału z eliminacją porównania do sygnału referencyjnego *42*

5. Autorska baza danych audio *44*
5.1. Wprowadzenie **44**
5.2. Parametry sygnału audio **44**
5.3. Przygotowanie bazy danych **45**
5.4. Przetwarzanie wstępne **47**
5.5. Skala melowa **50**
5.6. Parametr *Zero Crossing Rate* **51**
5.7. Parametr *Octave-Based Spectral Contrast* **51**
5.8. Głośność **52**
5.9. Głośność chwilowa **53**
5.10. Rzeczywista wartość szczytowa sygnału **54**
6. Projekt modelu *55*
6.1. Wprowadzenie **55**
6.2. Schemat działania modelu **55**
6.3. Hiperparametry modelu **56**
6.4. Konwolucyjne sieci neuronowe **57**
6.5. Podpróbkowanie **59**
6.6. Rekurencyjne sieci neuronowe **61**
6.7. Sieci typu *Long Short-Term Memory* **62**
6.8. Dwukierunkowe sieci rekurencyjne **64**
6.9. Funkcje aktywacji **66**
7. Ewaluacja modelu *69*
7.1. Wprowadzenie **69**
7.2. Miary klasyfikacji **69**
7.3. Wymagania sprzętowe **70**
8. Model sieci konwolucyjnych *72*
8.1. Wprowadzenie **72**
8.2. Architektura modelu **72**
8.3. Analiza wyników klasyfikacji **73**
8.4. Podsumowanie **73**
9. Model sieci konwolucyjno-rekurencyjnych *75*
9.1. Wprowadzenie **75**
9.2. Architektura modelu **75**
9.3. Wyniki klasyfikacji **76**
9.4. Wpływ liczby filtrów melowych dla sygnałów wejściowych na wyniki klasyfikacji ... **76**
9.5. Podsumowanie **76**

10. Model sieci konwolucyjno-rekurencyjnych z zastosowaniem dodatkowych parametrów wejściowych 78

10.1. Wprowadzenie 78

10.2. Zastosowanie parametru ZCR 78

10.2.1. Architektura modelu 78

10.2.2. Wyniki klasyfikacji 79

10.3. Zastosowanie parametru OBSC 79

10.3.1. Architektura modelu 79

10.3.2. Wyniki klasyfikacji 80

10.4. Zastosowanie parametrów głośności 80

10.4.1. Architektura modelu 81

10.4.2. Wyniki klasyfikacji 81

10.5. Podsumowanie 82

11. Podsumowanie 84

Bibliografia 88

Streszczenie

Niniejsza rozprawa prezentuje wyniki badań nad zastosowaniem sztucznych sieci neuronowych (konwolucyjno-rekurencyjnych) w analizie jakości sygnału audio, tj. automatycznej klasyfikacji wybranych zniekształceń w sygnałach muzycznych. Główną zaletą proponowanego rozwiązania, w porównaniu do innych dostępnych w literaturze dla tego typu sygnałów, jest brak konieczności porównania do tzw. sygnału referencyjnego, który podczas walidacji pozostaje często niedostępny. Znane aktualnie automatyczne metody oceny jakości dźwięku, nie wymagające takiego porównania, dotyczą głównie sygnałów mowy, podczas gdy analiza sygnałów muzycznych, ze względu na ich różnorodność, jest bardziej złożona.

Rozprawa zawiera opis badań przeprowadzonych przez autorkę, będących procesem opracowania prototypowego modelu automatycznej klasyfikacji wybranych zniekształceń sygnałów muzycznych. Na wstępie opisane zostały najważniejsze informacje i problemy związane z manualną i automatyczną analizą jakości dźwięku. Następnie przedstawiono etap projektowania, implementacji oraz ewaluacji modeli sieci neuronowych. Zawarto wyniki badań nad skutecznością zaimplementowanych modeli klasyfikacyjnych w zależności od zastosowanej architektury oraz wybranych danych wejściowych. Ostatnia część pracy zawiera omówienie uzyskanych wyników wraz z podsumowaniem.

Integralnym efektem pracy badawczej wykonanej w ramach niniejszej rozprawy jest także opracowanie własnej bazy danych, umożliwiającej ewaluację zaimplementowanego modelu sieci neuronowych, tj. zbioru sygnałów reprezentujących zniekształcenia kategorii wybranych na podstawie rekomendacji ITU BS.1284-2. Zbiór ten został utworzony na bazie niezniekształconych nagrań sygnałów muzycznych MUSDB18.

Głównym celem opracowywanej metody jest potencjalne usprawnienie kosztownych i czasochłonnych manualnych testów odsłuchowych, które ze względu na wysoką skuteczność, wciąż stanowią najpopularniejszą metodę oceny jakości dźwięku. Celem niniejszej pracy nie było natomiast ich zastąpienie – na ten moment nawet najlepsze metody referencyjne nie odzwierciedlają w pełni wyników subiektywnych testów odsłuchowych. Ograniczenie zakresu testowanego materiału metodą subiektywną poprzez wcześniejszą automatyczną detekcję i klasyfikację wybranych zniekształceń umożliwia przyśpieszenie i zmniejszenie kosztów procesu manualnego testowania oraz potencjalnej naprawy błędów badanego toru przetwarzania audio.

Według najlepszej wiedzy autorki, nie została jeszcze opracowana rekomendowana metoda oceny jakości rzeczywistych sygnałów muzycznych w sposób automatyczny, bez porównania do referencji, a proponowany w niniejszej pracy model jest pierwszym opublikowanym zastosowaniem konwolucyjno-rekurencyjnych sieci neuronowych do zadania automatycznej klasyfikacji zniekształceń sygnałów muzycznych bez porównania do sygnału referencyjnego.

Abstract

This thesis focuses on the usage of convolutional-recurrent neural networks in the field of audio quality analysis: automatic artefacts classification for a real music content. The key contribution in this approach, compared to the existing research, is that the examined model is evaluated in terms of detecting acoustic anomalies without the comparison to the known reference signal, since it is often unavailable at the time of validation. Currently, non-reference audio quality assessment methods are mainly developed for speech only. However, real music signals are more complex to analyze, since they may include instrumental sounds, speech, singing voice as well as various audio effects.

The thesis describes the research conducted by the author, as the process of developing the prototype model for automatic audio artefacts classification. First, we describe the most important information and problems related to the manual and automatic audio quality assessment. Then, we present all steps related to designing and implementing the artificial neural network model, which was evaluated in terms of various architectures and input parameters. The last section includes a discussion about the obtained results, and conclusions.

The integral part of the performed research is also the creation of the custom database, which was used for the process of model evaluation. The database is a set of signals with artefacts selected based on ITU Recommendation BS.1284-2. A publicly available dataset with clear (unprocessed) music signals MUSDB18 was used as a basis.

The main purpose of this thesis was to improve as much as possible the expensive and time-consuming process of manual listening tests. These are still the most popular way of assessing the audio quality, because of their high effectiveness for real music signals. The examined prototype model is not proposed to be a replacement for them – at this point, even the best known reference methods do not provide such accurate results. However, limiting the scope of manual tests cases by classifying the selected artefacts first, can speed up the process of manual testing and fixing issues in the examined audio processing chain.

To the best of the author's knowledge, there is no existing recommended method of automatic audio quality assessment for real music content without the comparison to the known reference signal, and the prototype model examined in this research is the first published usage of convolutional-recurrent neural networks for the task of automatic non-reference artefacts classification in real music signals.

Wykaz wybranych skrótów anglojęzycznych

AAC-LC	–	Advanced Audio Coding Low Complexity
ACC	–	Accuracy
ALAC	–	Apple Lossless Audio Codec
AMI	–	Amazon Machine Image
AWS	–	Amazon Web Service
BPTT	–	Back Propagation Through Time
CBRNN	–	Convolutional Bidirectional Recurrent Neural Network
CNN	–	Convolutional Neural Network
CQS	–	Continious Quality Scale
DFT	–	Discrete Fourier Transform
EBU	–	European Broadcast Union
EC2	–	Amazon Elastic Compute Cloud
FFT	–	Fast Fourier Transform
FLAC	–	Free Lossless Audio Codec
FNR	–	False Negative Ratio
FS	–	Full Scale
GRU	–	Gated Recurrent Unit
HAS	–	Human Auditory System
HE-AAC	–	High-Efficiency Advanced Audio Coding
IDCT	–	Inverse Discrete Cosine Transform
ISO	–	International Standards Organisation
ITU	–	International Telecommunication Union
LKFS	–	Loudness, K-weighted, relative to full scale
LRA	–	Loudness Range
LSB	–	Least Significant Bit
LSTM	–	Long-Short Term Memory
LU	–	Loudness Unit
MAF	–	Minimum Audible Field
MAP	–	Minimum Audible Pressure
MLP	–	Mulit-Layer Perceptron
MOS	–	Mean Opinion Square
MP3	–	MPEG Audio Layer 3
MRS	–	Music Recommendation Systems
ODG	–	Objective Difference Grade
PCM	–	Pulse Code Modulation
PEAQ	–	Perceptual Evaluation of Audio Quality
PSM	–	Perceptual Similarity Level
ReLu	–	Rectified Linear Unit
RNN	–	Recurrent Neural Network
S3	–	Amazon Simple Storage Service
SAR	–	Signal to Artifact Ratio
SDG	–	Subiective Difference Grade
SDR	–	Signal To Distortion
SIR	–	Signal To Interference
SNR	–	Signal-to-Noise Ratio
SPL	–	Sound Pressure Level
SPL	–	Sound Pressure Level
STFT	–	Short Time Fourier Transform
STFT	–	Short-Time Fourier Transform
TanH	–	Hyperbolic Tangent Function
THD	–	Total Harmonic Distortion
TN	–	True Negative
TNR	–	True Negative Ratio
TP	–	True Positive
VAD	–	Voice Activity Detection

Wykaz wybranych oznaczeń

N	–	całkowita liczba próbek;
n	–	indeks kolejnych próbek sygnału dyskretnego;
k	–	całkowita liczba kategorii sygnałów audio;
l	–	oczekiwane oznaczenie sygnału w bazie danych wejściowych (ang. *label*);
$x(n)$	–	wartość n-tej próbki sygnału w dziedzinie czasu;
f	–	częstotliwość sygnału;
ω	–	funkcja okienkowania
$\text{sign}(x(n))$	–	znak dla wartości chwilowej sygnału x
$\hat{y}$	–	wynik predykcji modelu sieci neuronowych
f_a	–	funkcja aktywacji
w	–	macierz wag wejściowych dla modelu sieci neuronowych
v	–	macierz wag rekurencyjnych dla modelu sieci neuronowych
f_s	–	częstotliwość próbkowania
T_s	–	okres próbkowania
S_n	–	przesunięcie/skok (ang. *stride*)
ρ	–	funkcja nieliniowa
b	–	wektor obciążeń dla modelu sieci neuronowych (ang. *bias*)
$X[j,t]$	–	składowa dyskretnej krótko-czasowej transformaty Fouriera o indeksie j i numerze ramki t
$X_{\text{mel}}[j,t]$	–	składowa spektrogramu w skali melowej o indeksie j i numerze ramki t
M	–	funkcja melowa
L	–	liczba składowych sinusoidalnych w sygnale wieloczęstotliwościowym
h	–	model sieci neuronowych
d	–	wymiar pojedynczej ramki sygnału
j	–	indeks pojedynczej ramki sygnału

Wprowadzenie

Cyfrowe przetwarzanie oraz transmisja sygnałów audio są nieustannie usprawniane w celu uzyskania jak największej prędkości strumieniowania danych, przy jednoczesnym zachowaniu wysokiej jakości dźwięku. Każda modyfikacja w łańcuchu przetwarzania dźwięku – rozpoczynając od etapu tworzenia ścieżki dźwiękowej, a kończąc na odsłuchu sygnału przy pomocy urządzenia konsumenckiego – może wprowadzać nieoczekiwane zniekształcenia.

Obecne metody oceny jakości sygnału audio podzielić można na dwie grupy: subiektywne i obiektywne. Pierwsza z nich – wciąż najbardziej popularna – polega na przeprowadzeniu testów odsłuchowych, zazwyczaj wśród grupy wyszkolonych słuchaczy. Metoda ta wyróżnia się najwyższą skutecznością, jednak ze względu na to, że jest to metoda manualna, jest ona czasochłonna, a do tego podatna na błędy wynikające z subiektywnego charakteru testów. Nie jest również możliwe sprawdzenie całego przetwarzanego materiału audio – do testów wybierane są określone sygnały, mające uwzględnić najbardziej popularne oraz skrajne przypadki testowe. Ograniczenie testowanego materiału do odsłuchu jest konieczne, jednakże wciąż nie rozwiązuje problemu analizy jakości ogromnej ilości, bardzo zróżnicowanych obecnie sygnałów audio.

Alternatywą testów subiektywnych mogłyby być metody w pełni automatyczne. Ze względu na to, że ich celem jest jak najdokładniejsze odwzorowanie subiektywnej oceny słuchacza, są one bardzo trudne w implementacji. Algorytmy oparte na prostej analizie widma częstotliwościowego lub też przebiegu czasowego sygnału audio, nie są w stanie zapewnić równie wysokiej skuteczności, jak subiektywne testy odsłuchowe, ponieważ nie uwzględniają one aspektów psychoakustycznych, a te mają istotny wpływ na odbiór dźwięku przez słuch ludzki. Dzięki poznanym dotąd mechanizmom, w jaki sposób odbierany jest dźwięk i jak wpływa na nie m.in. budowa ucha, możliwe było np. wprowadzenie kompresji stratnej, która pomimo istotnej redukcji wielkości kodowanego materiału, pozwala na zachowanie wysokiej jakości dźwięku. W takim przypadku, rezultatem prostego porównania ze sobą dwóch sygnałów (np. ich widma częstotliwościowego lub przebiegów czasowych), gdzie pierwszy z nich – sygnał „referencyjny" to sygnał oryginalny (np. przed kompresją), drugi – enkodowany za pomocą kodeka stratnego, a następnie dekodowany do podstawowego formatu PCM (ang. *Pulse Code Modulation*), byłyby zauważalne różnice, które jednocześnie mogłyby być zupełnie niesłyszalne podczas subiektywnych testów odsłuchowych. Z tego względu opracowane zostały inne metody, polegające na porównaniu sygnału testowego do referencyjnego na podstawie modelu psychoakustycznego (np. popularna metoda PEAQ, ang. *Perceptual Evaluation of Audio Quality*). Główną wadą takiego rozwiązania jest konieczność dostępu do sygnału referencyjnego, co często nie jest możliwe. Dla tego typu przypadków rozwiązaniem byłoby wprowadzenie obiektywnej metody w pełni automatycznej, która jednocześnie pozwalałaby na wykrycie zniekształceń sygnału bez konieczności porównania do innego sygnału, bez założeń odnośnie parametrów i zawartości audio, używając wyłącznie aktualnego sygnału testowego.

Większość znanych obecnie metod automatycznej analizy jakości sygnałów audio opiera się na porównaniu sygnału do referencji (ang. *intrusive metrics*), a metody bezreferencyjne (ang. *non-intrusive*, *non-reference* lub *single-ended metrics*) są dostępne przede wszystkim dla sygnału mowy. W niniejszej pracy natomiast głównym założeniem jest analiza i detekcja

zniekształceń dla rzeczywistych sygnałów muzycznych. Według najlepszej wiedzy autorki, nie została jeszcze opracowana rekomendowana metoda oceny jakości rzeczywistych sygnałów muzycznych w sposób automatyczny, bez użycia referencji. W niniejszej pracy została więc opisana proponowana obiektywna metoda, stanowiąca prototyp do dalszych badań, pozwalająca na wykrycie wybranych zniekształceń w sygnałach muzycznych, bez konieczności porównania do sygnału referencyjnego.

Zadanie detekcji zniekształceń wiąże się z kilkoma istotnymi problemami. Po pierwsze, niemożliwym jest zrekonstruowanie wszystkich potencjalnych zniekształceń sygnałów, podobnie jak zebranie próbek muzycznych pokrywających wszystkie możliwe konfiguracje, ze względu na różnorodność istniejących materiałów dźwiękowych. Po drugie, nawet w kontekście manualnych testów odsłuchowych występują przypadki, gdzie bez dostępu do sygnału referencyjnego, słuchacz nie jest w stanie określić, czy dany fragment dźwiękowy jest poprawny (np. jeśli zastosowane zostały specjalne efekty dźwiękowe, zwłaszcza w muzyce elektronicznej).

Proponowana w pracy metoda opiera się na modelu sieci neuronowych z wykorzystaniem warstw konwolucyjnych i rekurencyjnych dwukierunkowych, szczególnie popularnych obecnie w przetwarzaniu obrazów. Metoda ta została przeanalizowana pod kątem potencjalnego usprawnienia manualnych testów odsłuchowych. Celem niniejszej pracy nie było natomiast ich zastąpienie – na ten moment nawet najlepsze metody referencyjne nie odzwierciedlają w pełni wyników subiektywnych testów odsłuchowych.

Głównym założeniem pracy była możliwość detekcji zniekształceń z czterech wybranych kategorii według rekomendacji BS.1284-2, opracowanej przez Międzynarodowy Związek Telekomunikacyjny ITU (ang. *International Telecommunication Union*). Rekomendacja ta określa w sumie 11 kategorii, które mogą być użyte do analizy i klasyfikacji typów zniekształceń sygnałów audio, jednak część z nich dotyczy sygnałów wielokanałowych, gdzie problemy mogą wynikać z zależności pomiędzy poszczególnymi kanałami (np. przesłuchy lub też zniekształcenia percepcji przestrzenności dźwięku). Na tym etapie pracy, nie było celem badanie takich zależności, a każdy kanał audio analizowany był niezależnie, jest to jednak brane pod uwagę jako przedmiot dalszych badań. W pracy skupiono się na wyraźnych zniekształceniach, które w subiektywnych testach odsłuchowych ocenione zostałyby jako bardzo przeszkadzające. Celem było sprawdzenie, jak wybrana architektura modelu sieci neuronowych oraz typy danych wejściowych wpływają na skuteczność detekcji prostych zniekształceń oraz ocena, czy dalszy rozwój takiej metody (np. poprzez rozszerzenie bazy danych o kolejne zniekształcenia, mniej przeszkadzające lub prawie niesłyszalne) miałby w tym przypadku uzasadnienie.

Praca składa się z jedenastu rozdziałów. Na wstępie opisana została ogólna charakterystyka jakości sygnału audio wraz z omówieniem aspektów psychoakustycznych wpływających na jego ocenę oraz przykładowego procesu przetwarzania sygnału audio (rozdz. 1). W tej części wyjaśniono, dlaczego proste metody porównawcze nie znajdują zastosowania w ocenie jakości rzeczywistych sygnałów audio oraz uwzględniono przykłady wpływu cyfrowego przetwarzania i transmisji sygnału na jakość dźwięku. W kolejnym rozdziale (rozdz. 2) przedstawiona została manualna metoda oceny jakości dźwięku – subiektywne testy odsłuchowe. Opisana została procedura testowa, stosowane skale oceny jakości audio zalecane przez organizację ITU, wymagane zasoby do przeprowadzenia testów

oraz główne problemy wynikające z manualnego i subiektywnego charakteru tej metody. Następnie dokonano przeglądu istniejących automatycznych metod oceny jakości audio, ze szczególnym uwzględnieniem algorytmów do analizy rzeczywistych sygnałów muzycznych (rozdz. 3). Autorskie rozwiązania zawarte w pracy przestawiono w kolejnych rozdziałach (rozdz. 4 – 10). Rozdział czwarty zawiera opis opracowanej metody modelu sieci neuronowych – automatycznego klasyfikatora sygnału audio bez konieczności porównania do sygnału referencyjnego. Kolejne trzy rozdziały to odpowiednio: opis przygotowanej bazy danych oraz przetwarzania wstępnego sygnałów do analizy wraz z wprowadzeniem teoretycznym dla zastosowanych parametrów sygnałów audio; przedstawienie projektu modelu – opis wybranych rozwiązań oraz architektura modelu; miary ewaluacji modelu oraz wymagania sprzętowe. Rozdziały 8 – 10 zawierają wyniki ewaluacji dla zaimplementowanych w pracy modeli sieci neuronowych. Porównana została skuteczność ich klasyfikacji w zależności od zastosowanej architektury oraz danych wejściowych. W ostatnim rozdziale zawarte zostało podsumowanie, wnioski na podstawie wszystkich uzyskanych wyników klasyfikacji sygnałów audio oraz przedstawione zostały możliwe aspekty dalszych badań w celu poprawy aktualnej implementacji modelu lub zastosowania go jako model typu *transfer-learning* dla podobnych zadań.

Teza, cel i zakres pracy

Teza

Zastosowanie warstw dwukierunkowych rekurencyjnych w modelu sieci neuronowych wraz z odpowiednim doborem jego architektury, parametrów wejściowych oraz opracowaniem bazy zniekształceń do celów ewaluacji modelu, zwiększa znacząco skuteczność automatycznej klasyfikacji wybranych zniekształceń sygnałów muzycznych bez konieczności porównania do sygnału referencyjnego.

Cel pracy

Celem pracy jest opracowanie prototypowego modelu neuronowych sieci konwolucyjno-rekurencyjnych oraz zbadanie jego skuteczności w celu klasyfikacji zniekształceń rzeczywistych sygnałów muzycznych. Głównymi założeniami opracowywanego modelu jest możliwość wykrywania wybranych zakłóceń w sposób automatyczny, obiektywny oraz bez konieczności porównania sygnału testowego do referencyjnego. Uzyskany wynik powinien zapewnić podstawę do dalszych badań, w taki sposób, że może on być zastosowany w celu rozszerzenia bazy wykrywanych zniekształceń, lub też do innego pokrewnego problemu (*transfer learning*), co w przyszłości mogłoby usprawnić długotrwałe manualne testy odsłuchowe przez częściowe zastąpienie ich obiektywną metodą automatyczną.

Zakres pracy

Na zakres pracy składa się:

- analiza sposobu oceny jakości sygnałów muzycznych w praktycznych zastosowaniach (rozdz. 1);
- przegląd subiektywnych (rozdz. 2) i obiektywnych (rozdz. 3) metod wykorzystywanych do klasyfikacji cech sygnału muzycznego, umożliwiających wykrycie różnego rodzaju nieprawidłowości;
- opracowanie wstępnej metody, która umożliwiałaby detekcję wybranych zniekształceń w rzeczywistych sygnałach muzycznych w sposób obiektywny, automatyczny oraz bezreferencyjny (rozdz. 4);
- przygotowanie bazy sygnałów treningowych, testowych oraz walidacyjnych dla implementowanego modelu (rozdz. 5);
- projekt opracowanej metody z wykorzystaniem sieci neuronowych (rozdz. 6 – 7);
- wykonanie badań eksperymentalnych opracowanej metody wykorzystując chmurę obliczeniową AWS oraz analiza skuteczności modelu w zależności od zastosowanej architektury i danych wejściowych (rozdz. 8 – 10);
- analiza uzyskanych wyników w celu wykorzystania zaproponowanej metody w zastosowaniach praktycznych, szczególnie w kontekście usprawnienia manualnych subiektywnych testów odsłuchowych (rozdz. 11).

1. Ogólna charakterystyka oceny jakości sygnału audio

1.1. Główne trudności oceny jakości sygnału audio

Podstawą projektowania systemów przetwarzających sygnał audio jest poznanie, w jaki sposób działa słuch ludzki, a w szczególności jakie są jego ograniczenia. Najprostszym przykładem takich ograniczeń jest zakres częstotliwości słyszalnych przez człowieka (przyjmuje się, że jest to zakres ok. 20 – 20 000 Hz [1]). Te oraz inne ograniczenia opisane w poniższym rozdziale, mają istotny wpływ na wymagania stawiane podczas projektowania systemów audio. Jednocześnie mają też przełożenie na to, w jaki sposób oceniana jest jakość dźwięku przez danego słuchacza. Z tego względu subiektywne testy odsłuchowe są wciąż najbardziej wartościową metodą sprawdzania jakości sygnałów audio. Nie istnieje żadne inne narzędzie, które potrafiłoby skutecznie zastąpić wszystkie aspekty ludzkiej percepcji dźwięku [2]. Zrozumienie tego, w jaki sposób działa słuch ludzki jest istotnym elementem tworzenia m.in. cyfrowych kodeków percepcyjnych (ang. *perceptual audio codecs*), których działanie opiera się na zastosowaniu tzw. modelu psychoakustycznego. Na jego podstawie możliwa jest redukcja wymaganej ilości kodowanych danych, bez wprowadzania słyszalnych zniekształceń do sygnału [3]. Sygnał może być kodowany w sposób stratny (tj. część informacji zawartych w sygnale oryginalnym zostanie utracona), jednocześnie zapewniając bardzo dobrą jakość dźwięku, a różnice pomiędzy sygnałem oryginalnym oraz skompresowanym są często niesłyszalne podczas subiektywnych testów odsłuchowych.

Odwzorowanie, w jaki sposób działa słuch ludzki i odtworzenie subiektywnej oceny jest największym wyzwaniem dla algorytmów automatycznej oceny jakości sygnałów audio. Automatyczne wskazanie obiektywnych różnic pomiędzy dwoma sygnałami nie oznacza, że różnice te będą słyszalne przez słuch ludzki, a jeśli będą, to w jakim stopniu są one (subiektywnie) przeszkadzające. Nie mają tutaj zatem praktycznego zastosowania proste algorytmy, np. porównania bitowe przebiegów czasowych dwóch sygnałów audio lub ich widm częstotliwościowych. W idealnym przypadku, automatyczna ocena jakości sygnału powinna uwzględniać wszystkie aspekty psychoakustyczne, co wciąż stanowi wyzwanie dla aktualnie istniejących automatycznych metod.

Główne poznane dotąd aspekty psychoakustyczne, które mają istotny wpływ na odbiór dźwięku przez słuch ludzki to [3]:

- próg słyszalności;
- maskowanie dźwięku;
- głośność dźwięku i krzywe jednakowego słyszenia;
- pasma krytyczne;
- percepcja przestrzenności dźwięku;
- przyzwyczajenie do danego dźwięku i/lub zmęczenie.

Rozdz. 1.2 zawiera opis podstawowych ograniczeń słuchu ludzkiego i sposobu percepcji dźwięku, w celu przedstawienia głównych problemów, które muszą zostać uwzględnione podczas projektowania automatycznej obiektywnej metody oceny jakości sygnałów muzycznych. Stanowi to również uzasadnienie, dlaczego w niniejszej pracy wykorzystane zostały metody sztucznej inteligencji. Wyjaśnienie głównych aspektów kodowania i innych

najpopularniejszych zastosowań cyfrowego przetwarzania dźwięku zawarte zostało w rozdz. 1.3.

1.2. Wpływ psychoakustyki na percepcję dźwięku

Aspekty psychoakustyczne mają istotny wpływ na to, w jaki sposób dźwięk jest odbierany przez słuch ludzki oraz jak bardzo subiektywna jest jego percepcja w zależności od danego słuchacza. Na subiektywną ocenę jakości analizowanego dźwięku mają wpływ m.in.:

- zakres słyszalności dźwięku, który jest indywidualny dla każdej osoby, a jednocześnie zmienia się wraz z wiekiem;
- „rozdzielczość" słyszenia, czyli jak dobrze słuch ludzki jest w stanie słyszeć indywidualne składowe częstotliwościowe;
- odbierana głośność (krzywe jednakowego słyszenia);
- wysokość i barwa dźwięku;
- zjawisko maskowania;
- wrażenie przestrzenności;
- oraz pozostałe aspekty, takie jak zmęczenie czy nastrój.

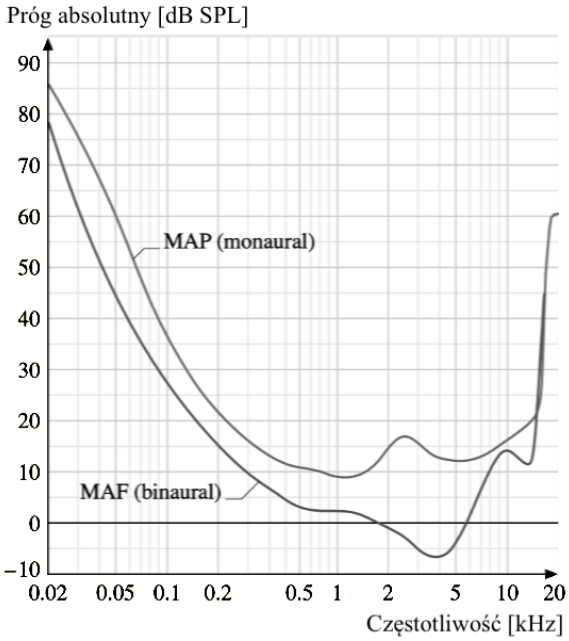
Poznanie właściwości słuchu ludzkiego było kluczowym czynnikiem pozwalającym stworzyć skuteczne algorytmy przetwarzania dźwięku, przede wszystkim na bazie kompresji stratnej, która mimo znacznej redukcji ilości kodowanych danych, wciąż pozwala na zachowanie stosunkowo dobrej jakości dźwięku (np. popularne do dziś kodeki AAC lub MP3). Automatyczne algorytmy weryfikujące jakość dźwięku wymagają więc uwzględnienia faktu, że zniekształcenia wykryte za pomocą prostej analizy widma częstotliwościowego, czy też przebiegu czasowego, będą odbiegać od subiektywnej oceny przez słuchacza. Manualne testy odsłuchowe są więc nadal istotne w procesie walidacji systemu przetwarzania i transmisji dźwięku. Celem niniejszej pracy nie jest więc ich całkowite zastąpienie, ale zbadanie czy metody oparte na sieciach neuronowych mogą umożliwić usprawnienie tych testów, poprzez ograniczenie ilości przypadków testowych niezbędnych do wykonania w sposób manualny. Sieci neuronowe, ze względu na to, że pozwalają na znaczną adaptację algorytmu w zależności od zastosowanej architektury oraz przygotowanej bazy danych (np. możliwość dostosowania wag na podstawie subiektywnej oceny słuchacza) wydają się być odpowiednim rozwiązaniem na problemy innych algorytmów, nie uwzględniających aspektów psychoakustycznych
w żaden sposób.

1.2.1. Obszar słyszalności

Jednym z aspektów psychoakustycznych wpływających na percepcję dźwięku przez słuch ludzki jest tzw. obszar słyszalności, ograniczony przez dwie krzywe progowe – próg słyszalności oraz próg słyszenia bolesnego [4]. Próg słyszalności definiowany jest jako dolna granica słyszalności dźwięku przy jednoczesnym braku innych zewnętrznych dźwięków [5]. Próg słyszenia bolesnego uznawany jest natomiast za górną granicę, określającą poziom ciśnienia akustycznego wywołującego wrażenie bólu podczas słuchania. Słuch ludzki jest najbardziej czuły dla częstotliwości około 3 kHz (rys. 1.1) i w tym obszarze dźwięki będą

słyszalne lepiej, nawet jeśli będą one miały niższy poziom ciśnienia akustycznego. Natomiast przy poziomie ciśnienia akustycznego około 120 dB pojawia się wrażenie bólu i ryzyko uszkodzenia słuchu.

Do wyznaczenia obszaru słyszalności stosowana jest jedna z dwóch metod. Pierwsza z nich, tzw. MAP (ang. *Minimum Audible Pressure*), polega na pomiarze poziomu ciśnienia akustycznego przy wejściu do kanału słuchowego za pomocą małego mikrofonu. Druga metoda, tj. MAF (ang. *Minimum Audible Field*), polega na pomiarze poziomu ciśnienia akustycznego wykorzystując zestaw głośnikowy oraz komorę bezechową, a poziom mierzony jest na pozycji głowy słuchacza [5]. Obie te metody wykorzystują jako źródło sygnał sinusoidalny, a długość trwania dźwięku jest nie krótszy niż 200 ms.



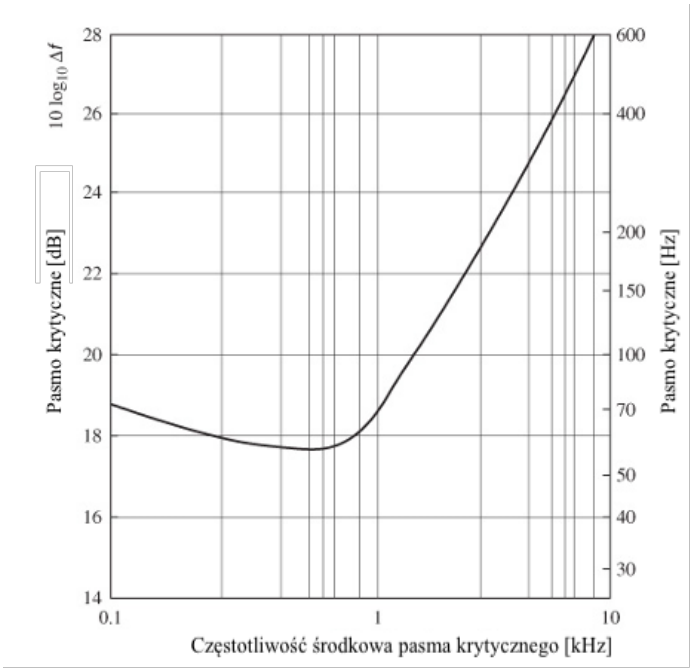
Rys. 1.1. Próg słyszalności [1].

Obszar słyszenia nie jest jednak jedynym aspektem, który wpływa na to, czy dany dźwięk będzie słyszalny dla danego słuchacza. Znane jest także zjawisko maskowania (rozdz. 1.2.3), które jest szczególnie istotne podczas implementacji stratnych kodeków audio. Kolejny rozdział (rozdz. 1.2.2) zawiera opis tzw. pasm krytycznych i stanowi on wstęp do przedstawienia zjawiska maskowania.

1.2.2. Pasma krytyczne

Koncepcja pasm krytycznych powstała dzięki badaniom zjawiska, że słuch ludzki nie jest jednakowo czuły na wszystkie dźwięki, a wynika to z jego budowy. Dźwięk dochodzący z przewodu słuchowego wywołuje drgania błony bębenkowej, które następnie przekazywane są do ślimaka. Ruch płynu w ślimaku oraz ruch błony podstawnej powodują wzbudzanie tzw. stereociliów [4]. Są to drobne rzęski, które stanowią zakończenie komórek słuchowych zewnętrznych (jest ich około 15 000 w narządzie Cortiego, na których znajduje się 140 stereocilia) oraz wewnętrznych (3 500, na których znajduje się około 40 stereocilia). W wyniku

pobudzania włoskowych komórek nerwowych do mózgu przekazywany jest sygnał w postaci impulsów nerwowych. Częstotliwość sygnału, który dociera do słuchacza decyduje o tym, w jaki sposób zostaną pobudzone stereocilia, a ich rozmieszczenie w stałych odległościach wpływa na percepcję wysokości dźwięku [6]. Dzięki zbadaniu tych zależności, wprowadzony został podział skali częstotliwościowej na 24 (rozłączne) części, zwane pasmami krytycznymi [7]. Szerokość pasm krytycznych w zakresie od 0 do 500 Hz jest w przybliżeniu stała, natomiast przy większych częstotliwościach rośnie w przybliżeniu liniowo (rys. 1.2).



Rys. 1.2. Pasma krytyczne dla odsłuchu monouralnego (Goldberg and Riek, 2000) [8].

Działanie wielu kodeków percepcyjnych oparte jest na maskowaniu amplitudy w pasmach krytycznych w celu redukcji rozdzielczości bitowej [8]. Pasma krytyczne mają też istotne znaczenie m.in. dla percepcji wysokości dźwięku, głośności, fazy sygnału i zrozumiałości mowy.

1.2.3. Maskowanie

Efekt maskowania definiowany jest jako proces, w wyniku którego próg słyszalności jednego dźwięku jest podniesiony w obecności innego dźwięku (maskującego) [1].

Pierwszym przykładem efektu maskowania jest sytuacja, gdzie sygnał maskujący oraz maskowany występują jednocześnie i są quasi-stacjonarne – jest to tzw. maskowanie jednoczesne (ang. *simultaneous masking*) [6]. To jak bardzo sygnał zostanie zamaskowany zależy od struktury sygnałów (maskującego i maskowanego). Dla przypadku, gdy sygnałem maskującym jest pojedynczy ton, a sygnałem maskowanym sygnał przypominający szum, wpływ na to jak bardzo sygnał zostanie zamaskowany ma częstotliwość maskera [6]:

$$M \approx \left(15.5 + \frac{c}{Bark}\right) \text{dB} \tag{1.1}$$

gdzie:
M – estymowana wielkość maskowania, wyrażona w dB;
c – częstotliwość pasma krytycznego maskera.

Efekt maskowania dla sygnałów o zbliżonych częstotliwościach tłumaczony jest tym, że podczas gdy błona podstawna ucha jest wystawiana na działanie głośniejszego dźwięku, jej reakcja będzie mniejsza na cichszy dźwięk o częstotliwości w tym samym paśmie krytycznym [1]. W innej sytuacji, gdy sygnał o pojedynczej częstotliwości byłby maskowany przez sygnał przypominający szum, to efekt maskowania jest prawie niezależny od częstotliwości [6]. Jeśli poziom ciśnienia akustycznego sygnału maskującego jest większy o 5 dB od maskowanego, to maskowany sygnał staje się niesłyszalny.

Innym przykładem zjawiska maskowania jest tzw. efekt „*cocktail party*", gdzie będąc otoczonym tłumem rozmawiających ze sobą ludzi, słuch ludzki jest w stanie skoncentrować się na konkretnym rozmówcy i zignorować pozostałe, mniej ważne w tym momencie komunikaty.

Efekt maskowania nie jest jednak ograniczony do sytuacji, gdzie sygnał maskowany oraz maskujący występują jednocześnie. Tzw. maskowanie niejednoczesne (ang. *non-simultaneous masking* lub *temporal masking*) polega na tym, że sygnał może być zamaskowany przez inny dźwięk, występujący na krótko przed pojawieniem się maskera (ang. *pre-masking*) lub tuż po nim (ang. *post-masking*) [1]. Oba zjawiska trwają stosunkowo krótko, dla efektu *pre-masking* jest to 20 ms, *post-masking* 150 ms, jednakże wciąż mogą mieć wpływ na ogólny odbiór sygnału.

Istnieją również inne zjawiska związane z efektem maskowania, jak np. maskowanie binauralne (ang. *binaural masking*). Maskowanie binauralne występuje w sytuacji, gdzie sygnały podawane na lewy i prawy kanał różnią się od siebie, np. przesunięciem fazowym, jednak każdy z nich zawiera sygnał maskowany oraz maskujący. Z przeprowadzonych badań wynika [5], że w przypadku gdy faza jednego z sygnałów jest odwrócona w stosunku do drugiego, próg detekcji może zostać obniżony nawet o 15 dB. Jest to więc ciekawy przykład, jak duży wpływ na ogólny odbiór sygnału ma słyszenie dwuuszne oraz sam aspekt przestrzenności dźwięku. W kontekście zaimplementowanego modelu sieci neuronowych – automatycznego klasyfikatora sygnału z eliminacją porównania do sygnału referencyjnego (rozdz. 4) – może to być istotnym tematem dalszych badań. Jednak na obecnym etapie pracy, ewaluacja modelu uwzględniała jedynie sygnały w postaci pojedynczej ścieżki dźwiękowej, bez uwzględnienia wzajemnej relacji pomiędzy poszczególnymi kanałami (w przypadku nagrań wielokanałowych). Nie zostały więc opisane szerzej aspekty psychoakustyczne z tym związane.

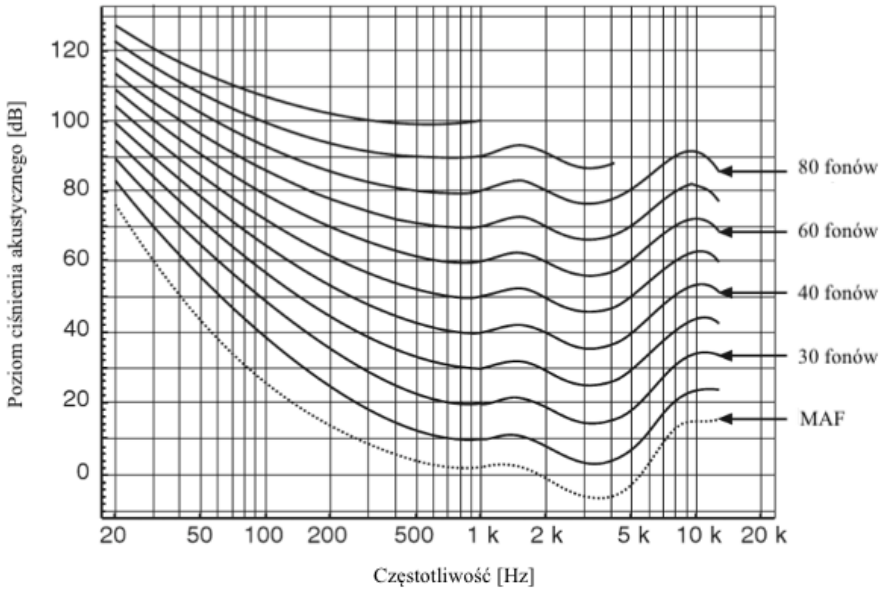
1.2.4. Głośność dźwięku

Na głośność dźwięku odbieraną przez słuch ludzki silnie wpływa częstotliwość, poziom ciśnienia akustycznego oraz czas jego trwania. W kontekście oceny jakości sygnałów audio istotnym jest również fakt, że głośność sygnału złożonego z kilku składowych będzie postrzegana jako mniejsza niż głośność każdej składowej, odsłuchiwanej osobno. Według

rekomendacji ITU BS.1387-1 [6], głośność zniekształcenia wprowadzonego do sygnału, może zostać zredukowana przez tzw. efekt „częściowego maskowania” (ang. *partial masking*).

1.2.5. Krzywe jednakowego słyszenia

Głośność dźwięku odbierana przez słuch ludzki zmienia się nie tylko w zależności od poziomu ciśnienia akustycznego, ale również w zależności od jego składowych częstotliwościowych [4]. Krzywe jednakowego słyszenia obrazują silną zależność pomiędzy głośnością dźwięku oraz jego częstotliwością. Poziom głośności 20 fonów może wystąpić przy poziomie ciśnienia akustycznego 20 dB dla częstotliwości 1 kHz (rys. 1.3). Natomiast by uzyskać ten sam poziom głośności (20 fonów) dla częstotliwości 10 kHz, potrzebny jest wyższy poziom ciśnienia akustycznego o ok. 10 dB. W kontekście sygnałów muzycznych, szczególnie istotnym jest fakt, że przy stosunkowo niskich poziomach głośności, słuch ludzki jest mniej czuły na mniejsze częstotliwości (mogą to być np. dźwięki basu) niż częstotliwości ze środkowego pasma. Poziom głośności ma więc istotny wpływ na to, jak odbierany jest dźwięk przez danego słuchacza, szczególnie w zakresie mniejszych częstotliwości [4].



Rys. 1.3. Krzywe jednakowego słyszenia według standardu ISO 226:2003 [9].

1.3. Wpływ cyfrowego przetwarzania i transmisji na jakość dźwięku

Każdy element łańcucha cyfrowego przetwarzania dźwięku, tj. jego produkcja, przetwarzanie wstępne (ang. *pre-processing*) i końcowe (ang. *post-processing*), kodowanie, transmisja oraz odtwarzanie, ma wpływ na jego jakość. Usługi związane z serwisami rozgłoszeniowymi oraz dostarczaniem materiałów multimedialnych stale ulepszają metody kodowania i przesyłania danych, w celu minimalizacji kosztów związanych przede wszystkim z czasem przetwarzania i transmisji, przy jednoczesnym zachowaniu jak najlepszej jakości dźwięku [10]. Każda modyfikacja łańcucha przetwarzania i transmisji audio, od momentu tworzenia materiału po urządzenie odbiorcze, może wprowadzać nieoczekiwane zmiany w sygnale, których wynikiem będą słyszalne zniekształcenia dźwięku.

Najskuteczniejszym sposobem oceny wpływu cyfrowego przetwarzania i transmisji na jakość dźwięku jest porównanie sygnału wynikowego do sygnału źródłowego (oryginalnego), jednakże w wielu przypadkach sygnał ten nie jest dostępny np. podczas walidacji sygnału po stronie odbiorczej lub też po zastosowaniu różnego rodzaju algorytmów typu *post-processing* m.in. znacznej ingerencji w zakres dynamiki sygnału lub zmiany docelowej konfiguracji kanałów (np. *downmix*).

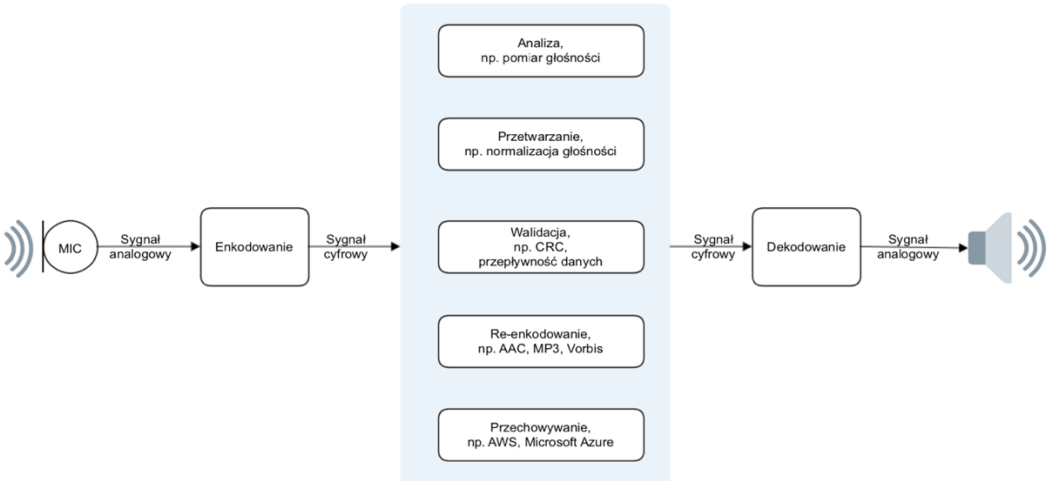
W kolejnych podrozdziałach (1.3.1 – 1.3.6) zawarto wprowadzenie do głównych elementów cyfrowego przetwarzania i transmisji sygnału audio wraz z nakreśleniem ich wpływu na ogólną jakość dźwięku.

1.3.1. Cyfrowe kodowanie dźwięku

Głównym celem cyfrowego kodowania dźwięku jest nie tylko transformacja sygnału dźwiękowego z postaci analogowej do cyfrowej, ale przede wszystkim możliwość reprezentacji sygnału przy wykorzystaniu minimalnej ilości danych, zachowując przy tym jak najwyższą jakość dźwięku [3]. Wyraz „kodek" łączy w sobie dwa elementy: „enkoder" oraz „dekoder" [11]. Etap enkodowania polega na przetworzeniu wejściowego sygnału analogowego do postaci cyfrowej i/lub skompresowanie danych zależnie od ustalonych parametrów (np. maksymalnej przepływności bitowej). Uzyskanie cyfrowej reprezentacji sygnału pozwala m.in. na:

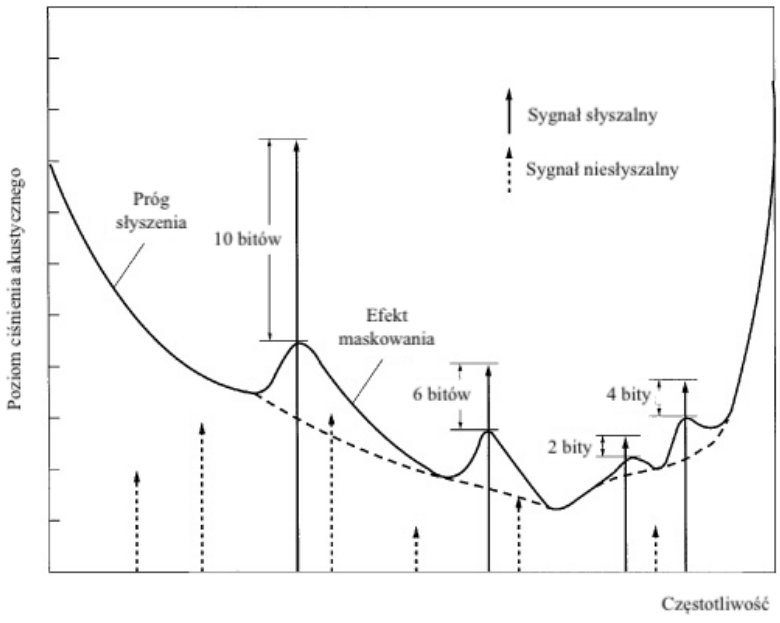
- dalszą konwersję sygnału za pomocą innych kodeków, np. w celu zwiększenia kompresji danych;
- dalsze cyfrowe przetwarzanie sygnału oraz jego analizę i walidację;
- transmisję sygnału i przechowywanie.

Końcowym etapem jest dekodowanie, tj. ponowne przetworzenie sygnału do postaci nieskompresowanej, a docelowo analogowej (rys. 1.4). Kodek audio jest więc urządzeniem sprzętowym lub algorytmem, którego wejściem jest sygnał audio, a wyjściem identyczny sygnał (lub bardzo podobny), ale jedynie pod względem percepcyjnym.



Rys. 1.4. Przykładowy tor cyfrowego przetwarzania dźwięku.

Wyróżnia się dwa rodzaje kodeków audio: bezstratne (ang. *lossless*) i stratne (ang. *lossy*). Pierwszy rodzaj to kodeki bazujące na modelach statystycznych obwiedni amplitudowej sygnału. Rekonstrukcja sygnału enkodowanego za pomocą kodeków bezstratnych pozwala na uzyskanie maksymalnej ilości danych, ograniczonych jedynie procesem próbkowania oraz kwantyzacją. Najpopularniejszym przykładem jest kodek PCM, gdzie w procesie enkodowania sygnał dźwiękowy próbkowany jest w regularnych odstępach czasu, a amplituda każdej próbki kwantowana w zależności od ustalonej rozdzielczości bitowej [3]. Kodeki tego typu pozwalają na wierne odtworzenie oryginalnego sygnału, jednakże główną ich wadą jest nadmiarowość danych. Z tego też względu wprowadzone zostały kodeki stratne, które mimo utraty części informacji, pozwalają zachować względnie dobrą jakość dźwięku (zależnie od zastosowanego algorytmu). Wykorzystują one zazwyczaj aspekty psychoakustyczne, których część została opisana w rozdz. 1.2.1 – 1.2.5. Dzięki wykorzystaniu informacji o sposobie percepcji dźwięku przez słuch ludzki możliwe jest uzyskanie wysokiej kompresji danych, a tym samym istotne zmniejszenie rozmiaru oraz wzrost prędkości transmisji kodowanych sygnałów. W przypadku kodeków percepcyjnych, długość słowa zależy od zastosowanego modelu HAS (ang. *Human Auditory System*). Oznacza to, że wykorzystanych jest jedynie tyle bitów, ile jest rzeczywiście niezbędne do reprezentacji słyszalnych składowych sygnału. Tony niesłyszalne, m.in. w wyniku efektu maskowania, nie będą więc enkodowane, a tony stosunkowo ciche będą wymagały mniejszej rozdzielczości bitowej (rys. 1.5). Dla porównania, stosując kodek bezstratny PCM, wszystkie próbki sygnału zapisywane są w jednakowej rozdzielczości bitowej, bez uwzględnienia powyższych aspektów.



Rys. 1.5. Przykład algorytmu alokacji bitów dla kodeka percepcyjnego [8].

Każda tego typu redukcja kodowanych danych wiąże się ze znacznym wzrostem ich przepływności (tab. 1.1), przy czym redukcja w skali 4:1 lub 6:1 może być zupełnie

niesłyszalna dla użytkownika [8]. Typowe wartości przepływności danych (ang. *data rate*), dla popularnych kodeków percepcyjnych mieszczą się w zakresie 64 – 128 kb/s.

Tab. 1.1. Wpływ redukcji rozdzielczości bitowej na przepływność danych dla częstotliwości próbkowania 48 kHz.

Skala redukcji	Liczba bitów na próbkę sygnału	Przepływność [kbps]
1:1	16	768
2:1	6	384
4:1	4	192
8:1	2	96

Obecnie znanych jest wiele kodeków, zarówno bezstratnych (np. FLAC, ALAC), jak i stratnych (np. AAC, MP3, Vorbis, Opus), które wyróżniają się przede wszystkim szybkością przetwarzania, rozmiarem wynikowego pliku i jakością dźwięku. Kodowanie stratne wykorzystywane jest obecnie znacznie częściej niż bezstratne, zwłaszcza przez systemy rozgłoszeniowe, platformy streamingowe, aplikacje mobilne itp. Wciąż wprowadzane są nowe rozwiązania w celu uzyskania coraz wyższej jakości dźwięku w połączeniu z wysoką kompresją danych. Kodowanie percepcyjne zrewolucjonizowało przetwarzanie i transmisję sygnałów audio, jednakże istotnym problemem jest trudność przeanalizowania jakości kodowanego w ten sposób dźwięku [12].

Podstawowe zniekształcenia sygnałów audio wynikające ze stosowania kodeków cyfrowych, również tych bezstratnych, to m.in. błędy kwantyzacji (np. gdy amplituda sygnału wejściowego przekracza maksymalny zakres kwantyzatora), problemy związane z niewłaściwym próbkowaniem (np. *aliasing*), a w przypadku kodeków stratnych, głównym źródłem zniekształceń może być sam algorytm kompresji danych, zwłaszcza jeśli ten sam sygnał jest enkodowany i dekodowany wielokrotnie.

1.3.2. Przetwarzanie typu downmix i upmix

Jednym z aspektów rozwoju technologii audio są badania nad jak najwierniejszym odwzorowaniem przestrzenności dźwięku. Od najprostszej konfiguracji pojedynczego kanału (mono), zaczęto stopniowo wprowadzać bardziej zaawansowane konfiguracje – początkowo dźwięk stereo, następnie dźwięk przestrzenny (ang. *surround sound*), w tym np. konfiguracja 5.1, po aktualnie popularny dźwięk typu *immersive*. Poza zwiększeniem liczby kanałów audio, wprowadzane algorytmy pozwalają na lepsze rozmieszczenie źródeł dźwięku oraz ich ruch w przestrzeni, dając wrażenie wielowymiarowości [13]. Jednakże ze względu na to, że wiele obecnych systemów dźwiękowych wciąż obsługuje jedynie podstawowe konfiguracje, część sygnałów dźwiękowych może wymagać dodatkowego przetwarzania w postaci tzw. *downmixu*.

Proces *downmixu* polega na zredukowaniu liczby kanałów audio, np. z konfiguracji 5.1 do stereo. Jest to szczególnie ważne dla materiału dźwiękowego przeznaczonego dla konfiguracji przestrzennych, który jednak ma zostać odtworzony jako stereo np. na słuchawkach. Przetwarzanie typu *upmix* natomiast jest procesem odwrotnym i polega na zwiększeniu ilości kanałów audio, np. utworzenie dźwięku przestrzennego z prostej konfiguracji stereo.

Zarówno proces typu *downmix* jak i *upmix* wpływają nie tylko na wynikową liczbę kanałów, ale także na głośność sygnału [12], a często też na jego ogólną jakość. Zależnie od zastosowanego algorytmu i przetwarzanego materiału mogą wprowadzać słyszalne zniekształcenia, np. znany jest problem występowania filtru grzebieniowego w przypadku zastosowania *downmixu* pasywnego dla sygnału z dużą liczbą kanałów [13].

1.3.3. Zmiany zakresu dynamiki sygnału

Kontrola zakresu dynamiki sygnału to proces automatycznego dostosowywania poziomu dźwięku, stosowany m.in. w telewizji czy produkcjach muzycznych [14]. Główne możliwości kontrolowania zakresu dynamiki uzyskuje się z pomocą m.in.: bramkowania, limitera, kompresora oraz ekspandera.

Bramkowanie (inaczej zastosowanie bramki szumów, ang. *noise gate*) polega na tym, że dla fragmentów sygnału, które nie przekraczają danego poziomu głośności, wyjście jest ustawiane na wartość 0 (czyli nieskończenie małe wzmocnienie). Dla pozostałych poziomów głośności, dane przekazywane są na wyjście bez zmian wzmocnienia.

Zadaniem limitera jest zapewnienie, że żaden fragment sygnału nie przekroczy zadanego progu poziomu głośności. Jeżeli poziom sygnału wejściowego będzie niższy od zadanego progu, to sygnał ten będzie przekazany na wyjście bez zmian. Jeżeli dany fragment będzie przekraczał tę wartość, algorytm limitera dostosowuje go do dopuszczalnej wartości maksymalnej.

Kompresor natomiast pozostawia sygnał niezmieniony, jeśli nie przekracza on zadanego poziomu głośności, natomiast dla fragmentów, które przekraczają ten próg, wyjściowy sygnał modyfikowany jest w taki sposób, żeby poziom sygnału wzrastał wolniej niż w sygnale wejściowym i w stałym tempie. Wynikiem działania kompresora jest redukcja zakresu dynamiki sygnału. Odwrotny efekt jest uzyskiwany za pomocą ekspandera, który powyżej zadanego poziomu głośności pozostawia sygnał niezmieniony, natomiast poniżej – poziom sygnału wzrasta szybciej niż w sygnale wejściowym, co wpływa na zwiększenie zakresu dynamiki.

Algorytmy modyfikujące zakres dynamiki sygnału mogą również wpływać na ogólny poziom głośności wyjściowego sygnału, stąd często wymagają zastosowania kolejnych algorytmów pozwalających na skompensowanie tych zmian, np. poprzez nałożenie dodatkowego wzmocnienia na sygnał (tzw. *makeup gain*). Przy nakładaniu różnego typu wzmocnienia należy mieć na uwadze konieczność zastosowania wygładzenia sygnału (ang. *gain smoothing*), ze względu na to, że fluktuacje wzmocnienia pomiędzy poszczególnymi próbkami sygnału mogą powodować słyszalne zniekształcenia [14].

1.3.4. Normalizacja poziomu sygnału

Normalizacja poziomu sygnału polega na dostosowaniu poziomu programu do ustalonej wartości, zazwyczaj wyznaczonej przez obowiązujący standard (np. w Europie obecnie obowiązującym standardem rozgłoszeniowym jest standard R128 zdefiniowany przez EBU, tj. *European Broadcast Union*). Standardy te powstają w celu zminimalizowania problemów powodowanych przez różnice w poziomach sygnałów pochodzących z różnych źródeł, np.

filmu i przerywającej go reklamy, ale przede wszystkim w celu zapewnienia jak najlepszej jakości i komfortu odsłuchu.

Normalizacja sygnału składa się zazwyczaj z dwóch kroków: pomiaru aktualnego poziomu dla całego sygnału, a następnie jego korekcji. Znane są dwie metody korekcji sygnału. Pierwsza z nich polega na modyfikacji materiału dźwiękowego w taki sposób, by odpowiadał on wyznaczonemu poziomowi. Druga – na ustawieniu odpowiedniego parametru metadanych w sygnale, co umożliwia dostosowanie poziomu sygnału na etapie końcowym, tj. podczas dekodowania [12]. W przypadku sygnałów kodowanych w sposób stratny (np. MP3), pierwsza metoda może wprowadzać znacznie więcej zniekształceń do sygnału, ze względu na to, że wymaga tzw. kaskadowego enkodowania i dekodowania, jednakże jest ona wciąż bardzo popularna.

Kolejnym ważnym aspektem normalizacji jest to, by zastosowany algorytm umożliwiał jak najmniejszą modyfikację sygnału. W idealnej sytuacji zmianie powinny ulec jedynie zadane parametry, np. docelowa głośność programu. Inne, jak np. zakres dynamiki, powinny pozostać bez zmian lub jak najbliższe oryginalnym wartościom.

1.3.5. Znakowanie wodne sygnałów fonicznych

Znakowanie wodne sygnałów fonicznych (ang. *audio watermarking*) to proces umieszczania specjalnych danych w istniejących cyfrowych sygnałach dźwiękowych. Dane te są zazwyczaj wykorzystywane po stronie odbiorczej i służą celom zabezpieczającym, np. ochronie praw autorskich utworów multimedialnych. W przypadku dźwięku, szczególnie ważnym jest, by tego typu proces nie wpływał na jakość dźwięku, a tym samym był możliwie niesłyszalny podczas odsłuchu.

Główne założenia systemu przetwarzania znaków wodnych dla sygnałów dźwiękowych to:

- niesłyszalność zawartego znaku wodnego;
- niezawodność algorytmu, tj. odporność na różnego typu zakłócenia, ataki, niekontrolowane modyfikacje sygnału;
- bezpieczeństwo, tj. zabezpieczenie materiału dźwiękowego przed nieautoryzowanym dostępem;
- odpowiedni niski limit ilości ukrytych danych, zazwyczaj podawana w jednostce bitów na sekundę lub na ramkę;
- odpowiednio mała złożoność obliczeniowa, co istotne jest szczególnie dla przetwarzania sygnałów w czasie rzeczywistym.

Istotnym problemem podczas projektowania tego typu algorytmów jest wzajemny wpływ poszczególnych kryteriów – próba poprawienia jednego z nich często wiąże się z pogorszeniem innego, np. zapisanie większej liczby bitów do danego sygnału, w celu zwiększenia niezawodności systemu na nieprzewidziane ataki, skutkować będzie wzrostem zniekształceń w sygnale [15].

Dostępnych jest wiele technik znakowania wodnego. Najpopularniejsze z nich opierają się na ograniczeniach wynikających ze sposobu percepcji dźwięku przez słuch ludzki [16]. Są to m.in. algorytm LSB (ang. *Least Significant Bits*), kodowanie fazowe (ang. *phase coding*),

widmo rozproszone (ang. *spread spectrum*), kodowanie echa (ang. *echo coding),* stosowanie bramek szumowych (ang. *noise gate technique*).

1.3.6. Transmisja sygnału audio

Transmisja sygnału audio pozwala na dostarczanie treści multimedialnych do odbiorcy. Łańcuch transmisji rozgłoszeniowej sygnału audio rozpoczyna się od producenta materiału dźwiękowego (w tym filmów, programów telewizyjnych, radiowych itd.), a kończy na urządzeniu odbiorczym po stronie słuchacza. Treści oraz usługi oferowane przez nadawców mogą być dostarczane za pomocą różnych mechanizmów transmisyjnych, które można podzielić na 2 kategorie: wykorzystujące sieci nadawcze lub sieci szerokopasmowe [17]. Pierwsze z nich stanowią tradycyjny sposób dystrybucji i stosowane są głównie dla stacji radiowych i telewizyjnych. Wymagają specjalistycznych urządzeń po stronie nadawcy i odbiorcy korzystających z sieci naziemnej, kablowej lub satelitarnej. Drugie, zazwyczaj związane są z sieciami opartymi na protokole IP i wykorzystywane przede wszystkim przez platformy streamingowe.

Przy zastosowaniu protokołu IP, sygnał wysyłany jest w postaci pakietów danych, a każdy z nich zawiera nagłówek (zawierający adres nadawcy i odbiorcy) oraz właściwe dane multimedialne. W tego typu transmisji głównym problemem jest opóźnienie czasowe lub utracenie danego pakietu. Z tego też względu częściej stosowany jest np. mechanizm *MPEG Transport Stream*, którego główną zaletą jest to, że zachowana jest kolejność strumieniowanych danych.

Dla programów oferowanych online stosowane są głównie dwie następujące metody udostępniania materiału dla użytkownika: serwis OTT (ang. *Over-The-Top*) lub serwis zarządzany (ang. *managed service*). Serwis OTT polega na udostępnieniu strony internetowej lub aplikacji, gdzie użytkownik ma możliwość odtworzenia materiałów multimedialnych (np. serwisy Netflix, YouTube). Ważnym aspektem dostarczania treści za pomocą OTT jest to, że jakość dostarczanych treści nie jest w pełni kontrolowana – użytkownicy otrzymują dane o zmiennej przepływności oraz czasie transmisji, zależnie od aktualnego obciążenia serwera oraz parametrów danego urządzenia odbiorczego (tzw. *best-effort service*). Serwisy zarządzane natomiast mają na celu usunięcie wad OTT i zapewnienie określonych parametrów, które gwarantować będą poprawną jakość audio i video po stronie odbiorcy, np. poprzez ustalenie minimalnej przepływności danych zależnie od sygnału.

2. Manualne metody oceny jakości audio

2.1. Wprowadzenie do manualnych testów odsłuchowych

Niniejszy rozdział ma na celu przedstawienie procedury manualnych, subiektywnych testów odsłuchowych. Uwzględnione zostały jedynie najpopularniejsze metody przeprowadzania testów, ustandaryzowane przez organizację ITU m.in. w dokumentach BS.1116-3 (2015) oraz BS.1284-2 (2019). Dokumenty te są stale uaktualniane wraz z pojawianiem się nowych wytycznych, a pierwsze ich wersje powstały w roku 1994 (BS.1116-0) i 1997 (BS.1284-0).

Zależnie od celu przeprowadzanego testu, procedura może różnić się od opisanej w poniższym rozdziale, a dokumenty opracowane przez ITU są jedynie rekomendacjami, nie ścisłymi formułami przeprowadzania testów. Istotnym jest też fakt, że podczas planowania i przeprowadzania testu należy zwrócić uwagę na wiele aspektów – poprzez odpowiednie sformułowanie celu testu, zebranie grupy wyszkolonych słuchaczy, dobranie odpowiedniej metody testowania, oprogramowania, urządzeń audio, po przeprowadzenie statystycznej analizy danych uzyskanych od słuchaczy i końcowych wyników. Ponadto, ważne jest również by przeprowadzane w taki sposób testy były jak najbardziej powtarzalne, a uzyskane wyniki dawały zbliżone wartości, co w manualnych testach subiektywnych wydaje się być szczególną trudnością, porównując do automatycznych testów obiektywnych.

2.2. Grupa słuchaczy

W przypadku manualnych testów odsłuchowych, rekomendacja BS.1284-2 określa, iż zawsze preferowani są słuchacze wyszkoleni, pomimo że wydawać by się mogło, że grupa niewyszkolona będzie lepszą reprezentacją ogólnej populacji. Z praktyki jednak wynika, że nawet osoby bez wcześniejszego przygotowania, w dłuższej perspektywie stają się słuchaczami wyszkolonymi, a to daje znacznie lepsze i szybsze rezultaty.

Celem szkolenia słuchaczy krytycznych jest przede wszystkim zwiększenie świadomości percepcji składowych dźwięku, poprawa szybkości rozpoznawania zmian lub nieoczekiwanych zniekształceń w sygnale oraz osiągnięcie powtarzalności wyników.

Kolejną kwestią jest liczba słuchaczy krytycznych. Według rekomendacji, minimalna liczba słuchaczy wyszkolonych na pojedynczy test to 10 osób, podczas gdy w przypadku słuchaczy niewyszkolonych jest ona aż dwukrotnie wyższa. Dodatkowo, jeśli system projektowany jest dla transmisji wysokiej jakości sygnału audio, rekomendowane jest, by testy wykonane zostały jedynie przez słuchaczy wyszkolonych. Słuchacze powinni również być zaznajomieni z procedurą testową, materiałami testowymi oraz ze środowiskiem testowym.

2.3. Procedura testowa

Zależnie od celu przeprowadzanego testu, procedura testowa składać się może z pojedynczej prezentacji sygnału, porównania dwóch sygnałów (gdzie zazwyczaj jeden z nich to sygnał referencyjny) lub też prezentacji wielu sygnałów (z lub bez referencji). Zastosowanie

mają tu skale przedstawione w rozdziale 2.4, a prezentacje sygnałów mogą być powtarzane słuchaczowi według potrzeby [18].

Główne aspekty tworzenia procedury testowej według rekomendacji ITU-R BS.1284-2 to m.in.:

- zalecane jest by czas trwania pojedynczego badanego sygnału nie przekraczał 15 – 20 sekund, jednocześnie może on być bardzo krótki (np. kilka sekund);
- jeśli badana sekwencja jest sygnałem muzycznym, to nie powinna ona być przerywana;
- przerwa pomiędzy kolejnymi prezentacjami (pierwszą i drugą oraz trzecią i czwartą) powinna wynosić od 0.5 do 1 sekundy, podczas gdy przerwa pomiędzy prezentacją drugą i trzecią powinna być dłuższa, np. 1.5 sekundy, jednakże zależy ona od typu przeprowadzanego testu;
- sekwencja sygnałów powinna być odtwarzana w losowej kolejności;
- urządzenie nie powinno wprowadzać żadnych słyszalnych zniekształceń podczas prezentacji i przełączania sygnałów.

2.4. Metoda testowa

Jedną ze stosowanych metod subiektywnej oceny jakości dźwięku (lub też wykrywania zakłóceń) jest stosowanie skali 5-stopniowej. Skala ta stosowana jest jako ogólna ocena jakości sygnału, jak i ocena występujących zniekształceń. Popularnym jej wariantem jest dyskretna skala jednobiegunowa (tab. 2.1 i tab. 2.2).

Tab. 2.1. Dyskretna skala jednobiegunowa oceny jakości sygnałów audio według ITU-R BS.1284-2 [18].

Ocena	Jakość sygnału audio
1	Znakomita
2	Dobra
3	Dostateczna
4	Słaba
5	Zła

Tab. 2.2. Dyskretna skala jednobiegunowa oceny zniekształceń sygnałów audio według ITU-R BS.1284-2 [18].

Ocena	Zniekształcenia sygnału audio
1	Niesłyszalne
2	Słyszalne, ale nieprzeszkadzające
3	Lekko przeszkadzające
4	Przeszkadzające
5	Bardzo przeszkadzające

Prócz powyższych, stosowane są również skale ciągłe CQS (ang. *Continious Quality Scale*) (zgodnie z rekomendacjami ITU BS.1534 oraz BT.500), mające zakres od 0 do 100 punktów i podzielone na pięć równych części (tab. 2.3). Stosowane są zazwyczaj w testach porównawczych.

Tab. 2.3. Ciągła skala jednobiegunowa oceny jakości sygnału audio według ITU-R BS.1284-2 [18].

Zakres ocen	Jakość sygnału audio
⟨80, 100⟩	Znakomita
⟨60, 80)	Dobra
⟨40, 60)	Dostateczna
⟨20, 40)	Słaba
⟨ 0, 20)	Zła

Dla testów porównawczych proponowana jest również skala dwubiegunowa 7-stopniowa (tab. 2.4).

Tab. 2.4. Dyskretna skala dwubiegunowa oceny jakości sygnału audio według ITU-R BS.1284-2 [18].

Ocena	Jakość sygnału audio
3	Znacznie lepsza
2	Lepsza
1	Trochę lepsza
0	Taka sama
-1	Trochę gorsza
-2	Gorsza
-3	Znacznie gorsza

Dla zamieszczonych powyżej skal, mogą być stosowane pewne punkty odniesienia (tzw. *anchor points*). Zwrócono jednak uwagę, że mogą one wprowadzać błędy przy interpretacji wyników [18]. Jeśli punkty odniesienia nie są więc stosowane, istotne jest, by wynik testu został znormalizowany, np.:

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s \tag{2.1}$$

gdzie:
x_i – wynik dla prezentacji i,
x_{si} – średni wynik prezentacji i w sesji s,
x_s – średni wynik wszystkich prezentacji w sesji s,
s_s – odchylenie standardowe wszystkich prezentacji podczas sesji s,
s_{si} – odchylenie standardowe prezentacji i dla sesji s.

2.5. Ocena jakości sygnału audio

Według rekomendacji ITU BS.1284-2 [18], w celu oceny jakości sygnałów audio powinny być uwzględnione następujące aspekty:

1) wrażenie przestrzenności dźwięku – subiektywne wrażenie wielowymiarowości dźwięku;
2) wrażenie stereofoniczności – subiektywne wrażenie, że dźwięk poszczególnych składowych propagowany jest z odpowiedniego kierunku w stronę słuchacza;
3) przezroczystość – subiektywna ocena, czy wszystkie składowe prezentowanego sygnału są odpowiednio słyszalne;

4) balans – subiektywna ocena równowagi poszczególnych składowych w ogólnym odbiorze prezentowanego sygnału;
5) barwa dźwięku – subiektywne wrażenie odpowiedniej barwy poszczególnych źródeł dźwięku;
6) brak szumu i zniekształceń – brak przeszkadzających składowych, takich jak szum środowiskowy, błędy rozdzielczości bitowej, zniekształcenia sygnału;
7) ogólne wrażenie – subiektywna średnia ważona ze wszystkich powyższych ocen, biorąc pod uwagę integralność dźwięku i zależności pomiędzy jego składowymi.

Wszystkie z wymienionych aspektów to subiektywne oceny dla danego słuchacza, silnie zależne od preferencji danej osoby. Ocena jakości dźwięku może się więc znacznie różnić wśród grupy słuchaczy. Najmniej uzależniony od indywidualnej oceny wydaje się być punkt szósty, dlatego też niniejsza praca skupia się właśnie na tym aspekcie. Jednakże i tutaj pojawić się mogą różnice związane z subiektywną granicą, gdzie zniekształcenia będą już słyszalne dla konkretnego słuchacza i na ile przeszkadzające w danym momencie.

Do analizy szumu i zniekształceń, które mogą wystąpić w cyfrowym przetwarzaniu sygnałów audio, proponowane jest w rekomendacji ITU-R BS.1284-2 11 kategorii zniekształceń:

1) defekty spowodowane kwantyzacją (ang. *quantisation defect*), tj. użyciem niewystarczającej rozdzielczości bitowej dla sygnału audio;
2) zniekształcenia charakterystyki częstotliwościowej (ang. *distortion of frequency characteristics*), np. brak składowych o dużych lub małych częstotliwościach w sygnale lub też nadmiar poszczególnych składowych, tj. *hissing* lub *comb-filter effect*;
3) zniekształcenia charakterystyki wzmocnienia (ang. *distortion of gain characteristics*), tj. m.in. zmiany (w tym skoki) poziomu dźwięku lub zakresu dynamiki sygnału;
4) efekty modulacji okresowej (ang. *periodic modulation effect*) tzn. okresowe zmiany amplitudy sygnału, np. *warbling*;
5) efekty modulacji nieokresowej (ang. *non-periodic modulation effect*), tj. przejściowe zniekształcenia sygnału np. *burst*;
6) zniekształcenia nieliniowe (ang. *non-linear distortion*) tj. zniekształcenia nieliniowe harmoniczne lub nieharmoniczne, jak np. *aliasing*;
7) zniekształcenia tempa (ang. *temporal distortions*) np. pre- lub post-echo, przesunięcie pomiędzy poszczególnymi kanałami audio;
8) dźwięki dodatkowe (ang. *extra sound*) niebędące elementem sygnału źródłowego, np. szum;
9) dźwięki brakujące (ang. *missing sound*), tj. utrata pewnych składowych materiału dźwiękowego;
10) efekty przesłuchu (ang. *correlations effect, crosstalk*), tj. liniowy lub nieliniowy efekt przesłuchu pomiędzy kanałami sygnału;
11) zniekształcenia przestrzenności (ang. *distortion of spatial image quality*), tj. wszystkie aspekty związane z wrażeniem przestrzenności dźwięku, jak balans, odpowiednia lokalizacja źródeł dźwięku, wrażenie stabilności dźwięku.

2.6. Najczęstsze błędy podczas testów odsłuchowych

Pomimo, iż umiejętność słuchania krytycznego może być wyuczona poprzez odpowiednią praktykę, testy odsłuchowe wciąż narażone są na błędy wynikające z subiektywnej oceny, w szczególności tzw. złudzenia słuchowe oraz błędy poznawcze [19, 20]. Wyróżnione są tutaj trzy kategorie błędów, tj. związane z: percepcją, oczekiwaniami oraz uprzedzeniami społecznymi. Pierwszy z nich wynika z faktu, że percepcja dźwięku skupiona jest na najważniejszych zdarzeniach akustycznych występujących wokół nas. Przykładem może być moment ustania dźwięku wentylatora powietrza – pomimo iż wentylator był wcześniej wyraźnie słyszalny w sposób ciągły, to stanowił tło, które nie było dostatecznie ważne dla naszej percepcji, dopóki stan ten nie został zmieniony. Przykład drugi, związany z oczekiwaniami, to sytuacja, gdy dany sygnał oceniany jest lepiej ze względu na dodatkowe informacje, jakie o nim posiadamy. Może to być sygnał oznaczony jako wersja o najwyższej częstotliwości próbkowania, mimo iż nie będzie różnił się od pozostałych, może on być oceniony najlepiej, jeśli informacja ta będzie jawna dla słuchaczy [19]. Ostatnie wymienione błędy związane są z sugestiami i opiniami od innych osób, które również mogą istotnie wpłynąć na ocenę jakości dźwięku [19]. Również osobiste preferencje mogą mieć wpływ na ocenę, a rozwiązaniem na tego typu problemy jest stosunkowo duża liczba badanych słuchaczy – różnice wynikające z błędów poznawczych, złudzeń słuchowych, preferencji itd. są redukowane poprzez uśrednienie większej liczby odpowiedzi [12].

2.7. Podsumowanie

Testy odsłuchowe są wciąż głównym sposobem oceny jakości rzeczywistych sygnałów audio. Głównymi problemami tej metody są:

- konieczność utworzenia i wyszkolenia grupy słuchaczy krytycznych;
- zapewnienie oprogramowania i sprzętu odsłuchowego;
- zapewnienie odpowiednich warunków do odsłuchu, tj. odpowiednia akustyka pomieszczenia odsłuchowego, dobrany poziom głośności, ustawienie zestawów głośnikowych, umiejscowienie punktu odsłuchowego;
- przygotowanie odpowiedniej procedury testowej, zwracając szczególną uwagę na długość sekwencji odsłuchowej;
- wysokie koszty oraz długi czas wykonania.

Ponadto, testy odsłuchowe są również podatne na błędy wynikające z subiektywnego charakteru oceny sygnału, w tym decyzje podjęte w wyniku zmęczenia lub według preferencji danego słuchacza, niezwiązane z samym celem badania oraz nieujęte w pytaniu testowym (np. różnice w poziomie sygnałów testowych – dla niektórych słuchaczy głośniejszy sygnał, pomimo braku innych różnic, będzie lepiej oceniony). Istotnym zatem elementem jest zaplanowanie procedury testowej w taki sposób, aby zminimalizować ryzyko występowania tego typu błędów.

3. Automatyczne metody oceny jakości audio

3.1. Wprowadzenie do automatycznych testów jakości dźwięku

Pomimo, iż opisane w poprzednim rozdziale testy odsłuchowe są wciąż najskuteczniejszą i najczęściej stosowaną metodą oceny jakości sygnału audio [12], ich główną wadą są przede wszystkim wysokie koszty (związane z zapewnieniem grupy słuchaczy, sprzętu i oprogramowania) oraz czasochłonność (w tym czas przygotowania procedury testowej, przeprowadzenia testu, analizy wyników, a co najważniejsze wyszkolenia grupy słuchaczy). Te problemy sprawiły, że rozpoczęto badania nad metodami w pełni automatycznymi, które mogłyby zastąpić subiektywne testy manualne. W niniejszym rozdziale opisane zostały najbardziej popularne metody stosowane obecnie.

Obiektywne metody oceny jakości dźwięku dzielą się zazwyczaj na algorytmy analizujące mowę oraz pozostałe sygnały audio. W związku z tym, że tematem niniejszej pracy jest badanie jakości rzeczywistych sygnałów muzycznych, w rozdziale uwzględniono metody dla drugiej kategorii, a metody stosowane wyłącznie do przetwarzania sygnału mowy nie zostały opisane.

Metody do oceny jakości sygnałów audio podzielić można na 3 kategorie [21]:

- referencyjne (porównawcze, ang. *intrusive*, *full-reference*, *comparison-based* lub też *input-to-output*);
- bezreferencyjne (ang. *nonintrusive*, *no-reference*, *output-based* lub *single-ended*);
- parametryczne (ang. *parametric*, znane też jako *glass box*).

Najważniejsze cechy wymienionych grup metod zostały przedstawione w rozdz. 3.2 – 3.5.

3.2. Metody referencyjne

Metody referencyjne opierają się na porównaniu sygnału badanego do znanego sygnału referencyjnego (zazwyczaj jest nim sygnał oryginalny, nieprzetworzony). Tradycyjnymi (a zarazem najprostszymi) miarami jakości sygnału audio w dziedzinie czasu są: SNR (ang. *Signal-To-Noise Ratio*) oraz THD (ang. *Total Harmonic Distortion*). Miary te są skuteczne w przypadku analizy prostych przebiegów czasowych, gdzie przetwarzany sygnał ma wiernie odwzorować sygnał referencyjny. Nie są one jednak odpowiednie m.in. dla dźwięku przetwarzanego za pomocą kodeków percepcyjnych [12]. Nie odzwierciedlają one również subiektywnej oceny jakości dźwięku, zwłaszcza jeśli badany sygnał zawiera większą ilość zniekształceń [21]. Mimo to, parametr SNR jest wciąż często stosowany w literaturze [22, 23, 24]. Powstało też wiele wariantów tej metryki, m.in. *segmental* SNR (stosunek sygnału do szumu wyliczany jest na krótkich segmentach sygnału), *frequency weighted segmental* SNR (do obliczenia SNR stosowane są wagi, których wartość jest zależna od pasma częstotliwościowego), SIR (ang. *signal-to-interference*), SDR (ang. *signal-to-distortion*) lub też SAR (ang. *signal-to-artifact ratio*).

3.2.1. Miara SNR

Przyjmując, że $s(n)$ jest sygnałem oryginalnym (nieprzetworzonym) oraz $s_d(n)$ jest sygnałem badanym (np. enkodowanym i dekodowanym), zniekształcenia wprowadzone przez przetwarzanie dźwięku jest definiowane za pomocą obiektywnej metryki SNR (*Signal-To-Noise Ratio*) następująco:

$$SNR = 10\log_{10}\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1}[s(n) - s_d(n)]^2} \tag{3.1}$$

gdzie N jest liczbą próbek sygnału audio.

Uśrednienie zastosowane podczas obliczania tej metryki powoduje, że wartość SNR może być mała, nawet jeśli w pewnym krótkim fragmencie sygnału wystąpią wyjątkowo duże różnice pomiędzy sygnałem testowym a referencyjnym. Jest to znaczna wada tej metody, powodująca, że jej wyniki nie pokrywają się z subiektywną oceną wynikającą z manualnych testów odsłuchowych [15]. Z tego powodu wprowadzony został nowy wariant tej metody – *segmental* SNR, który miał na celu poprawić adekwatność wyników poprzez podział badanego sygnału na mniejsze segmenty. Założeniem było podwyższenie skuteczności algorytmu w przypadku wystąpienia dużych zniekształceń jedynie na krótkim pojedynczym fragmencie sygnału.

3.2.2. Miara *Segmental* SNR

Miara *segmental* SNR polega na tym, że zamiast analizować cały sygnał jednocześnie, jest on dzielony na mniejsze fragmenty. W tym przypadku wskaźnik SNR jest obliczany najpierw dla poszczególnych segmentów, następnie na koniec wszystkie wyniki zostają uśredniane. Zaproponowane zostały dwa warianty *segmental* SNR – w dziedzinie czasu (*time-domain segmental* SNR) oraz częstotliwości (*frequency-weighted segmental* SNR).

Pierwszy z nich, *time-domain segmental* SNR obliczany jest następująco [15]:

$$segSNRt = \frac{1}{M}\sum_{m=0}^{M-1} \max\{\min\{35, SNR(m)\}, -10\} \tag{3.2}$$

gdzie: M jest całkowitą liczbą segmentów, a $SNR(m)$ to wynik SNR obliczony dla m-tego segmentu.

Z powyższej definicji wynika, iż wynik tej miary musi mieścić się w zakresie [–10, 35] dB. Jest to kolejny problem, ponieważ w przypadku analizy fragmentów ciszy z pojawiającym się szumem trudno słyszalnym w subiektywnych testach odsłuchowych, wynik ograniczony jest do –10 dB. Z drugiej strony, dla fragmentów znacznie różniących się od referencji, przekraczających wartość 35 dB, zostaną one obcięte do wartości granicznej. Brak jest zatem precyzjnej informacji, które fragmenty zostały najbardziej zniekształcone w wyniku przetwarzania sygnału.

Drugi wariant miary porównującej poziomy sygnałów to *frequency-weighted segmental* SNR. Miara SNR obliczana jest w tym przypadku dla każdego pasma krytycznego, a wynik mnożony przez wagę, odpowiednio do mocy sygnału w danym paśmie [15]:

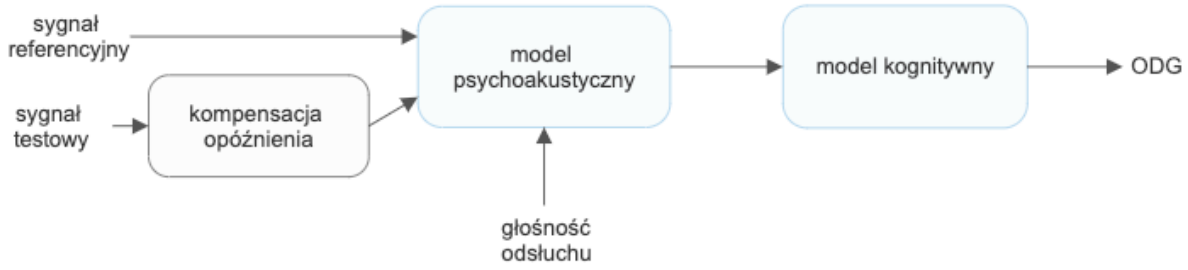
$$segSNRf = \frac{1}{M}\sum_{m=0}^{M-1}\left(\sum_{k=1}^{K} W_m(k)\cdot 10\log_{10}\frac{|X_m(k)|^2}{\left(|X_m(k)| - \left|X_{m,d}(k)\right|\right)^2}\right) \tag{3.3}$$

gdzie: K oznacza liczbę pasm krytycznych, $X_m(k)$ to widmo częstotliwościowe dla k-go pasma oraz m-tej ramki, uzyskane poprzez zsumowanie wszystkich widm częstotliwościowych obliczonych dla składowych w danym paśmie. $X_{m,d}(k)$ to odpowiednio widmo częstotliwościowe uzyskane w podobny sposób, ale dla sygnału testowanego. Ostatni element to znormalizowana waga $W_m(k)$, przez którą przemnażany jest wynik uzyskiwany dla każdego poszczególnego pasma częstotliwości, gdzie znormalizowana wartość pojedynczej wagi jest zawsze większa od 0, a suma wag dla wszystkich pasm częstotliwości krytycznych jest równa 1.

Pomimo tego, że algorytm może być optymalizowany, by uzyskać jak najlepszą korelację z wynikami subiektywnych testów odsłuchowych (np. przez zastosowanie odpowiednich wag dla poszczególnych pasm częstotliwości), to wartość wynikowa *segmental* SNR waha się od 13 do 90 dB [15], a to z kolei oznacza, że wciąż może znacznie odbiegać od wyników testów subiektywnych.

3.2.3. Metoda PEAQ

Główną wadą opisanych w rozdz. 3.2.1 i 3.2.2 metod działających w dziedzinie czasu jest to, że są bardzo wrażliwe na jakiekolwiek różnice na osi czasu lub przesunięcia fazowe sygnału. Problem ten jest mniej widoczny dla metod opartych na analizie sygnału w dziedzinie częstotliwości. Szeroko stosowanym przykładem takiej metody jest PEAQ (ang. *Perceptual Evaluation of Audio Quality*), która jest standardem oceny jakości audio według ITU-R BS.1387 [6] i polega na porównaniu sygnału testowanego z sygnałem referencyjnym. Dla każdego bloku sygnału audio, obliczane jest w jakim stopniu poziom zniekształceń sygnału przekracza próg maskowania, wyznaczony na podstawie analizy sygnału oryginalnego (tj. przed przetwarzaniem). Wyniki ze wszystkich pośrednich bloków są uwzględniane przy estymacji ogólnej jakości dźwięku [12]. Jak opisano w rekomendacji, wynik tej metody, tzw. ODG (ang. *Objective Difference Grade*), skutecznie odzwierciedla wyniki testów subiektywnych SDG (ang. *Subjective Difference Grade*), gdzie oceny ustalane są według jednobiegunowej skali pięciostopniowej (tab. 2.1). Wynik SDG obliczany jest jako różnica pomiędzy oceną dla sygnału testowanego a referencyjnego, przy czym 0 oznacza brak słyszalnych zniekształceń, a –4 zniekształcenia bardzo przeszkadzające.



Rys. 3.1. Diagram blokowy metody PEAQ na podstawie [6].

Pierwszym etapem opisywanej metody jest przekonwertowanie sygnałów do postaci DFT (jest to tzw. *Basic Version*, stosowana w przypadku zastosowań typu *real-time*) lub DFT wraz z zastosowaniem banku filtrów (*Advanced Version*, do zastosowań typu *file-based*). Następnie na podstawie składowych częstotliwościowych przeprowadzana jest analiza psychoakustyczna, a jej rezultat (głównie wzorzec wzbudzania oraz próg maskowania) przekazywany jest do modelu kognitywnego, który na podstawie dostarczonych danych oblicza wynik ODG (rys. 3.1).

Ograniczeniami tej metody są:

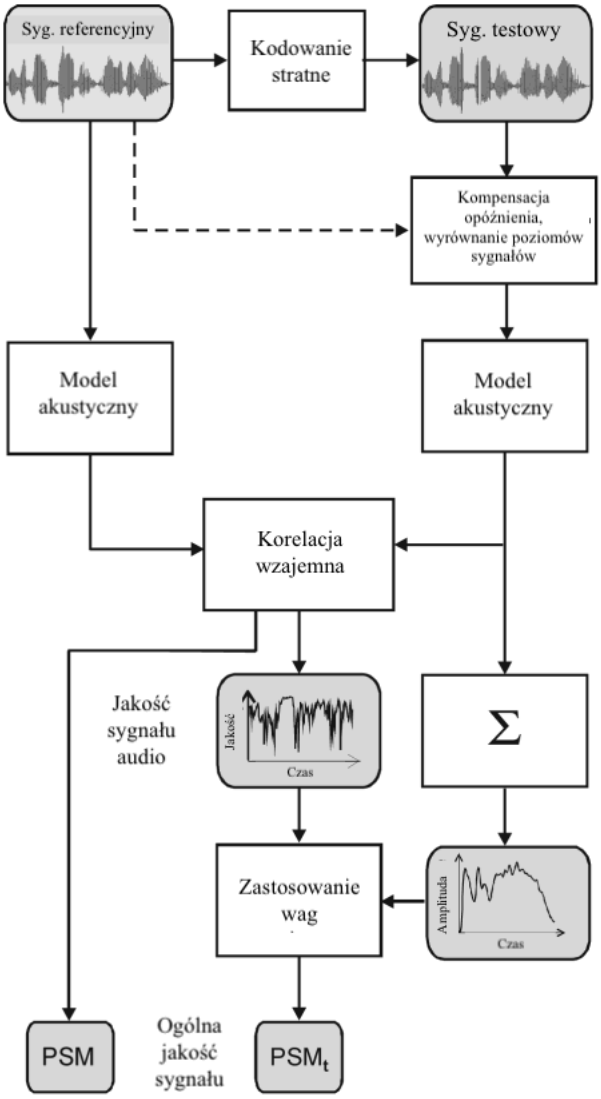
- wymagany jest sygnał referencyjny;
- oba sygnały – referencyjny oraz testowy – muszą być zsynchronizowane w czasie z dokładnością do 24 próbek [6].

3.2.4. Metoda PEMO-Q

Metoda PEMO-Q (*Perception Model-Based Quality*) [25] to metoda oceny jakości dźwięku bazująca na modelu psychoakustycznym i jest ona rozszerzeniem powstałej wcześniej metody oceny jakości sygnału mowy według autorów M. Hansen i B. Kollmeier [26]. Polega ona na porównaniu sygnału testowanego do referencyjnego i podobnie jak opisana wcześniej metoda PEAQ, wykorzystuje model kognitywny do określenia współczynnika korelacji, jednak w porównaniu do PEAQ wykazuje wyższą dokładność predykcji, z wyjątkiem sygnałów zawierających zniekształcenia liniowe. Metoda PEMO-Q charakteryzuje się dobrą korelacją w stosunku do wyników subiektywnych testów odsłuchowych [25]. Jest także odpowiednia do różnego typu sygnałów audio oraz zniekształceń.

Pierwszym etapem PEMO-Q (rys. 3.2) jest przetwarzanie wstępne, np. kompensacja opóźnienia wprowadzonego przez enkoder i/lub dekoder, wyrównanie poziomu głośności sygnału testowego w stosunku do sygnału referencyjnego, usunięcie fragmentów ciszy. Następnie oba sygnały analizowane są przez blok modelujący przetwarzanie sygnału dźwiękowego przez ucho ludzkie. Wyjściem tego modelu jest reprezentacja sygnału w dziedzinie czasu i częstotliwości [25]. Następnie obliczana jest wzajemna korelacja dla ramek o długości 10 ms uzyskanych wcześniej reprezentacji danych, a wynik tzw. PSM (*Perceptual Similarity Measure*) jest ograniczony zakresem $\langle -1, 1\rangle$, gdzie 1 oznacza, że sygnały są identyczne, natomiast mniejsze wartości wskazują na różnice lub zniekształcenia testowanego sygnału. W praktyce nie obserwuje się natomiast wartości ujemnych [25]. Wartości PSM konwertowane są na skalę SDG za pomocą funkcji regresji, tak by możliwe

było porównanie z wynikami subiektywnych testów odsłuchowych. Tak jak w przypadku techniki PEAQ (rozdz. 3.2.3), głównym ograniczeniem tej metody jest konieczność porównania badanego sygnału do referencji.



Rys. 3.2. Schemat blokowy metody PEMO-Q [25].

3.2.5. Metoda ViSQOLAudio

ViSQOLAudio [10] to metoda będąca adaptacją opracowanej wcześniej metody obiektywnej oceny jakości sygnału mowy ViSQOL [27] do oceny jakości innych sygnałów audio kodowanych za pomocą kodeków o stosunkowo niskiej przepływności danych (tj. HE-AAC, AAC-LC, MP3, Opus). Obie metody są w pełni referencyjne, tj. wymagają porównania dwóch sygnałów, w tym przypadku na podstawie analizy spektrogramów. Dostosowanie metody dla analizy sygnałów innych niż mowa polegało m.in. na usunięciu algorytmu VAD (*Voice Activity Detection*), zwiększeniu liczby analizowanych pasm częstotliwości (wcześniej

analizowano tylko częstotliwości z obszaru sygnału mowy tj. 50 – 8 000 Hz), zmiana interpretacji wyniku końcowego (zamiast stosowanej wcześniej metryki MOS, rezultat jest obliczany jako skala podobieństwa od 0 do 1). Ponadto, nowa metoda wykorzystuje model sieci neuronowych w celu uzyskania wyniku zbliżonego do subiektywnej oceny jakości audio. Jest też lepiej przystosowana do analizy sygnałów szerokopasmowych, w tym muzycznych, dzięki wykorzystaniu banku filtrów oraz wykorzystaniu informacji z obu kanałów (w przypadku sygnału stereo). Metoda ta jest wciąż udoskonalana, a ostatnią dostępną obecnie wersją jest ViSQOL v3 [28].

3.3. Metody bezreferencyjne

Metody bezreferencyjne (ang. *non-reference*, *non-instrusive* lub *singe-ended*) polegają na analizie sygnału testowego bez konieczności porównania go do innego sygnału (referencyjnego), np. w wersji nieskompresowanej. Dla zadań analizy muzyki wciąż jednak dominują metody referencyjne (opisane w rozdz. 3.2), a dostępne metody bezreferencyjne dotyczą głównie przetwarzania sygnałów mowy, np. ANIQUE (ang. *Auditory Non-Intrusive Quality Estimation*) [29], HASQI (ang. *Hearing Aid Speech Quality Index*) [30], POSQE (ang. *Perceptual Output-based Speech Quality Evaluation*) [31], SRMR (ang. *Standardized Root Mean Square*) [32], WaweNets [33], Quality-Net [34] i inne [35, 36, 37].

Dostępne są również prace dotyczące oceny wpływu separacji źródeł dźwięku na jakość sygnałów audio [38, 39].

3.4. Zastosowanie sztucznych sieci neuronowych do przetwarzania muzyki

Metody oparte na sztucznych sieciach neuronowych zyskały w ostatnim czasie bardzo dużą popularność w zadaniach przetwarzania sygnałów muzycznych. Liczba prac naukowych na ten temat wzrosła ponad dwukrotnie w ciągu ostatnich dwóch lat [40]. Główne problemy rozwiązywane za pomocą sztucznych sieci neuronowych (tab. 3.1), to:

- systemy do rekomendowania muzyki MRS (ang. *music recommendation systems*), których zadaniem jest dostosowanie polecanych materiałów audio na podstawie historii wcześniejszych odtworzeń dla danego użytkownika;
- klasyfikacja muzyki (ang. *music classification*), a dokładnie gatunku muzycznego lub też detekcji fragmentów muzycznych spośród sygnału mowy, dźwięków środowiskowych itp.;
- klasyfikacja i predykcja emocji (ang. *emotion classification and prediction*) – szczególnie użyteczne w systemach MRS oraz terapiach muzyką;
- identyfikacja instrumentu (ang. *instrument identification*) – szczególnie ważne w zadaniach separacji źródeł dźwięku w ścieżkach audio;
- identyfikacja i/lub separacja wokalu;
- transkrypcja muzyki, rozpoznawanie notacji muzycznej.

Tab. 3.1. Przykłady stosowanych modeli sieci neuronowych do zadań przetwarzania muzyki.

Model sieci neuronowych	Przykład zastosowania
FCDNN (*Fully Connected Deep Neural Networks*)	• Klasyfikacja emocji [41]; • MRS [42, 43];
RNN (*Recurrent Neural Network*)	• Klasyfikacja muzyki [44, 45]; • klasyfikacja emocji [46]; • MRS [47]; • generowanie muzyki [48];
LSTM (*Long Short-Term Memory Network*)	• Klasyfikacja muzyki [49]; • klasyfikacja emocji [50]; • identyfikacja/detekcja instrumentu [51]; • MRS [52]; • generowanie muzyki [53];
GRU-RNN (*Gated Recurrent Unit Recurrent Neural Network*)	• Detekcja wokalu [54];
CNN (*Convolutional Neural Networks*)	• Klasyfikacja muzyki [55]; • klasyfikacja emocji [56]; • identyfikacja/detekcja instrumentu [57]; • MRS [58]; • transkrypcja muzyki [59]; • aplikacja filtrów, np. equalizer [60]; • ekstrakcja/separacja wokalu [61];
GAN (*Generative Adversarial Networks*)	• Klasyfikacja emocji [62]; • rozpoznawanie notacji muzycznej [63]; • ekstrakcja/separacja wokalu [64]; • generowanie muzyki [65];
CRNN (*Convolutional Recurrent Neural Networks*)	• MRS [66]; • klasyfikacja muzyki [67];
CNN-LSTM	• Detekcja wokalu [68]; • MRS [69]; • klasyfikacja emocji [70];
CNN-GAN	• Rekonstrukcja sygnału muzycznego po zastosowaniu kompresji stratnej [71];

Analiza muzyki pod względem badania jakości dźwięku i wykrywania zniekształceń, bez konieczności porównania do sygnału referencyjnego, nie jest tematem popularnym wśród aktualnych publikacji naukowych [72, 73], a metody oparte na głębokich sieciach neuronowych do zastosowania w takich zadaniach wciąż wymagają badań, przede wszystkich dlatego, że potrzebna jest jak największa ilość danych audio wraz z rzetelnymi ocenami testów subiektywnych [72]. Uzyskanie modelu, który potrafiłby przeanalizować sygnał audio i ocenić go bez porównania z sygnałem referencyjnym, w taki sposób, by wynik ten odpowiadał subiektywnej ocenie manualnych testów odsłuchowych jest więc wciąż nierozwiązanym problemem.

3.5. Podsumowanie

Najpopularniejsze automatyczne metody, stosowane w pracach naukowych, z użyciem sygnału referencyjnego do oceny jakości sygnału muzycznego to PEAQ, PEMO-Q oraz ViSQOLAudio (tab. 3.2). Dostępnych jest wiele innych metod porównujących sygnał testowy z referencyjnym, jednakże są one stosowane głównie do oceny jakości sygnałów mowy (m.in. jeden ze standardów ITU – PEAQ).

Tab. 3.2. Podsumowanie automatycznych metod referencyjnych analizy sygnału audio. Kolorem zielonym zaznaczono metody stosowane dla sygnałów muzycznych.

Metoda	Główne cechy	Główne zastosowanie	Rok
PSQM (*Perceptual Evaluation of Speech Quality*) [74]	Zsynchronizowane w czasie widmowe gęstości mocy obliczane na segmentach sygnału	Ocena jakości sygnału mowy [74]	1994
PEAQ (*Perceptual Evaluation of Audio Quality*), standard ITU [75]	Zastosowanie DFT i banku filtrów	Ocena jakości kodeków audio [76]	2001
PESQ (*Perceptual Evaluation of Speech Quality*), standard ITU [77]	Zastosowanie modelu kognitywnego	Ocena jakości przetwarzania sygnału mowy [78, 79, 80], w tym np. rekonstrukcji sygnałów mowy [81]	2001
SDR (*Signal to Distortion Ratio*) **SIR** (*Signal to Artifacts Ratio*) **SNR** (*Signal to Noise Ratio*)	Miary porównujące energie sygnałów, wyrażone w dB. Precyzja wyniku zależy od rozdzielczości bitowej sygnałów.	Ocena dekwantyzacji sygnałów muzycznych [82], usuwanie zniekształceń wprowadzonych przez kompresję zakresu dynamiki audio [83]	Brak ścisłych danych
PEMO-Q [25]	Zastosowanie PSM (*Perceptual Similarity Measure*) oraz modelu psychoakustycznego	Analiza sygnałów mowy [84, 85], klasyfikacja dźwięków środowiskowych [86], analiza sygnałów syntetycznych MIDI [85], muzyki [85], dekwantyzacja sygnałów muzycznych [82], rekonstrukcja sygnałów mowy i muzyki [81], usuwanie zniekształceń wprowadzonych przez kompresję zakresu dynamiki audio [83]	2006
MSSIM (*Mean Structural Similarity Index Metric*) [87]	Miara podobieństwa jasności, kontrastu i struktury, np. porównując obrazy spektrogramów	Klasyfikacja dźwięków środowiskowych [86]	2008
POLQA (*Perceptual Objective Listening Quality Assessment*), standard ITU [88]	Zastosowanie modelu percepcyjnego	Ocena jakości przetwarzania sygnału, głównie mowy, w telekomunikacji wąskopasmowej (300–3400 Hz) i szerokopasmowej (50–14000 Hz), ocena jakości kodowania,	2011

Metoda	Główne cechy	Główne zastosowanie	Rok
		m.in. MP3, AAC, AAC-LD [89]	
ViSQOL (*Virtual Speech Quality Objective Listener*) [27]	Miara podobieństwa na podstawie cech spektralnych i czasowych	Ocena jakości sygnału mowy [27]	2012
ViSQOLAudio [10]	Zastosowanie modelu sieci neuronowych	Ocena przetwarzania sygnałów muzycznych [90], w tym dzielenia na segmenty sygnałów mowy i muzyki dla kompresji stratnej [91], analiza jakości kodeków audio [92]	2017
DPAM [93]	Zastosowanie konwolucyjnych sieci neuronowych	Klasyfikacja dźwięków środowiskowych [86]	2020
STOI (*Short-Time Objective Intelligibility* Metric) [94]	Miara zrozumiałości mowy na krótkich odcinkach czasu (400 ms) z zastosowaniem dekompozycji czasowo-częstotliwościowej (TF) opartej na DFT	Ocena zrozumiałości mowy, np. w algorytmach wzmacniania dialogu [95, 96] lub separacji sygnału mowy [97]	2010
SISDR (S*cale-Invariant SDR*) [98]	Miara oparta na metodzie SDR	Ocena przetwarzania wzmacniania mowy [99], separacji lub ekstrakcji źródeł dźwięku [100, 101], w tym muzyki [102]	2019
NISQA [103]	Miara ogólnej jakości mowy, wraz z analizą zaszumienia (ang. *noisiness*), barwy (ang. *coloration*), nieciągłości (ang. *discontinuity*) i głośności (ang. *loudness*)	Ocena jakości sygnału mowy [103]	2021

W przypadku metod bezreferencyjnych do oceny jakości dźwięku, najbardziej znane to: POSQE oraz HASQI, jednakże jak wiele innych metod (tab. 3.3) ich celem jest wyłącznie ocena jakości sygnału mowy i nie są one skuteczne w analizie innych sygnałów. Brak rekomendacji odnośnie automatycznej analizy jakości lub detekcji zniekształceń w sygnałach muzycznych był powodem zaimplementowania w niniejszej pracy prototypowego modelu i zbadanie jego skuteczności do tego zadania.

Tab. 3.3. Podsumowanie metod bezreferencyjnych analizy sygnału audio. Kolorem zielonym zaznaczono metody stosowane dla sygnałów muzycznych.

Metoda	Główne cechy metody	Główne zastosowanie	Rok
PLP (*Perceptual Linear Prediction*) [104]	Ekstrakcja cech predykcji liniowej	Ocena jakości sygnału mowy [104]	1995
ANIQUE+ (A*uditory Non-Intrusive Quality Estimation Plus*) [105]	Zastosowanie modelu percepcyjnego	Ocena jakości sygnału mowy [105]	2005

Metoda	Główne cechy metody	Główne zastosowanie	Rok
WADA-SNR (*Waveform Amplitude Distribution Analysis*) [106]	Zastosowanie informacji statystyczne uzyskanych z rozkładu amplitudy przebiegu czasowego sygnału mowy	Ocena jakości sygnału mowy [106]	2008
POSQE (*Perceptual Output-based Speech Quality Evaluation*) [107]	Zastosowanie modelu percepcyjnego	Ocena jakości sygnału mowy [107]	2010
HASQI (*Hearing-Aid Speech Quality Index*) [108]	Liniowe i nieliniowe pomiary obwiedni i czasowych modyfikacji struktury sygnału	Ocena jakości sygnału mowy [108]	2010
NIST-STNR [109]	Metoda oparta o algorytm SNR	Ocena jakości sygnału mowy [109]	~2011
SNRVAD [110]	Metoda oparta o algorytm SNR	Ocena jakości sygnału mowy [110]	~2011
SRMR (S*peech to Reverberation Modulation Energy Ratio*) [33]	Analiza za pomocą banku filtrów modulacji, inspirowana modelowaniem percepcyjnym	Ocena jakości sygnału mowy [111, 112, 113]	2014
AutoMOS [114]	Metoda oparta o rekurencyjne sieci neuronowe	Ocena jakości sygnału mowy [114]	2016
MOSNET [115]	Metoda oparta o konwolucyjno-rekurencyjne sieci neuronowe (CNN-BLSTM), zastosowane do predykcji MOS (*Mean Opinion Score*)	Ocena jakości sygnału mowy, np. ocena kowersji głosu [115]	2019
WaveNets [116]	Metoda oparta o konwolucyjne sieci neuronowe; analiza przebiegu czasowego sygnału	Ocena jakości sygnału mowy [117]	2020
Autorska metoda opisana w niniejszej pracy (częściowo opublikowana w [118])	**Metoda oparta o konwolucyjno-rekurencyjne sieci neuronowe**	**Detekcja i klasyfikacja zniekształceń dla sygnałów muzycznych** [118]	**2020**
GRU [119]	Metoda oparta o rekurencyjne sieci neuronowe	Ocena jakości sygnałów audio tzw. UGM (*User Generated Content*), tj. rozmowy, dźwięki otoczenia, muzyka itp. [119]	2022

Według najlepszej wiedzy autorki, najnowszą publikacją o tematyce automatycznej analizy jakości dźwięku (listopad 2022) jest metoda oparta również o rekurencyjne sieci neuronowe, jednakże z wykorzystaniem warstw GRU (bez zastosowania warstw konwolucyjnych) [119]. Celem tej metody była analiza sygnałów typu *User Generated Multimedia* (UGM), tj. rozmów, dźwięków otoczenia, ale również muzyki. Metoda ta charakteryzuje się wyższą skutecznością w porównaniu do innych metod oceny jakości mowy, zarówno referencyjnych (STOI, PESQ) jak i bezreferencyjnych (m.in. WaveNets, SRMR, NIST-STNR). Analizując różnice pomiędzy metodą z wykorzystaniem GRU oraz autorską metodą opisaną w niniejszej pracy (tab. 3.4), można wnioskować, że są one komplementarne.

Tab. 3.4. Różnice pomiędzy autorską metodą opisaną w pracy oraz ostatnią dostępną publikacją bezreferencyjnej metody oceny jakości audio.

	Metoda	
	Autorska metoda	Metoda GRU [119]
Typ analizowanego sygnału	Muzyka (w tym muzyka w jakości studyjnej poddawana przetwarzaniu cyfrowemu, np. normalizacji lub kompresji).	UGM (materiał dźwiękowy tworzony przez użytkownika, np. telefonem komórkowym, jak mowa, dźwięki otoczenia, muzyka).
Modelowane typy zniekształceń	4 typy zniekształceń: problemy kwantyzacji, zniekształcenia charakterystyki wzmocnienia, dodatkowe (np. szum) lub brakujące dźwięki.	2 typy zniekształceń: szum tła, kompresja o niskiej przepływności danych.
Model	CNN-LSTM	GRU-RNN
Baza danych	Autorska baza danych utworzona w oparciu o bazę sygnałów muzycznych MUSDB18 (łącznie 57 120 próbek audio).	Autorska baza danych (łącznie 1 150 nagrań audio).
Rezultat	Klasyfikacja do jednej z pięciu kategorii ze wskazaniem typu zniekształcenia (sygnał poprawny lub zawierający co najmniej jedno zniekształcenie).	Obiektywny wynik jakości audio (ang. *objective score*).

4. Autorska prototypowa metoda automatycznej klasyfikacji sygnału z eliminacją porównania do sygnału referencyjnego

Rekomendacja organizacji ITU BS.1284-2 [18] stworzona w celu ustandaryzowania oceny zniekształceń sygnałów audio, definiuje dla nich 11 kategorii, które mogą być użyte do zadań analizy lub klasyfikacji jakości cyfrowego przetwarzania lub transmisji (rozdz. 2.5). W niniejszej pracy jednak liczbę docelowych kategorii do wykrycia ograniczono do następujących:

1. problemy kwantyzacji (ang. *quantisation defect*) związane z użyciem niewystarczającej rozdzielczości bitowej dla sygnału audio;
2. zniekształcenia charakterystyki wzmocnienia (ang. *distortion of gain characteristics*), tj. m.in. zmiany (w tym skoki) poziomu dźwięku lub zakresu dynamiki sygnału;
3. dodatkowy dźwięk (ang. *extra sound*) niebędący elementem sygnału źródłowego, np. szum;
4. dźwięk brakujący (ang. *missing sound*), tj. utrata pewnych składowych materiału dźwiękowego.

Rekomendacja zawiera również zniekształcenia typowe dla sygnałów wielokanałowych (np. zniekształcenia przestrzenności dźwięku lub też korelacji pomiędzy kanałami typu *crosstalk*). W niniejszej pracy jednak zbadane zostały jedynie kanały audio w formie mono, bez uwzględnienia wzajemnej relacji pomiędzy nimi w przypadku nagrań wielokanałowych. Uwzględnienie tych zależności może być tematem dalszych badań w tej dziedzinie.

Badany w pracy problem detekcji zniekształceń może być interpretowany jako wielomianowy problem klasyfikacji, gdzie funkcja klasyfikatora parametryzowana jest za pomocą modelu sztucznych sieci neuronowych. Wykorzystując spektrogramy jako dane wejściowe (oznaczone jako x, gdzie $x \in \mathbb{R}_{+}^{d \times t}$, a t i d to odpowiednio długość oraz wymiary poszczególnej ramki) oraz kategorie sygnałów $y \in \{1, \dots, k\}$ (gdzie k jest liczbą kategorii), problem detekcji zniekształceń może być sformułowany jako znalezienie modelu $h : \mathbb{R}_{+}^{d \times t} \rightarrow \{1, \dots, k\}$, który dla każdej instancji x przypisze wartość prawdopodobieństwa $P(y|x)$ przynależności do danej klasy y:

$$h(x) = \arg\max_{y \in C} P(y|x) \tag{4.1}$$

Badany w pracy model sieci neuronowych opiera się na warstwach konwolucyjnych i rekurencyjnych. Oba te typy warstw sieci neuronowych charakteryzuje wysoka skuteczność w zadaniach przetwarzania sygnałów audio, szczególnie detekcji zdarzeń akustycznych [120, 121, 122]. Podczas gdy warstwy konwolucyjne są bardzo efektywne w detekcji różnego typu wzorców, jednocześnie zmniejszając złożoność obliczeniową poprzez przechowywanie znacznie mniejszej liczby parametrów analizy, sieci rekurencyjne pozwalają na analizę zdarzeń sekwencyjnych, uwzględniając informacje o danych przeszłych i przyszłych, co w kontekście przetwarzania rzeczywistych sygnałów muzycznych, powinno znacznie zwiększyć skuteczność detekcji zniekształceń.

Warstwy te oraz ich znaczenie dla przypadku badanego w niniejszej pracy wraz z całą architekturą zaimplementowanego modelu, zostały szerzej opisane w dalszej części pracy (rozdz. 6 – 10).

Implementacja modelu składa się z następujących etapów:

1) przygotowanie bazy sygnałów audio oraz ich podział na zestaw treningowy, walidacyjny oraz testowy (odpowiednio 60%, 20%, 20% wstępnie przygotowanej bazy sygnałów);
2) projekt modelu oraz implementacja na podstawie wstępnych założeń.
3) wstępne testowanie zaimplementowanego modelu na niewielkiej bazie sygnałów w celu weryfikacji poprawności działania;
4) ulepszenia modelu w kolejnych iteracjach wykonania treningu modelu sieci neuronowych. Do uczenia modelu wykorzystane zostały zestawy sygnałów: treningowy oraz walidacyjny. Sprawdzanie skuteczności działania programu i dopasowanie hiperparametrów (np. liczbę epok, parametr *batch size*);
5) zapisanie optymalnego modelu uzyskanego w procesie uczenia;
6) walidacja uzyskanego modelu dla zestawu sygnałów testowych (które nie były wcześniej użyte podczas uczenia).

Zaimplementowany model sieci neuronowych trenowany był początkowo z wykorzystaniem obrazów spektrogramowych w skali melowej wygenerowanych dla każdego sygnału muzycznego z bazy danych. Następnie sprawdzono, jaki wpływ na skuteczność modelu mają dodatkowe parametry ekstraktowane z sygnałów wejściowych.

5. Autorska baza danych audio

5.1. Wprowadzenie

W celu przeprowadzenia procesu uczenia oraz ewaluacji prototypowego modelu sieci neuronowych, przygotowana została baza danych zawierająca sygnały muzyczne wraz z ich oznaczeniem o przynależności do jednej z pięciu kategorii. Zanim jednak omówione zostanie bardziej szczegółowo tworzenie bazy danych oraz wstępne przetwarzanie sygnałów przed dostarczeniem ich do modelu sieci neuronowych, poniżej przedstawione zostały parametry sygnałów audio, które były rozważane pod kątem zadania automatycznej klasyfikacji zniekształceń w rzeczywistych sygnałach muzycznych.

5.2. Parametry sygnału audio

Wyróżnić można następujące cechy sygnałów audio do zadań detekcji i klasyfikacji zdarzeń akustycznych [123, 124]:

1) **cechy czasowe** (ang. *temporal features*) – cechy sygnałów obliczane bezpośrednio z przebiegów czasowych sygnałów audio. Do tego typu cech zaliczyć można m.in.:
   a) obwiednia sygnału w dziedzinie czasu (ang. *time domain envelope*);
   b) gęstość przejść przez zero ZCR (ang. *Zero Crossing Rate*);
   c) momenty przebiegu czasowego sygnału (ang. *temporal waveform moments*), np. środek ciężkości (ang. *temporal centroid*);
2) **cechy widmowe** (ang. *spectra shape features*) – bazujące na informacjach pozyskiwanych z częstotliwościowej reprezentacji sygnału, np.:
   a) energia sygnału (ang. *energy*);
   b) spektrogram (ang. *spectrogram*);
   c) kontrast widmowy (ang. *spectral contrast*);
   d) obwiednia widmowa (ang. *spectral envelope*);
   e) entropia widmowa (ang. *spectral entropy*);
   f) momenty widmowe (ang. *spectral moments*);
   g) płaskość widmowa (ang. *spectral flatness*);
   h) nachylenie widmowe (ang. *spectral slope*);
   i) strumień widmowy (ang. *spectral flux*);
3) **cechy cepstralne** (ang. *cepstral features*) – bazujące na dekompozycji sygnału zgodnie z modelowaniem uwzględniającym źródło sygnału oraz nałożony filtr (źródłem może być np. wzbudzenie głośni w przypadku mowy ludzkiej, natomiast filtrem będzie trakt głosowy i pozycja języka). Wśród cech cepstralnych można wyróżnić:
   a) współczynniki mel-cepstralne MFCC (ang. *Mel Frequency Cepstral Coefficients*);
   b) współczynniki cepstralne predykcji liniowej LPCC (ang. *Linear Prediction Cepstral Coefficients*);
   c) współczynniki gammatone-cepstralne GTCC (ang. *Gammatone Cepstral Coefficients*);

4) **cechy motywowane percepcyjne** (ang. *perceptually motivated features*) – bazujące na wiedzy w jaki sposób działa słuch ludzki. Prócz wspomnianego wyżej MFCC, przykładem są:
   a) głośność (ang. *loudness*);
   b) ostrość (ang. *sharpness*);
   c) rozpiętość percepcyjna (ang. *perceptual spread*);
5) **cechy oparte na obrazie spektrogramu** (ang. *spectrogram image-based features*) – cechy ekstraktowane z obrazu spektrogramu, który uwzględnia zarówno czas, jak i widmo badanego sygnału. Przykłady:
   a) histogram zorientowanych gradientów HOG (ang. *Histogram of Oriented Gradients*);
   b) lokalny wzorzec binarny LBP (ang. *Local Binary Pattern*).

W niniejszej pracy zbadano wpływ wybranych parametrów:
- spektrogram z wykorzystaniem filtrów melowych – cecha widmowa;
- parametr OBSC – cecha widmowa;
- parametr ZCR – cecha czasowa;
- głośność chwilowa oraz rzeczywiste wartości szczytowe sygnału – cechy motywowane percepcyjnie;

Zostały one szerzej opisane w rozdz. 5.6 – 5.10.

5.3. Przygotowanie bazy danych

Według najlepszej wiedzy autorki, nie jest znana oficjalna baza danych, reprezentująca rzeczywiste sygnały muzyczne wraz z kategoriami odpowiadającymi różnego typu zniekształceniom. Do celów niniejszej pracy została więc przygotowana specjalna baza danych, na podstawie MUSDB18 [125]. MUSDB18 to zestaw rzeczywistych sygnałów muzycznych – łącznie około 10 godzin nagrań różnych gatunków, a także izolowanych instrumentów (np. bębny, bas, wokal). Baza ta jest dostępna w postaci sygnałów nieskompresowanych o częstotliwości próbkowania 44 100 Hz, co w porównaniu z innymi dostępnymi bazami sygnałów, charakteryzuje się stosunkowo wysoką jakością dźwięku. Często stosowana w publikacjach naukowych baza sygnałów GTZAN [126], pochodząca z 2002 roku, zawiera sygnały próbkowane z częstotliwością jedynie 22 050 Hz. Inna, „Million Song Dataset" (2011) [127], o tej samej częstotliwości próbkowania, umożliwia dostęp jedynie do wybranych cech wyekstraktowanych z sygnałów wejściowych oraz ich metadanych, bez dostępu do właściwych nieprzetworzonych próbek audio. Natomiast baza sygnałów FMA (2017) [128] zawiera próbki audio skompresowane do stratnego formatu MP3, co istotnie wpływa na jakość dźwięku.

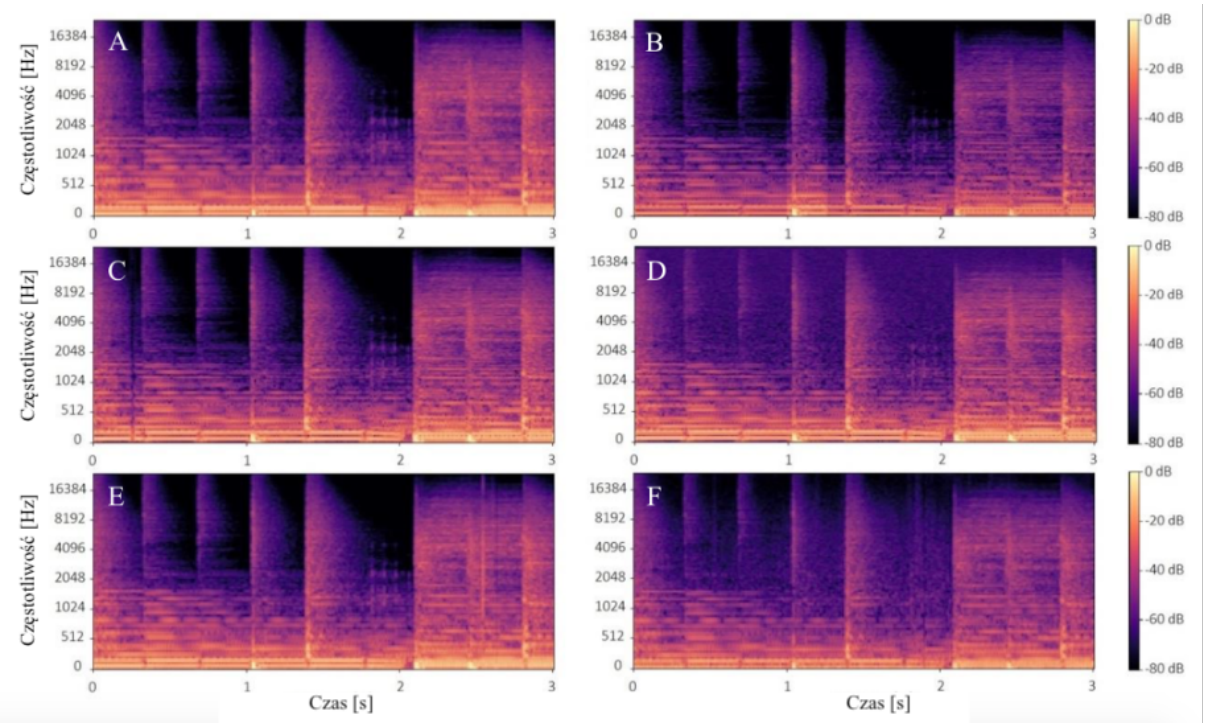
Bazując na zestawie MUSDB18, przygotowanych zostało 5 kategorii sygnałów muzycznych. Pierwsza grupa zawierała sygnały niezniekształcone, w ich oryginalnej formie bez kompresji. Kolejne cztery grupy zawierały sygnały wraz z odpowiednio wygenerowanymi wcześniej zniekształceniami. Zostały one utworzone poprzez zmodyfikowanie oryginalnego sygnału, zależnie od wybranej kategorii zniekształcenia:

- kategoria 1: zniekształcenia głośności – reprezentowane przez sygnały z modyfikacją głośności polegającą na zmianie zakresu dynamicznego wybranych ich fragmentów, przy czym każdy skok poziomu sygnału trwał co najmniej 20 ms;
- kategoria 2: brakujące składowe sygnału – reprezentowane przez sygnały, gdzie występuje problem brakujących składowych lub poszczególnych fragmentów. Modyfikacje sygnałów w tym przypadku polegały na dodaniu fragmentów z niskopoziomowym szumem, symulującym brakujące ramki lub też powtórzeniu ostatniej poprawnej ramki (co może wystąpić w rzeczywistym systemie przetwarzającym audio/video) o zmiennej długości trwania ramki w zakresie 20 – 100 ms;
- kategoria 3: problem kwantyzacji sygnału – reprezentowana przez sygnały, gdzie zastosowana została niewystarczająca rozdzielczość bitowa lub też niepoprawna konwersja pomiędzy poszczególnymi typami zmiennych;
- kategoria 4: dodatkowe zniekształcenia (szum) – reprezentowane przez sygnały, do których dodane zostały różnego typu zniekształcenia, tj. wygenerowany szum kolorowy (biały, różowy, niebieski, brązowy, fioletowy) o SNR (*Signal-to-Noise Ratio*) nie przekraczającym 20 dB, o zmodyfikowanych pojedynczych bitach lub dłuższych fragmentów sygnału audio lub też łączonych z rzeczywistymi próbkami zniekształceń (np. szumu addytywnego oraz typu *burst*) z publicznie dostępnej bazy danych *Freesound* [129].

Na rys. 5.1 przedstawione zostały przykładowe sygnały dla każdej kategorii w postaci spektrogramu w skali melowej:

- spektrogram A to oryginalny, niezniekształcony 3-sekundowy fragment sygnału z bazy danych MUSDB18 (oznaczony jako *Leaf* „*Summerghost*”);
- spektrogram B reprezentuje poprzedni sygnał z modyfikacją wzmocnienia w pierwszej sekundzie trwania (długość zniekształcenia to 250 ms);
- spektrogram C to przykład brakującej pojedynczej ramki o długości 20 ms na początku sygnału (w czasie 240 ms);
- spektrogram D to sygnał reprezentujący niewystarczającą rozdzielczość bitową (próbki sygnału niepoprawnie zapisane na 8 bitach);
- spektrogram E zawiera dodatkowe zniekształcenie (trzask pomiędzy 2 a 3 sekundą),
- spektrogram F zawiera również dodatkowe zniekształcenie (szum nałożony na całą długość trwania sygnału).

Oś pozioma reprezentuje czas (sekundy), oś pionowa – częstotliwość (Hz), a każda wartość spektrogramu (reprezentowana w kolorze) odpowiada mocy w skali decybelowej.



Rys. 5.1. Przykładowe sygnały wejściowe przygotowane dla zaimplementowanego modelu sieci neuronowych: spektrogramy w skali melowej. Spektrogram A: sygnał czysty; B: zniekształcenie głośności (w czasie 1sek.), C: brakująca ramka o długości 20ms (w czasie 240ms), D: problem kwantyzacji, E, F: trzask (w czasie ok. 2.5sek) oraz szum nie będące częścią oryginalnego sygnału.

Baza danych została podzielona na trzy rozłączne podzestawy (żadna z próbek nie znalazła się w więcej niż w jednej grupie): treningową, walidacyjną oraz testową (tab. 5.1). W procesie treningu wykorzystane są 2 podzestawy: treningowy oraz walidacyjny. Zestaw testowy użyty został w procesie końcowej oceny skuteczności wynikowego modelu.

Tab. 5.1. Przygotowana baza sygnałów audio do ewaluacji modelu sieci neuronowych.

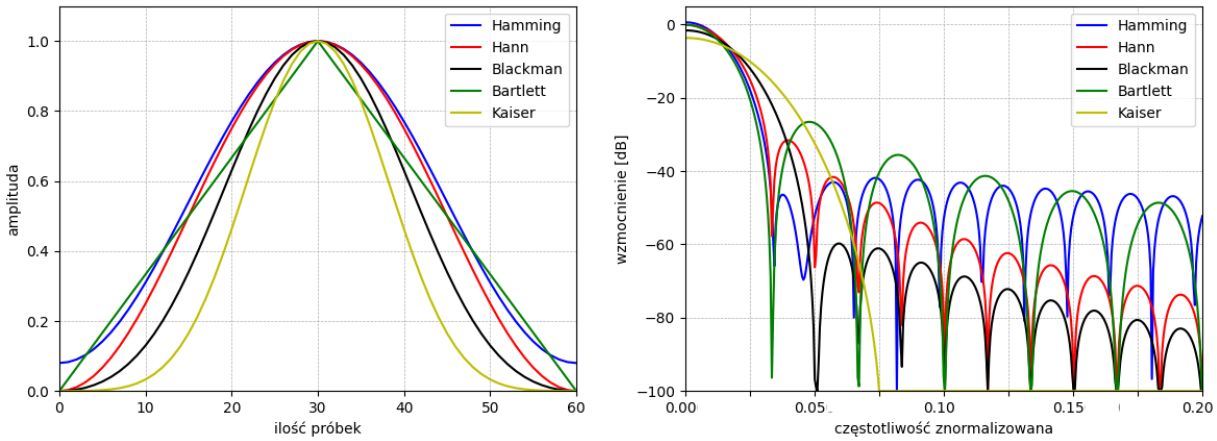
Ilość kategorii	5
Długość trwania pojedynczego sygnału (sek.)	3.0
Ilość sygnałów w bazie treningowej (na kategorię)	6 766
Ilość sygnałów w bazie testowej/walidacyjnej (na kategorię)	2 329

5.4. Przetwarzanie wstępne

Pierwszym etapem działania zaimplementowanego programu było wstępne przetwarzanie sygnałów wejściowych, w taki sposób żeby mogły one zostać przeanalizowane przez model sieci neuronowych, w tym warstwy konwolucyjne oraz rekurencyjne.

Według dostępnej literatury, w problemie klasyfikacji sygnałów audio stosuje się najczęściej fragmenty sygnału o długości w zakresie 1 – 10 sekund [130, 131, 132]. Użycie takiej wielkości ramki jest kompromisem pomiędzy możliwą redukcją złożoności obliczeniowej i tym samym czasu przetwarzania, a zapewnieniem wystarczającej ilości danych do uczenia implementowanego modelu. W niniejszej pracy przygotowana baza danych zawierała sygnały o długości 3 sekund (w przypadku gdy sygnał nie był wystarczająco długi

był dopełniany zerami), częstotliwości próbkowania 44 100 Hz oraz rozdzielczości bitowej 16. Ekstrakcja cech sygnału w dziedzinie częstotliwości składała się z następujących kroków: podziału sygnału na ramki, okienkowania oraz obliczenia widma częstotliowościowego. W pierwszym etapie wybór wielkości ramki ma znaczący wpływ na rozdzielczość widma częstotliwościowego – im większa długość ramki, tym wyższa rozdzielczość analizy, przy czym wiąże się to również z dłuższym czasem przetwarzania. Z tego względu w zadaniach wykrywania zdarzeń akustycznych najczęściej stosowana jest ramka o długości 20 – 60 ms z nakładkowaniem co najmniej 50% [133]. W niniejszej pracy każdy sygnał został podzielony więc na ramki o długości 2048 próbek (dla częstotliwości próbkowania 44 100 Hz jest to czas trwania ok. 46 ms) z nakładkowaniem 75%. Dla każdej ramki zastosowano funkcję nieprostokątnego okna czasowego, która minimalizuje efekt tzw. wycieku widma częstotliwościowego. Porównując różne okna czasowe (rys. 5.2), istotnym jest uzyskanie odpowiednio dużego tłumienia poziomu listków bocznych przy jednoczesnym niezbyt dużym poszerzeniu listka głównego. Do takich okien należy okno Hanna, które jest najczęściej wybierane, ponieważ poszerza listek główny względem okna prostokątnego zaledwie dwukrotnie, przy jednoczesnym dużym tłumieniu listków bocznych. Zostało ono zastosowane również w tym przypadku.



Rys. 5.2. Porównanie okien czasowych w dziedzinie czasu i częstotliwości (z parametrem $\beta = 14$ dla okna Keisera).

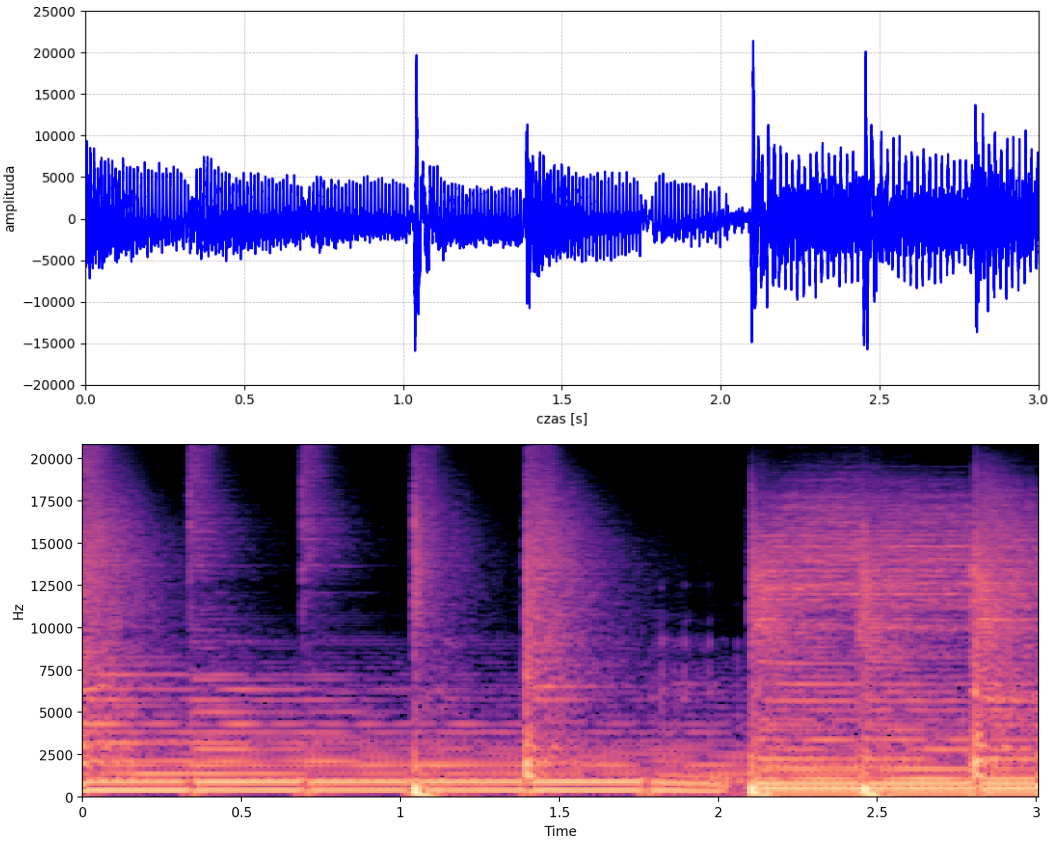
Podczas wstępnego przetwarzania sygnałów audio popularne jest stosowanie operacji zwanej *pre-emphasis*, która ma na celu podwyższenie energii wysokich częstotliwości sygnału wejściowego, zanim sygnał ten zostanie poddany dalszej analizie i ekstrakcji parametrów [123]. Operacja ta stosowana jest głównie ze względu na to, że energia w sygnale audio zazwyczaj koncentrowana jest na mniejszych częstotliwościach – jej zadaniem jest więc zbalansowanie składowych widma. W przypadku różnego typu zadań analizy sygnału audio, szczególnie przy rozpoznawaniu mowy, może ona znacznie polepszyć efektywność modelu, dzięki poprawie współczynnika SNR. W niniejszej pracy jednak, gdzie głównym celem jest wstępna ocena jakości sygnału (tj. wykrycie zniekształceń), ważniejsze wydaje się zachowanie jak największej ilości informacji o oryginalnej wartości SNR oraz wzmocnieniu składowych o dużych częstotliwościach, dlatego też operacja ta nie została zastosowana.

Dla każdej okienkowanej ramki obliczana jest dyskretna transformata Fouriera (DFT, ang. *Discrete Fourier Transform*), która jest efektywnym narzędziem w przypadku analizy sygnałów niestacjonarnych [134], w tym sygnałów muzycznych. Obliczanie DFT na podstawie okienkowanych ramek określane jest również jako krótko-czasowa transformata Fouriera (ang. *Short-Time Fourier Transform*, STFT), a jej składowa obliczana jest jako [123]:

$$X[t,j] = \sum_{m=0}^{N-1} \omega[m]x[tN+m]e^{\frac{-i2\pi mj}{N}} \tag{5.1}$$

gdzie:
j – numer składowej DFT,
t – numer ramki,
$\omega[m]$ – funkcja okienkowania (np. prostokątne, Hanna, Hamminga),
N – długość ramki dla sygnału wejściowego $x[n]$.

Uzyskany wynik STFT to widmo częstotliwościowe, które jest dwuwymiarową reprezentacją sygnału audio, tj. amplituda składowych częstotliwości w funkcji czasu (rys. 5.3). Następnie uzyskane wyniki zostały przekonwertowane do skali melowej opisanej w rozdz. 5.5.



Rys. 5.3. Przykładowy przebieg czasowy (A) oraz spektrogram (B) sygnału wejściowego.

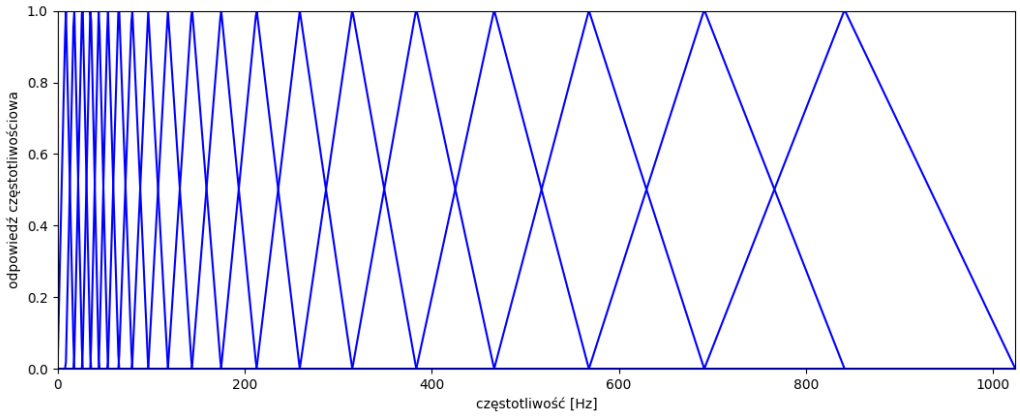
5.5. Skala melowa

Zastosowanie skali melowej podczas analizy widmowej sygnału audio pozwala na dokładniejszą symulację rzeczywistej percepcji słuchu ludzkiego [135]. Funkcja melowa $M(f)$ jest funkcją nieliniową, mającą na celu uwzględnienie wrażliwości słuchu ludzkiego na poszczególne częstotliwości:

$$M(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.2}$$

gdzie f oznacza częstotliwość sygnału, a wynik powyższej funkcji jest wyrażony w jednostce Hz.

Analiza spektrogramów przy zastosowaniu skali melowej jest przeprowadzana za pomocą tzw. banku filtrów melowych (ang. *Mel-frequency filter bank* lub inaczej *Mel filter bank*) [124]. Bank ten jest zestawieniem nakładających się na siebie filtrów trójkątnych, których częstotliwości centralne rozłożone są równomiernie na osi reprezentującej wartości w skali melowej. Na rys. 5.4 przedstawiony został przykładowy bank filtrów, zawierający ich jedynie 20, jednak w zaimplementowanym w pracy programie zastosowano ich znacznie więcej (256).



Rys. 5.4. Bank filtrów melowych.

Współczynniki dla spektrogramu w skali melowej obliczane są następująco [124]:

$$X_{\text{mel}}[j,t] = \sum_{l=0}^{L-1} m_j[l]|S[l,t]|^2 \tag{5.3}$$

gdzie:
X_{mel} – współczynnik wektora spektrogramu melowego o indeksie j oraz ramki t,
L – liczba składowych częstotliwości oryginalnego sygnału STFT $X[j,t]$,
$m_j[l]$ – filtr trójkątny o indeksie l z banku filtrów melowych.

5.6. Parametr *Zero Crossing Rate*

Parametr *Zero Crossing Rate* (ZCR) jest wciąż jednym z najczęściej używanych parametrów sygnału z dziedziny czasu dla zadań klasyfikacji sygnałów muzycznych [136]. Parametr ten jest wynikiem obliczenia, ile razy wartości chwilowe sygnału przecinają wartość zerową w określonym odcinku czasu. Informacja ta może wskazywać, jakie jest zaszumienie badanego sygnału – wyższa wartość oznaczać będzie większe zaszumienie oraz mniejszą periodyczność sygnału [137] (przykładowo, szum biały charakteryzuje się wysoką wartością ZCR). Na wartość ZCR bardzo duży wpływ mają zewnętrzne zakłócenia [124], w związku z tym parametr ten ma częstsze zastosowanie dla sygnałów nagrywanych w warunkach zbliżonych do studyjnych, mniej zaszumionych, co w przypadku niniejszej pracy i badanych próbek muzycznych powinno być zaletą.

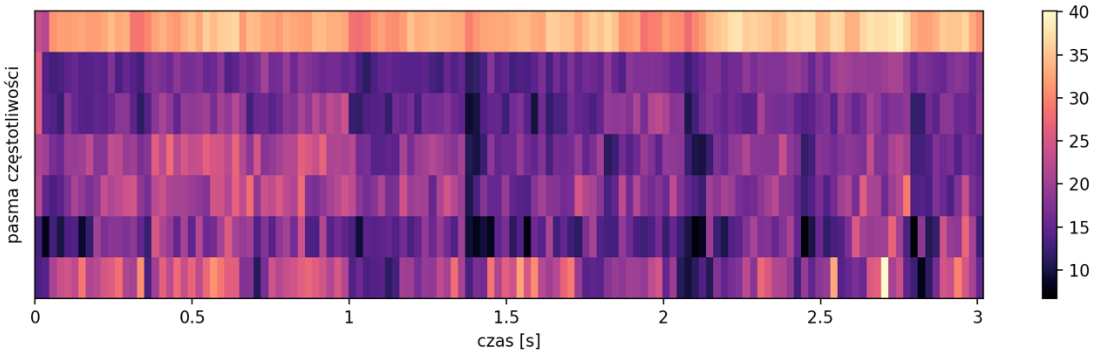
Dla ramki t o rozmiarze N, parametr ZCR definiowany jest jako [123]:

$$ZCR(t) = \frac{1}{2}\sum_{n=n_t}^{n_t+N} |\text{sign}(x[n]) - \text{sign}(x[n-1])| \tag{5.4}$$

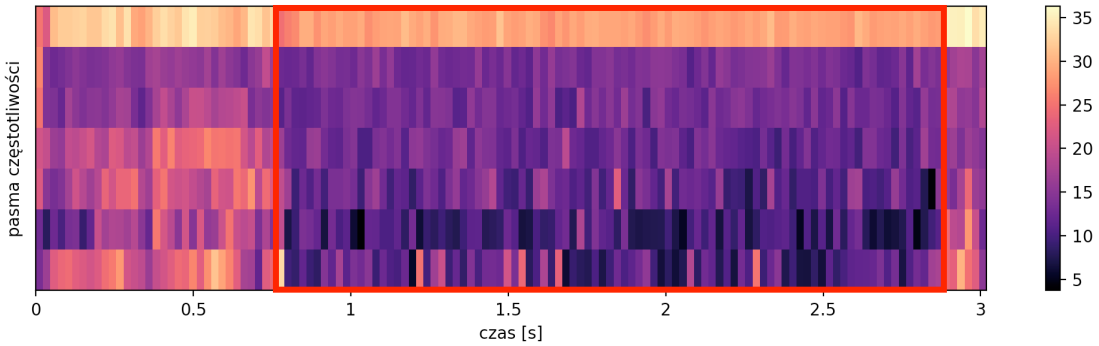
gdzie $\text{sign}(x[n])$ oznacza znak dla wartości chwilowej sygnału $x[n]$.

5.7. Parametr *Octave-Based Spectral Contrast*

Parametr *Octave-Based Spectral Contrast* (OBSC) to parametr szeroko wykorzystywany w zadaniach klasyfikacji muzyki [138] (np. wykrywanie tła muzycznego w nagraniach audio lub detekcja gatunków muzycznych). Pozwala na znalezienie różnic pomiędzy wartościami maksymalnymi i minimalnymi w widmie częstotliwościowym sygnału dla każdego pasma. W niniejszej pracy, do obliczenia tego parametru zastosowano sześć pasm (rys. 5.5, rys. 5.6). Wartości szczytowe widma częstotliwościowego reprezentują element harmoniczny, natomiast doliny w widmie – element nieharmoniczny (szum) [139].



Rys. 5.5. Kontrast spektralny z zastosowaniem sześciu pasm częstotliwości dla sygnału przedstawionego na przebiegu czasowym rysunku 5.3.



Rys. 5.6. Kontrast spektralny dla sygnału z nałożonym szumem na zaznaczonym fragmencie.

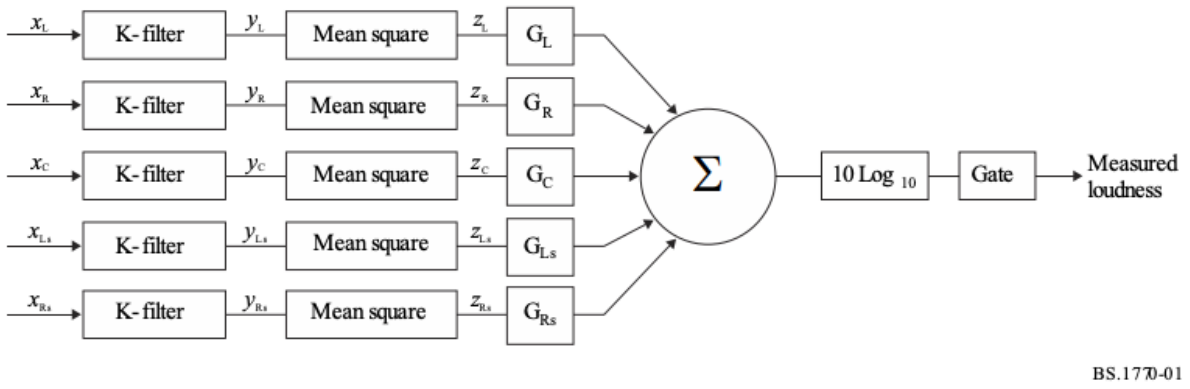
5.8. Głośność

Głośność sygnału definiowana jest jako subiektywna ocena poziomu dźwięku [53]. Ze względu na to, że ocena taka jest indywidualna dla każdego słuchacza, stworzenie urządzenia pomiarowego, którego wynik odpowiadałby percepcji sygnału przez słuch ludzki, jest zadaniem bardzo trudnym. Organizacja ITU stworzyła więc rekomendację dla pomiaru zintegrowanej głośności programu i zaproponowała algorytm, który działałby równie skutecznie dla sygnałów mono, stereo jak i wielokanałowych. W praktycznych implementacjach pomiaru poziomu głośności wykorzystywane są jednostki relatywne – LU (ang. *Loudness Unit*) oraz absolutne – LKFS (równoznaczne jednostce LUFS dla standardu *European Broadcast Union,* EBU). LKFS oznacza poziom sygnału w odniesieniu do pełnej skali (ang. *Full Scale*) z nałożeniem krzywej ważonej K. Przy pomiarze głośności stosowane są następujące miary [140]:

1) głośność uśredniona (ang. *Integrated Loudness*) to pomiar głośności obejmujący całość trwania sygnału. Wykorzystywane jest bramkowanie sygnału, zależnie od specyfikacji;
2) głośność krótko-czasowa (ang. *Short-Term Loudness*) to pomiar głośności bez wykorzystania algorytmów bramkowania, uśredniany na 3-sekundowych blokach sygnału;
3) głośność chwilowa (ang. *Momentary Loudness*) to pomiar głośności bez wykorzystania algorytmów bramkowania. Pomiar dokonywany jest na blokach o długości 400 ms.

Algorytm pomiaru głośności według najnowszej obecnie rekomendacji ITU BS.1770-4 [141] składa się z następujących kroków (rys. 5.7):

1) zastosowanie krzywej ważonej K;
2) obliczenie średniej arytmetycznej kwadratów wartości próbek każdego kanału;
3) przemnożenie każdego kanału audio przez odpowiednią wagę (np. kanały tylne mają większą wagę), a następnie zsumowanie wszystkich wartości;
4) bramkowanie o długości ramki 400 ms z nakładkowaniem 75% z zastosowaniem dwóch progów: pierwszy na poziomie –70 LKFS, drugi na poziomie –10 relatywnie do poziomu pomiaru otrzymanego po nałożeniu pierwszego progu.

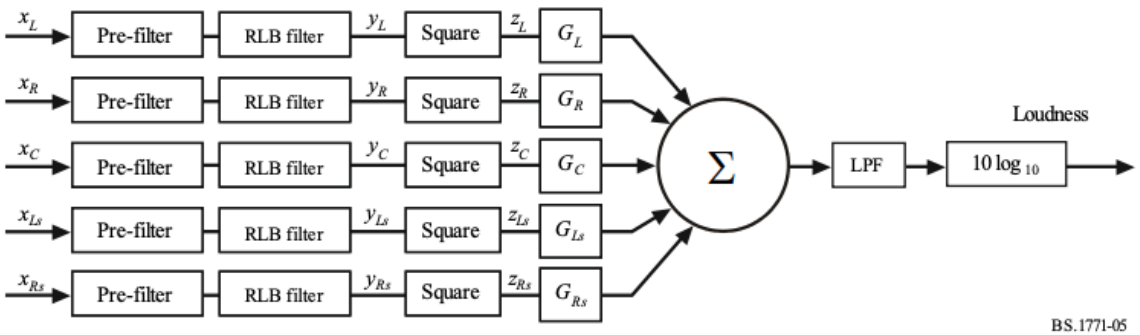


Rys. 5.7. Uproszczony schemat blokowy algorytmu pomiaru głośności dla programu zawierającego 1 – 5 kanałów według rekomendacji ITU BS.1770-4 [141].

5.9. Głośność chwilowa

Głośność chwilowa (ang. *Momentary Loudness*) to parametr obliczany na podstawie przesuwnego prostokątnego okna czasowego o długości 400 ms. W przypadku rekomendacji ITU, algorytm pomiaru jest podobny do opisanego w rozdz. 5.8, z tą różnicą, że zastosowany jest filtr dolnoprzepustowy (ang. *Low-Pass Filter*, LPF) przed obliczeniem końcowego wyniku, a wynik ten nie jest bramkowany (rys. 5.8). Dwa początkowe etapy filtrowania sygnału, tj. filtracja wstępna (ang. *pre-filter*) uwzględniająca własności akustyczne modelowanej głowy słuchacza oraz filtr górnoprzepustowy o charakterystyce krzywej RLB (ang. *Revised Low-frequency B-curve*) składają się na operację tzw. „ważenia K" (ang. *K-weighting*), która ma na celu odwzorowanie subiektywnych wrażeń odsłuchowych na obiektywny wynik pomiaru.

Pomiar ten stosowany jest głównie przy produkcji, post-produkcji oraz prezentacji sygnału [142]. Jest on również rekomendowany jako sprawdzenie, czy program nie przekracza założonego limitu przy docelowym odsłuchu [143].



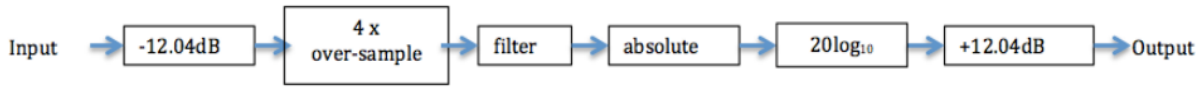
Rys. 5.8. Schemat blokowy pomiaru głośności chwilowej [142].

5.10. Rzeczywista wartość szczytowa sygnału

Parametr *True-Peak* polega na estymacji rzeczywistej wartości szczytowej dla danego sygnału audio. Jest to maksymalna wartość próbki, jaka może wystąpić w przebiegu sygnału w czasie ciągłym, tj. po przetworzeniu na sygnał analogowy. Jeszcze do niedawna urządzenia pomiarowe w cyfrowych systemach przetwarzania sygnałów audio miały możliwość rejestrowania jedynie parametru *Sample-Peak*, zamiast *True-Peak*. Pomiar taki polegał na porównaniu wartości bezwzględnej każdej kolejnej próbki dla sygnału cyfrowego, a wynikiem była najwyższa zmierzona wartość. Algorytm ten jest bardzo prosty w implementacji, jednakże występuje tu istotny problem – rzeczywiste wartości szczytowe dla próbkowanego sygnału występują zazwyczaj pomiędzy poszczególnymi próbkami, a nie dokładnie w momencie zarejestrowanej próbki [141]. Tym samym rzeczywista wartość szczytowa może znacznie różnić się od maksymalnej wartości *Sample-Peak*. Ponadto zauważono również, że pomiar parametru *Sample-Peak* prowadzi do zróżnicowanych wyników, np. jeśli sygnał w postaci cyfrowej jest kilkukrotnie konwertowany do różnych częstotliwości próbkowania. Nie chroni on również skutecznie przed przeciążeniami, które w wynikowym sygnale powodowałyby obcięcie fragmentów sygnału wykraczających poza skalę.

Algorytm pomiaru parametru *True-Peak* według rekomendacji ITU-R BS.1770-4 [141] składa się z następujących kroków (rys. 5.9):

1) zastosowanie tłumienia sygnału o wartości 12.04 dB;
2) zastosowanie 4-krotnego nadpróbkowania sygnału;
3) nałożenie filtru niskoczęstotliwościowego;
4) obliczenie wartości bezwzględnej;
5) konwersja uzyskanej wartości do skali dBTP.



Rys. 5.9. Schemat blokowy pomiaru true-peak [141].

6. Projekt modelu

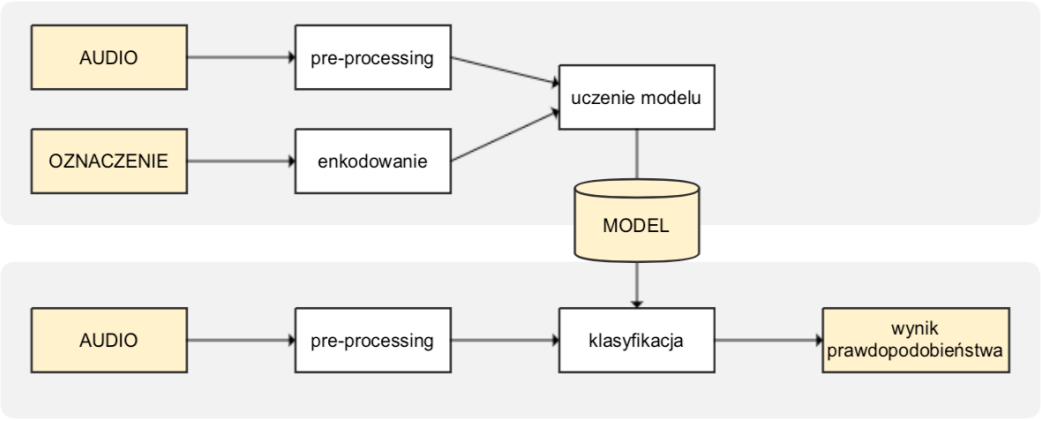
6.1. Wprowadzenie

W niniejszym rozdziale opisany został etap projektowania prototypowego modelu klasyfikacji zniekształceń dla sygnałów muzycznych. Na początku przedstawiony został ogólny schemat działania modelu oraz jego podstawowe założenia. Następnie opisane zostały parametry zastosowane do kontrolowania procesu uczenia modelu, tj. jego hiperparametry (rozdz. 6.3), główne warstwy sieci neuronowej i operacje z nimi związane (konwolucyjne: rozdz. 6.4 – 6.5 i rekurencyjne: rozdz. 6.6 – 6.8) oraz funkcje aktywacji (rozdz. 6.9).

6.2. Schemat działania modelu

Projektowany model opiera się na sieciach neuronowych, co wiąże się z istotną zaletą, że jego działanie polega na przetwarzaniu danych pozyskanych jedynie z sygnałów wejściowych, bez konieczności wstępnych założeń odnośnie ich zawartości czy też spodziewanych zakłóceń lub stacjonarności sygnału. Istotny wpływ na skuteczność modelu ma zarówno jego struktura, jak i przygotowana baza danych (treningowa, testowa oraz walidacyjna, opisane szerzej w rozdz. 5).

W modelu zastosowana została klasyfikacja typu *single-label* (tzn. każdy z segmentów audio został przypisany do pojedynczej kategorii). Oznaczenia segmentów zostały dołączone do danych wejściowych i na tej podstawie wagi modelu były stopniowo dostosowywane w procesie uczenia (tzw. *supervised learning*). W przygotowanej bazie sygnałów, kategorie są rozłączne (tj. żadna próbka nie należy do więcej niż jednej kategorii), a rozszerzenie bazy danych oraz modelu o obsługę klasyfikacji typu *multi-label*, gdzie pojedyncza próbka zawierałaby więcej niż jedno zniekształcenie, będzie przedmiotem dalszych badań.



Rys. 6.1. Procedura uczenia i testowania modelu sieci neuronowych klasyfikatora.

Procedura uczenia, walidacji oraz testowania zaimplementowanego modelu wygląda następująco (rys. 6.1): każdy sygnał z przygotowanej bazy danych poddawany jest przetwarzaniu wstępnemu (rozdz. 5.4), podczas którego następuje ekstrakcja wybranych cech

(np. spektrogramu). Uzyskane cechy wraz z oznaczeniem oczekiwanej kategorii sygnału, są przekazywane na wejście zaimplementowanego modelu, gdzie następuje proces uczenia w przypadku procedury treningowej lub końcowa klasyfikacja, w przypadku procedury walidacyjnej i testowej. Wyjściem modelu jest wynik prawdopodobieństwa przynależności danego sygnału do każdej z kategorii, a klasyfikacja następuje na podstawie wybrania wartości maksymalnej.

6.3. Hiperparametry modelu

Podczas implementacji sztucznych sieci neuronowych zostały dostosowane następujące hiperparametry modelu:

- liczba epok (ang. *epochs*);
- wielkość wsadu *(ang. batch size)*;
- funkcja kosztów (ang. *cost function*);
- funkcja optymalizująca (ang. *optimizer*);
- funkcja aktywacji (ang. *activation function*).

Pozostałe parametry sieci – ilość warstw, ich typ oraz połączenia ze sobą, ilość komórek dla każdej z nich, filtry oraz ich rozmiar – zostały opisane w rozdz. 8 – 10.

Liczba epok to ilość kompletnych cykli treningowych, podczas których dostosowywane są wagi modelu na podstawie wyniku funkcji kosztów [144]. W każdym cyklu wszystkie próbki z bazy treningowej są przetwarzane przez uczony model sieci neuronowych tylko raz. Mogą one być przekazywane do sieci neuronowej pojedynczo lub partiami, a rozmiar tych partii to tzw. *batch size*.

Model trenowany był z wielkością wsadu równą 100 oraz liczbą epok 50, przy czym po każdej epoce następowała walidacja modelu (z wykorzystaniem osobnej – walidacyjnej – bazy sygnałów) i to na jej podstawie wybrane zostały parametry docelowego modelu.

Wybór funkcji kosztów uzależniony jest od zastosowanej funkcji aktywacji [144]. Do zadań klasyfikacji wieloklasowych, zastosowanie ma wyjście z funkcją aktywacji *softmax*, w której zwracana jest wartość probabilistyczna. Wartość taka nie może być analizowana w identyczny sposób jak np. wartość numeryczna, wymaga ona użycia odpowiedniej wersji funkcji kosztów. Do zadań predykcji probabilistycznych, używane są dwa typy funkcji kosztów. Dla klasyfikacji binarnej stosowana jest funkcja tzw. *logistic regression*, natomiast dla klasyfikacji wielodecyzyjnej – *cross-entropy loss.* W pracy zastosowano więc funkcję *cross-entropy loss,* ponieważ celem jest przypisanie każdego wyjścia modelu do jednej z pięciu kategorii. Funkcja ta definiowana jest następująco [145, 146]:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k}^{5}(l_{nk})\log(f_a(\hat{y}_{nk})), \tag{6.1}$$

gdzie:

N – całkowita liczba próbek w bazie danych wejściowych,

k – kategoria sygnału,
l – oczekiwane oznaczenie dla n-tej sekwencji w bazie danych wejściowych,
$\hat{y}$ – wynik predykcji modelu sieci neuronowych,
f_a – to funkcja aktywacji, w niniejszej pracy *softmax*.

Funkcje aktywacji zostały opisane szerzej w rozdz. 6.9.

6.4. Konwolucyjne sieci neuronowe

Konwolucyjne sieci neuronowe (ang. *Convolutional Neural Networks,* CNN) to specjalny typ sieci wyspecjalizowany do przetwarzania danych opartych na topologii siatki (ang. *grid-structured*), o silnych zależnościach przestrzennych [144, 147], z tego względu szczególnie popularny w zadaniach przetwarzania obrazów.

CNN to sieć, zawierająca co najmniej jedną warstwę konwolucyjną, która pozwala na wyodrębnianie wzorców z obrazu niezależnie od ich lokalizacji. Działanie tej warstwy opiera się na liniowej operacji matematycznej – splotu, który polega na obliczeniu iloczynu skalarnego pomiędzy zbiorem wag (filtrów) w postaci macierzy w, a danymi wejściowymi x o strukturze siatkowej:

$$s(t) = (x * w)(t) = \int_{-\infty}^{\infty} x(a)w(t-a)da = -\frac{1}{N_T}\sum_{n=1}^{N_T}\sum_{i=0}^{2}(t_n)_i \ln\sigma(z_{in}), \tag{6.2}$$

Wynikiem operacji splotu jest tzw. mapa atrybutów (ang. *feature map*). W przypadku cyfrowego przetwarzania danych stosowana jest konwolucja dyskretna (np. przy założeniu, że filtry oraz dane wejściowe są zdefiniowane jedynie dla wartości całkowitych czasu t, np. w regularnych odstępach czasu):

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \tag{6.3}$$

Operacja splotu jest zazwyczaj używana wielokrotnie w modelach konwolucyjnych sieci neuronowych, w celu wykrycia jak największej ilości zależności przestrzennych dla danych wejściowych. Może ona być również wykorzystywana w większej liczbie wymiarów równocześnie. W przypadku przetwarzania obrazu *I* o dwóch wymiarach (np. spektrogramu), operacja splotu z użyciem filtru dwuwymiarowego definiowana jest następująco [147]:

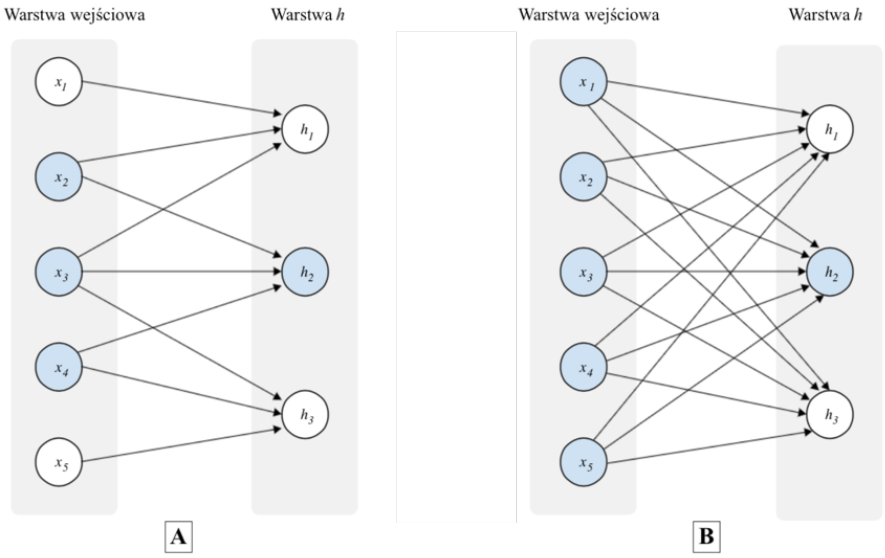
$$s(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(m,n)K(i-m, j-n) \tag{6.4}$$

Jednak w przypadku zastosowań praktycznych, w wielu bibliotekach programistycznych stosowana jest tzw. korelacja wzajemna (ang. *cross-correlation*). Odpowiada ona operacji

splotu, jednak jest prostsza w implementacji, gdyż nie wymaga odwracania kierunku przetwarzania:

$$s(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n) \tag{6.5}$$

Konwolucyjne sieci neuronowe wyróżniają się trzema istotnymi zaletami: zastosowaniem połączeń splotowych (ang. *sparse interactions*), dzieleniem parametrów (ang. *parameter sharing*) oraz równoważnością reprezentacji (ang. *equivariant representation*). W porównaniu do tradycyjnych warstw sieci neuronowych, które wymagają interakcji każdego neuronu wyjściowego z wejściowym, warstwy konwolucyjne wykorzystują tzw. połączenie splotowe (rys. 6.2). Pozwalają one na znaczną redukcję parametrów, a tym samym zmniejszenie wielkości potrzebnej pamięci oraz poprawę wydajności modelu. Uzyskiwane jest to dzięki zastosowaniu tzw. filtrów (ang. *filters*, zwane inaczej *kernels*), których rozmiar jest mniejszy niż danych wejściowych. W przypadku obrazu składającego się z milionów pikseli, filtry są w stanie wykryć najistotniejsze wzorce, redukując tym samym ilość wyjściowych danych.



Rys. 6.2. Schemat połączeń splotowych z parametrami współdzielonymi (A) oraz połączeń gęstych z parametrami niezależnymi (B).

Współdzielenie parametrów pozwala na ponowne użycie danego parametru przez więcej niż jedną funkcję w modelu, podczas gdy w tradycyjnych warstwach, każdy element z macierzy wag jest wykorzystany dokładnie raz do obliczenia danych wyjściowych [147]. Powodem dla którego parametry mogą zostać dzielone w przypadku sieci konwolucyjnych jest to, że dany wzorzec wykryty w dowolnym fragmencie obrazu powinien być przetwarzany w ten sam sposób, niezależnie od jego położenia w przestrzeni. Wiąże się to z pojęciem równoważności reprezentacji. Jeśli dany piksel zostanie przesunięty na obrazie w dowolnym kierunku o jedną jednostkę, a następnie zastosowana zostanie operacja splotu, uzyskany atrybut zostanie również odpowiednio przesunięty. Pozwala to na znaczącą poprawę efektywności modelu oraz zmniejszenie wymaganej pamięci.

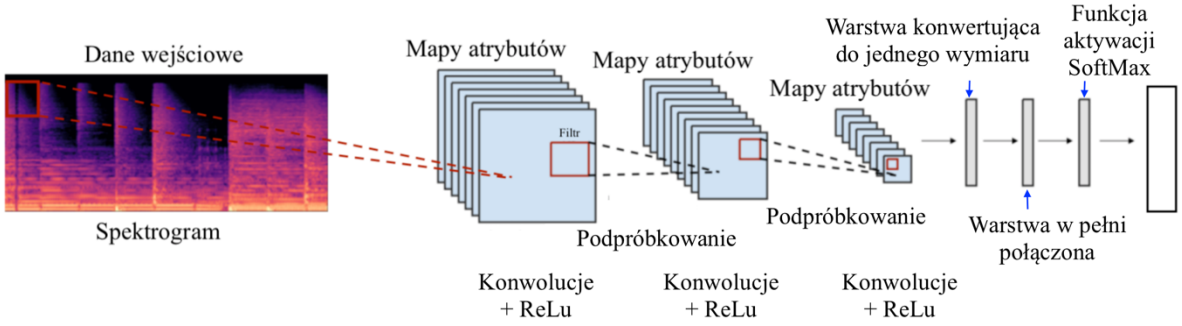
Elementami konwolucyjnych sieci neuronowych są tzw. filtry (ang. *filters* lub też *kernels*). Są to macierze o wymiarach $m \times n$, gdzie m i n są liczbami całkowitymi, zazwyczaj stosunkowo małymi (np. 3 lub 5) [148]. Filtry określają liczbę pikseli, które mają zostać wspólnie przeanalizowane, jednocześnie umożliwiając wykrycie różnego typu wzorców, np.:

1) krawędzi poziomych: $M = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$,

2) krawędzi pionowych: $M = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$,

3) lub fragmentów, gdzie znacząco zmienia się jasność obrazu: $M = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$.

Warstwa konwolucyjna składa się zazwyczaj z trzech etapów. Pierwszy etap to przeprowadzenie równoległych operacji konwolucji. Następnie dla wyniku każdej z nich zastosowana jest nieliniowa funkcja aktywacji, pozwalająca na wykrycie różnego typu wzorców z danych wejściowych. Ostatnim etapem jest operacja tzw. podpróbkowania [149]. Przykładowy schemat przetwarzania cech sygnału audio w postaci spektrogramów został przedstawiony na rys. 6.3.



Rys. 6.3. Przykład przetwarzania spektrogramu audio za pomocą konwolucyjnej sieci neuronowej.

6.5. Podpróbkowanie

Celem podpróbkowania jest nie tylko redukcja liczby parametrów przetwarzanych przez sieć konwolucyjną, ale również umożliwienie detekcji obiektów, które potencjalnie mogą zajmować większy obszar niż fragment analizowany przez dany filtr (którego wymiary są zazwyczaj stosunkowo małe) [149]. Podpróbkowanie zatem jest sposobem zmniejszania wymiarów analizowanego wewnątrz sieci obrazu. Najczęściej stosowanymi technikami są: konwolucje kroczące oraz operacja grupowania.

Konwolucje kroczące (ang. *strided convolutions*) polegają na pominięciu wybranej liczby pikseli podczas przesuwania filtra konwolucyjnego na obszarze analizowanego obrazu, a efektem takiej operacji jest zmniejszenie wielkości wynikowego obrazu.

Grupowanie (ang. *pooling*), zwane też inaczej redukcją lub odpytywaniem, to kolejna operacja stosowana w konwolucyjnych sieciach neuronowych, pozwalająca na znaczną redukcję złożoności obliczeniowej. Zazwyczaj jest ona ostatnim elementem warstwy konwolucyjnej i polega na modyfikacji wyjścia tej warstwy poprzez zastąpienie danego fragmentu danych wyjściowych zbiorczymi statystykami, uzyskanymi wspólnie dla sąsiadujących danych wyjściowych [147].

Metodą grupowania zastosowaną w modelu zaimplementowanym w niniejszej pracy jest tzw. *max-pooling*. Oznacza to, że dla każdego n-tego fragmentu danych wyjściowych (tzw. *activation maps*) o rozmiarze $p_n \times p_n$, uzyskanych po zastosowaniu nieliniowych funkcji aktywacji, operacja grupowania zwraca wartość maksymalną [144]. W przypadku zastosowania operacji grupowania o skoku S_n (ang. *stride*) równym 1, wynikowym rozmiarem utworzonej warstwy wyjściowej będzie:

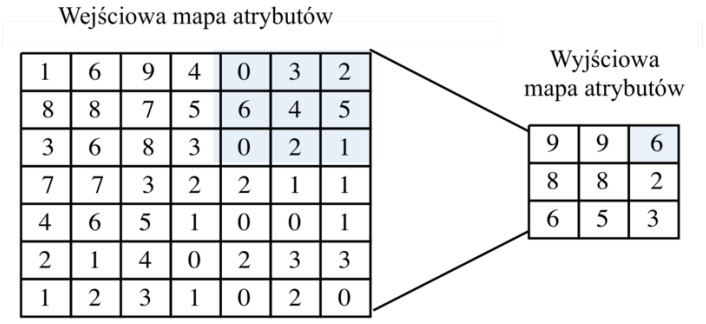
$$(L_n - p_n + 1) \times (B_n - p_n + 1) \times d_n \tag{6.6}$$

gdzie L_n oraz B_n to wymiary, a d_n to głębokość danych wejściowych dla n-tej warstwy.

Natomiast dla skoku S_n większego niż 1, rozmiar wynikowej warstwy zostanie znacznie zredukowany:

$$\frac{(L_n - p_n)}{S_n} + 1 \tag{6.7}$$

W przeciwieństwie do operacji konwolucji, grupowanie przetwarza każdą mapę atrybutów (ang. *feature maps*) osobno. Oznacza to, że liczba wyjściowych map atrybutów pozostaje bez zmian, a redukcji ulegają jedynie ich wymiary (rys. 6.4).



Rys. 6.4. Przykład grupowania typu max-pooling na blokach o rozmiarze 3x3 ze skokiem 2 dla wejściowej mapy atrybutów o rozmiarze 7x7.

Znane są również inne typy operacji grupowania, np. *average-pooling*, w tym jego wariant LeNet-5. Nie zostały jednak opisane w niniejszej pracy ze względu na to, iż nie były one zastosowane w zaimplementowanym modelu, nie są również tak popularne w praktycznych aplikacjach, jak opisane grupowanie typu *max-pooling*.

6.6. Rekurencyjne sieci neuronowe

Rekurencyjne sieci neuronowe (ang. *Recurrent Neural Networks*, RNN) to specjalny rodzaj architektury sieci neuronowych, który dzięki zastosowaniu sprzężeń zwrotnych, został przystosowany do przetwarzania danych sekwencyjnych [147]. Inne typy sieci neuronowych, w tym opisana wcześniej sieć konwolucyjna, nie są przygotowane do modelowania danych rozłożonych w dłuższych sekwencjach lub zależnościach o zmiennej długości czasu trwania, podczas gdy jest to szczególnie potrzebne w przypadku analizy i przetwarzania sygnałów audio, np. w zadaniach rozpoznawania tekstu z sygnału mowy, czy detekcji wybranych fragmentów w sygnałach muzycznych [123].

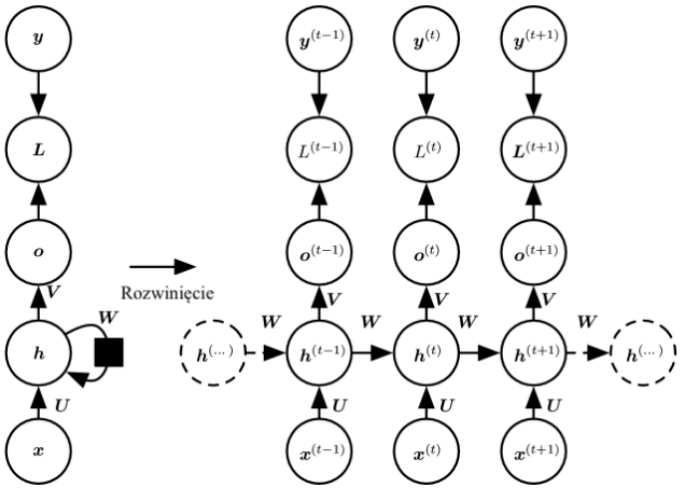
Podobnie jak sieci konwolucyjne lub MLP, sieć rekurencyjna jest zbudowana z warstw neuronów, jednak w tym przypadku, stan neuronu w czasie t jest zależny od stanu poprzedniego $(t-1)$, a ten również zależny od poprzedzającego itd. W najprostszej postaci warstwa rekurencyjna definiowana jest jako wektor stanów $h[t] \in \mathbb{R}^{d_i}$ dla czasu t, który jest obliczany na podstawie bieżących danych wejściowych $x[t] \in \mathbb{R}^{d_{i-1}}$ oraz poprzedniego wektora stanów $h[t-1]$ [123]:

$$h[t] \coloneqq \rho(w^t x[t] + v^t h[t-1] + b) \tag{6.8}$$

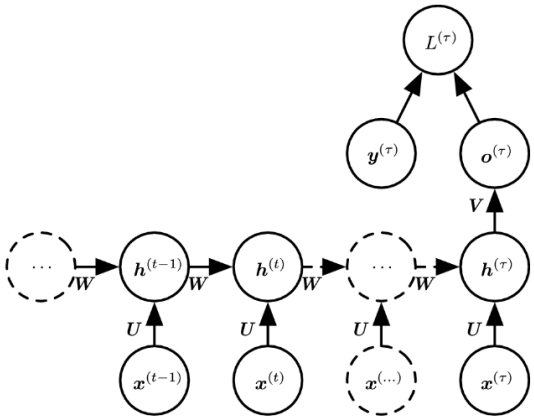
gdzie: ρ oznacza funkcję nieliniową, a warstwa sparametryzowana jest za pomocą macierzy wag wejściowych $w \in \mathbb{R}^{d_{i-1} \times d_i}$, macierzy wag rekurencyjnych $v \in \mathbb{R}^{d_i \times d_i}$ oraz wektora obciążeń (ang. *bias*) $b \in \mathbb{R}^{d_i}$.

Obliczanie wag warstwy rekurencyjnej w celu uwzględnienia informacji o poprzednich stanach stanowi problem, gdyż w przypadku, gdy wagi są stosunkowo małe, ich rekursywne mnożenie sprawia, że będą one bardzo szybko maleć. W innej sytuacji, gdy wagi będą stosunkowo duże, obliczany gradient będzie gwałtownie rosnąć. Sytuacja ta nazywana jest problemem znikającego i eksplodującego gradientu (ang. *vanishing and exploding gradient problem*) i pojawia się ona w szeroko stosowanej metodzie obliczania wag dla warstw rekurencyjnych, zwanej BPTT (ang. *Back-Propagation Through Time*).

Charakterystyczną cechą rekurencyjnej sieci neuronowej jest współdzielenie parametrów pomiędzy poszczególnymi warstwami sieci. Wiąże się ono z pojęciem rozwijania (ang. *unfolding*), gdzie obliczenia rekurencyjne są rozwijane na graf obliczeniowy. Poniżej przedstawione zostały dwa wybrane przykłady najważniejszych wzorców projektowych dla rekurencyjnych sieci neuronowych (rys. 6.5 i rys. 6.6).



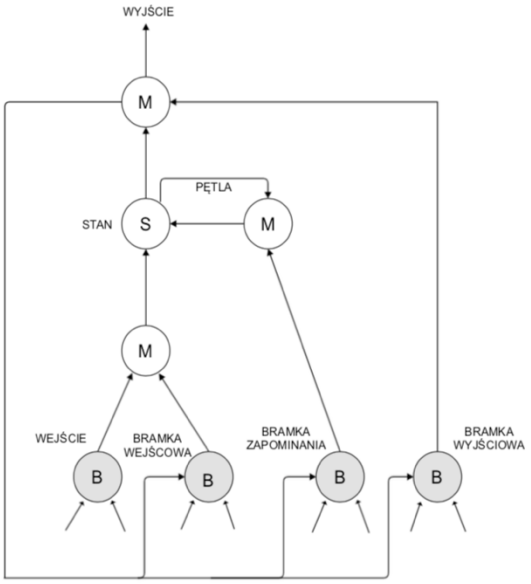
Rys. 6.5. Graf obliczeniowy dla przykładowej sieci rekurencyjnej, generującej wartości wynikowe w każdym kroku [147]. Parametry o, L oraz y oznaczają kolejno: bieżącą wartość wyjściową, stratę (ang. loss) oraz wyjściową wartość oczekiwaną.



Rys. 6.6. Graf obliczeniowy dla rozwiniętej w czasie rekurencyjnej sieci neuronowej, generującej pojedynczą wartość wynikową po przeanalizowaniu całej sekwencji danych [147]. Parametr τ oznacza długość sekwencji danych wejściowych.

6.7. Sieci typu *Long Short-Term Memory*

Sieci typu *Long Short-Term Memory* (LSTM) to sieci rekurencyjne bazujące na tzw. długiej pamięci krótkoterminowej. Obecnie są one oceniane jako jedne z najbardziej efektywnych wraz z sieciami opartymi na bramkowanej jednostce rekurencyjnej (ang. *Gated Recurrent Unit*, GRU) [147]. Warstwa LSTM składa się z trzech wektorów bramkujących, tj. bramki wejściowej, bramki zapominania oraz bramki wyjściowej oraz komórki pamięci i wektora stanów (rys. 6.7).



Rys. 6.7. Schemat blokowy przedstawiający przetwarzanie dla pojedynczego neuronu rekurencyjnej sieci LSTM. S oznacza sumę, M – multiplikację, B – bramkę.

Poszczególne elementy warstwy LSTM dla czasu t oraz komórki i są obliczane według następujących równań propagacji progresywnej [147]:

- bramka wejściowa $g_i^{(t)}$:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j u_{i,j}^g x_j^{(t)} + \sum_j w_{i,j}^g h_j^{(t-1)}\right) \tag{6.9}$$

- bramka zapominania $f_i^{(t)}$:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j u_{i,j}^f x_j^{(t)} + \sum_j w_{i,j}^f h_j^{(t-1)}\right) \tag{6.10}$$

- bramka wyjściowa $q_i^{(t)}$:

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j u_{i,j}^o x_j^{(t)} + \sum_j w_{i,j}^o h_j^{(t-1)}\right) \tag{6.11}$$

Zastosowanie funkcji sigmoidalnej w powyższych równaniach pozwala na uzyskanie wartości wag z zakresu od 0 do 1. Na ich podstawie obliczany jest wewnętrzny stan komórki LSTM:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma \left(b_i + \sum_j u_{i,j} x_j^{(t)} + \sum_j w_{i,j} h_j^{(t-1)} \right) \tag{6.12}$$

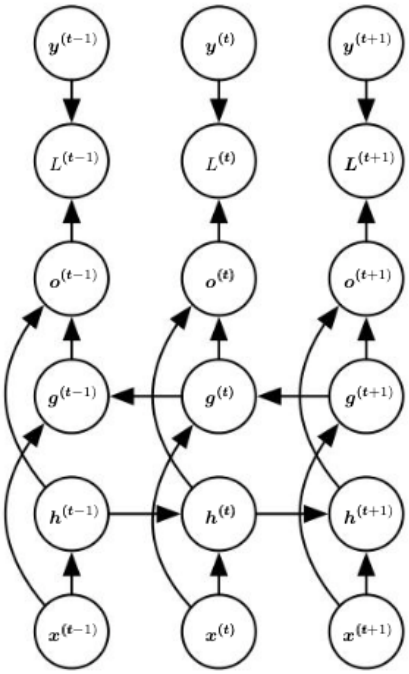
6.8. Dwukierunkowe sieci rekurencyjne

Standardowe architektury opisanych sieci RNN (rozdz. 6.6) i LSTM (rozdz. 6.7) pozwalają na przetwarzanie danych wejściowych w jednym kierunku, tzn. że stan neuronu w czasie t zależny jest jedynie od stanu poprzedniego. Jednak dla wielu praktycznych aplikacji korzystne jest wykorzystanie informacji z sekwencji danych w obu kierunkach – przeszłych i przyszłych. Najpopularniejszym przykładem jest zadanie rozpoznawania mowy, gdzie interpretacja fonemów może być bardziej skuteczna znając informację o kilku następnych fonemach, a wyraz może być zinterpretowany inaczej zależnie od tego w jakim kontekście – zdaniu – został użyty [147].

Przetwarzanie sekwencji danych w obu kierunkach możliwe jest dzięki rekurencyjnym sieciom dwukierunkowym (ang. *Bidirectional Recurrent Neural Networks*, BRNN), których każda warstwa składa się z dwóch warstw rekurencyjnych (rys. 6.8), jedna przekazująca informację do przodu w czasie (ang. *forward layer*), druga wstecz (ang. *backward layer*). Warstwa BRNN jest ich konkatenacją [123]:

$$f_i(x|\theta) \coloneqq \begin{bmatrix} f_i^{\rightarrow}(x|\theta) \\ \mathrm{rev}(f_i^{\leftarrow}(\mathrm{rev}(x)|\theta)) \end{bmatrix} \tag{6.13}$$

gdzie:
$f_i^{\rightarrow}$ to warstwa przekazująca informacje do przodu,
$f_i^{\leftarrow}$ to warstwa przekazująca informacje wstecz, rev oznacza zastosowanie wartości odwrotnej,
$f_i(x|\theta)$ jest wyjściem modelu integrującym informacje uzyskane z całej dostępnej sekwencji.

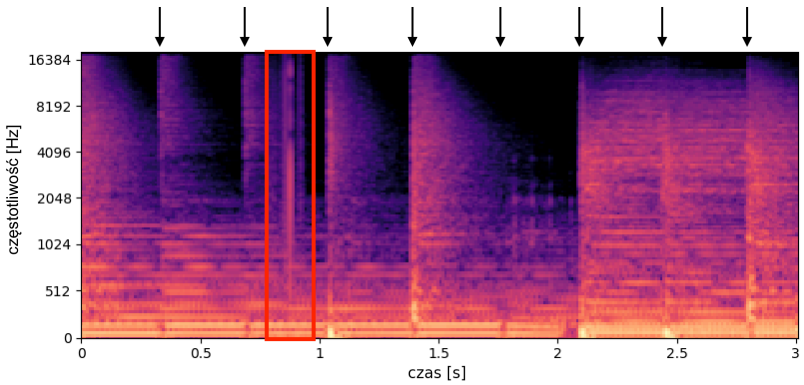


Rys. 6.8. Przykładowy graf dwukierunkowej sieci rekurencyjnej [147].

Zastosowanie sieci BRNN okazało się być efektywnym rozwiązaniem w zadaniach analizy sygnałów audio, m.in. w klasyfikacji zdarzeń akustycznych, tj. dźwięków środowiskowych [150], rozpoznawaniu emocji dla sygnałów muzycznych [151], klasyfikacji gatunków muzycznych [152]. W związku z tym, w niniejszej pracy ich skuteczność została również zbadana dla zadania wykrywania zniekształceń w rzeczywistych sygnałach muzycznych. Główna zaleta sieci dwukierunkowych, tj. zastosowanie informacji o stanie przyszłym, może być wadą w przypadku aplikacji działających w czasie rzeczywistym, gdzie te informacje mogą jeszcze nie być dostępne podczas przetwarzania bieżących danych. Jednak jednym z założeń dla zaimplementowanego modelu jest przetwarzanie typu *file-based* i dostarczenie do klasyfikatora jak największej liczby danych wejściowych w celu poprawnej detekcji zniekształceń. Analiza sygnałów w czasie rzeczywistym będzie przedmiotem dalszych badań. Porównując sieci RNN i BRNN do sieci konwolucyjnych, są one bardziej złożone obliczeniowo, jednak są też znacznie skuteczniejsze w rozpoznawaniu wzorców dla dłuższej sekwencji danych [153].

Głównym problemem w zadaniu klasyfikacji (lub też detekcji) zniekształceń w rzeczywistych sygnałach muzycznych, jest duże zróżnicowanie powstającego materiału dźwiękowego, a tym samym trudność znalezienia wzorca, który powinien być uznany za defekt. Przykładem mogą być nagrania instrumentów perkusyjnych – ich widmo częstotliwościowe ma szerokie spektrum, które podczas analizy może przypominać szum, czy trzaski, które w sygnale znaleźć się nie powinny. Decydując się na zastosowanie warstw rekurencyjnych dwukierunkowych, głównym celem było zwiększenie skuteczności implementowanego modelu dla tych właśnie przypadków. Zastosowanie sieci rekurencyjnej, która umożliwia zachowanie informacji o tym, co działo się w sygnale przed i po aktualnie

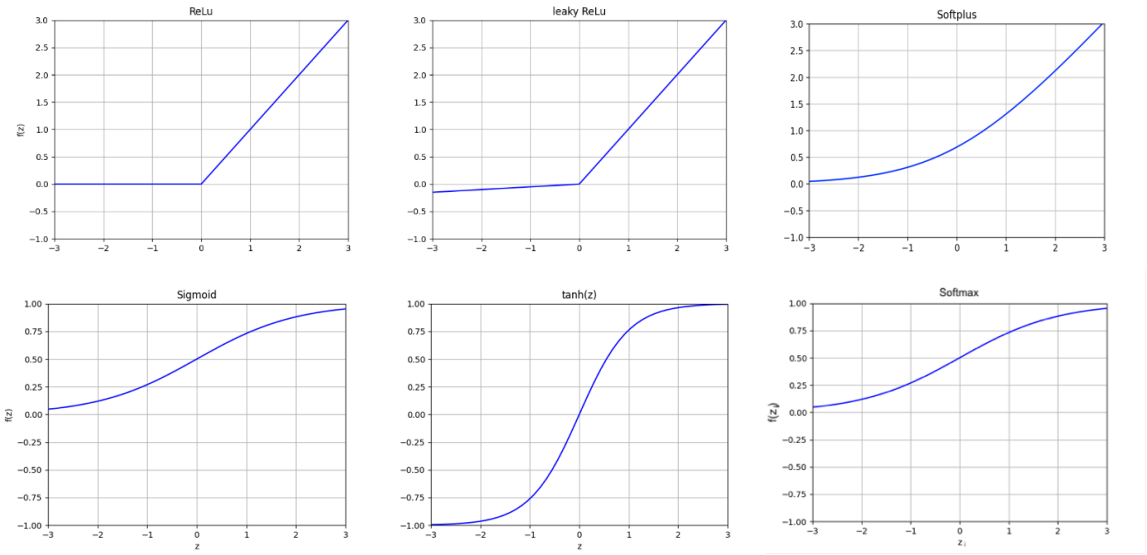
analizowanej próbce, może zwiększyć skuteczność wykrywania zniekształceń np. poprzez analizę czy dźwięki perkusji pojawiają się w stałych odstępach czasu (rys. 6.9).



Rys. 6.9. Widmo częstotliwościowe sygnału z oznaczonymi rytmicznymi uderzeniami perkusji (czarne strzałki) oraz zniekształceniem nie będącym częścią oryginalnego sygnału (czerwona ramka).

6.9. Funkcje aktywacji

Funkcja aktywacji (ang. *activation function* lub też *transfer function* lub *non-linearity*) ma bezpośredni wpływ zarówno na proces uczenia sieci neuronowej, jak i uzyskane wyniki, a wybór tej funkcji zależny jest od rozwiązywanego problemu. Część funkcji aktywacji nieliniowych (rys. 6.10, tab. 6.1), może ulec nasyceniu zwracając wartość stałą (np. funkcja tanh lub sigmoid) przy wzroście wartości modułu parametru *z*. W pewnych sytuacjach może to być zaletą, np. w przypadku końcowej klasyfikacji. W innych, może utrudniać proces uczenia i wpływać na propagację błędu, gdyż w momencie nasycenia do wartości stałej, pochodna w tym obszarze będzie równa 0 [123]. Rozwiązaniem tego problemu może być skalowanie funkcji aktywacji lub też użycie innej funkcji, nieulegającej nasyceniu np. *leaky* ReLu [154].



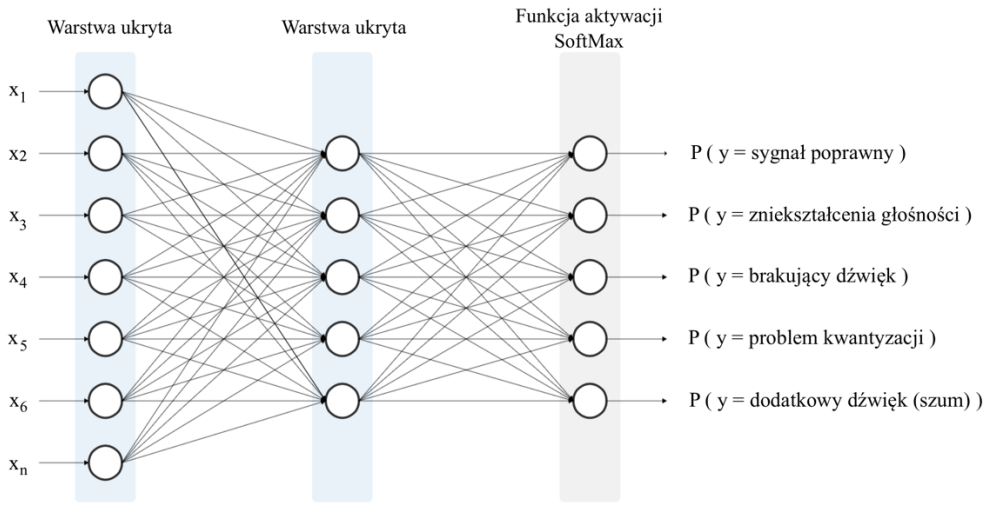
Rys. 6.10. Przykładowe funkcje aktywacji.

Tab. 6.1. Definicje wybranych funkcji aktywacji [155, 144].

Funkcja aktywacji	Definicja	Pierwsza pochodna
ReLu	$\sigma(z) = \max(0, z)$	$\varphi(z) = \begin{cases} 1\ dla\ z \geq 0 \\ 0\ dla\ z < 0 \end{cases}$
leaky ReLu	$\sigma(z) = \max(\alpha z, z)$	$\varphi(z) = \begin{cases} 1 \quad dla\ z \geq 0 \\ 0.01\ dla\ z < 0 \end{cases}$
sigmoid	$\sigma(z) = \frac{1}{1 + \exp(-z)}$	$\varphi(z) = \frac{\exp(-z)}{(1 + \exp(-z))^2}$
softplus	$\sigma(z) = \ln(1 + \exp(z))$	$\varphi(z) = \frac{1}{1 + \exp(-z)^2}$
tanh	$\sigma(z) = \frac{\exp(2z) - 1}{\exp(2z) + 1}$	$\varphi(z) = 1 - \sigma(z)^2$
softmax	$\sigma(z_i) = \frac{\exp(z_i)}{\sum_{p=1}^{K} \exp(z_p)}, \quad (i = 1, \ldots, k)$	$\varphi(z_i) = \sigma(z_i) - y_i\ , \quad (y \in \{0, 1\})$

W zaimplementowanym w pracy modelu zastosowane zostały następujące dwie funkcje aktywacji: ReLu (ang. *Rectified Linear Unit*) oraz *softmax*. Pierwsza z wymienionych funkcji jest funkcją aktywacji nasycającą się do wartości stałej tylko w przypadku ujemnych wartości wejściowych. Funkcja ReLu pozwala na ograniczenie problemu znikającego i eksplodującego gradientu (ang. *vanishing and exploding gradient problem*) opisanego w rozdz. 6.6, a to z kolei znacznie przyśpiesza proces uczenia, jednocześnie poprawiając skuteczność modelu [156]. W zaimplementowanym modelu, funkcja ta została zastosowana dla każdej warstwy konwolucyjnej, a także dla przedostatniej warstwy MLP. Zarówno w przypadku tradycyjnych warstw sieci neuronowych, jak i konwolucyjnych, funkcja ReLu nie zmienia wymiarów przetwarzanych danych.

Dla badanego w pracy problemu klasyfikacji typu *multi-class* zastosowana została funkcja aktywacji *softmax*. Pozwala ona na uzyskanie wektora o znormalizowanych, nieujemnych wartościach prawdopodobieństwa przynależności do każdej z kategorii, a suma wartości tego wektora jest równa 1 [144]. Funkcja *softmax* zastosowana została w ostatniej warstwie modelu (rys. 6.11), tj. warstwie klasyfikującej dany sygnał audio do jednej z pięciu kategorii.



Rys. 6.11. Ostatnia warstwa klasyfikująca modelu z funkcją aktywacji softmax.

7. Ewaluacja modelu

7.1. Wprowadzenie

Proponowany w niniejszej pracy model automatycznego klasyfikatora sygnału z eliminacją porównania do sygnału referencyjnego jest wynikiem eksperymentów z różnymi architekturami sieci neuronowych, hiperparametrów oraz zastosowania wybranych informacji ekstraktowanych z sygnałów wejściowych. Poniższy rozdział zawiera podsumowanie przeprowadzonych badań, dzielących się na trzy etapy:

1) zaimplementowanie podstawowego modelu sieci neuronowych z warstwami konwolucyjnymi, służącego za podstawę do dalszych modyfikacji;
2) zmodyfikowanie modelu bazowego przez dodanie warstw rekurencyjnych dwukierunkowych i sprawdzenie skuteczności klasyfikacji;
3) zastosowanie zaimplementowanego i wytrenowanego modelu w poprzednim etapie jako *transfer learning* i dodanie kolejnych informacji z sygnałów wejściowych w celu poprawy skuteczności klasyfikacji.

W każdym etapie sprawdzano skuteczność aktualnie zaimplementowanego modelu na podstawie miar klasyfikacji opisanych poniżej.

7.2. Miary klasyfikacji

Do oceny poprawności działania zaimplementowanego modelu oraz porównania poszczególnych parametrów modelu, użyte zostały następujące metryki: *accuracy*, *specificity*, *precision*, *recall*, *F1-score* [157, 158].

Miara *accuracy* to stosunek liczby próbek poprawnie sklasyfikowanych do łącznej liczby badanych przypadków [159]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.1}$$

gdzie:
TP (ang. *True Positive*) – liczba próbek z klasy pozytywnej, które zostały poprawnie sklasyfikowane jako przynależące do danej kategorii;
FP (ang. *False Positive*) – liczba próbek z klasy negatywnej, błędnie sklasyfikowane jako należące do danej kategorii;
TN (ang. *True Negative*) – liczba próbek z klasy negatywnej, poprawnie sklasyfikowanych jako nienależące do danej kategorii;
FN (ang. *False Negative*) – liczba próbek z klasy pozytywnej, błędnie odrzuconych jako nienależące do danej kategorii.

W dalszej części pracy zmierzone wartości *accuracy* podawane są jako wynik ogólny (wspólny dla wszystkich kategorii). Natomiast dla każdej z kategorii sygnałów podawane są bardziej szczegółowe metryki opisane poniżej.

Miara *specificity* lub inaczej *True Negative Rate* (*TNR*) jest kolejną miarą klasyfikatora, wskazującą jaka jest poprawność klasyfikacji dla próbek z klasy negatywnej (nieprzynależących do danej kategorii):

$$Specificity\ (TNR) = \frac{TN}{TN + FP} \tag{7.2}$$

Miara *precision* to miara wskazująca jaka jest poprawność klasyfikacji dla próbek z klasy pozytywnej – ile próbek sklasyfikowanych jako pozytywne, rzeczywiście przynależą do danej kategorii:

$$Precision = \frac{TP}{TP + FP} \tag{7.3}$$

Miara *recall* lub inaczej *sensitivity* – miara wskazująca jaka jest poprawność klasyfikacji dla próbek z klasy pozytywnej – ile próbek należących do danej kategorii zostało poprawnie sklasyfikowanych jako pozytywne:

$$Recall = \frac{TP}{TP + FN} \tag{7.4}$$

Miara *F1-score* – łączy wyniki dwóch poprzednich metryk *precision* oraz *recall* jako:

$$Fscore = \frac{2 \cdot Precision\ \cdot Recall}{Precision + Recall} \tag{7.5}$$

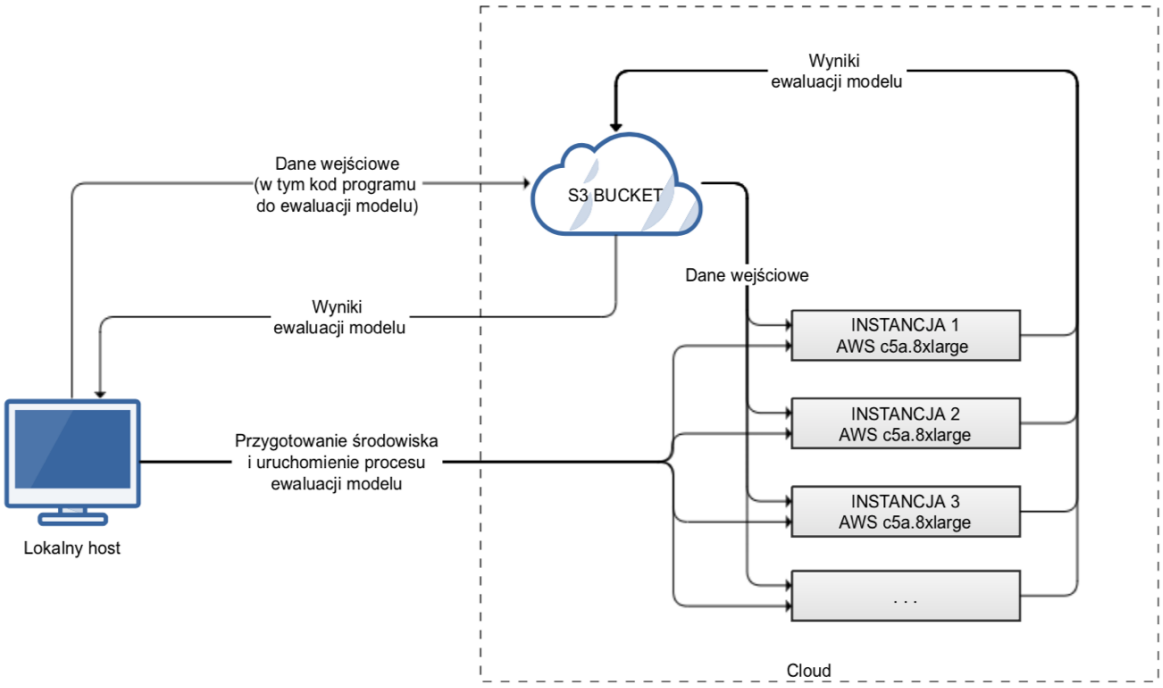
7.3. Wymagania sprzętowe

Proces ewaluacji różnych wariantów architektury modelu sieci neuronowych przeprowadzony został na instancji Amazon EC2 (*Amazon Elastic Compute Cloud*) dostępnej poprzez serwis *Amazon Web Services* (AWS) *Cloud*, który oferuje m.in. wirtualne środowiska obliczeniowe o konfigurowalnych parametrach. Wybrano typ maszyny to c5a.4xlarge i c5a.8xlarge [160], zoptymalizowany dla zadań uczenia maszynowego. Podstawowe parametry instancji zostały podane w tab. 7.1. Użyty został również prekonfigurowany obraz systemu *Deep Learning AMI GPU TensorFlow 2.11.0 (Ubuntu 20.04)*. Dane wejściowe oraz wyjściowe modelu były dostarczane do/z wirtualnej maszyny za pomocą serwisu Amazon S3 (*Amazon Simple Storage Service*). W przypadku przetwarzania dużej ilości danych (w niniejszej pracy sygnałów muzycznych o łącznym czasie trwania prawie 10 godzin), dzięki wykorzystaniu serwisu S3 dane te nie musiały być transferowane ani przechowywane na lokalnym dysku.

Zastosowanie instancji Amazon EC2 serwisu AWS pozwoliło na zrównoleglenie procesu treningu oraz testowania różnych wariantów architektury modelu sieci neuronowych (rys. 7.1).

Tab. 7.1. Specyfikacja instancji EC2 AWS użytej do ewaluacji modelu sieci neuronowych.

Typ instancji	c5a.4xlarge / c5a.8xlarge
Procesor	AMD EPYC 7R32
Częstotliwość pracy procesora (GHz)	3,3
Ilość CPU	16 / 32
Pamięć (GB)	32 / 64
Dysk (GB)	EBS wielkość konfigurowalna
Przepustowość sieci (Gbps)	10
Przepustowość EBS (Mbps)	3170



Rys. 7.1. Środowisko przygotowane do ewaluacji modelu sieci neuronowej.

8. Model sieci konwolucyjnych

8.1. Wprowadzenie

Głównym celem rozprawy było zbadanie skuteczności zastosowania sieci neuronowych w automatycznej detekcji zniekształceń w sygnałach muzycznych, bez konieczności porównania do innego sygnału (referencyjnego). W związku z popularnością konwolucyjnych sieci neuronowych oraz ich wysokiej skuteczności zarówno w analizie obrazów jak i sygnałów audio, pierwszym badanym w pracy modelem była sieć neuronowa z zastosowaniem warstw konwolucyjnych.

Opisane w niniejszym rozdziale eksperymenty mogą być traktowane jako jedne z pierwszych prób zastosowania sztucznych sieci neuronowych do problemu automatycznej klasyfikacji (i detekcji) zniekształceń sygnałów audio (innych niż mowa), bez porównania do sygnału referencyjnego. Według najlepszej wiedzy autorki, nie została jeszcze wprowadzona rekomendowana automatyczna metoda do tego typu zadań, która mogłaby stanowić model referencyjny dla badań niniejszej pracy. W związku z szerokim zastosowaniem sieci konwolucyjnych do pokrewnych problemów analizy i klasyfikacji zdarzeń akustycznych [123], głównym celem implementacji modelu, wykorzystującego właśnie te warstwy, było sprawdzenie ich skuteczności dla zadanego w pracy problemu oraz potencjalne wykorzystanie uzyskanych w ten sposób wyników klasyfikacji jako referencji dla dalszych eksperymentów.

8.2. Architektura modelu

Zaimplementowany model sieci neuronowych (tab. 8.1) składa się z trzech warstw konwolucyjnych, każda z 256 filtrami o rozmiarze 3 [161] (wraz z zastosowaniem operacji *maxpooling* i *dropout*) oraz następującej po niej warstwie typu *dense*, składającej się z 64 komórek (oraz *dropout*). Celem zastosowania tak małego rozmiaru filtru (jednocześnie typowego dla sieci konwolucyjnych) jest detekcja obiektów odporna na przesunięcia sygnału w dziedzinie czasu, a redukcja typu *maxpooling* oraz *dropout*, ma na celu umożliwienie detekcji wzorców, które obejmują większy obszar niż fragment analizowany przed dany filtr.

Tab. 8.1. Architektura zaimplementowanego modelu z zastosowaniem konwolucyjnych sieci neuronowych.

Layer	# of filters	Kernel / pool size	Stride	Activation function
Conv2D	256	(3, 3)	(1, 2)	ReLu
Maxpooling 2D	N/A	(3, 2)	N/A	N/A
Dropout	0.3			
Conv2D	256	(3, 3)	N/A	ReLu
Maxpooling 2D	N/A	(1,2)	N/A	N/A
Dropout	0.3			
Conv2D	256	(3, 3)	N/A	ReLu
Maxpooling 2D	N/A	(1, 4)	N/A	N/A
Dropout	0.3			
Dense	64 nodes			ReLu
Dropout	0.3			
Dense	5 nodes			softmax

Dodatkowo, dla pierwszej warstwy konwolucyjnej zastosowany został skok 1 × 2, w celu dalszej redukcji złożoności obliczeniowej. Funkcją aktywacji użytą dla każdej warstwy splotowej była funkcja ReLu.

Ostatni element modelu to warstwa wyjściowa – klasyfikacyjna – składająca się z 5 komórek przy zastosowaniu funkcji aktywacji *softmax*. Wynikiem sieci jest wartość prawdopodobieństwa przynależności badanego sygnału do jednej z 5 kategorii.

Na wejście modelu podawane były spektrogramy przekonwertowane do skali melowej z zastosowaniem 256 filtrów, wyekstraktowane dla wszystkich sygnałów z przygotowanej bazy danych, co opisane zostało szerzej w rozdz. 5.3 – 5.5.

8.3. Analiza wyników klasyfikacji

Do oceny skuteczności zaimplementowanego modelu sieci neuronowych z warstwami konwolucyjnymi (CNN) zastosowano następujące metryki: *precision*, *recall*, *F1*, *TNR* (rozdz. 7.2). Podczas gdy na etapie uczenia modelu użyte zostały zbiory danych: treningowy oraz walidacyjny, końcowa ocena ewaluacji modelu dla wyżej wymienionych metryk obliczana była wykorzystując odrębny zbiór – testowy.

Na podstawie analizy uzyskanych wyników (tab. 8.2), najwyższą wartość *F1* uzyskano dla kategorii sygnałów z błędami kwantyzacji (0.946), jednakże dla sygnałów oryginalnych (bez zniekształceń) oraz zniekształceń wzmocnienia wynik *F1* to odpowiednio 0.612 oraz 0.633, co nie jest zadowalającym wynikiem z perspektywy potencjalnego zastosowania badanej metody do celów automatycznej detekcji zniekształceń i usprawnienia manualnych testów odsłuchowych, ponieważ jak opisano w rozdz. 5.3, przygotowana baza danych zawiera zniekształcenia, które w manualnych testach odsłuchowych byłyby wyraźnie słyszalne i ocenione jako „bardzo przeszkadzające". W związku z tym, uzyskany na tym etapie model sieci konwolucyjnych nie jest modelem proponowanym w niniejszej pracy jako potencjalne rozwiązanie do celów usprawnienia testów odsłuchowych, stanowił jednak podstawę do dalszych badań omówionych w rozdz. 9 i 10.

Tab. 8.2. Wyniki ewaluacji modelu sieci neuronowych z warstwami konwolucyjnymi (CNN).

KATEGORIA	CNN			
	precision	*recall*	*F1*	*TNR*
Czyste sygnały	0.542	0.703	0.612	0.851
Błędy kwantyzacji	0.997	0.900	0.946	0.999
Zniekształcenia wzmocnienia	0.657	0.611	0.633	0.920
Dodatkowy dźwięk (szum)	0.984	0.729	0.838	0.997
Brakujący dźwięk	0.657	0.611	0.876	0.951

8.4. Podsumowanie

W niniejszym rozdziale opisane zostały eksperymenty przeprowadzone w celu zbadania skuteczności konwolucyjnych sieci neuronowych do zastosowania w automatycznej klasyfikacji sygnałów muzycznych z podstawowymi zniekształceniami. W związku z popularnością zastosowania tego typu sieci do zadań klasyfikacji zdarzeń akustycznych, zaimplementowany model uznany został za bazę do kolejnych modyfikacji architektury oraz

parametrów, a wyniki uzyskane w tej części stanowią referencję dla dalszych etapów tej pracy. Ogólna skuteczność klasyfikacji uzyskanego modelu CNN, tj. metryka *accuracy* mierzona jako wartość wspólna dla wszystkich kategorii, wynosi 77.5%. Przy czym najniższe wartości metryk F1 oraz TNR uzyskano dla sygnałów oryginalnych bez zniekształceń (odpowiednio 0.612 i 0.851), co jest istotnym problemem w kontekście automatycznej klasyfikacji sygnałów muzycznych. Z perspektywy automatycznej oceny jakości dźwięku, najistotniejszym zadaniem jest poprawne odróżnienie sygnału oryginalnego od sygnału zaszumionego lub zniekształconego. Klasyfikacja zniekształcenia do konkretnej podkategorii, jak np. błędy kwantyzacji, jest tematem drugorzędnym, jednakże wciąż pomocnym z perspektywy dalszej analizy uzyskanych wyników, tj. znalezienia błędu w łańcuchu przetwarzania sygnału i jego potencjalnej naprawy. Kolejny opisany w pracy model sieci neuronowych (rozdz. 9), wraz z zastosowaniem dodatkowych parametrów wejściowych (rozdz. 10) ma na celu poprawę uzyskanych wyników w niniejszym rozdziale, zwłaszcza w kontekście poprawnej klasyfikacji sygnałów oryginalnych.

9. Model sieci konwolucyjno-rekurencyjnych

9.1. Wprowadzenie

Kolejnym badanym w pracy modelem była sieć konwolucyjna rozszerzona o zastosowanie dwukierunkowych sieci rekurencyjnych (CBRNN) z użyciem warstw LSTM. Zastosowanie warstw rekurencyjnych pozwala na znalezienie zależności sekwencyjnych w analizowanych sygnałach audio, przez co mogą one znacząco poprawić skuteczność sieci dla detekcji składowych niewystępujących w oryginalnym sygnale. Połączenie hybrydowe tych dwóch architektur – konwolucyjnej i rekurencyjnej – umożliwia wykorzystanie lokalnej detekcji wzorców z obrazu (w tym przypadku spektrogramu w skali melowej), wraz z modelowaniem sekwencyjnych zależności pomiędzy dostarczonymi danymi. Zastosowanie tego typu warstw sieci neuronowych okazało się efektywne w zadaniach przetwarzania sygnałów audio, jak rozpoznawanie mowy lub transkrypcja muzyki [123], w związku z tym niniejszy rozdział poświęcony jest analizie ich skuteczności dla badanego w pracy problemu. Według najlepszej wiedzy autorki, proponowany w niniejszej pracy model jest pierwszym opublikowanym zastosowaniem tego typu architektury do zadania automatycznej klasyfikacji zniekształceń sygnałów muzycznych, bez porównania do sygnału referencyjnego [118].

9.2. Architektura modelu

Proponowany w niniejszej pracy model automatycznego klasyfikatora sygnału z eliminacją porównania do sygnału referencyjnego składa się z trzech części (tab. 9.1). Na początku zostały użyte trzy warstwy konwolucyjne składające się z 256 filtrów o rozmiarze 3 [161], podobnie jak w opisanym wcześniej modelu CNN (rozdz. 8). Po każdej warstwie konwolucyjnej dodane zostały operacje typu *maxpooling* oraz *dropout*, w celu redukcji złożoności obliczeniowej. Następnie użyte zostały trzy warstwy rekurencyjne dwukierunkowe typu LSTM, z których każda składała się ze 128 ukrytych komórek.

Tab. 9.1. Architektura zaimplementowanego modelu CBRNN.

Layer	# of filters	Kernel / pool size	Stride	Activation function
Conv2D	256	(3, 3)	(1, 2)	ReLu
Maxpooling 2D	N/A	(3, 2)	N/A	N/A
Dropout	0.3			
Conv2D	256	(3, 3)	N/A	ReLu
Maxpooling 2D	N/A	(1,2)	N/A	N/A
Dropout	0.3			
Conv2D	256	(3, 3)	N/A	ReLu
Maxpooling 2D	N/A	(1, 4)	N/A	N/A
Dropout	0.3			
Reshaping	(86, 4 * 256)			
B-LSTM	128 nodes			
B-LSTM	128 nodes			
B-LSTM	128 nodes			
Dense	64 nodes			ReLu
Dropout	0.3			

Na końcu dodana została warstwa gęsta (ang. *dense layer*) składająca się z 64 komórek wraz z metodą regularyzacji *dropout* oraz warstwa wyjściowa składająca się z 5 komórek z zastosowaniem funkcji aktywacji *softmax*. Wynikiem sieci jest wartość prawdopodobieństwa przynależności badanego sygnału do jednej z 5 kategorii.

9.3. Wyniki klasyfikacji

W celu porównania skuteczności z poprzednio badanym modelem konwolucyjnych sieci neuronowych CNN (rozdz. 8), uczenie oraz walidacja nowego modelu CBRNN została przeprowadzona używając dokładnie tej samej bazy danych oraz tej samej techniki ekstrakcji cech akustycznych sygnału – spektrogramów w skali melowej (256 filtrów). Na podstawie uzyskanych wyników ewaluacji nowego modelu (tab. 9.2), zauważono znaczną poprawę klasyfikacji w przypadku zastosowania dodatkowych warstw rekurencyjnych. Analizując współczynnik *F1*, model CBRNN jest znacznie skuteczniejszy w przypadku klasyfikacji zniekształceń wzmocnienia (wynik wyższy o 0.121). Odnotowano również poprawę dla innych kategorii: sygnałów czystych (poprawa o 0.075), dodatkowego dźwięku (szumu) (poprawa o 0.037) oraz brakującego dźwięku (poprawa o 0.026).

Tab. 9.2. Wyniki ewaluacji modelu sieci neuronowych z warstwami konwolucyjno-rekurencyjnymi (CBRNN).

KATEGORIA	CBRNN			
	precision	*recall*	*F1*	*TNR*
Czyste sygnały	0.635	0.748	0.687	0.893
Błędy kwantyzacji	0.999	0.899	0.946	0.999
Zniekształcenia wzmocnienia	0.764	0.744	0.754	0.943
Dodatkowy dźwięk (szum)	0.990	0.784	0.875	0.998
Brakujący dźwięk	0.841	0.973	0.902	0.954

9.4. Wpływ liczby filtrów melowych dla sygnałów wejściowych na wyniki klasyfikacji

Na tym etapie pracy sprawdzono również wpływ liczby zastosowanych filtrów melowych dla sygnałów wejściowych na wyniki klasyfikacji zaimplementowanego modelu CBRNN. Według dostępnej literatury [162], dla zadania klasyfikacji zdarzeń akustycznych użycie 40 filtrów melowych dla spektrogramu audio skutkuje zadowalającymi wynikami w procesie ewaluacji modelu. Na podstawie przeprowadzonych eksperymentów wynika jednak, że w przypadku oceny jakości sygnałów muzycznych (znajdowania nieprawidłowości w sygnale), wymagana jest większa ilość informacji. Porównując wyniki skuteczności modelu, gdzie danymi wejściowymi były spektrogramy z nałożonymi filtrami melowymi w liczbie 40, 128 oraz 256, jedynie ostatnia wartość pozwalała na uzyskanie wyników, które mogłyby stanowić podstawę do dalszych eksperymentów (tj. ogólna skuteczność powyżej 70%).

9.5. Podsumowanie

W niniejszym rozdziale opisano badania przeprowadzone dla konwolucyjno-rekurencyjnych sieci neuronowych dla zadania automatycznej klasyfikacji zniekształceń

sygnałów muzycznych. Porównując skuteczność zaimplementowanego modelu CBRNN do omówionego w rozdziale 9 modelu CNN (tj. bez zastosowania warstw dwukierunkowych rekurencyjnych LSTM), aktualne wyniki są równe lub wyższe dla wszystkich kategorii badanych sygnałów, przy czym największą poprawę (o wartość 0.121 dla metryki F1) uzyskano dla sygnałów zawierających zniekształcenia charakterystyki wzmocnienia, a ogólna skuteczność modelu CBRNN, opisanego w niniejszym rozdziale, wynosi 83.0%. W przypadku kategorii sygnałów oryginalnych bez zniekształceń, uzyskano nieznaczną poprawę wartości F1 (o 0.075) w stosunku do poprzedniego modelu CNN. W kolejnym rozdziale opisane zostały eksperymenty przeprowadzone w celu dalszej poprawy skuteczności modelu.

10. Model sieci konwolucyjno-rekurencyjnych z zastosowaniem dodatkowych parametrów wejściowych

10.1. Wprowadzenie

W dalszej ewaluacji modelu sieci konwolucyjno-rekurencyjnych, opisanego w poprzednim rozdziale, sprawdzony został wpływ zastosowania dodatkowych parametrów na skuteczność klasyfikacji zniekształceń sygnałów audio. Uwzględnione zostały następujące parametry:

1. gęstość przejść przez zero, tj. parametr ZCR,
2. kontrast spektralny, tj. parametr OBSC,
3. parametry głośności: głośność chwilowa oraz rzeczywista wartość szczytowa sygnału.

Powyższe parametry zastosowane zostały w celu rozszerzenia zaimplementowanego wcześniejszego (rozdz. 9) modelu CBRNN (jako tzw. *transfer learning*). W związku z tym architektura sieci została odpowiednio zmodyfikowana, tak by możliwa była analiza dodatkowych parametrów. Zaimplementowane w pracy architektury sieci neuronowych wraz z analizą skuteczności klasyfikacji każdej z nich zostały przedstawione w kolejnych podrozdziałach.

10.2. Zastosowanie parametru ZCR

Pierwszym dodatkowym parametrem, który został sprawdzony w celu poprawy skuteczności proponowanego modelu CBRNN był parametr ZCR. Pozwala on określić jak bardzo zaszumiony jest sygnał – wyższy współczynnik może wskazywać na większą zawartość szumu, a tym samym niższą jakość sygnału [137] (parametr ten został omówiony szerzej w rozdz. 5.6).

10.2.1. Architektura modelu

Parametr ZCR przekazywany był do sieci neuronowej wraz ze spektrogramem w skali melowej (256 filtrów) dla danego sygnału. W celu zastosowania dodatkowego parametru, opisana wcześniej architektura sieci neuronowych została rozszerzona (tab. 10.1), w taki sposób, że uzyskany model CBRNN (rozdz. 9) zastosowano w tym przypadku jako ekstraktor parametrów (ang. *features extractor*) w celu analizy sygnałów w postaci spektrogramów w skali melowej, natomiast parametry ZCR przetwarzane były przez nową warstwę dwukierunkową rekurencyjną LSTM.

Tab. 10.1. Architektura zaimplementowanego modelu sieci neuronowych dla dodatkowego parametru: ZCR.

Warstwa	Komórki	Funkcja aktywacji	CBRNN (ekstraktor parametrów)
B-LSTM	128 komórek		
Konkatenacja			
Dense	64 komórek	ReLu	ReLu
Dropout	0.3	-	
Dense	5 komórek	softmax	softmax

Następnie na podstawie wszystkich dostępnych informacji – zarówno z nowej warstwy LSTM, jak i ekstraktora danych CBRNN – następowała klasyfikacja sygnału. Ostatni etap klasyfikacji składał się z dwóch warstw gęstych, z których pierwsza zawierała 64 komórki, natomiast druga 5 komórek z funkcją aktywacji typu *softmax*.

10.2.2. Wyniki klasyfikacji

Na podstawie wyników ewaluacji modelu CBRNN z zastosowaniem dodatkowego parametru ZCR (tab. 10.2) uzyskana została jedynie nieznaczna poprawa metryk *precision*, *recall* oraz *F1* porównując do poprzedniego modelu CBRNN z zastosowaniem wyłącznie spektrogramów w skali melowej jako danych wejściowych. Analizując metrykę *F1*, dla sygnałów z kategorii „sygnały czyste" uzyskano poprawę o wartość 0.013, natomiast dla „dodatkowy dźwięk (szum)" o 0.007. Największa różnica wystąpiła dla kategorii „brakujący dźwięk", tj. o 0.022.

Ogólna skuteczność klasyfikacji w tym przypadku wynosi 83.8%. Jest to poprawa o jedynie 0.8 punktu procentowego w stosunku do poprzedniego modelu CBRNN.

Tab. 10.2. Wyniki ewaluacji modelu sieci neuronowych z warstwami konwolucyjno-rekurencyjnymi (CBRNN) z zastosowaniem dodatkowego parametru wejściowego: ZCR.

KATEGORIA	Spektrogram w skali melowej + ZCR			
	precision	*recall*	*F1*	*TNR*
Czyste sygnały	0.637	0.777	0.700	0.889
Błędy kwantyzacji	0.999	0.899	0.946	0.999
Zniekształcenia wzmocnienia	0.765	0.752	0.758	0.942
Dodatkowy dźwięk (szum)	0.989	0.796	0.882	0.998
Brakujący dźwięk	0.886	0.967	0.924	0.969

10.3. Zastosowanie parametru OBSC

Kolejnym parametrem sprawdzanym pod kątem poprawy skuteczności zaimplementowanego modelu CBRNN był parametr OBSC, szczególnie często stosowany w zadaniach klasyfikacji sygnałów muzycznych. Parametr został opisany szerzej w rozdz. 5.7 i tak jak w przypadku parametru ZCR, przekazywany on był do sieci neuronowej wraz ze spektrogramem w skali melowej (256 filtrów).

10.3.1. Architektura modelu

Podobnie jak w przypadku analizy spektrogramów w skali melowej, parametr OBSC analizowany był jako 2-wymiarowy obraz, gdzie oś pozioma reprezentuje czas (sekundy), natomiast oś pionowa – pasma częstotliwości, a wartości dla każdego podpasma (kolor) to kontrast energii szacowany przez porównanie średniej energii w górnym (tzw. *peak energy*) i dolnym kwantylu (tzw. *valley energy*). Do analizy takiego obrazu zastosowano pojedynczą warstwę konwolucyjną z liczbą filtrów 256 wraz z regularyzacją typu *maxpooling* i *dropout* oraz skokiem 1×2, a jej wynik przetworzony został przez pojedynczą warstwę dwukierunkową LSTM.

Wyjście z warstwy B-LSTM połączone zostało z informacjami uzyskanymi z zaimplementowanego wcześniej modelu CBRNN (rozdz. 9), zastosowanego jako ekstraktor parametrów dla spektrogramów w skali melowej. Ostatnia warstwa klasyfikacyjna pozostała niezmieniona (tab. 10.3).

Tab. 10.3. Architektura zaimplementowanego modelu sieci neuronowych dla dodatkowego parametru: OBSC.

Warstwa	Ilość filtrów	Rozmiar	Skok	Funkcja aktywacji	
Conv2D	256	(3, 3)	(1, 2)	ReLu	CBRNN (ekstraktor parametrów)
Maxpooling 2D	-	(3, 2)	-	-	
Dropout	0.3				
Reshaping	(86, 7 * 256)				
B-LSTM	128 nodes				
Konkatenacja					
Dense	64 komórki			ReLu	ReLu
Dropout	0.3				
Dense	5 komórki			softmax	softmax

10.3.2. Wyniki klasyfikacji

Na podstawie wyników ewaluacji modelu CBRNN z zastosowaniem dodatkowego parametru OBSC (tab. 10.4) uzyskana została znaczna poprawa klasyfikacji dla większości kategorii, porównując do modelu CBRNN z zastosowaniem wyłącznie spektrogramów w skali melowej jako danych wejściowych. Analizując wyniki F1, dla sygnałów czystych – wartość ta wzrosła o 0.18, zniekształceń wzmocnienia – 0.188, dodatkowy dźwięk (szum) – 0.009, brakujący dźwięk – 0.04.

Ogólna skuteczność tak rozbudowanego modelu to 91.4% i jest to wzrost o 8.4 punkty procentowe w stosunku do poprzedniego modelu CBRNN (rozdz. 9). Najbardziej istotna jest w tym przypadku poprawa skuteczności klasyfikacji dla sygnałów oryginalnych bez zniekształceń, co było najistotniejszym problemem na początkowych etapach pracy.

Tab. 10.4. Wyniki ewaluacji modelu sieci neuronowych z warstwami konwolucyjno-rekurencyjnymi (CBRNN) z zastosowaniem dodatkowego parametru wejściowego: OBSC.

KATEGORIA	Spektrogram w skali melowej + OBSC			
	precision	*recall*	*F1*	*TNR*
Czyste sygnały	0.775	0.969	**0.861**	0.930
Błędy kwantyzacji	0.999	0.899	0.946	0.999
Zniekształcenia wzmocnienia	0.949	0.935	**0.942**	0.988
Dodatkowy dźwięk (szum)	0.987	0.800	**0.884**	0.997
Brakujący dźwięk	0.918	0.967	**0.942**	0.978

10.4. Zastosowanie parametrów głośności

Dla zaimplementowanego modelu automatycznego klasyfikatora sygnału z eliminacją porównania do sygnału referencyjnego zbadano również wpływ krótko-czasowych parametrów głośności jako dodatkowych danych wejściowych. Zostały sprawdzone następujące parametry: głośności chwilowe oraz rzeczywiste wartości szczytowe sygnału.

Wartości głośności chwilowych zostały przekazane na wejście modelu jako wektory wartości, mierzone na ramkach sygnału o długości 400 ms. Podobnie, dla rzeczywistych poziomów szczytowych sygnału – wektor danych zawierał pomiary dla ramek o długości 100 ms. Oba parametry zostały opisane szerzej w rozdz. 5.9 (głośność chwilowa) oraz 5.10 (rzeczywista wartość szczytowa sygnału). Wraz z nowymi parametrami, do modelu przekazywane były spektrogramy w skali melowej (256 filtrów) jako dane wejściowe.

10.4.1. Architektura modelu

Oba dodatkowe parametry opisywane w niniejszym rozdziale mierzone są na krótkich fragmentach sygnału, a parametrem wejściowym dla modelu sieci neuronowych jest wektor wartości – dla głośności chwilowej są to wartości mierzone dla ramek długości 400 ms, dla wartości szczytowych – dla ramek długości 100 ms. Do przeanalizowania nowych danych zastosowano dodatkową warstwę dwukierunkową rekurencyjną LSTM.

Zmodyfikowana architektura sieci składa się z trzech części: 1) analizy nowych danych za pomocą pojedynczej warstwy B-LSTM, 2) zastosowania wytrenowanego modelu CBRNN (rozdz. 9) jako ekstraktor parametrów dla spektrogramów w skali melowej, 3) połączenia uzyskanych z poprzednich kroków wyników oraz przekazanie ich do warstwy klasyfikacyjnej (tab. 10.5). Warstwa klasyfikacyjna pozostała niezmieniona w stosunku do poprzednio opisywanych modeli.

Tab. 10.5. Architektura zaimplementowanego modelu sieci neuronowych dla dodatkowych parametrów głośności.

Layer	Nodes	Activation function	Pre-trained CBRNN (features extractor)
B-LSTM	24 nodes		
Concatenate			
Dense	64 nodes	ReLu	ReLu
Dropout	0.3		
Dense	5 nodes	softmax	softmax

10.4.2. Wyniki klasyfikacji

Na podstawie uzyskanych wyników ewaluacji modelu CBRNN z zastosowaniem dodatkowych parametrów głośności (tab. 10.6 oraz tab. 10.7), w obu przypadkach zauważono poprawę skuteczności klasyfikacji w porównaniu do wyników modelu CBRNN z zastosowaniem wyłącznie spektrogramu w skali melowej jako danych wejściowych.

Tab. 10.6. Wyniki ewaluacji modelu sieci neuronowych z warstwami konwolucyjno-rekurencyjnymi (CBRNN) z zastosowaniem dodatkowego parametru wejściowego: głośności chwilowych.

KATEGORIA	CBRNN + głośność chwilowa			
	precision	*recall*	*F1*	*TNR*
Czyste sygnały	0.715	0.815	0.762	0.919
Błędy kwantyzacji	0.999	0.899	0.946	0.999
Zniekształcenia wzmocnienia	0.778	0.817	0.797	0.942
Dodatkowy dźwięk (szum)	0.987	0.810	0.890	0.997
Brakujący dźwięk	0.891	0.970	0.930	0.970

W przypadku zastosowania wektora głośności chwilowych, analizując wartość metryki F1, dla sygnałów czystych odnotowano wzrost o 0.075, zniekształceń wzmocnienia – o 0.043, dodatkowy dźwięk (szum) – o 0.015, brakujący dźwięk – o 0.028. Ogólna skuteczność tego modelu to 86.2%, co stanowi poprawę o 3.2 punkty procentowe w stosunku do podstawowego modelu CBRNN.

Tab. 10.7. Wyniki ewaluacji modelu sieci neuronowych z warstwami konwolucyjno-rekurencyjnymi (CBRNN) z zastosowaniem dodatkowego parametru wejściowego: rzeczywiste wartości szczytowe sygnału.

KATEGORIA	CBRNN + rzeczywista wartość szczytowa (mierzona na ramkę o długości 100ms)			
	precision	*recall*	*F1*	*TNR*
Czyste sygnały	0.743	**0.839**	**0.788**	0.928
Błędy kwantyzacji	0.999	0.899	0.946	0.999
Zniekształcenia wzmocnienia	**0.797**	0.839	**0.817**	**0.947**
Dodatkowy dźwięk (szum)	0.992	0.830	0.904	0.998
Brakujący dźwięk	**0.904**	**0.979**	**0.940**	**0.974**

W przypadku zastosowania wektora rzeczywistych wartości szczytowych sygnału, poprawa wyników dla metryki F1 jest większa niż uzyskano poprzednio dla głośności chwilowych. Dla sygnałów czystych jest to poprawa o 0.101, zniekształceń wzmocnienia – o 0.063, dodatkowy dźwięk (szum) – o 0.029, brakujący dźwięk – o 0.038. Ogólna skuteczność tego modelu to 87.7%, co stanowi poprawę o 4.7 punktów procentowych w stosunku do podstawowego modelu CBRNN.

10.5. Podsumowanie

Głównym celem niniejszej pracy było zaimplementowanie prototypowego modelu oraz zbadanie jego skuteczności dla zadania oceny jakości rzeczywistych sygnałów muzycznych. Z perspektywy testowania oraz monitorowania sygnałów audio, najistotniejszą kwestią jest decyzja, czy badany sygnał jest poprawny (tj. bez jakichkolwiek słyszalnych zniekształceń). Natomiast sama klasyfikacja poszczególnych kategorii traktowana jest jako zadanie drugorzędne, pomocne przy ewaluacji modelu oraz analizie jakie fragmenty sygnału mogą być najbardziej problematyczne dla zaprojektowanego modelu.

W przypadku proponowanego modelu sieci neuronowych z zastosowaniem warstw konwolucyjno-rekurencyjnych oraz spektrogramów w skali melowej jako danych wejściowych, najwyższy współczynnik FPR (*False Positive Ratio*) odnotowany został dla sygnałów czystych (tj. 0.1074) oraz zniekształceń wzmocnienia (0.0573). Wykorzystanie dodatkowego parametru ZCR nie wpływa znacząco na wyniki, jednak przy zastosowaniu parametru OBSC, błędy te zostały zredukowane, odpowiednio do wartości 0.0703 oraz 0.0124. Poprawę zauważono również w przypadku dodania parametrów głośności. Przy zastosowaniu pomiarów rzeczywistych wartości szczytowych błędy te zostały zredukowane odpowiednio do wartości 0.0725 oraz 0.0535. Natomiast w przypadku głośności chwilowych FPR dla sygnałów czystych uległa poprawie (0.08126), jednak dla zniekształceń wzmocnienia – nieznacznemu pogorszeniu (0.0583).

Najwyższy współczynnik FNR (False Negative Ratio) występuje w niniejszej pracy dla zniekształceń wzmocnienia (0.2559). Błędy są również dostrzegalne dla kategorii sygnałów czystych (0.2525) oraz dodatkowego dźwięku, tj. szumu (0.2164). Również w tym przypadku błędy te zostały zredukowane poprzez zastosowanie dodatkowego parametru OBSC do wartości odpowiednio 0.2482, 0.2233 oraz 0.2040. Jednakże, biorąc pod uwagę fakt, że model idealny charakteryzowałby się błędami jak najbliższymi wartości 0, uzyskane błędy powinny być przedmiotem dalszych badań w tym temacie.

W celu dalszej ewaluacji modelu zidentyfikowane zostały następujące możliwości poprawy:

1. wyekstraktowanie dodatkowych cech z sygnałów muzycznych oraz w miarę potrzeby – rozszerzenie modelu o dodatkowe warstwy;
2. zastosowanie klasyfikacji typu *multi-label* (obecnie klasyfikacja ograniczona jest do jednej konkretnej kategorii, przy czym rzeczywiste sygnały mogą zawierać kilka różnych zniekształceń jednocześnie);
3. rozszerzenie bazy danych o dodatkowe rzeczywiste sygnały muzyczne (szczególnie o sygnały problematyczne, np. muzyka elektroniczna).

Porównując wyniki wszystkich eksperymentów opisanych w niniejszym rozdziale, najwyższą skuteczność klasyfikacji sygnałów uzyskano dla modelu sieci neuronowych z wykorzystaniem warstw konwolucyjno-rekurencyjnych (CBRNN) wraz ze spektrogramami w skali melowej oraz parametru OBSC jako danych wejściowych (tj. ogólna skuteczność równa 91.4%). Model ten jest zatem proponowany w niniejszej pracy jako prototyp automatycznego klasyfikatora sygnału z eliminacją porównania do sygnału referencyjnego i stanowi podstawę do dalszych badań w temacie automatycznej analizy jakości sygnałów muzycznych.

Stosunkowo wysoką skuteczność uzyskano również przy zastosowaniu pomiarów rzeczywistych wartości szczytowych (87.7%) oraz głośności chwilowych (86.2%). W przypadku dodatkowego parametru ZCR, skuteczność modelu nie uległa istotnej poprawie (83.8%).

11. Podsumowanie

Tematem niniejszej pracy była klasyfikacja podstawowych zniekształceń w rzeczywistych sygnałach muzycznych, jako pierwszy etap do automatycznej oceny jakości sygnałów niebędących mową, bez porównania do sygnału referencyjnego. Celem opisanych badań było sprawdzenie skuteczności zastosowania do tego zadania sieci neuronowych, w tym konwolucyjnych i rekurencyjnych, które charakteryzują się wysoką skutecznością w podobnych zadaniach przetwarzania sygnałów audio. W pracy został zaprojektowany oraz zaimplementowany autorski prototypowy model sieci neuronowych, a jego skuteczność została zweryfikowana na podstawie stworzonej specjalnie do tego celu bazy danych sygnałów muzycznych zawierającej ponad 10 godzin nagrań.

Przedstawiony w pracy prototypowy model automatycznego klasyfikatora sygnałów muzycznych bez porównania do sygnału referencyjnego, rozwijany był w kontekście potencjalnego usprawnienia subiektywnych testów odsłuchowych. Jest to wciąż najpopularniejsza oraz najskuteczniejsza metoda oceny jakości sygnałów audio, jednakże – poprzez swój manualny charakter – bardzo czasochłonna, kosztowna oraz podatna na błędy, ze względu na subiektywny charakter oceny. Rozwiązaniem mogą być obiektywne metody oceny jakości sygnału, jednak znane obecnie metody w zdecydowanej większości opierają się na porównaniu sygnału do referencji, a metody bezreferencyjne rozwijane są głównie dla sygnału mowy. Ze względu na to, że sygnał referencyjny często jest niedostępny podczas badania aktualnie testowanego sygnału, istotnym usprawnieniem byłaby automatyczna obiektywna metoda oceny jakości dźwięku, bez konieczności porównania go do innego znanego sygnału oraz bez wstępnych założeń odnośnie zawartości badanego sygnału lub jego zniekształceń. Na podstawie przeprowadzonego przeglądu literatury, nie jest obecnie znana zalecana automatyczna metoda oceny jakości sygnałów muzycznych bez porównania do referencji. W niniejszej pracy opracowano więc autorski prototypowy model umożliwiający wstępną ocenę sygnału poprzez klasyfikację go do jednej z pięciu wybranych kategorii: sygnał czysty (bez słyszalnych zniekształceń), błędy kwantyzacji, zniekształcenia wzmocnienia, dodatkowy dźwięk (szum) oraz brakujący dźwięk. Kategorie te zostały wybrane na podstawie rekomendacji organizacji ITU dla ogólnych metod subiektywnej oceny jakości dźwięku BS.1284-2. Rekomendacja ta określa w sumie 11 kategorii, które mogą być użyte do analizy i klasyfikacji typów zniekształceń sygnałów audio, jednak część z nich dotyczy sygnałów wielokanałowych oraz dźwięku przestrzennego, gdzie problemy mogą wynikać z zależności pomiędzy poszczególnymi kanałami (np. przesłuchy lub też zniekształcenia percepcji przestrzenności dźwięku). Jednak na tym etapie pracy nie było celem badanie takich zależności, a każdy kanał audio analizowany był niezależnie. Analiza sygnału wielokanałowego i badanie zależności pomiędzy kanałami jest przedmiotem dalszych badań, nie uwzględnionych w niniejszej pracy.

Pracę można podzielić na trzy główne części. Pierwsza z nich (rozdz. 1 – 3) stanowi wstęp teoretyczny. Dokonano tu przeglądu literatury istniejących obecnie rozwiązań dla problemu oceny jakości sygnałów audio. Opisano najpopularniejsze metody ze szczególnym uwzględnieniem metod umożliwiających przetwarzanie rzeczywistych sygnałów muzycznych, zarówno manualne jak i automatyczne, wraz z omówieniem problemów jakie się z nimi wiążą. W drugiej części pracy (rozdz. 4 – 7) przedstawiono proponowaną metodę detekcji

i klasyfikacji zniekształceń w rzeczywistych sygnałach muzycznych, która pozwalałaby na analizę sygnału w sposób automatyczny, obiektywny i bez konieczności porównania do sygnału referencyjnego. Opisano zaprojektowaną architekturę, wybór hiperparametrów modelu, stworzenie autorskiej bazy danych sygnałów wejściowych reprezentujących pięć wybranych kategorii dla zbioru treningowego, testowego oraz weryfikacyjnego, wybór typu danych wejściowych oraz przetwarzanie wstępne danych do modelu. Ostatnia – trzecia część pracy (rozdz. 8 – 10) to zbiór wyników badań zaprojektowanej i zaimplementowanej metody, w celu określenia jej skuteczności w zależności od architektury modelu oraz wybranych typów danych wejściowych.

Badania proponowanej metody podzielone zostały na dwie części. Pierwsza z nich polegała na znalezieniu wstępnej architektury modelu, druga na jego dalszej ewaluacji z wykorzystaniem różnego typu danych wejściowych ekstraktowanych z sygnałów muzycznych. Na podstawie przeprowadzonych eksperymentów, mających na celu dopasowanie architektury modelu oraz jego hiperparametrów, zamieszczone zostały wyniki dla dwóch modeli – CNN oraz CBRNN. Drugi z nich, ze względu na wyższą skuteczność poddany był dalszej ewaluacji. Podstawowym parametrem wejściowym dla modelów sieci neuronowych, obliczanym dla każdego sygnału z przygotowanej bazy danych, był spektrogram przekonwertowany do skali melowej (256 filtrów). Dodatkowo zbadany został wpływ na skuteczność klasyfikacji zniekształceń przy zastosowaniu następujących parametrów wejściowych: ZCR, OBSC, głośności chwilowe (mierzone na fragmentach 400 ms) oraz rzeczywiste wartości szczytowe sygnału (mierzone na fragmentach 100 ms). Zgodnie z prezentowanymi wynikami (tab. 11.1), najwyższą skuteczność uzyskano przy zastosowaniu spektrogramów w skali melowej wraz z parametrem OBSC (tj. skuteczność 91.4%). Jednakże znacząca poprawa klasyfikacji, szczególnie dla kategorii sygnału z dodatkowym dźwiękiem (zaszumionego), została zauważona również przy zastosowaniu informacji o rzeczywistych wartościach szczytowych sygnału. Uzyskano w tym przypadku poprawę ogólnej skuteczności o 4.7% oraz największą redukcję FNR dla kategorii dźwięku z dodatkowym szumem (z wartości 0.216 do 0.170), porównując do modelu z zastosowaniem wyłącznie spektrogramów jako danych wejściowych.

Tab. 11.1. Porównanie ogólnej skuteczności klasyfikacji badanych sieci neuronowych.

Model	ACC [%]
CNN	77.5
CBRNN	83.0
CBRNN + parametr ZCR	83.8
CBRNN + parametr OBSC	**91.4**
CBRNN + pomiary głośności chwilowych sygnału	86.2
CRBNN + pomiary rzeczywistych wartości szczytowych sygnału	87.7

Wyniki przeprowadzonych badań pokazały, że największą skuteczność uzyskano dla modelu z wykorzystaniem konwolucyjno-rekurencyjnych sieci neuronowych przy zastosowaniu spektrogramów w skali melowej wraz z dodatkowym parametrem OBSC. Metoda oparta na zastosowaniu uzyskanego w ten sposób modelu może być potencjalnym usprawnieniem dla zadań detekcji i/lub klasyfikacji podstawowych zniekształceń

w rzeczywistych sygnałach muzycznych, jednocześnie stanowiąc bazę dla dalszych badań w tym temacie.

Przeprowadzone badania potwierdzają tezę postawioną w niniejszej pracy, że *zastosowanie warstw dwukierunkowych rekurencyjnych w modelu sieci neuronowych wraz z odpowiednim doborem jego architektury, parametrów wejściowych oraz opracowaniem bazy zniekształceń do celów ewaluacji modelu, zwiększa znacząco skuteczność automatycznej klasyfikacji wybranych zniekształceń sygnałów muzycznych bez konieczności porównania do sygnału referencyjnego*.

Największe **zalety opracowanego prototypowego modelu** automatycznego klasyfikatora sygnałów muzycznych to:

- brak konieczności zastosowania sygnału referencyjnego do oceny jakości sygnału badanego (w odróżnieniu do metod referencyjnych);
- brak wstępnych założeń odnośnie zawartości badanego sygnału oraz jego zniekształceń (w odróżnieniu do metod parametrycznych);
- powtarzalność wyników oraz brak podatności na błędy związane z oceną subiektywną, które są problemem w tradycyjnych, manualnych testach odsłuchowych;
- uniwersalność zastosowania – metoda stanowi podstawę do dalszych badań, a prezentowany model może zostać zastosowany do innego pokrewnego problemu jako *transfer learning*;
- potencjalne usprawnienie długotrwałych manualnych testów odsłuchowych.

Na tym etapie pracy celem było wykrywanie i klasyfikacja jedynie wybranych zniekształceń, jednak zarówno model jak i baza treningowa może być dowolnie rozszerzana w celu wykrywania większej ilości zniekształceń. Obecnie czas automatycznej klasyfikacji fragmentu sygnału o długości 3 sekund za pomocą zaimplementowanego modelu nie przekracza 50 ms. Skrócenie czasu automatycznej klasyfikacji jest przedmiotem dalszej ewaluacji modelu.

Dalsze badania opracowanej metody będą skoncentrowane przede wszystkim na: 1) ekstrakcji oraz zbadaniu dodatkowych parametrów sygnałów audio; 2) dodaniu klasyfikacji typu *multi-label* dla przypadku, gdy dany sygnał zawiera kilka różnych zniekształceń jednocześnie; 3) rozszerzeniu bazy danych. Aktualna baza danych została przygotowana na podstawie zbioru MUSDB18 i zawiera obecnie ok. 10 godzin nagrań sygnałów muzycznych. Zgodnie z literaturą [121], jest to uznawane za wystarczającą ilość materiału na potrzeby ewaluacji prototypowego modelu, jednakże do uzyskania skuteczności na poziomie produktu końcowego, preferowana byłaby baza danych przekraczająca 5000 godzin nagrań [163]. Rozszerzenie bazy danych pozwoliłoby również na dalszą redukcję błędów klasyfikacji typu *False-Positive* oraz *False-Negative*, a przede wszystkich umożliwiło rozpoznawanie większej ilości zniekształceń. Na tym etapie pracy, badania skoncentrowane były na klasyfikacji zniekształceń łatwo słyszalnych w manualnych testach odsłuchowych, np. takich które byłyby ocenione jako „bardzo przeszkadzające" według rekomendacji ITU-R BS.1284-2. Kolejnym ważnym usprawnieniem byłoby więc rozszerzenie bazy danych oraz ewaluacja modelu dla zniekształceń mniej wyraźnych. Na tym etapie pracy nie został również przewidziany podział sygnałów ze względu na gatunek muzyczny, jednakże może to być celem dalszych badań nad efektywnością zaimplementowanego modelu.

Za **najważniejszy autorski wkład** w dziedzinę analizy cyfrowych sygnałów audio można uznać ideę wykorzystania konwolucyjno-rekurencyjnych sieci neuronowych do automatycznej klasyfikacji zniekształceń sygnałów muzycznych bez konieczności porównania do sygnału referencyjnego, gdyż autorska publikacja (2020) [118] jest pierwszą znaną mi publikacją zawierającą tę koncepcję i prezentującą pierwsze wyniki z jej realizacji. Druga z publikacji z tego zakresu (2022) [119] prezentuje podobne podejście, z tą różnicą, że oceniany jest materiał dźwiękowy tworzony przez użytkownika np. telefonem komórkowym (tj. rozmowy, dźwięki otoczenia itp.), modelowane są 2 typy zniekształceń (zamiast 4), a wyniki oceniane były na podstawie porównania do wybranych metod oceny jakości mowy, brak jest natomiast dokładniejszych danych odnośnie analizy sygnałów muzycznych.

Do szczegółowych **autorskich elementów badań** zaliczam również:

- opracowanie własnej bazy zniekształceń, reprezentujących cztery kategorie zniekształceń wybranych na podstawie rekomendacji ITU BS.1284-2. Zbiór ten został utworzony na bazie niezniekształconych nagrań sygnałów muzycznych MUSDB18 (rozdz. 5.3);
- dobór architektury badanych modeli sieci neuronowych, w tym m.in. rodzaju, liczby warstw oraz jego hiperparametrów (rozdz. 6);
- selekcja parametrów sygnałów audio jako danych wejściowych w celu ewaluacji modeli sieci neuronowych (rozdz. 5.4 – 5.10);
- zaimplementowanie oraz ewaluacja modelu sieci konwolucyjnych CNN z zastosowaniem spektrogramów w skali melowej jako danych wejściowych. Na tym etapie uzyskano ogólną skuteczność klasyfikacji 77.5% (rozdz. 8);
- zaimplementowanie oraz ewaluacja modelu sieci konwolucyjno-rekurencyjnych (dwukierunkowych) CBRNN, również z zastosowaniem spektrogramów w skali melowej jako danych wejściowych, gdzie uzyskano ogólną skuteczność 83.0% (rozdz. 9);
- rozszerzenie opracowanego modelu CBRNN (rozdz. 9) w celu zbadaniu wpływu dodatkowych parametrów wejściowych (rozdz. 10.2.1, 10.3.1, 10.4.1);
- zbadanie wpływu parametru ZCR na skuteczność klasyfikacji modelu CBRNN, gdzie uzyskano jedynie nieznaczną poprawę (tj. o 0.8%) w stosunku do zastosowania wyłącznie spektrogramów w skali melowej jako danych wejściowych (rozdz. 10.2.2);
- zbadanie wpływu parametru OBSC na skuteczność klasyfikacji modelu CBRNN, gdzie uzyskano najwyższą skuteczność modelu, tj. 91.4 % (rozdz. 10.3.2);
- zbadanie wpływu parametrów głośności chwilowych (mierzonych na ramkach 400 ms) oraz rzeczywistych wartości szczytowych (mierzonych na ramkach 100 ms) na skuteczność klasyfikacji modelu CBRNN. Na tym etapie uzyskano poprawę odpowiednio o 3.2% oraz 4.7% w stosunku do zastosowania wyłącznie spektrogramów w skali melowej jako danych wejściowych (rozdz. 10.4.2);
- zbadanie wpływu liczby filtrów melowych stosowanych dla danych wejściowych (spektrogramów w skali melowej) na skuteczność klasyfikacji modelu CBRNN (rozdz. 9.4);
- pozostałe autorskie eksperymenty nieujęte w niniejszej pracy to badania pośrednie, w których nie uzyskano zadowalających wyników. Były to m.in.: implementacja i ewaluacja modelu sieci kownolucyjno-rekurencyjnych przy zastosowaniu uczenia pół-nadzorowanego (ang. *semi-supervised learning*) oraz modelu SVM.

Bibliografia

[1] F. Dunn, T. Rossing, W. M. Hartmann, D. M. Campbell i N. H. Fletcher, *Springer Handbook of Acoustics*, New York: Springer New York, 2015.

[2] E. B. Brixen, *Audio Metering: Measurements, Standards and Practice*, New York, NY: Routledge, 2020.

[3] R. E. Goldberg, M. Bosi, *Introduction to digital audio coding and standards*, Springer, 2012.

[4] F. A. Everest, *Podręcznik akustyki*, Katowice: Wydawnictwo Sonia Draga, 2014.

[5] J. Zjalic, Digital Audio Forensics Fundamentals, Focal Press, 2020.

[6] *Method for objective measurements of perceived audio quality*, ITU-R BS.1387-2, International Telecommunications Union, Geneva, Switzerland, 2023.

[7] E. Zwicker, R. Feldtkeller, *Das Ohr als Nachrichtenempfänger*, Stuttgart: Hirzel Verlag, Federal Republic of Germany, 1967.

[8] K. C. Pohlmann, *Principles of Digital Audio*, 6th Edition, New York: McGraw-Hill, 2011.

[9] M. Kleiner, *Acoustics and Audio Technology: Third Edition*, Ft. Lauderdale: J. Ross Publishing, Inc, 2012.

[10] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, N. Harte, „ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *J. Acoust. Soc. Amer.,* p. Vol. 137, No. 6, 2015.

[11] F. F. Li, T. J. Cox, *Digital signal processing in audio and acoustical engineering*, Milton: CRC Press, 2019.

[12] G. C. Cavell, *National Association of Broadcasters Engineering Handbook*, London: Taylor and Francis, 2017.

[13] Y. Jia, „Research Status and the Development of Audio Downmix in Convergence Media," *IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS),* p. p.474-477, 2019.

[14] M. G. Christensen, *Introduction to Audio Processing*, Cham: Springer International Publishing, 2019.

[15] Y. Xiang, G. Hua, B. Yan, *Digital Audio Watermarking: Fundamentals, Techniques and Challenges*, Singapore: Springer Singapore, 2017.

[16] M. Nematollahi, C. Vorakulpipat, H. Rosales, „Audio Watermarking," w *Digital Watermarking*, Singapore, Springer Topics in Signal Processing, Vol. 11, 2017.

[17] R. Beutler, *Evolution of Broadcast Content Distribution*, Cham: Springer International Publishing, 2017.

[18] *General methods for the subjective assessment of sound quality*, ITU-R BS.1284-2, International Telecommunications Union, Geneva, Switzerland, 2019.

[19] J. Corey, D. H. Benson, *Audio Production and Critical Listening: Technical Ear Training*, Abingdon: Routledge, 2017.

[20] H. Burton, *Believing Your Ears: Examining Auditory Illusions. A Conversation with Diana Deutsch*, Open Agenda Publishing, 2020.

[21] Z. Akhtar, T. H. Falk, „Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey," *IEEE access,* Vol. 5, p. 21090-21117, 2017.

[22] H. Becerra Martinez, A. Hines, M. C. Q. Farias, „Perceptual Quality of Audio-Visual Content with Common Video and Audio Degradations," *Applied sciences,* Vol. 11 (13), p. 5813, 2021.

[23] J.-H. Flesner, T. Biberger, S. Ewert, „Subjective and Objective Assessment of Monaural and Binaural Aspects of Audio Quality," *IEEE/ACM transactions on audio, speech, and language processing*, Vol. 27 (7), p. 1112-1125, 2019.

[24] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, A. C. Bovik, „Study of Subjective and Objective Quality Assessment of Audio-Visual Signals," *IEEE transactions on image processing, Vol.29, p.6054-6068*, 2020.

[25] R. Huber, B. Kollmeier, „PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio Speech Lang. Process.,* p. Vol. 14, No. 6, p. 1902-1911, 2006.

[26] M. Hansen, B. Kollmeier, „Objective modeling of speech quality with a psychoacoustically validated auditory model," *J. Audio Eng. Soc,* Vol. 48, No. 5, p. 395-409, 2000.

[27] A. Hines, J. Skoglund, A. Kokaram, N. Harte, „ViSQOL: The virtual speech quality objective listener," *Proc. 13th Int. Workshop Acoust. Signal Enhancement,* p. 1-4, 2012.

[28] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, A. Hines, „ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX),* p. 1-6, 2020.

[29] K. Doh-Suk, „ANIQUE:Anauditory model for single-ended speech quality estimation," *Speechand Audio Processing, IEEE Transactions.,* Vol. 13, p. 821 - 831, 2005.

[30] J. M. Kates, K. H. Arehart, „The hearing-aid speech quality index (HASQI)," *J. Audio Eng. Soc,* p. 58(5), 363–381, 2010.

[31] A. E. Mahdi, D. Picovici, „New single-ended objective measure for non-intrusive speech quality evaluation," *Signal, Image Video Process,* p. 4(1), 23–38., 2010.

[32] T. H. Falk, C. Zheng, W.-Y. Chan, „A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transaction on Audio, Speech, Language Processing,* p. 18(7), 1766–1774, 2010.

[33] A. A. Catellier, S. D. Voran, „WaweNets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process,* p. 331–335, 2020.

[34] S.-W. Fu, Y. Tsao, H.-T. Hwang, H.-M. Wang, „Quality-Net: An endto-end non-intrusive speech quality assessment model based on BLSTM," *Proc. Interspeech,* p. 1873–1877, 2018.

[35] L. Orcik, M. Voznak, J. Rozhon, F. Rezac, J. Slachta, H. Toral-Cruz, J. C.-W. Lin, „Prediction of Speech Quality Based on Resilient Backpropagation Artificial Neural Network.," *Wireless Personal Communications,* Vol. 96(4), p. 5375-5389., 2017.

[36] T. Falk, W.-Y. Chan, „Single-Ended Speech Quality Measurement Using Machine Learning Methods.," *IEEE Transactions on Audio, Speech, and Language Processing,* p. 14(6), 1935-1947, 2006.

[37] W. A. Jassim, M. S. Zilany, „NSQM: A non-intrusive assessment of speech quality using normalized energies of the neurogram," *Computer Speech & Language,* Vol. 58, p. 260-279, 2019.

[38] E. Manilow, P. Seetharaman, F. Pishdadian, B. Pardo, „Predicting algorithm efficacy for adaptive multi-cue source separation," *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.,* p. 274–278, 2017.

[39] E. M. Grais, H. Wierstorf, D. Ward, R. Mason, M. D. Plumbley, „Referenceless performance evaluation of audio source separation using deep neural networks," *Proc. 27th Eur. Signal Process. Conf.,* p. 1–5, 2019.

[40] L. Moysis, L. A. Iliadis, S. P. Sotiroudis, A. D. Boursianis, M. S. Papadopoulou, K.-I. D. Kokkinidis, C. Volos, P. Sarigiannidis, S. Nikolaidis, S. K. Goudos, „Music Deep Learning: Deep Learning Methods for Music Signal Processing—A Review of the State-of-the-Art," *IEEE access,* Vol. 11, p. 17031-17052, 2023.

[41] V. R. Revathy, A. S. Pillai, „Binary emotion classification of music using deep neural networks," *Proc. Int. Conf. Soft Comput. Pattern Recognit.,* p. 484–492, 2021.

[42] D. Martin-Gutierrez, G. Hernandez Penaloza, A. Belmonte-Hernandez, F. Alvarez Garcia, „A multimodal end-to-end deep learning architecture for music popularity prediction," *IEEE Access,* Vol. 8, p. 39361–39374, 2020.

[43] F. Fessahaye, L. Perez, T. Zhan, R. Zhang, C. Fossier, R. Markarian, C. Chiu, J. Zhan, L. Gewali, P. Oh, „T-RECSYS: A novel music recommendation system using deep learning," *Proc. IEEE Int. Conf. Consum. Electron. (ICCE),* p. 1–6, 2019.

[44] J. R. Castillo, M. J. Flores, „Web-based music genre classification for timeline song visualization and analysis," *IEEE Access,* Vol. 9, p. 18801–18816, 2021.

[45] F. Zhang, „Research on music classification technology based on deep learning," *Secur. Commun. Netw.,* p. 1–8, 2021.

[46] S. Rajesh i N. J. Nalini, „Musical instrument emotion recognition using deep recurrent neural network," *Procedia computer science,* Vol. 167, p. 16–25, 2020.

[47] A. Vall, M. Quadrana, M. Schedl i G. Widmer, „The importance of song context and song order in automated music playlist generation," Web., 2018.

[48] C. De Boom, S. Van Laere, T. Verbelen, B. Dhoedt, „Rhythm, Chord and Melody Generation for Lead Sheets Using Recurrent Neural Networks", Springer, Web., 2020.

[49] S. K. Prabhakar, S.-W. Lee, „Holistic approaches to music genre classification using efficient transfer and deep learning techniques," *Exp. Syst. Appl.,* Vol. 211, No. 118636, 2023.

[50] X. Li, H. Xianyu, J. Tian, W. Chen, F. Meng, M. Xu, L. Cai, „A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP),* p. 544–548, 2016.

[51] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, Y. Mitsufuji, „Improving music source separation based on deep neural networks through data augmentation and network blending," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP),* p. 261–265, 2017.

[52] W. Gong, Q. Yu, „A deep music recommendation method based on human motion analysis," *IEEE Access,* Vol. 9, p. 26290–26300, 2021.

[53] S. Chikkamath, N. S R, „Melody Generation Using LSTM and BI-LSTM Network," *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA),* p. 1–6, 2021.

[54] W. Li, S. Li, X. Shao, Z. Li, „A Practical Singing Voice Detection System Based on GRU-RNN," *Proceedings of the 6th Conference on Sound and Music Technology (CSMT),* p. 15–25, 2019.

[55] L. Qiu, S. Li, Y. Sung, „3D-DCDAE: Unsupervised music latent representations learning method based on a deep 3D convolutional denoising autoencoder for music genre classification," *Mathematics,* Vol. 9, No. 18, p. 2274, 2021.

[56] T. Ciborowski, S. Reginis, D. Weber, A. Kurowski, B. Kostek, „Classifying emotions in film music—A deep learning approach," *Electronics,* Vol. 10, No. 23, p. 2955, 2021.

[57] A. Wise, A. S. Maida, A. Kumar, „Attention augmented CNNs for musical instrument identification," *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO),* p. 376–380, 2021.

[58] M. Sheikh Fathollahi, F. Razzazi, „Music similarity measurement and recommendation system using convolutional neural networks," *International journal of multimedia information retrieval,* Vol. 10, No. 1, p. 43–53, 2021.

[59] K. W. Cheuk, H. Anderson, K. Agres, D. Herremans, „NnAudio: An on-the-fly GPU audio to spectrogram conversion toolbox using 1D convolutional neural networks," *IEEE Access,* Vol. 8, p. 161981–162003, 2020.

[60] J. Iriz, M. A. Patricio, A. Berlanga i J. M. Molina, „CONEqNet: convolutional music equalizer network," w *Multimedia tools and applications*, Vol. 82, p. 3911-3930, New York, Springer US, 2023.

[61] L. Su, „Vocal Melody Extraction Using Patch-Based CNN," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 371–375, 2018.

[62] Huang, I.-S. et al., „A generative adversarial network model based on intelligent data analytics for music emotion recognition under IoT," *Mobile information systems,* Vol. 2021, p. 1-8, 2021.

[63] N. Li, „Generative adversarial network for musical notation recognition during music teaching," *Comput. Intell. Neurosci.,* p. 1–9, 2022.

[64] Z.-C. Fan, Y.-L. Lai, J.-S. R. Jang, „SVSGAN: Singing voice separation via generative adversarial network," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP),* p. 726–730, 2018.

[65] Y. Yu, Z. Zhang, W. Duan, A. Srivastava, R. Shah, Y. Ren, „Conditional Hybrid GAN for Melody Generation from Lyrics," *Neural computing & applications,* p. 3191–3202, 2023.

[66] Adiyansjah, A. A. S. Gunawan, D. Suhartono, „Music recommender system based on genre using convolutional recurrent neural networks," *Procedia computer science,* Vol. 157, p. 99–109, 2019.

[67] K. Choi, G. Fazekas, M. Sandler, K. Cho, „Convolutional Recurrent Neural Networks for Music Classification," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 2392–2396, 2017.

[68] R. Romero-Arenas, A. Gómez-Espinosa, B. Valdés-Aguirre, „Singing voice detection in electronic music with a long-term recurrent convolutional network," *Applied sciences,* Vol. 12, No. 15, p. 7405, 2022.

[69] S. Joshi, T. Jain, N. Nair, „Emotion Based Music Recommendation System Using LSTM - CNN Architecture," *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT),* p. 01–06, 2021.

[70] Y. Yu, „Research on Music Emotion Classification Based on CNN-LSTM Network," *2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT),* p. 473–476, 2021.

[71] Y. Liu, „Recovery of Lossy Compressed Music Based on CNN Super-Resolution and GAN," *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC),* p. 623–629, 2021.

[72] M. Torcoli, T. Kastner, J. Herre, „Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence," *IEEE/ACM transactions on audio, speech, and language processing,* Vol. 29, 2021.

[73] L. A. Iliadis, S. P. Sotiroudis, K. Kokkinidis, P. Sarigiannidis, S. Nikolaidis, S. K. Goudos, „Music Deep Learning: A Survey on Deep Learning Methods for Music Processing," *2022 11th International Conference on Modern Circuits and Systems Technologies (MOCAST),* p. 1-4, 2022.

[74] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for end-to-end Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, International Telecommunications Union, Geneva, Switzerland, 2001.

[75] *Method for Objective Measurements of Perceived Audio Quality*, ITU-R Rec. BS.1387-2, International Telecommunications Union, Geneva, Switzerland, 2023.

[76] P. M. Delgado, J. Herre, „Can We Still Use PEAQ? A Performance Analysis of the ITU Standard for the Objective Assessment of Perceived Audio Quality," *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX),* p. 1-6, 2020.

[77] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for end-to-end Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, International Telecommunications Union, Geneva, Switzerland, 2001.

[78] D. Mukhutdinov, A. Alex, A. Cavallaro, L. Wang, „Deep Learning Models for Single-Channel Speech Enhancement on Drones," *IEEE Access,* Vol. 11, p. 22993-23007, 2023.

[79] M. Pashaian, S. Seyedin, S. M. Ahadi, „A Novel Jointly Optimized Cooperative DAE-DNN Approach Based on a New Multi-Target Step-Wise Learning for Speech Enhancement," *IEEE Access,* Vol. 11, p. 21669-21685, 2023.

[80] D. Lee, C. Jung-Woo, „DeFT-AN: Dense Frequency-Time Attentive Network for Multichannel Speech Enhancement," *IEEE Signal Processing Letters,* Vol. 30, p. 155-159, 2023.

[81] P.-H. Vial, P. Magron, T. Oberlin, C. Fevotte, „Phase Retrieval With Bregman Divergences and Application to Audio Signal Recovery," *IEEE Journal of Selected Topics in Signal Processing,* Vol. 15, No. 1, p. 51-64, 2021.

[82] P. Zaviska, P. Rajmic, O. Mokry, „Audio Dequantization Using (Co)Sparse (Non)Convex Methods," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 701-705, 2021.

[83] P. Zaviska, P. Rajmic, A. Ozerov, L. Rencker, „A Survey and an Extensive Evaluation of Popular Audio Declipping Methods," *IEEE Journal of Selected Topics in Signal Processing,* Vol. 15, No. 1, p. 5-24, 2021.

[84] D. K. Tran, M. Unoki, „Matching Pursuit and Sparse Coding for Auditory Representation," *IEEE Access,* Vol. 9, p. 167084-167095, 2021.

[85] A. Marafioti, N. Holighaus, P. Majdak, „Time-Frequency Phase Retrieval for Audio—The Effect of Transform Parameters," *IEEE Transactions on Signal Processing,* Vol. 69, p. 3585-3596, 2021.

[86] A. Madhu et al., „SiamNet: Siamese CNN Based Similarity Model for Adversarially Generated Environmental Sounds," *IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP),* p. 1-6, 2021.

[87] S. Kandadai, J. Hardin, C. Creusere, „Audio quality assessment using the mean structural similarity measure," *IEEE International Conference on Acoustics, Speech and Signal Processing,* p. 221–224, 2008.

[88] *Perceptual Objective Listening Quality Prediction*, ITU-T Rec. P.863, International Telecommunications Union, Geneva, Switzerland, 2018.

[89] T. Bäckström, *Speech Coding With Code-Excited Linear Prediction*, Springer International Publishing, 2017.

[90] Y. Li, L. Lei, Y. Wang, J. Jing, Q. Zhou, „TrustSAMP: Securing Streaming Music Against Multivector Attacks on ARM Platform," *IEEE Transactions on Information Forensics and Security,* Vol. 17, p. 1709-1724, 2022.

[91] C. Tarjano, V. Pereira, „An Efficient Algorithm for Segmenting Quasi-Periodic Digital Signals Into Pseudo Cycles: Application in Lossy Audio Compression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* Vol. 30, p. 1694-1703, 2022.

[92] J. A. Baloch, A. K. Jumani, A. A. Laghari, V. V. Estrela, R. T. Lopes, „A Preliminary Study on Quality of Experience Assessment of Compressed Audio File Format," *2021 IEEE URUCON,* p. 161-165, 2021.

[93] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, Z. Jin, „A Differentiable Perceptual Audio Metric Learned from Just Noticeable Differences," *CoRR,* Vol. abs/2001.04460, 2020.

[94] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, „A short-time objective intelligibility measure for time-frequency weighted noisy speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* p. 4214–4217, 2010.

[95] D. Lee, C. Jung-Woo, „DeFT-AN: Dense Frequency-Time Attentive Network for Multichannel Speech Enhancement," *IEEE Signal Processing Letters,* Vol. 30, p. 155-159, 2023.

[96] R. Soleymanpour, M. Soleymanpour, A. J. Brammer, M. T. Johnson, I. Kim, „Speech Enhancement Algorithm Based on a Convolutional Neural Network Reconstruction of the Temporal Envelope of Speech in Noisy Environments," *IEEE Access,* Vol. 11, p. 5328-5336, 2023.

[97] S. Maiti, Y. Ueda, S. Watanabe, C. Zhang, M. Yu, S.-X. Zhang, Y. Xu, „EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers," *2022 IEEE Spoken Language Technology Workshop (SLT),* p. 480-487, 2023.

[98] J. L. Roux, S. Wisdom, H. Erdogan, J. R. Hershey, „SDR – Half-baked or Well Done?," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 626-630, 2019.

[99] D. Lee, C. Jung-Woo, „DeFT-AN: Dense Frequency-Time Attentive Network for Multichannel Speech Enhancement," *IEEE Signal Processing Letters,* Vol. 30, p. 155-159, 2023.

[100] A. Kovalyov, K. Patel, I. Panahi, „DSENet: Directional Signal Extraction Network for Hearing Improvement on Edge Devices," *IEEE Access,* Vol. 11, p. 4350-4358, 2023.

[101] S. Soni, R. N. Yadav, L. Gupta, „State-of-the-Art Analysis of Deep Learning-Based Monaural Speech Source Separation Techniques," *IEEE Access,* Vol. 11, p. 4242-4269, 2023.

[102] K. Schulze-Forster, G. Richard, L. Kelley, C. S. J. Doire, R. Badeau, „Unsupervised Music Source Separation Using Differentiable Parametric Source Models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* Vol. 31, p. 1276-1289, 2023.

[103] G. Mittag, B. Naderi, A. Chehadi, S. Möller, „NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," *Proc. Interspeech Conf.,* p. 2127–2131, 2021.

[104] H. Salehi, V. Parsa, „Nonintrusive speech quality estimation based on Perceptual Linear Prediction," *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE),* p. 1-4, 2016.

[105] D.-S. Kim, A. Tarraf, „ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal,* Vol. 12, No. 1, p. 221-236, 2007.

[106] C. Kim, R. M. Stern, „Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis," *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.,* p. 2598–2601, 2008.

[107] A. E. Mahdi, D. Picovici, „New single-ended objective measure for non-intrusive speech quality evaluation," *Signal, Image Video Process.,* Vol. 4, No. 1, p. 23–38, 2010.

[108] Z. Zhang, Y. Shen, D. S. Williamson, „Objective Comparison of Speech Enhancement Algorithms with Hearing Loss Simulation," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 6845-6849, 2019.

[109] D.Ellis, https://labrosa.ee. columbia.edu/dpwe/tmp/nist/doc/stnr.txt, [Online], dostęp: maj 2023.

[110] D.Ellis, https://labrosa.ee. columbia.edu/projects/snreval, [Online], dostęp: sierpień 2021.

[111] Y. Zhu, T. H. Falk, „Fusion of Modulation Spectral and Spectral Features with Symptom Metadata for Improved Speech-Based Covid-19 Detection," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 8997-9001, 2022.

[112] G. Li, J. Yu, J. Deng, X. Liu, H. Meng, „Audio-Visual Multi-Channel Speech Separation, Dereverberation and Recognition," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 6042-6046, 2022.

[113] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, H. Cucu, „A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," *IEEE Access,* Vol. 10, p. 47628-47642, 2022.

[114] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, D. Sculley, „AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," Web., 2016.

[115] L. Chen-Chou, S.-W. Fu, H. Wen-Chin, X. Wang, J. Yamagishi, Y. Tsao, H.-M. Wang, „MOSNet: Deep learning-based objective assessment for voice conversion," *Proc. Interspeech Conf.,* p. 1541–1545, 2019.

[116] A. A. Catellier, S. D. Voran, „WaweNets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* p. 331–335, 2020.

[117] I. H. Parmonangan, J. Santoso, „Prediction of Perceived Synthesized Speech Quality with Wav2Vec2 Features on Small Dataset," *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI),* p. 497-502, 2022.

[118] K. Organiściak, J. Borkowski, „Single-Ended quality measurement of a music content via convolutional recurrent neural networks," *Metrology and Measurement Systems,* Vol. 27, No. 4, p. 721-733, 2020.

[119] D. Mumtaz, V. Jakhetiya, K. Nathwani, B. N. Subudhi, S. C. Guntuku, „Nonintrusive Perceptual Audio Quality Assessment for User-Generated Content Using Deep Learning," *IEEE Transactions on Industrial Informatics,* Vol. 18, No. 11, p. 7780-7789, 2022.

[120] H. Zhao, S. Zarar, I. Tashev, C.-H. Lee, „Convolutional- Recurrent Neural Networks for Speech Enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2401-2405.,* p. 2401-2405, 2018.

[121] J. Sang, S. Park, J. Lee, „Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms," *26th European Signal Processing Conference (EUSIPCO),* p. 2444 - 2448, 2018.

[122] R. J. Portsev, A. V. Makarenko, „Convolutional Neural Networks for Noise Signal Recognition," *IEEE International Workshop on Machine Learning for Signal Processing,* p. 1-6, 2018.

[123] T. Virtanen, M. D. Plumbley, D. Ellis, *Computational Analysis of Sound Scenes and Events*, Cham: Springer International Publishing, 2018.

[124] R. Singh, Profiling Humans from their Voice, Singapore: Springer Singapore : Imprint: Springer, 2019.

[125] Z. Rafii, A. Liutkus, F. Stöter, S. Mimilakis, R. Bittner, „MUSDB18 dataset," 2018. https://sigsep.github.io/datasets/musdb.html, [Online], dostęp: czerwiec 2023.

[126] B. L. Sturm, „The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval," *Journal of New Music Research,* Vol. 43.2, p. 147–172, 2014.

[127] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, P. Lamere, „The Million Song Dataset," 2017.

[128] M. Defferrard, K. Benzi, P. Vandergheynst i X. Bresson, „FMA: A Dataset For Music Analysis," 2017, https://github.com/mdeff/fma, [Online], dostęp: listopad 2021.

[129] F. Font, G. Roma i X. Serra, „Freesound technical demo," *Association for Computing Machinery,* 2013.

[130] S. Jung, J. Park, S. Lee, „Polyphonic sound event detection using convolutional bidirectional LSTM and synthetic data-based transfer learning," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 885-889, 2019.

[131] J. Sang, S. Park, J. Lee, „Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms," *26th European Signal Processing Conference (EUSIPCO),* p. 2444-2448, 2018.

[132] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, A. C. Bovik, „Study of Subjective and Objective Quality Assessment of Audio-Visual Signals," *IEEE transactions on image processing,* Vol. 29, p. 6054-6068, 2020.

[133] T. Heittola, E. Çakır, T. Virtanen, „The Machine Learning Approach for Analysis of Sound Scenes and Events," w *Computational Analysis of Sound Scenes and Events*, Springer International Publishing, 2018.

[134] T.-T.-H. Phan, H.-D. Nguyen, D.-D. Nguyen, „Evaluation of Feature Extraction Methods for Bee Audio Classification," *Intelligence of Things: Technologies and Applications . ICIot 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 148. Springer,* 2022.

[135] O. P. Jena, *Computational Intelligence and Healthcare Informatics*, New Jersey: Scrivener Publishing: Wiley, 2021.

[136] M. Bhattacharjee, S. R. M. Prasanna, P. Guha, „Time-Frequency Audio Features for Speech-Music Classification," 2018.

[137] S. R. Gulhane, S. D. Shirbahadurkar, S. Badhe, „Cepstral (MFCC) Feature and Spectral (Timbral) Features Analysis for Musical Instrument Sounds.," *IEEE Global Conference on Wireless Computing and Networking (GCWCN),* p. 109-113, 2018.

[138] B. K. Khonglah, S. R. M. Prasanna, „Clean speech/speech with background music classification using HNGD spectrum," *International Journal of Speech Technology,* Vol. 20(4), p. 1023-1036, 2017.

[139] J.-M. Ren, M.-J. Wu, J.-S. R. Jang, „Automatic Music Mood Classification Based on Timbre and Modulation Features," *236 IEEE Transactions on Affective Computing*, Vol. 6, No. 3, 2015.

[140] E. B. Brixen, *Audio metering : measurements, standards and practice*, NY: Routledge, New York, 2020.

[141] *Algorithms to measure audio programme loudness and true-peak audio level*, ITU-R BS.1770-4, International Telecommunications Union, Geneva, Switzerland, 2015.

[142] *Requirements for loudness and true-peak indicating meters*, ITU-R BS.1771-1, International Telecommunications Union, Geneva, Switzerland, 2012.

[143] *Loudness Normalisation and Permitted Maximum Lefel of Audio Signals*, EBU R 128, European Broadcast Union, Geneva, Switzerland, 2020.

[144] C. C. Aggarwal, *Neural Networks and Deep Learning. A Textbook*, Springer International Publishing, 2018.

[145] R. J. Portsev, A. V. Makarenko, „Convolutional Neural Networks for Noise Signal Recognition," *IEEE International Workshop on Machine Learning for Signal Processing,* p. 1-6, 2018.

[146] H. Shih-Chia, L. Trung-Hieu, „Principles and Labs for Deep Learning," Elsevier, 2021.

[147] I. Goodfellow, Y. Bengio, A. C. Courville, Deep Learning., London: The MIT Press., 2017.

[148] U. Michelucci, *Advanced Applied Deep Learning*, Berkeley, CA: Apress L. P, 2019.

[149] D. Osinga, *Deep Learning Receptury*, Helion S.A., 2019.

[150] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu, X. Feng, „Acoustic Scene Classification Using Deep Audio Feature and BLSTM Network," *2018 International Conference on Audio, Language and Image Processing (ICALIP),* 2018.

[151] N. HE, S. Ferguson, „Multi-view Neural Networks for Raw Audio-based Music Emotion Recognition," *2020 IEEE International Symposium on Multimedia (ISM),* p. 168-172, 2020.

[152] H. Yuan, W. Zheng, Y. Song, Y. Zhao, „Parallel Deep Neural Networks for Musical Genre Classification: A Case Study," *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC),* p. 1032-1035, 2021.

[153] N. Kalchbrenner, I. Danihelka, A. Graves, „Grid Long Short-Term Memory," arXiv preprint arXiv:1507.01526, 2015.

[154] S. Ioffe, C. Szegedy, „Batch normalization: accelerating deep network training by reducing internal covariate shift," *Proceedings of the 32nd International Conference on Machine Learning (ICML-15),* p. 448–456, 2015.

[155] G.-R. Liu, *Machine Learning With Python: Theory And Applications*, World Scientific Publishing Company, 2022.

[156] A. L. Caterini, *Deep Neural Networks in a Mathematical Framework*, Springer International Publishing, 2018.

[157] A. Mesaros, T. Heittola, T. Virtanen, „Metrics for Polyphonic Sound Event Detection," *Applied sciences,* Vol. 6, p. 162-162, 2016.

[158] B. Dave, K. Srivastava, „Convolutional Neural Networks for Audio Classification: An Ensemble Approach," *Proceedings of the 6th International Conference on Advance Computing and Intelligent Engineering,* 2023.

[159] K.-L. Du, M. N. S. Swamy, *Neural Networks and Statistical Learning*, London: Springer London, 2019.

[160] *AWS EC2 Instance Types*, https://aws.amazon.com/ec2/instance-types/ [Online]. dostęp: maj 2023.

[161] S. Jung, J. Park i S. Lee, „Polyphonic sound event detection using convolutional bidirectional LSTM and synthetic data-based transfer learning," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 885-889, 2019.

[162] S. Jung, J. Park, S. Lee, „Polyphonic sound event detection using convolutional bidirectional LSTM and synthetic data-based transfer learning," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* p. 885-889, 2019.

[163] L. T. Wu Y., *IEEE International Conference on Acoustics, Speech and Signal Processing,* p. 331-335, 2018.