GDAŃSK UNIVERSITY OF TECHNOLOGY

The author of the doctoral dissertation: mgr inż. Damian Koszewski
Scientific discipline: **Information and communication technology**

DOCTORAL DISSERTATION

Title of doctoral dissertation: **Automatic audio signal mixing system based on one-dimensional Wave-U-Net autoencoders**

Title of doctoral dissertation (in Polish): **Automatyczne miksowanie sygnałów fonicznych z użyciem jednowymiarowych autoenkoderów typu Wave-U-Net**

Supervisor *signature*	Second supervisor *signature*
Prof. dr hab. inż. Bożena Kostek	
Auxiliary supervisor *signature*	Cosupervisor *signature*

Gdańsk, year 2022

GDAŃSK UNIVERSITY OF TECHNOLOGY

RESEARCH UNIVERSITY
EXCELLENCE INITIATIVE

DOCTORAL DISSERTATION

Automatic audio signal mixing system based on one-dimensional Wave-U-Net autoencoders

mgr inż. Damian Koszewski

Supervisor: prof. dr hab. inż. Bożena Kostek

Gdańsk, 2022

GDAŃSK UNIVERSITY OF TECHNOLOGY

STATEMENT

The author of the doctoral dissertation: mgr inż. Damian Koszewski

I, the undersigned, declare that I am aware that in accordance with the provisions of Art. 27 (1) and (2) of the Act of 4th February 1994 on Copyright and Related Rights (Journal of Laws of 2021, item 1062), the university may use my doctoral dissertation entitled: Automatic audio signal mixing system based on one-dimensional Wave-U-Net autoencoders may be used for scientific or didactic purposes.[1]

Gdańsk, April '2022 .....

signature of the PhD student

Aware of criminal liability for violations of the Act of 4th February 1994 on Copyright and Related Rights and disciplinary actions set out in the Law on Higher Education and Science(Journal of Laws 2021, item 478), as well as civil liability, I declare, that the submitted doctoral dissertation is my own work.

I declare, that the submitted doctoral dissertation is my own work performed under and in cooperation with the supervision of prof. dr hab. inż. Bożena Kostek.

This submitted doctoral dissertation has never before been the basis of an official procedure associated with the awarding of a PhD degree.

All the information contained in the above thesis which is derived from written and electronic sources is documented in a list of relevant literature in accordance with Art. 34 of the Copyright and Related Rights Act.

I confirm that this doctoral dissertation is identical to the attached electronic version.

Gdańsk, April '2022 .....

signature of the PhD student

I, the undersigned, agree/~~do not agree~~* to include an electronic version of the above doctoral dissertation in the open, institutional, digital repository of Gdańsk University of Technology.

Gdańsk, April '2022 .....

signature of the PhD student

* delete where appropriate.

[1] Art 27. 1. Educational institutions and entities referred to in art. 7 sec. 1 points 1, 2 and 4–8 of the Act of 20 July 2018 – Law on Higher Education and Science, may use the disseminated works in the original and in translation for the purposes of illustrating the content provided for didactic purposes or in order to conduct research activities, and to reproduce for this purpose disseminated minor works or fragments of larger works.

2. If the works are made available to the public in such a way that everyone can have access to them at the place and time selected by them, as referred to in para. 1, is allowed only for a limited group of people learning, teaching or conducting research, identified by the entities listed in paragraph 1.

GDAŃSK UNIVERSITY OF TECHNOLOGY

DESCRIPTION OF DOCTORAL DISSERTATION

The Author of the doctoral dissertation: mgr inż. Damian Koszewski
Title of doctoral dissertation: Automatic audio signal mixing system based on one-dimensional Wave-U-Net autoencoders
Title of doctoral dissertation in Polish: Automatyczne miksowanie sygnałów fonicznych z użyciem jednowymiarowych autoenkoderów typu Wave-U-Net
Language of doctoral dissertation: English
Supervisor: Prof. dr hab. inż. Bożena Kostek
Date of doctoral defense:
Keywords of doctoral dissertation in Polish: automatyczny miks sygnałów fonicznych, automatyczne przetwarzanie muzyki, autoenkodery, Wave-U-Net, tetsy odsłuchowe, analiza statystyczna, macierze podobieństwa, chromagram
Keywords of doctoral dissertation in English: automatic audio mixing, music information retrieval, autoencoders, Wave-U-Net, subjective tests, statistical analysis, similarity matrices, chromagram

Summary of doctoral dissertation in Polish: Celem pracy jest stworzenie systemu do automatycznego miksu utworów, który jest w stanie w sposób automatyczny zmiksować utwór z dobrą jakością w dowolnym gatunku muzycznym. W pracy w pierwszej kolejności przywołano metody przetwarzania sygnałów fonicznych używane w procesie miksowania dźwięku oraz opisano wybrane metody automatycznego miksowania dźwięku. Zaproponowano nowatorską architekturę zbudowaną z jednowymiarowych autoenkoderów Wave-U-Net do automatycznego tworzenia miksów muzycznych. Modele zostały wytrenowane na specjalnie przygotowanej bazie danych. Miksy stworzone za pomocą proponowanego systemu porównano z miksami amatorskimi, stworzonymi przy pomocy aktualnych metod znanych z literatury oraz profesjonalnymi, wykonanymi przez inżynierów dźwieku. Osiągnięte wyniki dowodzą, że miksy tworzone automatycznie przez sieć Wave-U-Net mogą być obiektywnie oceniane tak samo wysoko jak miksy stworzone profesjonalnie. Potwierdza to również analiza statystyczna wyników przeprowadzonych testów odsłuchowych. Osiągnięte wyniki wskazują na silną korelację między doświadczeniem słuchaczy w miksowaniu a prawdopodobieństwem wyższej oceny miksu Wave-U-Net i miksu profesjonalnego niż miksu amatorskiego czy przygotowanego z wykorzystaniem uznanego oprogramowania. Wyniki te zostały również potwierdzone za pomocą analizy wykorzystującej macierze podobieństwa.

Summary of doctoral dissertation in English: The purpose of this dissertation is to develop an automatic song mixing system that is capable of automatically mixing a song with good quality in any music genre. This work recalls first the audio signal processing methods used in audio mixing, and it describes selected methods for automatic audio mixing. Then, a novel architecture built based on one-dimensional Wave-U-Net autoencoders is proposed for automatic music mixing. Models are trained on a custom-made database. Mixes created using the proposed system are compared with amateur, state-of-the-art software and professional mixes prepared by audio engineers. The achieved results prove that mixes created automatically by Wave-U-Net can objectively be evaluated as highly as mixes created professionally. This is also confirmed by the statistical analysis of the results of the conducted listening tests. The results show a strong correlation between the experience of the listeners in mixing and the likelihood of a higher rating of the Wave-U-Net mix and the professional mix than the amateur mix or the mix prepared using state-of-the-art software. These results are also confirmed by the results of the similarity matrix-based analysis.

ABSTRACT

The purpose of this dissertation is to develop an automatic song mixing system that is capable of automatically mixing a song with good quality in any music genre. This work recalls first the audio signal processing methods used in audio mixing, and it describes selected methods for automatic audio mixing. Then, a novel architecture built based on one-dimensional Wave-U-Net autoencoders is proposed for automatic music mixing. Models are trained on a custom-made database. Mixes created using the proposed system are compared with amateur, state-of-the-art software and professional mixes prepared by audio engineers. The achieved results prove that mixes created automatically by Wave-U-Net can objectively be evaluated as highly as mixes created professionally. This is also confirmed by the statistical analysis of the results of the conducted listening tests. The results show a strong correlation between the experience of the listeners in mixing and the likelihood of a higher rating of the Wave-U-Net mix and the professional mix than the amateur mix or the mix prepared using state-of-the-art software. These results are also confirmed by the results of the similarity matrix-based analysis.

Keywords: automatic audio mixing, music information retrieval, autoencoders, Wave-U-Net, subjective tests, statistical analysis, similarity matrices, chromagram

Field of science and technology in accordance with OECD requirements: Information and Communication Technology

STRESZCZENIE

Celem pracy jest stworzenie systemu do automatycznego miksu utworów, który jest w stanie w sposób automatyczny zmiksować utwór z dobrą jakością w dowolnym gatunku muzycznym. W pracy w pierwszej kolejności przywołano metody przetwarzania sygnałów fonicznych używane w procesie miksowania dźwięku oraz opisano wybrane metody automatycznego miksowania dźwięku. Zaproponowano nowatorską architekturę zbudowaną z jednowymiarowych autoenkoderów Wave-U-Net do automatycznego tworzenia miksów muzycznych. Modele zostały wytrenowane na specjalnie przygotowanej bazie danych. Miksy stworzone za pomocą proponowanego systemu porównano z miksami amatorskimi, stworzonymi przy pomocy aktualnych metod znanych z literatury oraz profesjonalnymi, wykonanymi przez inżynierów dźwieku. Osiągnięte wyniki dowodzą, że miksy tworzone automatycznie przez sieć Wave-U-Net mogą być obiektywnie oceniane tak samo wysoko jak miksy stworzone profesjonalnie. Potwierdza to również analiza statystyczna wyników przeprowadzonych testów odsłuchowych. Osiągnięte wyniki wskazują na silną korelację między doświadczeniem słuchaczy w miksowaniu a prawdopodobieństwem wyższej oceny miksu Wave-U-Net i miksu profesjonalnego niż miksu amatorskiego czy przygotowanego z wykorzystaniem uznanego oprogramowania. Wyniki te zostały również potwierdzone za pomocą analizy wykorzystującej macierze podobieństwa.

Słowa kluczowe: automatyczny miks sygnałów fonicznych, automatyczne przetwarzanie muzyki, autoenkodery, Wave-U-Net, testy odsłuchowe, analiza statystyczna, macierze podobieństwa, chromagram

Dziedzina nauki i techniki, zgodnie z wymogami OECD: Informatyka techniczna i telekomunikacja

STRESZCZENIE ROZSZERZONE W J. POLSKIM

Proces produkcji utworu muzycznego może być podzielony na następujące etapy: kompozycja, nagranie, edycja, miks oraz mastering. Wzrastająca potrzeba automatyzacji poszczególnych kroków procesu produkcji muzycznej zainspirowała tę pracę, która skupia się na fazie miksowania utworu muzycznego. Definicyjnie miks można określić jako proces obróbki pojedynczych ścieżek dźwiękowych prowadzący do połączenia nagrania wielośladowego w jeden, stereofoniczny plik muzyczny. Wybór tego właśnie etapu w badaniach został podyktowany tym, że miksowanie jest istotnym etapem produkcji muzycznej i prawdopodobnie najważniejszym krokiem w zapewnieniu tego, że bez względu na sposób, w jaki muzyka została nagrana, będzie ona oddziaływać z odpowiednią estetyką i przekazywać artystyczną wypowiedź zgodną z intencją muzyka. Jest to etap zarówno techniczny, jak i kreatywny. Ponadto ten właśnie etap jest obecnie najczęściej poddawany próbom automatyzacji.

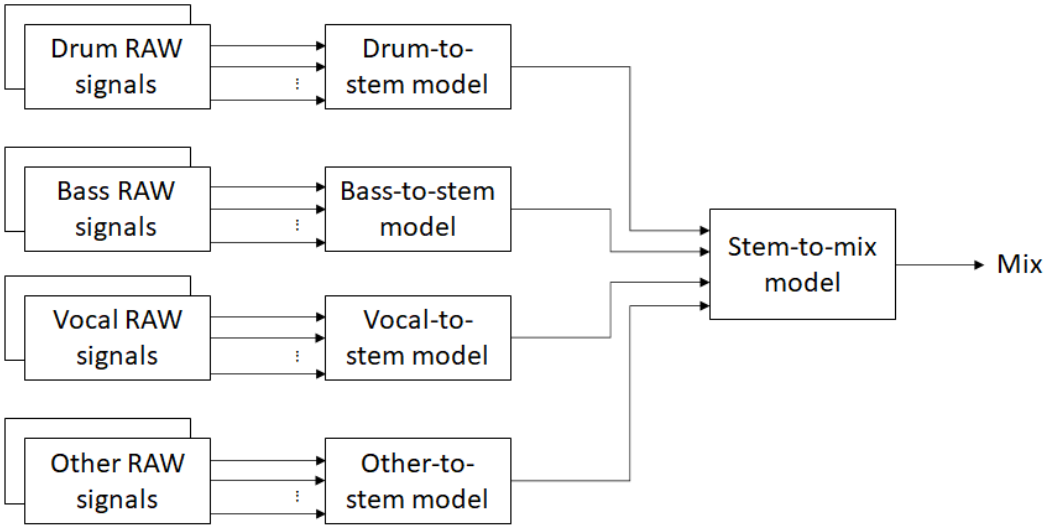
W rozprawie − w pierwszej kolejności − dokonano przeglądu modułów przetwarzania sygnału używanych podczas miksowania dźwięku z naciskiem na procesy związane z techniczną stroną miksowania. Opisano wybrane, znane metody automatycznego miksowania dźwięku.

Celem pracy jest stworzenie systemu do automatycznego miksu utworów, który otrzymując na wejściu nagrane i wyedytowane ścieżki, jest w stanie w sposób automatyczny (bez ingerencji użytkownika) zmiksować utwór z dobrą jakością w dowolnym gatunku muzycznym. Aby sformalizować ten cel, zdefiniowane zostały następujące tezy, które następnie zostały udowodnione w dalszej części rozprawy:

1. Możliwe jest miksowanie utworu muzycznego składającego się z wielu ścieżek z użyciem jednowymiarowej adaptacji autoenkodera Wave-U-Net, które może być obiektywnie oceniane jako porównywalne jakościowo do miksu stworzonego przez profesjonalnego inżyniera dźwięku

2. Miksy utworów uzyskane przy pomocy zaproponowanego systemu można subiektywnie ocenić jako lepsze jakościowo niż miksy amatorskie i miksy wykonane metodami znanymi ze stanu wiedzy (ang. *state-of-the-art*), jak również porównywalne z miksami przygotowanymi przez profesjonalnego inżyniera miksu.

W pracy zaproponowano nowatorską architekturę zbudowaną z jednowymiarowych autoenkoderów Wave-U-Net. System złożony jest z pięciu modeli. Wszystkie modele zostały wytrenowane osobno, a następnie połączone w jeden system. Modele różnią się liczbą wejść i wyjść. System imituje sposób, w jaki miksy są tworzone przez inżyniera dźwięku, tzn. pojedyncze ścieżki są miksowane razem do grupy instrumentów (np. ścieżki wokali do grupy wokali itd.), a następnie grupy te są miksowane ze sobą, tworząc końcowy efekt, co pokazano na rys. 1. Modele zostały wytrenowane na specjalnie przygotowanej bazie danych. Do znanej bazy MUSDB18-HQ [89] dograno pięć autorskich utworów i zmiksowano je na potrzeby niniejszej rozprawy.



Rysunek. 1. Schemat blokowy zaproponowanego system automatycznego miksu

Aby sprawdzić słuszność postawionych tez badawczych, przeprowadzono szereg eksperymentów −zarówno obiektywnych, jak i subiektywnych. Miksy stworzone za pomocą proponowanego systemu porównano z miksami amatorskimi − takimi, które zostały stworzone przy pomocy metod i oprogramowania zgodnych ze stanem wiedzy (ang. *state-of-the-art.*) [44] oraz przez profesjonalnych inżynierów dźwięku. Pierwsza z tez dotyczy obiektywnych cech otrzymanych miksów. Opracowany system powinien automatycznie miksować ścieżki wejściowe w taki sposób, aby miks uzyskany na wyjściu był obiektywnie jakościowo lepszy niż znane metody/oprogramowanie *state-of-the-art* i porównywalny (lub nie do odróżnienia od) miksu stworzonego przez profesjonalnego inżyniera miksu. W rozprawie udowodniono, że możliwe jest automatyczne miksowanie ścieżek wejściowych dostarczonych przez użytkownika, z wykorzystaniem wcześniej wytrenowanych modeli, w taki sposób, aby efekt końcowy był obiektywnie bardzo zbliżony do miksów przygotowanych przez profesjonalnego inżyniera miksu.

Podczas prowadzenia badań poddano obiektywnej analizie próbki sygnałów muzycznych, które nie zostały znormalizowane, tj. zbadano przebiegi poziomu takie jak: poziom średni głośności (ang. *RMS level*), zintegrowaną głośność (ang. *Integrated Loudness*), zakres głośności (ang. *Loudness Range*) i poziom szczytowy (ang. *True Peak* level) oraz wyznaczono wartości niskopoziomowych parametrów, m.in. deskryptorów MPEG-7 [133] (*Odd-to-Even Harmonic Ratio*, *RMS-Energy Envelope* i *Harmonic Energy*). Na podstawie uzyskanych wyników przeprowadzono analizę statystyczną i porównano uzyskane wyniki z parametrami wyznaczonymi dla porfesjonalnycy miksów. Uzyskane wyniki analizy statystycznej potwierdzają tezę nr 1, tj.: **„Możliwe jest miksowanie utworu muzycznego składającego się z wielu ścieżek z użyciem jednowymiarowej adaptacji autoenkodera Wave-U-Net, które może być obiektywnie oceniane jako porównywalne jakościowo do miksu stworzonego przez profesjonalnego inżyniera dźwięku”.**

W celu udowodnienia drugiej tezy przeprowadzono testy odsłuchowe na znormalizowanych próbkach miksów, w których słuchacze oceniali każdą próbkę w wielu

kategoriach oceny jakości dźwięku: balans (ang. *balance*), przejrzystość (ang. *clarity*), panoramowanie (ang. *panning*), przestrzeń (ang. *space*) i dynamika (ang. *dynamics*). Wyniki testów subiektywnych poddano analizie statystycznej.

Dodatkowo przeprowadzono analizę porównawczą wyników oceny obiektywnej i subiektywnej, bazującą na macierzach podobieństwa, obrazującą graficznie podobieństwo w parach obiektów, przedstawionych w formie chromagramów. Przeprowadzone analizy stanowią autorską metodologię oceny przygotowanych miksów.

Uzyskane wyniki potwierdzają tezę nr 2, tj.: **„Miksy utworów uzyskane przy pomocy zaproponowanego systemu można subiektywnie ocenić jako lepsze jakościowo niż miksy amatorskie i miksy wykonane metodami znanymi ze stanu wiedzy (ang. *state-of-the-art*), jak również porównywalne z miksami przygotowanymi przez profesjonalnego inżyniera miksu".**

W niniejszej rozprawie można wyróżnić oryginalne dokonania autora:

- Zaproponowano automatyczną metodę miksowania sygnałów fonicznych przy użyciu autoenkoderów Wave-U-Net.
- Przygotowano niestandardową bazę danych na potrzeby wytrenowania modeli.
- Zaproponowano autorską metodologię oceny przygotowanych miksów, której elementami są analizy obiektywne, testy subiektywne oraz korelacje pomiędzy wynikami tych ocen.
- Zaproponowano szereg obiektywnych parametrów i testów, które pozwoliły obiektywnie ocenić jakość otrzymanych miksów.
- Przygotowano test odsłuchowy, który pozwolił przetestować wiele charakterystyk otrzymanych miksów.
- Przeprowadzono liczne eksperymenty w formie testów subiektywnych, w których słuchacze oceniali jakość otrzymanych miksów.
- Zaproponowano metodę porównania przygotowanych miksów za pomocą graficznej reprezentacji, tj. macierzy podobieństwa, co pozwoliło na dodatkową weryfikację wyników oceny jakości obiektywnej i subiektywnej.
- Zbadano korelację wyników obiektywnych i subiektywnych.

Dalsze kierunki badań

W planach rozwoju proponowanej metody przewiduje się uwzględnienie w proponowanym systemie dodatkowego modułu, jakim jest integracja modułu automatycznej klasyfikacji instrumentów na wejściu systemu. W ten sposób użytkownik nie musiałby przypisywać odpowiednich ścieżek muzycznych do odpowiednich wejść w systemie. W obecnej formie, aby system działał poprawnie, użytkownik musi wskazać ścieżki basowe do modelu basu, ścieżki perkusyjne do modelu perkusji, itp. Automatyczna klasyfikacja instrumentów jest możliwa i poprawiłaby wydajność systemu. Takie próby zostały przeprowadzone przez autora

rozprawy na wstępnym etapie badań [7][57]. Poprawiłoby to również wrażenia użytkownika i łatwość użytkowania dla początkujących użytkowników, którzy nie są profesjonalnymi inżynierami dźwięku.

Zaproponowany przez autora system można by dodatkowo rozbudować, wprowadzając automatyczny moduł wykrywania gatunku muzycznego utworu. Może być zaimplementowany na wejściu systemu (podczas predykcji), jak i na wyjściu (podczas treningu). Takie rozwiązanie przyniosłoby dodatkowy zysk podczas uczenia modeli, które są później wykorzystywane do predykcji (miksu). System mógłby pozyskiwać informacje o gatunku wejściowym podczas uczenia bezpośrednio z bazy danych, dzięki czemu modele uczyłyby się miksowania utworów w danym gatunku. Podczas predykcji automatyczna klasyfikacja gatunkowa pozwoliłaby na lepsze miksowanie utworów w różnych gatunkach, ponieważ modele byłyby trenowane w analogiczny sposób. Co więcej, możliwe byłoby miksowanie tych samych zestawów utworów w różnych gatunkach muzycznych. Interesującym rozwiązaniem byłaby możliwość jednoczesnego miksowania utworu w wielu różnych gatunkach, gdzie użytkownik mógłby wybrać preferowany miks lub łączenia gatunków (np. 60% Rock i 40% Electronica). Obecnie modele są trenowane na bazie danych składającej się z czterech gatunków muzycznych – Pop, Alternative, Rock, Electronica. Kolejnym proponowanym kierunkiem dalszych prac badawczo-rozwojowych jest dodatkowy moduł umożliwiający edycję poszczególnych utworów. Taki moduł pozwalałby na automatyczną synchronizację utworów ze sobą (np. w wielościeżkowych nagraniach perkusji) oraz automatyczne usuwanie (lub zmniejszanie głośności) niepożądanych dźwięków (takich jak oddech wokalisty czy przypadkowe uderzenia mikrofonu pomiędzy wartościowym sygnałem). Taki moduł powinien zostać zaimplementowany na wejściu systemu, aby wszystkie ścieżki można było edytować przed miksowaniem. Obecnie użytkownik musi ręcznie synchronizować wszystkie utwory i edytować niepożądane lub przypadkowe dźwięki.

Szczególnie interesującym kierunkiem badawczym, będącym zakresem odrębnym od badań prowadzonych na potrzeby niniejszej rozprawy, jest automatyczne rozpoznawanie rytmu i tempa. Znajomość tempa danego utworu w jednostkach uderzeń na minutę (ang. *beats per minute*) jest kluczowa przy korzystaniu z efektów takich jak pogłos czy echo (ang. *delay)*. Jeśli pogłos nałożony na ścieżkę jest zbyt długi, może wystąpić efekt maskowania. Znajomość rytmu i tempa utworu może również pozwolić na ustawienie idealnego tempa echa. Synchronizowanie efektu echa z tempem utworu to bardzo powszechna procedura stosowana przez inżynierów miksu. Na przykład, wiedząc, że tempo utworu wynosi 120 BPM, mikser jest w stanie ustawić tempo odtwarzania kolejnych odbić echa jako ćwierćnut. Większość popularnych wtyczek programowych, które oferują ten efekt, pozwala użytkownikowi wybrać opcję „1/4” lub w przypadku konieczności podania wartości w milisekundach użytkownik może ją obliczyć (w podanym przykładzie będzie to 500 ms).

Automatyzacja procesu miksowania sygnałów fonicznych może być szczególnie przydatna w obszarze produkcji gier komputerowych czy tworzenia tzw. sygnatury muzycznej firm czy sieci sklepowych (ang. audio/music branding).

ACKNOWLEGMENTS

I would like to acknowledge and thank my supervisor, ***prof. dr hab. inż. Bożena Kostek,*** *for giving me the opportunity to focus on a very specific topic that is of great interest to me and is important for my personal experience of music. I would also like to express my gratitude for solving all potential issues, supporting my professional involvement, internships and traveling ideas which enabled me to broaden my horizons and gain experience.*

I want to thank my colleagues at Multimedia Systems Department for all the helpful discussions on the issues related to my Ph.D. work. Special thanks to ***dr inż. Michał Lech*** *for countless debates on this very topic. At the same time, I would like to express gratitude to all the listeners who participated in numerous subjective tests.*

I would like to thank ***prof. Thomas Görne*** *from Hamburg University of Applied Sciences for his help in organizing my internship and realizing the program as well as facilitating the subjective test performance.*

I would like to thank ***mgr inż. Łukasz Pindor*** *for his interest and support with the programming aspect of this work.*

I am very grateful to my ***mother*** *for her support during all stages of my education, her understanding and respect for my choices.*

Special thanks to my fiancée, ***Ada Durzyńska****, for the extra help with translations, being with me during hard times, and for making me believe that anything is possible.*

This dissertation was partially funded by InterPhD-2 project POWR.03.02.00-IP.08-00-DOK/16.

LIST OF THE MOST IMPORTANT SYMBOLS AND ABBREVIATIONS

ADC – Analog-to-digital converter
AES – Audio Engineering Society
ANN – Artificial Neural Network
ASE – Audio Spectrum Envelope
AT – Attack Time
ATR – Analog Tape Recorder
BPM – Beats Per Minute
BP – Bandpass
BR – Bandreject
CR – Compressor Ratio
CS – Compressor Slope
CT – Compressor Threshold
DAC – Digital to Analog Converter
DAW – Digital Audio Workstation
DFT – Discrete Fourier Transform
DI – Direct Input
EQ – Equalization
ER – Expander Ratio
ES – Expander Slope
ET – Expander Threshold
FIR – Finite Impulse Response
GUI – Graphical User Interface
HC – Highcut
HE – Harmonic Envelope
HP – Highpass
IIR – Infinite Impulse Response
LC – Lowcut
LP – Lowpass
LSI – Large-scale integrated
LT – Limiter Threshold
LU – Loudness Units
LUFS – Loudness Units relative to Full Scale
MIDI – Musical Instrument Digital Interface
MIR – Music Information Retrieval
MUSHRA – Multiple Stimuli with Hidden Reference and Anchor
NT – Noise Gate Threshold
OEHR – Odd to Even Harmonic Ratio
RMS – Root Mean Square
RMSE – Root Mean Square Error

RMSEE – Root Mean Square Energy Envelope
RT60 – Reverb Time
RT – Release Time
SNR – Signal-to-Noise Ratio
SFX – Sound effects
SSIM – Structural Similarity Index
SSM – Self-Similarity Matrix
VIF – Visual Information Fidelity
VST – Virtual Studio Technology

TABLE OF CONTENTS

1. INTRODUCTION

The realm of modern music and sound production is complex and diverse. To achieve a single end product – music that reaches the world – it takes effort, immense commitment, money, and the combined creative talents of all kinds of experts. The music world consists of artists, engineers, producers, managers, executives, manufacturers, and marketing strategists. All of whom are experts in their fields, such as music, recording, acoustics, production, electronics, law, media, marketing, sales, and graphics. They work together to capture a spark of creativity and transform it into a product that can be marketable. What drives these teams of people, and what has driven them throughout the entire history of recorded sound, are changes in the industry, the cultural tastes, the art of music, and the ever-present changes and challenges in production technology.

The process of a musical piece production can be divided into the following steps: composition, recording, editing (sometimes done just after recording or during the mixing stage), mixing, and mastering (Fig. 1.1). The composition step can take on many forms. It can be creating a song in MIDI (Musical Instrument Digital Interface) in any DAW (Digital Audio Workstation), writing down the composition on a five-line staff, or just having a music piece in the songwriter's head (Fig. 1.2). The recording step can also vary. Nowadays, it rarely happens to rent a big studio with an engineer and a producer. More commonly, the artists record their songs track by track in a home studio. Regardless of how a song is recorded, the result is a recorded song where each instrument is given a separate mono track and, in some cases, multichannel. An example of a multichannel recording setup is shown in Fig. 1.3. After an artist decides to record a musical piece or song, the sound engineer uses the correct microphones, records the multitrack material, and edits it. The mixer's role is to set up proper proportions between the elements and adapt their properties (i.e., time and frequency-based properties) [135]. A more than adequate mix can emphasize the artistic character of a song or even define the music genre [17][20][46]. Mixing techniques have existed ever since people learned to record music. Mixing was first introduced as physically adjusting the instrument and microphone setup. When a multitrack recording became possible, the mixing process was performed using analog hardware and − later − digital tools.

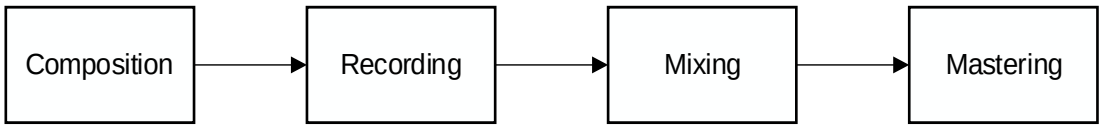


Fig. 1.1. Process of a musical piece production

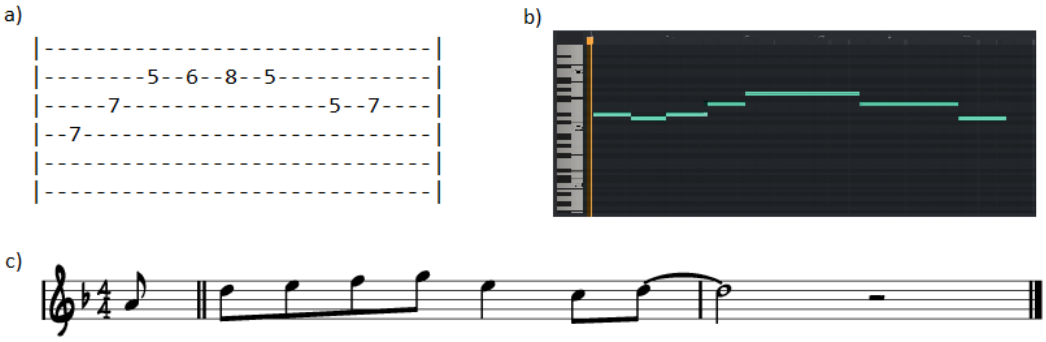


Fig. 1.2. Various types of composing a song: a) writing the guitar part as a tablature, b) composing using MIDI, c) writing down notes on a five-line staff

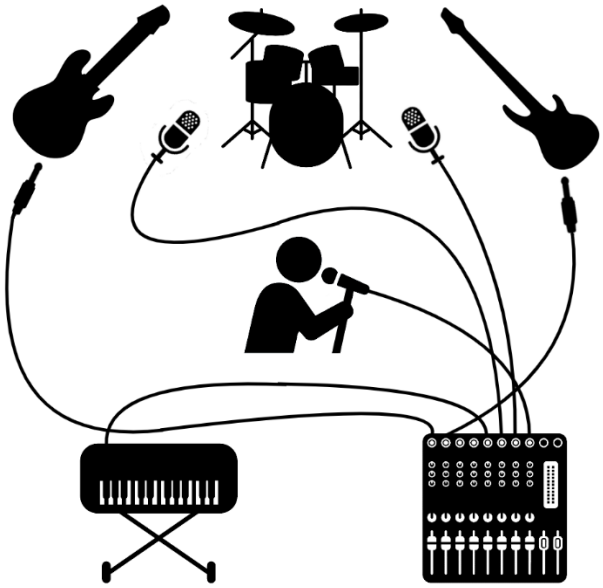


Fig. 1.3. Example of a multitrack recording setup

After recording, assembling, and editing all tracks of a given project, the tracks should then be mixed into the final media form. The mixing process can be done in several ways:

- Routing analog tape recorder (ATR) tracks through an analog console.
- Routing Digital Audio Workstation (DAW) tracks to an analog console, mixing in the analog domain.
- Mixing DAW tracks within the software mixer (“in-the-box” mixing).

Nowadays, fully analog studios are very rare. Analog equipment is expensive, requires special care and effort to upkeep, and restoring a session to mix is complex and requires the work of multiple people. However, the so-called *analog sound* is what every mixing engineer is looking for, regardless of how they work. The second mentioned mixing approach is hybrid mixing, where songs from the DAW are routed to an analog mixing console or where single tracks are routed into outboard hardware, e.g., compressor, equalizer, or reverb. This approach is precisely in-between in the cost/effect category. The engineer does not have to buy

expensive tapes or analog mixing boards – instead, a few channel strips will suffice. The least expensive and the easiest method of fully mixing a song is the fully digital approach, called *in-the-box.* Many renowned engineers, for example, Andrew Scheps, entirely changed their approach to mixing to digital [2][86]. The *in-the-box* way of mixing has various advantages, such as being relatively inexpensive, the fact that there are a lot of plugins that emulate the analog equipment more and more faithfully and being able to go back to a project with one click of the mouse.

Regardless of the approach chosen, mixing is used to shape the character, tone, and intention of the production in relation to:

- Relative level.
- Spatial processing or panning (placement of the sound within the stereo or surround field).
- Equalization (altering the relative frequency balance of a track).
- Dynamics processing (adjusting the dynamic range of a single track, a group, or an output bus to optimize the levels or allow it to fit better within a mix).
- Effects processing (adding delay-, pitch- or reverb-related effects to a mix to alter or augment the piece in an attractive, natural, or unnatural way).

Capturing music and turning it into a product used to be far less accessible in the early decades of music production. It required visiting a commercial recording studio equipped expensively and hiring an expertly skilled team. Another option appeared after the large-scale integrated (LSI) circuit was introduced and then mass-produced. It became possible for musicians, producers, and engineers to create and record music at home or any other facility of their choosing. This technology achievement originates from the idea and possibility of constructing a personal audio production studio that is affordable and easier to master by almost anybody, while professional engineers are constantly under pressure to produce high-quality mixes quickly and at a low cost [85].

The human way of creating “a good mix” is always superior to computer-based artificial intelligence. Still, there are areas where much faster, more intelligent, and more powerful algorithms are needed and can be implemented [97], e.g., game development or music branding may be mentioned here.

Aims of the study

The growing need for automation of every step during music production inspired this work, which focuses on the mixing stage of the process. Therefore, the primary aim of the presented dissertation is to propose a framework for automatic music mixing (see Fig. 1.4).

As a result of the doctoral dissertation, the following theses are expected to be proven:

1. **It is possible to mix music consisting of separate raw recordings using a one-dimensional adaptation of the Wave-U-Net autoencoder that can objectively be evaluated similarly to a professional mix.**
2. **The prepared mixes may subjectively be evaluated as better ones than recordings created by an amateur engineer or mixes produced using state-of-**

the-art methods and can be comparable to mixes produced by a professional mixer.

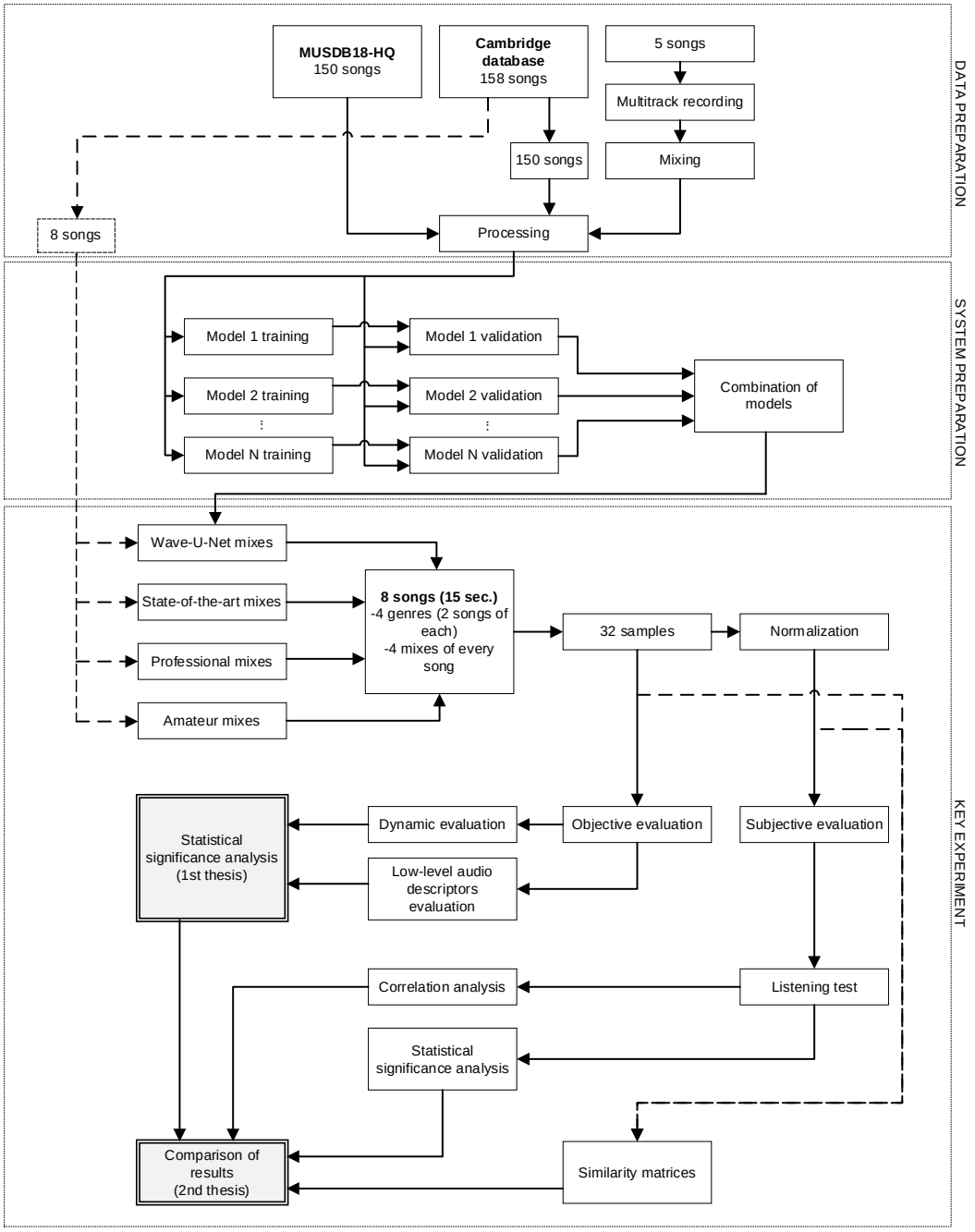


Fig. 1.4. Stages of analysis executed in the course of the dissertation

In Fig. 1.5, the organization of the dissertation work is shown, in which all chapters summarizing their content are introduced. The theoretical background of the work is provided in the following two chapters (Chapters 2 and 3). It includes signal processing methods, technology applied to audio mixing, and selected machine learning methods applied to audio processing. Chapters 4 and 5 illustrate the system architecture and information about the key experiment, where proprietary engineered models are introduced and all mixes are prepared.

The experimental results and their objective, subjective and comparative analyses are presented in Chapter 6. Finally, Chapter 7 provides conclusions, summarizes the main findings, and discusses further research perspectives.

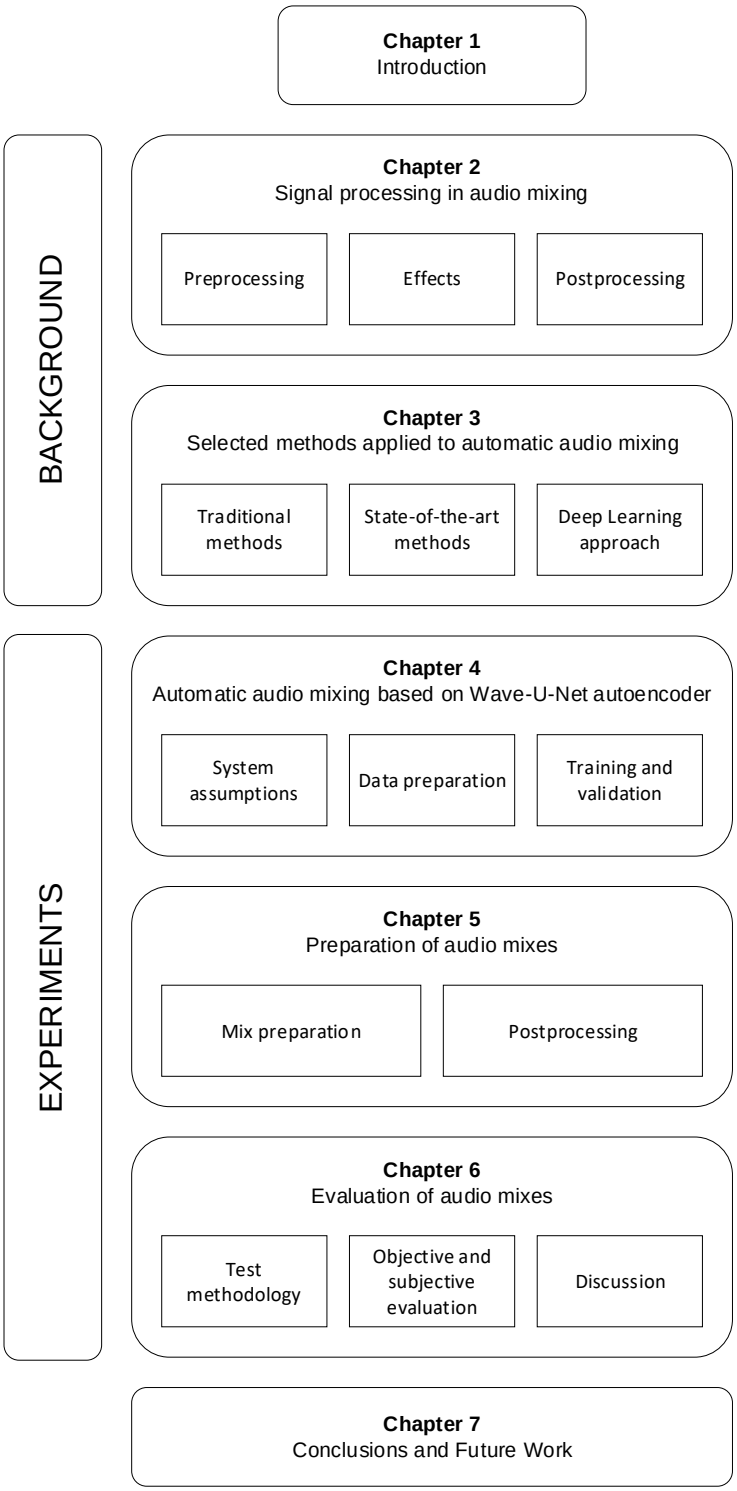


Fig. 1.5. Organization of the dissertation. Chapters are presented along with their content

2. SIGNAL PROCESSING IN AUDIO MIXING – AN OUTLINE

In this chapter, elements of audio signal manipulation are to be presented shortly. The methods shown are directly connected to the audio mixing task. However, basics such as signal sampling are not described here. In contrast, the most important signal operations related to sound manipulation are highlighted.

Audio signal processing is a subfield of signal processing that is concerned with the electronic manipulation of audio signals. Processing methods and application areas include storage, data compression, music information retrieval [50][82][103][141], speech processing, localization, acoustic detection [55], music transcription [112], noise cancellation, acoustic fingerprinting, sound recognition [54], synthesis [25], and enhancement (e.g., equalization, filtering, level compression, echo and reverberation removal or addition, etc.), and, of course, music mixing [92] and mastering.

There are several definitions of audio mixing. It seems that the best suited was the definition proposed by Ramirez and Reiss in [92], which says: **"Audio mixing essentially tries to solve the problem of unmasking by manipulating the dynamics, spatialization, timbre or pitch of multitrack recordings"** [92].

This definition combines all of the mixer's essential tasks into one problem – unmasking. One can look at the mixer's job to do just that, but the reality is more complicated. There are multiple tools to achieve the desired effect. In the subsequent sections, the author of this dissertation focuses on signal processing devices, applications, and techniques that can be divided into the following areas:

- Amplitude level processing – in terms of dynamic range processing and spatialization (subsections 2.1, 2.2, 2.4, and 2.6);
- Spectral content of sound – in terms of equalization and timbre quality (subsection 2.3);
- Time-based effects – augmentation or re-creation of room ambiance, delay, time or pitch alterations, and other effects – referring to timbre quality (subsection 2.5).

As already mentioned, one of the main phenomena, such as masking occurring in the mixing of multitrack recordings, should be brought here, even though this does not constitute the main subject of this dissertation directly [61][138]. However, this issue will be discussed in Chapter 3.

2.1. Pre-processing and level adjustment

Even the best-recorded music source requires some pre-processing [14]. It could be achieved by adequately preparing the tracks for mixing, bringing them to the same frequency response and bit depth or initial level setting (gain staging). When mixing using analog equipment, gain staging is an important step because of how analog gear operates. Traditionally, this is done by an assistant mixer or the mix engineer themselves.

Based on the ITU-R BS.1770-2 standard [42], the loudness (level) of a single channel track may be represented by RMS level (2.1) and peak level (2.2):

$$L_{rms} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}|x(n)|^2} \tag{2.1}$$

$$L_{peak} = \max(x) \tag{2.2}$$

where x is the amplitude vector of a mono track. For a stereo track $x = [x_L\ x_R]$, these equations become:

$$L_{rms} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|x_L(n)|^2} + \sqrt{\frac{1}{N}\sum_{n=1}^{N}|x_R(n)|^2}}{2} \tag{2.3}$$

$$L_{peak} = \max(\max(x_L), \max(x_R)) \tag{2.4}$$

Additionally, for more precise level measurement, a simple hysteresis gate can be applied (Fig. 2.1) to a signal that determines which part of the signal is active and calculates the level purely on those snippets (Fig. 2.2):

$$a(n) = \begin{cases} 0, & if\ a(n-1) = 1\ and\ \tilde{x}(n) \leq T_1 \\ 1, & if\ a(n-1) = 0\ and\ \tilde{x}(n) > T_2 \\ a(n-1), & otherwise \end{cases} \tag{2.5}$$

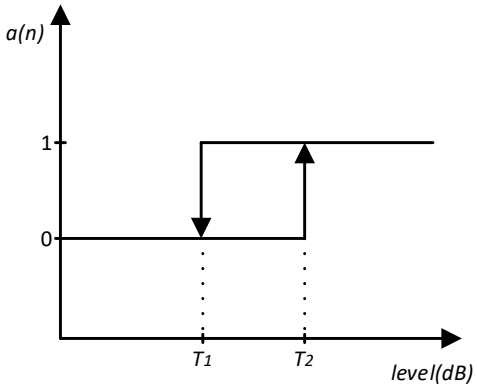


Fig. 2.1. Hysteresis gate

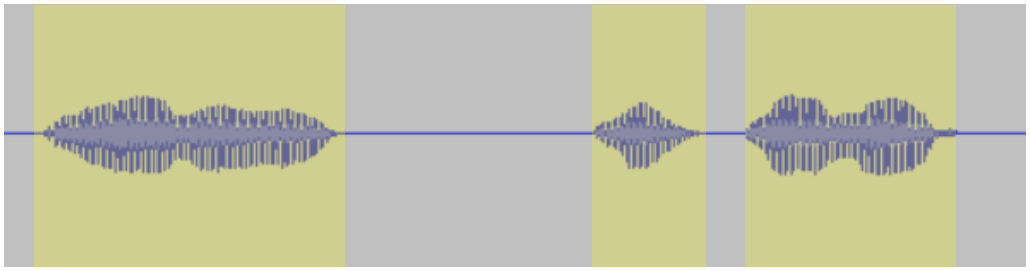


Fig. 2.2. Highlighted audio regions are active

where a is the binary vector that indicates the activity of the track, $\tilde{x}$ is the audio track smoothed version, T_1 is the level threshold when audio is active, T_2 is the threshold when audio is inactive, and $T_1 \leq T_2$. x is downmixed (summed) to mono and divided by two for stereo tracks.

2.2. Spatial processing

The distribution of an audio signal into a new sound field (stereo or multi-channel) determined by a panning control setting is known as *panning*. In a standard recording console, there is a panning control, known as a *taper* or *pan law*, for each incoming source channel. On-screen virtual knobs and sliders replace a panning potentiometer in an audio mixing software. This is to determine how much of a source signal is sent to the left and right channels.

Pan law is a principle in recording and mixing [29]. It states that, provided a perfect response in the loudspeaker system is present and the room is acoustically perfect, any signal of equal phase and amplitude played in both channels of a stereo system will increase in loudness up to 6.02 dBSPL. The acoustic summing of a room and system can often be less than perfect, so as the mono signal is panned from center to hard left or right, the specific relative loudness level will increase from -3 to 0 dB without any interference to the signal. The *pan law* prevents this from happening. There are many different pan law rules, however, the 3 dB *pan law* rule is the most commonly used in DAWs (Fig. 2.3) [29].

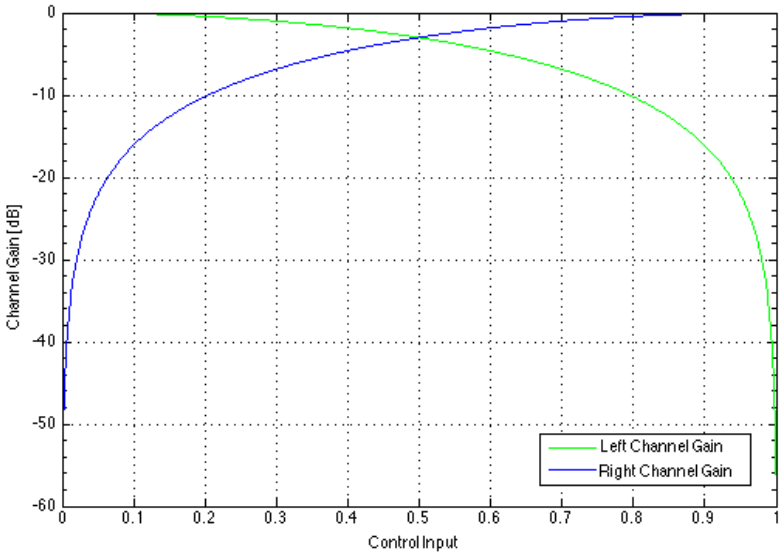


Fig. 2.3. 3 dB Equal Power Pan [29]

2.3. Equalization

An audio equalizer is a circuit, plug-in, or device that controls the timbral or harmonic content of a recorded sound. Equalization (EQ) can be defined as a process of amplifying or suppressing different frequency components within an electronic signal. In the digital domain, it is usually implemented with the use of filters. A digital signal can be described as a set of varying partials with frequency and amplitude differences. A selection of partials is to be performed by the filter to modify their amplitude to reject, retain or emphasize selected frequencies. Different filter types are classified as follows (see Fig. 2.4) [150]:

- Low-pass (LP) filter when low frequencies up to the particular cut-off frequency f_c are selected and transmitted, whereas frequencies above f_c are suppressed. A resonance

could also occur around f_c. High-cut (HC) and low-pass filtering may be treated as two interchangeable terms.

- High-pass (HP) filter when high frequencies (higher than f_c) are selected, whereas frequencies below f_c are suppressed. A probable resonance may occur around f_c. Low-cut (LC) filtering is a synonym for high-pass filtering.
- Band-pass (BP) filter when frequencies between a lower cut-off (f_{cl}) and a higher cut-off (f_{ch}) are picked and transmitted; those below f_{cl} and higher than f_{ch} are suppressed.
- Band-reject (BR) filter when frequencies between a lower cut-off (f_{cl}) and a higher cut-off (f_{ch}) are suppressed. Frequencies below f_{cl} and higher than f_{ch} are passed.
- All-pass when all frequencies are transmitted, but the phase of the input signal is altered.

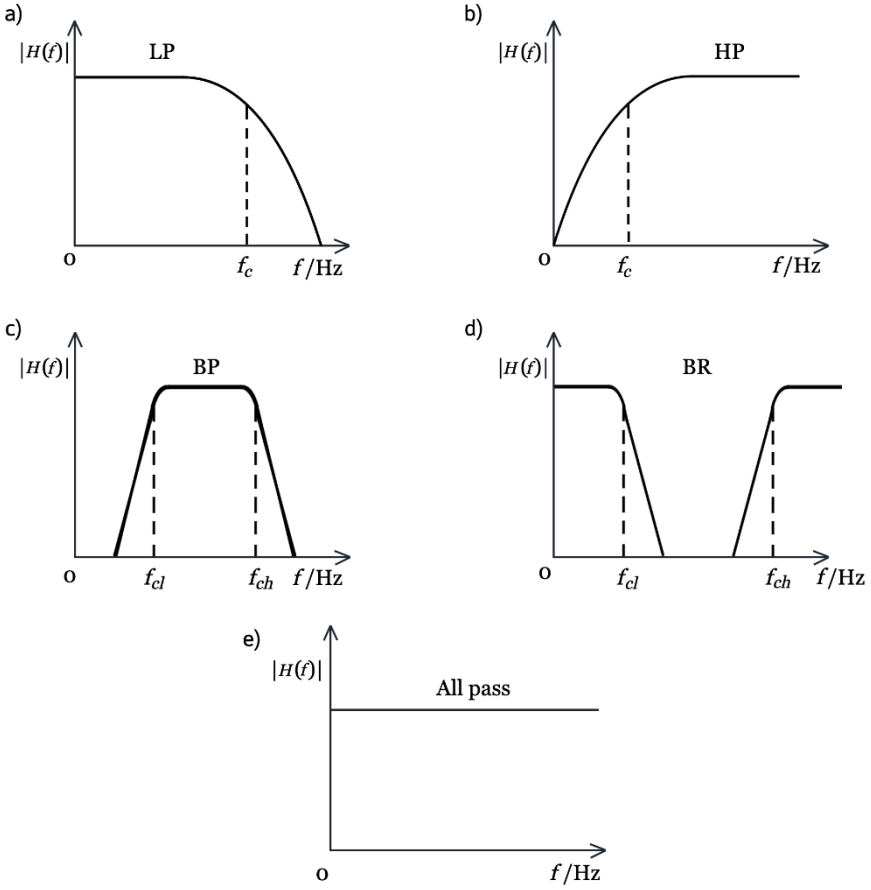


Fig. 2.4. Response characteristics of the selected filter types: a) low-pass, b) high-pass, c) band-pass, d) band-reject, e) all-pass [150]

A filter can be designed in a number of different ways. However, the canonical filter is the easiest to build, as shown by the difference equations (Eqs. (2.6) and (2.7)) [150]:

$$x_n(n) = x(n) - a_1 x_h(n-1) - a_2 x_h(n-2) \tag{2.6}$$

$$y(n) = b_0 x_h(n) + b_1 x_h(n-1) + b_2 x_h(n-2) \tag{2.7}$$

which leads to the transfer function:

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \quad (2.8)$$

When $a2 = b2 = 0$, the filter is reduced to the first order. These types of filters can be implemented as a low-cut, high-cut, or all-pass based on the coefficients contained in Table 2.1, where K depends on the cut-off frequency f_c by:

$$K = \tan\left(\frac{\pi f_c}{f_s}\right) \quad (2.9)$$

When a phase shift of $-90°$ is reached, the coefficient K controls f_c the all-pass.

Table 2.1. Coefficients for first-order filters

	b_0	b_1	a_1
Low-pass	$K/(K+1)$	$K/(K+1)$	$(K-1)(K+1)$
High-pass	$1/(K+1)$	$-1(K+1)$	$(K-1)(K+1)$
All-pass	$(K-1)(K+1)$	1	$(K-1)(K+1)$

Table 2.2 contains the second order filter coefficients. Also, the Q factor is defined, which determines the behavior for different filter types [150], i.e.:

- For low-pass and high-pass, the height of the resonance is controlled. The filter is maximally flat up to the cut-off frequency for the value of Q equals $\frac{1}{\sqrt{2}}$. For lower Q, higher pass-band attenuation occurs; for higher Q, frequencies around f_c are amplified.
- For band-pass and band-reject Q factor is related to the bandwidth f_b by $Q = \frac{f_c}{f_b}$.
- For all-pass Q factor controls the bandwidth and depends on the locations where the $\pm 90°$ phase shift at f_c appears in relation to the $-180°$ phase shift.

Table 2.2. Coefficients for second order filters

	b_0	b_1	b_2	a_1	a_2
Low-pass	$\frac{K^2Q}{K^2Q+K+Q}$	$\frac{2K^2Q}{K^2Q+K+Q}$	$\frac{K^2Q}{K^2Q+K+Q}$	$\frac{2Q(K^2-1)}{K^2Q+K+Q}$	$\frac{K^2Q-K+Q}{K^2Q+K+Q}$
High-pass	$\frac{Q}{K^2Q+K+Q}$	$-\frac{2Q}{K^2Q+K+Q}$	$\frac{Q}{K^2Q+K+Q}$	$\frac{2Q(K^2-1)}{K^2Q+K+Q}$	$\frac{K^2Q-K+Q}{K^2Q+K+Q}$
Band-pass	$\frac{K}{K^2Q+K+Q}$	0	$-\frac{Q}{K^2Q+K+Q}$	$\frac{2Q(K^2-1)}{K^2Q+K+Q}$	$\frac{K^2Q-K+Q}{K^2Q+K+Q}$
Band-reject	$\frac{Q(1+K^2)}{K^2Q+K+Q}$	$\frac{2Q(K^2-1)}{K^2Q+K+Q}$	$\frac{Q(1+K^2)}{K^2Q+K+Q}$	$\frac{2Q(K^2-1)}{K^2Q+K+Q}$	$\frac{K^2Q-K+Q}{K^2Q+K+Q}$
All-pass	$\frac{K^2Q-K+Q}{K^2Q+K+Q}$	$\frac{2Q(K^2-1)}{K^2Q+K+Q}$	1	$\frac{2Q(K^2-1)}{K^2Q+K+Q}$	$\frac{K^2Q-K+Q}{K^2Q+K+Q}$

Depending on the desired effect, EQ may be applied to a single instrument (channel), to a group, or even to a master bus (output) of a DAW. Equalizers may be constructed fully analog or made digitally as VST (Virtual Studio Technology) plugins.

2.4. Dynamic range control

Dynamics processing is executed by amplifying devices where the level of the input signal automatically controls gain. The processing is formed on an envelope follower (an amplitude/level detection scheme), a static curve which, from the result of the amplitude/level detection scheme, derives a gain factor, a smoothing filter that prevents gain changes that are too abrupt, and a multiplier that weights the input signal (Fig. 2.5) [150]. The input signal can be (optionally) delayed, compensating for the delay in the side chain (lower path in Fig. 2.5). The gain factor is conventionally derived from the input signal. To control the gain factor of the input signal, the side chain path can also be connected to another signal.

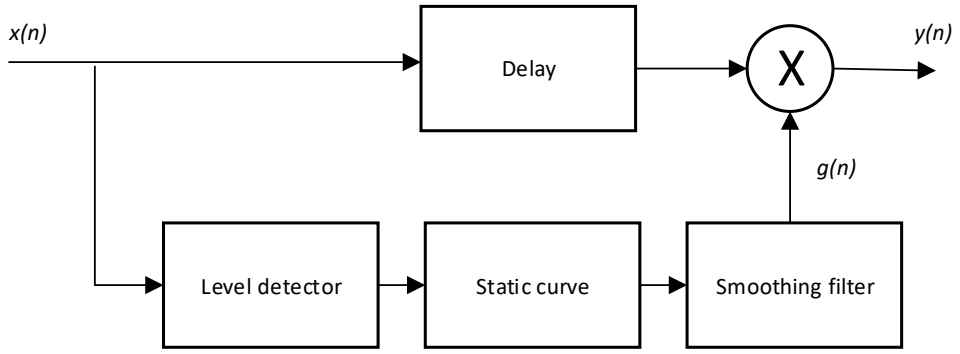


Fig. 2.5. Block diagram of a dynamic range controller [150]

The most dynamic range controllers have the following parameters that can be set up:

- Input gain: controls how much signal will be processed.
- Threshold: controls at which level signal will be affected.
- Output gain: controls “how much” of the signal is to be sent to the output of the processor.
- Slope ratio: controls the ratio of the input to output gain.
- Attack: controls how fast (usually in milliseconds) signal is to be processed after reaching the threshold level.
- Release: controls how slow (usually in milliseconds) signal level comes back to its original value after being processed.

Dynamic range control tools consist of:

- Limiters.
- Compressors and expanders.
- Noise gates.
- De-essers.

2.4.1. Limiter

The function of a limiter is to provide control over the highest peaks in the signal. In the process, it should change the dynamics of the signal as little as possible. This is accomplished by implementing a characteristic curve with an infinite ratio $R = \infty$ above LT (a limiter threshold) (Equation (2.10)).

$$G = \begin{cases} 0\ dB & if\ X < LT \\ LT - X & else \end{cases} \tag{2.10}$$

Consequently, the output level $Y = X + G$ (where X is the input signal and G is gain) should not exceed the limiter threshold (LT). Lowering the peaks can boost the overall signal. Limiting can be performed on single instrument signals, as well as the final mixes of multichannel applications. A limiter utilizes the peak level measurement and should react quickly when the input signal exceeds the LT. Parameters typical for a limiter are attack time: $t_{AT} = 0.02 \dots 10\ ms$ and release time $t_{RT} = 1 \dots 5000\ ms$, for both the smoothing filter and the peak measurement. An implementation, such as in Fig. 2.6, may perform the computation of gain in linear values by:

$$g(n) = \min\left(1, \frac{l_t}{x_{PEAK}(n)}\right) \tag{2.11}$$

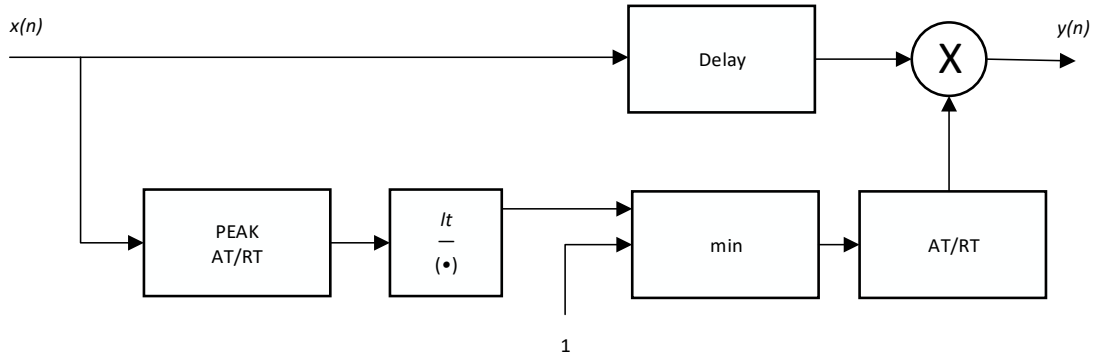


Fig. 2.6. Block diagram of a limiter [150]

where $l_t = 10^{\frac{LT}{20}}$ is the threshold measured on the linear scale.

2.4.2. Compressor and expander

In contrast to a limiter, which purpose is to eliminate any dynamics above a threshold by keeping the output level constant, a compressor only reduces the dynamics and compresses the dynamic range. The dynamics, reduced by a compressor, can be utilized to increase the overall level, boosting the loudness while staying within the allowed amplitude range. The opposite of a compressor is an expander. The expander increases the dynamics by mapping the input's small-level changes to the output signal at larger levels. The expander, when applied to low-level signals, produces a lively sound characteristic. The corresponding ratios (ER, CR) and slopes (ES, CR) of the characteristic curve are as follows: $0 < ER < 1$ and $ES < 0$ for the expander and $CR > 1$ and $0 < CS < 1$ for the compressor.

Typically, RMS level detectors are employed by expanders and compressors with an averaging time in the range $t = 5 \dots 130\ ms$ and a smoothing filter with $t_{AT} = 0.1 \dots 2600\ ms$ and $t_{RT} = 1 \dots 5000\ ms$. Combining an expander for low signal levels with a compressor for high signal levels leads to the gain computation:

$$G = \begin{cases} CS(CT - X) & if\ X > CT \\ 0\ db & if\ ET \leq X \leq CT \\ ES(ET - X) & if\ X < ET \end{cases} \tag{2.12}$$

$$= \min(0, CS(CT - X), ES(ET - X)) \tag{2.13}$$

CT (compressor threshold) denotes the threshold above which the compressor affects the signal. ET (expander threshold) represents the threshold below which the expander affects the signal [150]. The resulting combined system is depicted in Fig. 2.7.

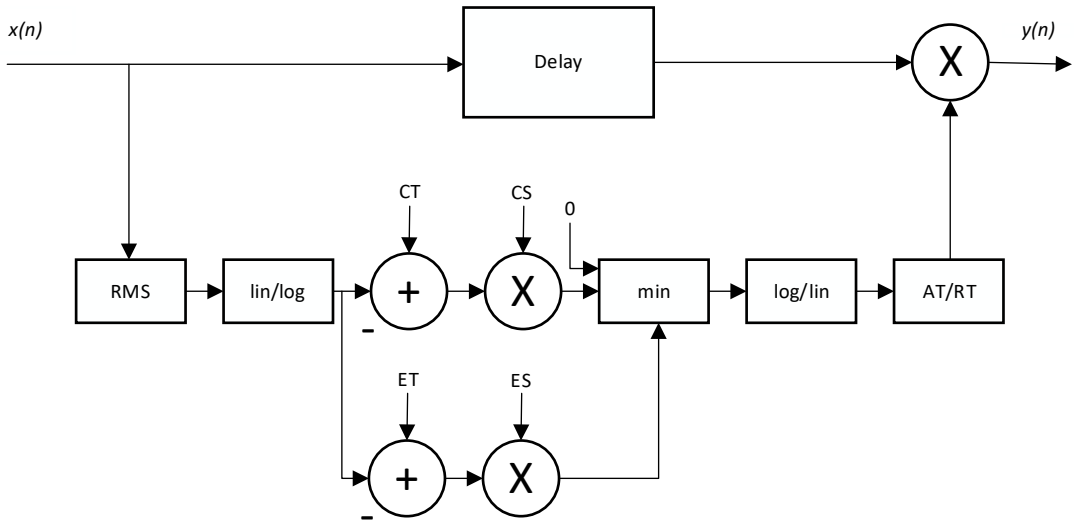


Fig. 2.7. Block diagram of a compressor/expander [150]

Gain can also be calculated without the use of logarithmic values by:

$$g(n) = \min\left(1, \left(\frac{x_{RMS}(n)}{ct^2}\right)^{-\frac{CS}{2}}, \left(\frac{x_{RMS}(n)}{et^2}\right)^{-\frac{ES}{2}}\right) \tag{2.14}$$

where the square root of the RMS is taken to the powers by halving the respecting slopes. This approach requires exponentiation and makes the conversion to and from the logarithmic domain unnecessary.

Compressors are usually applied to reduce the amplitude variations above a set threshold. However, in particular circumstances, e.g., for acoustic guitar or percussion instruments, high-amplitude transients should be left unaffected, while the weaker parts of the signal should be enhanced. Parallel compression is recommended for such applications. A parallel compression means adding a heavily compressed version of the signal to the unaffected signal. Typically, the softer parts of the sound are heavily compressed using a low threshold, a short attack time, and a high ratio. Then, the compressed region is amplified and mixed with the original signal. The sound acquired, as a result, retains the original sound's transparency because the transients are unaffected. The above kind of processing can be referred to as side-chain compression or New-York compression [13][41][43][47].

2.4.3. Multiband compressor

Compression might be required only on a specific frequency band; this depends on the spectral content of the sound to process. The multiband compressor has been developed for

when several frequency bands of the spectrum must be processed individually. The input signal is split into several bands, typically three to five, by a filter bank. Then, the bands are individually processed. From a band-limited version of the input signal, the side-chain signal of each compressor is derived. Usually, the filter bank is both the same for the side-chain and the input signal. The side-chain is sometimes implemented as an independent processing unit. This helps to set up intricate relations, in which the content of the input signal in a limited frequency band conditions the compression in another frequency band.

Multiband dynamics processors can be used in mastering (the process is described in detail in Chapter 2.7). No direct correction of the individual musical parts is possible when the mix is completed and the music is available as a stereo track. If each musical part or instrument occupies mainly a particular frequency band, minor corrections are still possible if they are band-limited.

A significant advantage of multiband compression is the ability to limit the unwanted side-effects of compression, e.g., distortions or tonal balance alteration, to a given frequency band. In individual frequency bands, the reduction of the ratio of peak to RMS amplitude can be implemented more effectively than at the full audio bandwidth, which allows the increase of the track's overall loudness. When presented with two musical works, a casual listener often prefers the louder one. Considering this, a higher average levels trend has developed. However, moderation is necessary since loudness can induce degradation of quality in both the audio and musical quality [47][62].

2.4.4. Noise gate

A noise gate can be considered an extreme expander with a slope of $-\infty$, resulting in the complete muting of signals below the chosen noise gate threshold (NT). The noise gate is used to gate out noise by setting the threshold just above the background noise level so that the gate opens only when the desired signal with a level above the threshold is present. When recording a drum set, a specific application is found. A different decay time occurs for each element of a standard drum set. When not damped manually, their sounds mix. The result is not distinguishable. However, every sound can be faded out automatically after the sound's attack part, when each element is processed by a noise gate. The result is an overall cleaner sound. The noise gate functional units are shown in Fig. 2.8.

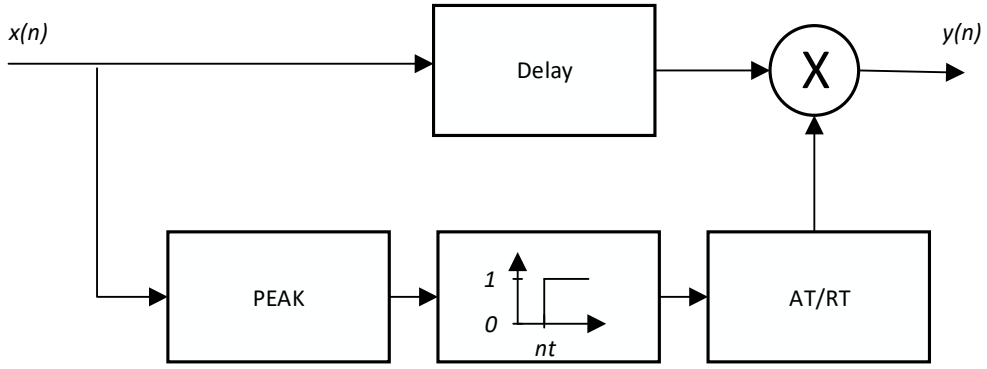


Fig. 2.8. Block diagram of a noise gate [150]

A peak measurement is usually the deciding factor for activating the noise gate. This leads to fade in/fade out of the gain factor $g(n)$ with an appropriate attack (AT) and release (RT) times. There are further refinements possible. A hold time has to be implemented to avoid a stuttering effect (when the input signal is close to the threshold) [5].

2.4.5. De-esser

A de-esser is a signal processing device typically used to process speech and vocals. It consists of a bandpass filter typically between 2 and 6 kHz (the main range of a human voice.) The bandpass filter detects the level of the signal in this frequency band. If a certain threshold is exceeded, the gain factor is applied to the same frequency as the peak/notch filter is tuned to, as shown in Fig. 2.9. Applying de-essers to speech or vocal signals helps avoid high-frequency sibilance.

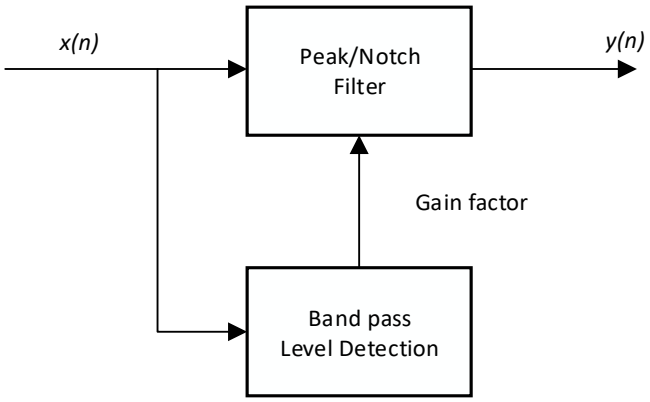


Fig. 2.9. Block diagram of a de-esser [150]

High-pass and shelving filters are used with good results as an alternative to the bandpass/notch filters. The threshold should depend on the overall level of the signal (a relative threshold) to make the de-esser more effective against input level changes [78]. A de-esser is essentially a one-band compressor within a specific frequency range, typically used to turn down the harsh sibilance in vocal performance.

2.5. Time-based effects

An important category of effects that one can use to augment or alter an audio signal revolves around delaying signal and replicating sound over time. These effects are time-based and add a perceived depth to signals or change the perception of the dimensional space of the recorded audio signal. Although a range of time-based effects can be applied during the mixing session, the main area of focus of this work will be the use of:

- Time-delay.
- Reverberation (reverb).

2.5.1. Time-Delay

FIR comb filter

A Finite Impulse Response (FIR) comb filter (Fig. 2.10) is a structure that simulates a single delay of the input signal by a given time duration. The effect can only be audible after the processed and input signals are combined, with the input signal present in the output signal first (as a reference). There are two tuning parameters included in this effect: τ – the amount of time delay and the relative amplitude of the delayed signal to the reference signal.

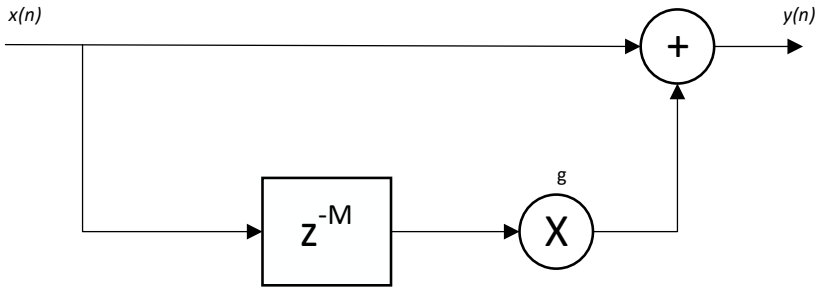


Fig. 2.10. Block diagram of an FIR comb filter [150]

The difference equation and the transfer function are given by:

$$y(n) = x(n) + gx(n - M) \tag{2.15}$$

$$\text{with } M = \frac{\tau}{f_s} \tag{2.16}$$

$$H(z) = 1 + gz^{-M} \tag{2.17}$$

In the above filter, the direct signal and the delayed version make up the time response. This type of time-domain behavior creates unique frequency-domain patterns. For negative values of g, frequencies that are multiples of $1/\tau$ are attenuated by the filter, while the frequencies in-between are amplified. The opposite happens for the positive g values – the filter amplifies the frequencies that are multiples of $1/\tau$ and the frequencies in-between are attenuated. The transfer function of this type of filter shows a series of spikes. The gain may vary between $1 + g$ and $1 - g$ [81].

Similarly, as with acoustical delays, the Finite Impulse Response comb filter has an effect on the time and frequency domains. Depending on the range set for the time delay, the human ear is more sensitive to either one of the aspects. For larger τ values, an echo distinct from the direct signal can be heard. The relative closeness of the frequencies that the comb amplifies causes the spectral effect to be barely identifiable. The time events can no longer be segregated by the human ear for smaller τ values, but the spectral effect of the comb can be noticed.

IIR comb filter

The Infinite Impulse Response (IIR) comb filter (Fig. 2.11) produces an endless series of responses $y(n)$ to an input $x(n)$. The input signal circulates in a delay line and with each circulation, the signal is attenuated by g. The line is fed back to the input. In many cases, it may be crucial for the input signal to be scaled by c to compensate for the structure-produced high amplification.

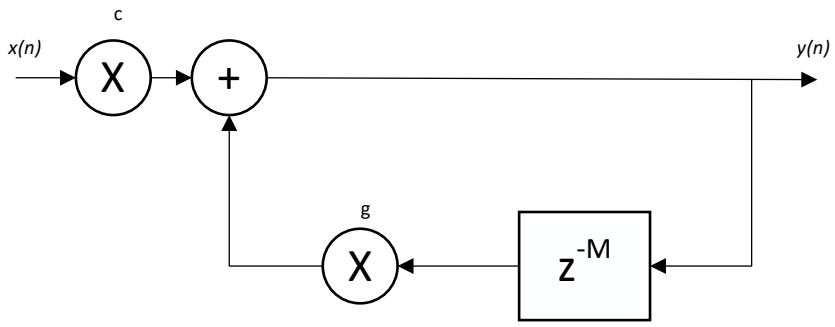


Fig. 2.11. Block diagram of an IIR comb filter [150]

The difference equation and the transfer function are given by:

$$y(n) = cx(n) + gy(n - M) \tag{2.18}$$

$$\text{with } M = \frac{\tau}{f_s} \tag{2.19}$$

$$H(z) = c/(1 - gz^{-M}) \tag{2.20}$$

The response time of this filter is infinite due to the feedback loop. Following each time delay τ, a copy of the input signal with a g^p amplitude will come out, where p notes the number of cycles the signal has cycled through the delay line. The signal does not grow, meaning that $|g| \leq 1$ is the stability condition. The gain varies between $1/(1-g)$ and $1/(1+g)$. The frequencies affected by the IIR or FIR comb filters are similar. However, in the IIR comb filter, the gain grows very high, and as $|g|$ comes closer to 1, the frequency peaks get narrower.

2.5.2. Reverb

The first efforts to create electronic devices that could simulate the effects of sound propagation in enclosed spaces were taken up in the second half of the twentieth century. Although many pioneers of artificial reverberation have impacted the field throughout the years, the most important work has been done by Manfred Schroeder, who operated at the Bell Laboratories in the early sixties [104-108]. The most groundbreaking innovations from Schroeder were recursive comb filters and delay-based all-pass filters introduced as computational structures, which are suitable for simulating complex patterns of echoes in an inexpensive way. The all-pass filter based on the recursive delay line has the form:

$$y(n) = -gx(n) + x(n - m) + gy(n - m) \tag{2.21}$$

where m is the length of the delay in samples. Presented in Fig. 2.12 is the filter structure where $A(z)$ is most often replaced by the delay line.

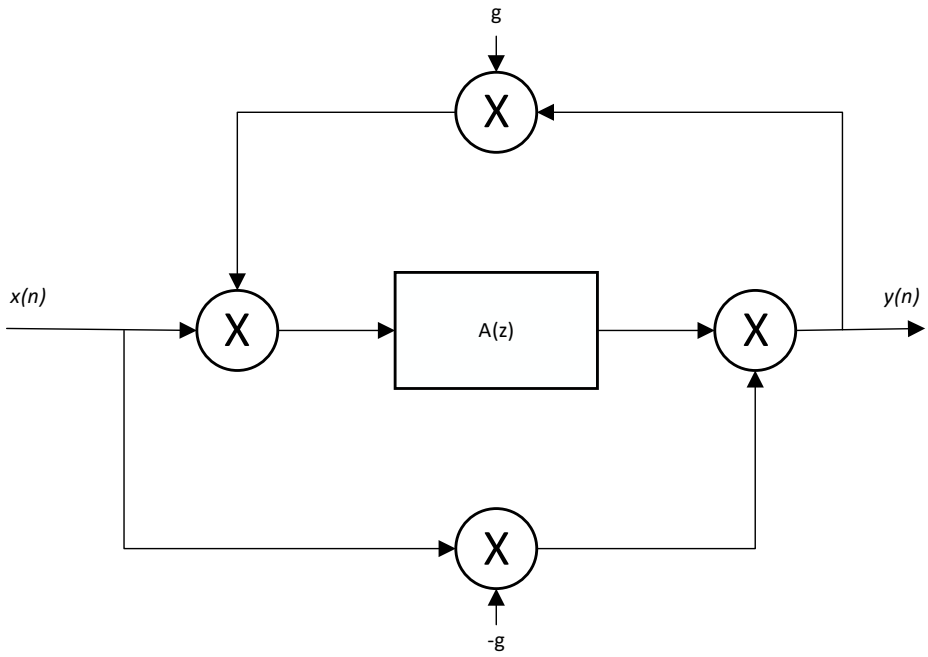


Fig. 2.12. The all-pass filter structure [150]

The filter shown above (Fig. 2.12) allows obtaining a flat frequency response and a dense impulse response. This structure is, to this day, utilized in almost every artificial reverberator as a standard component [75]. All-pass filters are often assumed not to introduce coloration to the input sound. This assumption can be true from a perceptual point of view only when the delay line is about 50 ms [152]. Otherwise, the incoming signal's timbre is affected in a significant way and the time-domain effects become more relevant.

The single-input single-output all-pass filter was generalized to a multi-input multi-output structure, where an order-N unitary network replaced the delay line of m samples by Michael Gerzon [32]. Parallel connections of delay lines or all-pass filters and orthogonal matrices are examples of trivial unitary networks. The generalization increases the impulse response's complexity without any appreciable coloration being introduced in the frequency. Gerzon's generalization suggests that all-pass filters can be nested within all-pass structures. This telescopic embedding is realizable with at least one delay element in the block $A(z)$ of Fig. 2.16 and is equivalent to lattice all-pass structures [30].

During the process of professional audio production, natural reverb is a vital tool. A correctly recreated natural acoustic space adds depth and authenticity to digitally recorded sources. In cases when the recorded instrument (group of instruments) is placed in a room with small reverberation, the artificial reverb expands the space. Reverb consists of many very closely spaced reflections from all surfaces of the room. The larger the room, the longer the reverb time (RT60). One can listen to only the reverb and figure out the size, density, and nature of a given space. In general, a room impulse response (reverb) can be separated into three parts (Fig. 2.13):

- Direct signal.
- Early reflections.

- Reverberation.

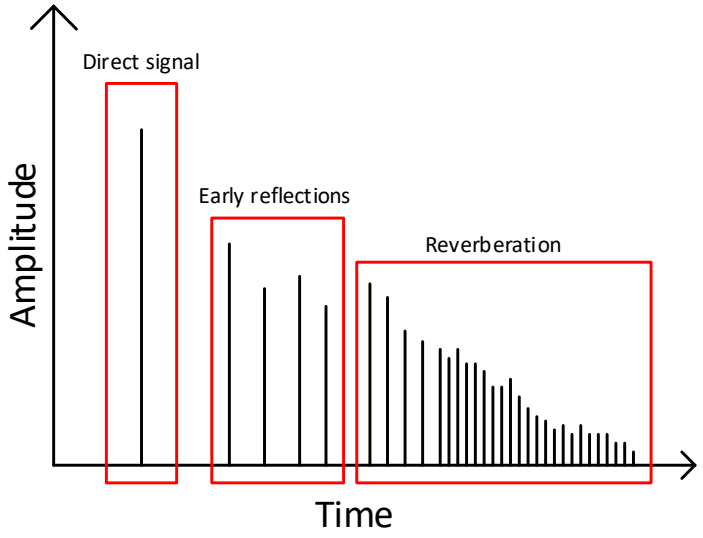


Fig. 2.13. Components of a room impulse response

A direct signal is a signal which travels directly from the source to the listener through the shortest line. The first few signals (except for the direct signal) that reach the listener right after reflecting are called early reflections. The reflecting surfaces are usually the floor, the ceiling, or a particular surface (depending on the circumstances). In general, those reflections allow people to categorize spaces in the context of size or depth subconsciously. The last set of reflections is usually the longest, and it is what creates the reverberation characteristic of a room. This set consists of all possible reflections from all surfaces in all possible angles from the source to the listener in the entire room (space). Those reflections are spaced so densely that humans cannot identify singular reflections. This leads to the brain recognizing the reflections as a one, slowly fading signal.

Changing the settings of parameters allows a digital reverb to be used to simulate various acoustic spaces. Some popular reverb categories are:

- Hall simulates concert hall acoustics; often a dispersed and rich setting with a longer reverb time (RT60).
- Chamber is used to simulate live echo chamber acoustics. Chamber effect settings often simulate the bright reflectivity of surfaces such as cement or tile.
- Room settings aim to simulate the distinctive acoustics of mid- to large-sized rooms. These settings are most appropriate to use for an intimate solo instrument or to achieve sound of a chamber atmosphere.
- Live is used for live stage performance simulation. The Live settings differ, but in most cases, they simulate long early-delay reflections.
- Spring is a simulation of the spring reverb effect.
- Plate effects simulate metallic plate reverb devices and are characterized by their bright diffuse; mostly used for vocal tracks and percussive instruments.

- Reverse effects are achieved by reversing the envelope of the decay trail. The gradual level increase of decay ends with a quick cut-off at the end trail.
- Gate effect cuts off a reverb signal's decay trail and is used to emphasize drums and percussion instruments.

2.6. Audio normalization

Filters are expected to affect the signal frequency content primarily, as they are, more often than not, designed in the frequency domain. The modification of the signal's loudness level is a side effect that should not be overlooked. While the desired effect might be achieved by the filter, the sound may become too strong or too weak for the result to be useful. As an example, an input signal (pink noise), its spectrogram and initial level (top), and the same signal with a high pass filter (1000 Hz, 6db/octave, q=1), its spectrogram and final level (bottom) are presented in Fig. 2.14. Without normalization, did not only the frequency domain change but also the output level of the signal (showing a 5.40 dB RMS difference).

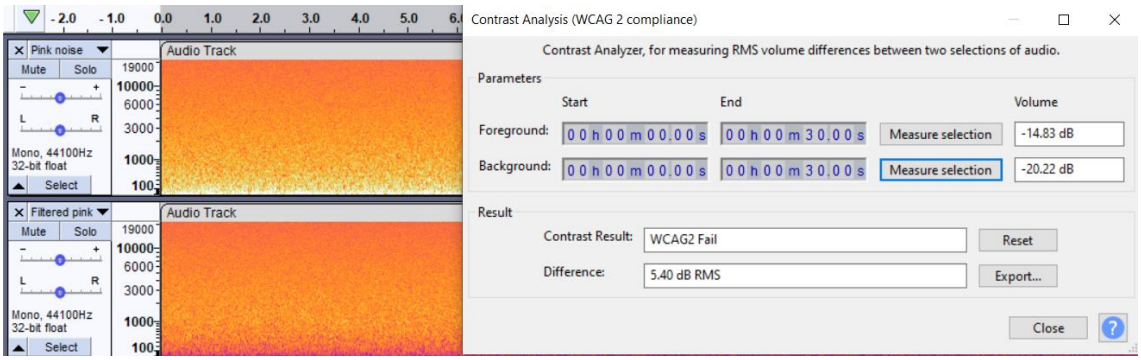


Fig. 2.14. Pink noise with the RMS level (top) and filtered pink noise with the RMS level (bottom)

Normalization is the method of compensating for these variations in amplitude and is performed by scaling the filter:

$$L_p = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})d\omega|^p\right)^{\frac{1}{p}} = 1 \tag{2.22}$$

where typically L_2 or $L_\infty = max(|H(e^{j\omega}|)$ are used [150]. The L_2 norm is accurate for broadband signals and is employed so that the signal's loudness is normalized. To avoid the filter overloading and to normalize the maximum of the frequency response, L_∞ is used. Whether the normalization scheme is suitable or not can be determined if the filter is accepted by musicians. The filter can be rejected because of its difficulty in operating when the normalization is wrong.

Several effective implementation schemes are proposed in [24], where the normalization of the state variable filter has been studied. In practice, the first-order lowpass filter which processes the input signal performs normalization if the f_c frequency and amplitude correction in $\sqrt{\varsigma}$ can be normalized in ς (Fig. 2.15). This normalization scheme makes it possible to operate the filter with damping factors as low as 10^{-4} while the filter gain reaches approximately 74 dB at f_c.

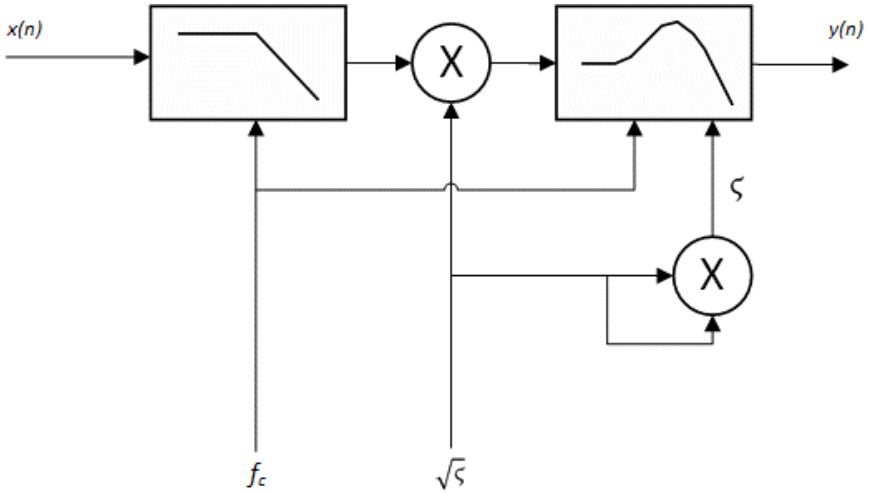


Fig. 2.15. L_2-normalization in f_c and ς for the state variable filter

To set a desired level of the recording amplitude, a gain can be applied to the audio signal, which is an alternative perspective of audio normalization. To enable the Signal-to-Noise Ratio (SNR) and the relative dynamics to stay the same as they were before the operation, the same amount of gain is applied to the entire recording. This is a process that can be performed in any DAW. The two types of audio normalization are peak and loudness normalization.

In peak normalization, the gain is set based on the highest PCM sample value. The difference between the abovementioned value and the normalization level (usually 0 dBFS) is calculated and applied to the whole signal. The 0 dBFS level is the most common level as it is the loudest level allowed in a digital system without introducing extra distortion to the signal [129]. The apparent loudness of the content is not accounted for by peak normalization alone, as it searches only for the highest level. Therefore, peak normalization is used to change the volume so that, during the mastering process of a digital recording, it ensures the available dynamic range for optimal use. However, a loudness advantage over material that is non-peak normalized can be achieved when peak normalization is combined with limiting/compression. The so-called *loudness war* was the side effect of excessive use of limiting and peak normalization.

Another type of normalization is based on calculating the loudness level where the output level is set to the target level of mean amplitude. The mean can be a regular measurement of the average power in the signal, as it is in the case of the RMS level, or it can be a measurement of the human-perceived loudness, such as EBU R128 [129]. For example, the reference level for the platform YouTube is -14 LUFS, so if the input signal is analyzed and its level is -10 LUFS, YouTube's algorithm lowers the level by 4 dB. Loudness normalization was created to combat the everchanging loudness level of recordings played consecutively. It is not a major issue when one CD album of a given artist is being listened to, but it increases in importance when listening to a playlist on, for example, a platform like Spotify. Without such normalization, one song in the playlist could be much quieter than the rest, which would require the user to change the playback level multiple times [47]. Depending on the dynamic scope of a

song and the target level, the normalization result can cause the peaks in a recording to go over a 0 dBFS scale (in the digital world). Software that offers such normalization usually has the option to use a compressor or limiter to avoid clipping. In this case, SNR and the dynamic scope in a song would change.

2.7. Mastering

Although mastering and mixing are separate processes, it is worthy of mention. Mastering is the last step of the music production process and means preparing audio files to be recorded on a compact disc, a vinyl record, or any other medium. The mastering process can be conducted on a single song and on a full album. Good mastering should make a song sound much better than it did before the process. When mastering a full album, the main goal is to produce a uniform, cohesive and full sound throughout all the recordings. Mastering is performed on a previously mixed file, which means that the mastering engineer does not have direct access to the individual channels (instruments) of the song.

When analog devices are used, a mastering engineer transforms the material from the tape to a vinyl record mold. The engineer's tasks involve matching the sound of the audio material to the physical capabilities of the medium, usually by applying appropriate corrections. Mastering processes are conducted in rooms with better acoustic properties than the rooms used for mixing.

The process of mastering includes:

- Equalization.
- Dynamic range control.
- Noise reduction.
- Setting the width of the stereo base.
- Sample rate conversion.
- Adding ISRC codes.
- Dithering.

3. SELECTED METHODS APPLIED TO AUTOMATIC MIXING APPROACH – RELATED WORK

In the following Sections, methods of audio mixing automatization are presented. The methods are divided into traditional, knowledge-based, technology-based, and deep learning approaches.

The Merriam-Webster dictionary defines “automatic” as “largely or wholly involuntary.” [67]. In the automatic mixing approach, the author will describe techniques and methods that help achieve the end result without or with the least possible amount of user interference.

3.1. Traditional approach to automatic audio mixing

Traditional methods of automatic mixing regard mainly the easier tasks, such as setting the maximum level of the microphone in a live situation in a way that does not allow for the feedback of the system or distortion of the speakers. The more manageable tasks which the methods can perform are also automatic mixing of audio elements in cases where it is not the most important aspect (for example, in video games) or even in audio/music branding (for instance, in stores) where the songs are automatically mixed together one after the other. In the last case, the mixing happens not in the context of multiple tracks in one song but rather in entire songs where the previous song is smoothly mixed with the next one.

Back in 2008, Gonzalez and Reiss proposed automatic level normalization of a system that changed linearly based on a mathematical model [34]. The proposed normalization technique worked in real-time, but unfortunately, the final level was always set at maximum before the feedback threshold; thus, it was only usable in live situations.

Kolasinski [51] proposed a framework that used the Euclidean distance between modified Spectral Histograms to calculate the distance between a mix and a target sound and applied a genetic optimization algorithm to figure out the best coefficients for that mix. The system, however, resulted in good quality for only four different tracks in the mix.

Next, Gonzalez et al. [36] developed a method that aimed to achieve optimal mixing levels by optimizing the ratios between the loudness of each individual input channel and the overall loudness contained in a stereo mix. This method lacked taking into account psychoacoustic bases of loudness; moreover, subjective tests were not performed. Later in 2012, Mansbridge et al. [65], based on previous findings, developed an autonomous multitrack fader control. Mixing levels were determined using the EBU R-128 loudness measure.

In 2009, Terell et al. [130] developed a framework used to optimize the monitor mix (level-wise) experienced by each musician. This approach was purely live-situation related; thus, it contains a lot of extra variables such as room dependence, in-air level, the response of the speakers, etc. Terell continued his work [131] by introducing mixing in different locations within the performance stage, and the algorithm used was proven to perform better than the brute-force approach.

Scott et al. [110-111] proposed a model for determining the relative gain levels of each track. Still, in reality, “each track” was a stem built from individual tracks, e.g., drums were

treated as one track, but in fact, drums are often recorded with at least four microphones, thus four different tracks. Authors trained their system to predict the time-varying parameters that produce a perceptually coherent mixture of unknown source content using minimal prior information. The results were promising, so naturally, machine-learning-based systems were gaining popularity.

Ward et al. [140] proposed multitrack (stems) recording mixing based on the concept that all instrument loudness should be equal. The idea was rather elementary and involved dividing each track into frames (15-25 seconds long) and finding desired gains on each track to achieve overall loudness.

Another level-matching solution introduced by Terrel et al. [132] treated the mixing problem as a numerical optimization dilemma. The authors found that the optimization-theory approach offers several advantages over the human process, but they did not achieve satisfying results. The masking issue was mentioned, but it was not fully resolved.

Wichern et al. [143] proposed three different automatic level-based mixing algorithms. The first one was based on a simple energy-based loudness model, the second one used a psychoacoustical model, and the third one incorporated masking effects into a psychoacoustical model. The authors conducted MUSHRA subjective listening tests, and − surprisingly − listeners preferred the simple energy-based loudness model. However, it was discussed that listening conditions were not ideal, and other processing steps were omitted.

Wilson et al. [144] presented a more abstract approach by allowing the efficient generation of random mixes. The genetic algorithm was used; it learned how to set inter-channels volume ratios from the expertise of the user. This approach, however, needed an expert/experts and a lot of training.

Gonzalez et al. [35] presented a simple cross-adaptive method for averaging perceptual loudness on all frequencies amongst a multitrack recording. Five first-order filters with a flat frequency response were used in the system. A set of eight-channel live recordings, as well as the white noise, were used for testing. Results − produced using a five-band spectral decomposition implementation − indicated that the Fourier-based spectral decomposition, together with a corresponding Fourier-based equalizer, could dramatically improve results. No subjective tests were conducted.

Ma et al. [63] introduced a new approach for automatically equalizing an audio signal towards a target frequency spectrum. Equalization (eq) curves were extracted and used as a reference for spectral balance. Matching spectral distribution of an input signal to the target curve was based on the Yule-Walker algorithm. The objective evaluation showed that the algorithm was capable of matching the eq curve to the target, but the experiment lacked subjective evaluation.

Hafezi [37] implemented an autonomous multitrack equalization system for reducing masking in multitrack audio that works both offline and in real-time. The system was tested both subjectively and objectively, and the results showed slight changes in the amount of masking. The authors prepared four different models (two offline and two online) presented to the

listeners. The so-called “Offline Semi” and “Real-time Unconstrained” models achieved poor results, and subjective evaluation did not confirm the reduction of masking. However, the other two models achieved not only better results than RAW (anchor) material but also better than the “Amateur” mix. The authors admitted that the implementations left a lot of room for improvement, but this experiment showed a clear path for the future, that is – toward the unmasking problem.

While in the mixing area, the topic of automatic equalization is very broad, it is easier to look from the perspective of a mastering engineer; i.e., one source instead of many is easier to deal with. Mimilakis et al. [70] introduced a novel method of fundamental frequency tracking. The pitch tracking subsystem was presented as follows: a copy of an unmastered audio material was summed up into mono, then low- and high-pass filters were applied, as well as envelope processing; finally, pitch estimates were found using frequency demodulation based on a third-order phase-lock loop. The same equalization filters were used on both channels. The authors performed a series of subjective and objective tests where the enhanced audio material was preferred within the range of 79%.

All of the above experiments and methods were designed to solve a level-, eq- or reverb-matching problem during either live or offline mixing situations. None of the above algorithms were matched with the genre or mood of the desired mix, e.g., different levels should be set differently for a specific genre. All “separate” tracks used were, in fact, combined stems of individual instruments. Almost all of them were based on either simple loudness comparison between channels or incorporated psychoacoustical elements into it.

3.2. Knowledge-based audio mixing

Knowledge-based audio mixing can be described as a kind of departure from the standard automation methods described in the previous Chapter. Many of those methods use much larger databases that are applied for training the models. The methods themselves are utilized to change multiple parameters (e.g., level, panning, and equalization) at the same time. The most commonly used databases are Open Multitrack Testbed [18] and MUSDB-18 [89]. The methods cited below use expert knowledge during training or creating a specific model or application.

Chourdakis et al. [15] proposed an adaptive digital audio effect (artificial reverb) that learns from the user in a supervised way. The publication listed the features of the audio files, and these features were spatially reduced for training purposes. Additionally, the user should provide examples of reverb parameters for the database to train. Next, a group of classifiers was trained and compared using 10-fold cross-validation to compare the success ratios and MSE (Mean Square Error). The Open Multitrack database was used for training the model. In addition, this research was presented in work by Chourdakis et al. [16], where a new design of the same effect was proposed. The new design is controlled directly by the desired reverberation characteristics. The learning was also conducted in a supervised way, and the same database was used. The trained models are able to replicate a “characterized” reverb

from one track to another. The models were evaluated using the F1-score-based classifier, mean squared errors (MSE), and multi-stimulus listening tests.

Ramirez et al. [91] proposed analyzing low-level features to mix individual audio tracks to stems automatically. For this purpose, 1812 features were extracted from guitar, bass, vocal, and keys tracks. Random forest classifiers were used to find the features that were most distorted by the mixing process. On this basis, the authors trained various multi-output regression models. Next, the number of features that could be used for such a transformation was reduced. Mapping the selected audio features described the characteristics of changes taking place in individual tracks after mixing them into stems. The authors did not conduct subjective tests supporting this research hypothesis.

Wilson et al. [145] presented an innovative method of “mix-space,” i.e., a parameter space that contains all possible mixes that use a finite set of tools and parametric methods of creating artificial mixes (in this space.) In their work, mixes that only took into account level, panorama, and equalization changes were used. The authors applied statistical methods such as Monte Carlo and the population-based optimization method to investigate the accuracy of tempo-estimation algorithms and examined the distribution of spectral centroid values in all mixes.

Everardo in [26] utilized Answer Set Programming (ASP) to make an audio mix. A kind of dictionary was created with rules for audio engineers to follow, proposed by professional music producers and audio engineers. As a result of using the complete dictionary, the program was capable of returning a mixed file or output − a kind of a mix plan in human-readable format, which should be used as a starting point. During subjective testing, the listeners were asked a few basic questions, such as whether all the instruments used in the mix were audible? Unfortunately, there was no comparison made between these mixes and professional mixes, neither subjectively nor objectively.

Moffat et al. [74] developed a web-based tool that utilized a logical constraint solver to apply real-time rules to audio tracks. The system uses OWL (Web Ontology Language) reasoning inference on sets of mixing rules to determine which subset should be handled. The Rule Interchange Format (RIF) was proposed for presenting the rules. It was required that such mix rules could be transferred between different systems (and shared online). Rules can be created by an expert, learned from existing datasets, or a mixture of both.

Ronan et al. [99] proposed a multitrack masking metric derived from the MPEG psychoacoustic model. Various sound processing techniques were examined to manipulate the frequency and the dynamics of the signal in order to reduce masking based on the proposed metric. The authors examined whether automatic mixing with the use of subgroups is beneficial to the perceived quality and clarity of the mix. The results suggested that using subgrouping during automatic mixing improved the perceived mix clarity and quality. Also, the results indicated that the proposed masking metric (during automatic mixing) could be used to reduce inter-channel masking.

All of the above experiments used expert knowledge during training or the automatic mixing process. Few of them have been tested by conducting extensive objective and subjective tests. However, the abovementioned research proves that the bottom-up approach to automatic mixing (the tracks are mixed to stems, and the stems to the finished mix) is correct and provides good results during the automated mixing task. To properly train the aforementioned models, feature sets are used, either calculated or collected from existing databases.

3.3. Technology-based automatic audio mixing

The productization of technology and giving it user-friendly interfaces influence the growth of technology and allow for more advanced automatic sound manipulation. Plugins available on the market are digital audio processors that can not only be the digital equivalents (simulations) of analog devices but also can exceed traditional boundaries. One plugin can act as a substitute for a few ones or even a few dozens of analog devices. Moreover, modern plugins are actively helping the user and executing tasks that would be unachievable otherwise.

The first group of advanced plugins consists of plugins that are capable to auto-equalize sound (called matchEQ or autoEQ). In matchEQ, a reference is needed to produce a set of appropriate filters as the output so that the plugins can match the frequency response of the input file to the reference as closely as possible. The user can provide the reference (in plugins such as Izotope Ozone 9 Equalizer [45] or FabFilter Pro Q3 [27]) or choose from available presets supplied by the software creator. Another approach to auto-equalizing sound, autoEQ, is the automatic creation of equalization presets. The user inputs any audio signal into the plugin, and the plugin creates a custom preset while analyzing the spectral content of the signal. The preset allows the output audio to sound better (for example, the signal is corrected for tonal imbalances or devoided of any undesired resonance effects) or have the mood of the sound resembling some previously prepared styles (e.g., normal, speech, aggressive, etc.). Examples of plugins that can perform this task are Izotope Neutron Pro [44], Sonible smart EQ+ [119], Soundtheory Gullfoss [123], or oeksound Soothe 2 [79].

Another group of plugins is capable of performing the task of compression. Usually, the compression process consists of two steps: learn and apply. First, the user provides any audio signal as the input to the plugin. Next, the plugin, while “listening” to the signal, automatically adapts its settings to create an automatic preset for the given audio signal. In most cases, the user is able to choose the style in which the compression would be applied (applying the so-called profile), for example: standard, drums, kick, snare, bass, guitar, keys, vocal female, and vocal male. Some plugins have additional capabilities. For example, the Sonible smart:comp [120] plugin offers spectral compression, which acts as an intelligent ultra-high-resolution multiband compressor that dynamically smooths out tonal imbalances. When active, smart:comp applies compression only where its built-in artificial intelligence thinks it is needed. Other plugins with additional capabilities offer to adapt the compression level to the music genre and automatic classification of instruments. The plugins provide not only the basic compression algorithms but also multiband compression, noise gate, de-essing, or even limiting.

The following step during the mixing process is commonly creating the sound of individual tracks by using saturation or transient shapers. The Neutron Pro [44] plugin can, in an automatic way, adjust its settings based on an excerpt of a signal provided by the user (in two analogous steps: learn and apply). Saturation (this block in the plugin is called exciter) consists of three “colors” (full, defined, clean) and four types of saturation (tube, warm, tape, and retro). On the other hand, some plugins offer transient shapers – an extremely effective tool when producing music. One of the primary uses for these plugins is applying them to drum and percussion elements. The transient shapers can be useful, for example, when trying to make drums cut through a mix and stand out by increasing the attack. Decreasing the attack, however, can reduce the start of the transient, helping the applied sound blend into the background. The sustain or release section in these plugins can give a more sustained volume to the end of the sound or, conversely, shorten the sound making it “snappier.”

Izotope plugins can simultaneously classify instruments and, based on that (and a short fragment of the signal input by the user), create a custom preset that consists of multiple elements. In Table 3.1, there are companies presented that produce plugins. This is shown along with the functionalities they offer – more precisely, the elements that can be mixed automatically.

Table 3.1. List of companies producing plugins that allow for automatic mixing with a breakdown of their capabilities

	Auto-detect instruments	Balance	Equalization	Dynamic range control	Saturation	Transient shaping
Izotope [44-45]	X	X	X	X	X	X
FabFilter [27]		X	X			
Sonible [119-120]			X	X		
Soundtheory [123]			X			
Oeksound [79]			X			

Although plugin producers do not publicly disclose any information on exactly how their plugins work, with the most recent state-of-the-art knowledge, one can assume with great probability that all methods of automatic preset creation − based only on the audio signal provided by the user − are made by appropriate training of models on a vast database and consequently utilize those models to create new presets. The current state-of-the-art methods that use Deep Learning are presented in the following Chapter along with an explanation of how the models used later, e.g., in the abovementioned plugins, are created.

3.4. Deep Learning approach

An Artificial Neural Network (ANN) works to imitate the human brain in virtual reality. The definition of an artificial neural network describes it as a group of elements – simple neurons that process input data. The communication between individual neurons happens in parallel. Each neuron has its own weight [38][52][127].

An artificial neuron consists of $(x_1 - x_n)$ inputs, which correspond to synapses in a biological neuron model. The main goal of an input is to collect data and transfer it to the kernel of the neuron, where the signal is subjected to $(w_0 - w_n)$ weighing. Next, the signal is processed by the activation function $(f(e))$, where the processing of input information into output information happens according to the used mathematic formula. The most commonly used function is the sigmoidal function, which is presented in Eq. (3.1) [40].

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3.1}$$

The simplest ANN model consists of one neuron. In practice, however, such a construction is not used due to the low entropy of such a model [40]. Thus, neurons are connected into multidimensional networks, which amplify their ability to process data. The number of layers and neurons is virtually unlimited. However, performance can be an issue when building large neural networks. The human brain consists of approximately 100 billion neurons. Such a size of a neural network is impossible to achieve in practice, and the development in this field suggests that it will not be achieved any time soon. There are reports suggesting that the currently conducted research aims to replicate the brain of a mouse in a digital form [128].

Neural networks can be divided into:

- Feedforward networks, where information passes smoothly from one end of the network to the other [114][139][151].
- Recurrent networks, where information can return into previous layers [21][87][95].
- Unsupervised networks, where the network functions without the supervision of its results [109][117].
- Supervised networks, where a dedicated block exists to control the way the network functions and the quality of the results [8][146][149].

Similar to other decisive algorithms, a key step that impacts the method's success (i.e., high precision of the results) is the training step [148]. In artificial neural networks, the learning process consists of assigning weights for neurons and inputs used in the network. In the case of supervised networks, there is a block called a "critic," which determines parametrically if an improvement of the classification effectiveness occurred. For better results, it is necessary to provide extensive training datasets for the network's input. Training a model on very large databases (e.g., 100 000 images) is called deep learning [40][77][128]. Recently, research of this type has risen in popularity thanks to increased resources for training and the constant growth of computing power. A significant disadvantage of such solutions is the length of the model training process, which can take multiple weeks for large testing datasets. The use of extensive training bases prevents overtraining, which is harmful to ANN.

As De Man stated in his work [19], the problem of mixing is multidimensional. Engineers must decide whether the source is too loud or too quiet, the frequency range is set correctly, the panning of an instrument complements the whole mix, the reverb is fixed correctly, etc. This said, the various types of processing cannot be done separately; instead, this challenge should be to set an all-in-one task. Isolating one problem will lead to another unresolved issue.

According to new research [6], [15], [16], [69], [71], and deep learning [144] is gaining such popularity, and its unmatched ability to connect the puzzle pieces can be a way to determine whether artificial intelligence is capable of mixing a song in a fully automated way.

Reiss et al. proposed a deep learning solution for the task of equalization [98]. Previously known automatic equalization systems (matchEQ task) required the implementation of the creation logic and types of filters. Additionally, the previous methods were limited by the need to apply filter banks in the system. The authors proposed a new end-to-end architecture based on Convolutional Neural Networks. Thus, the creator of the network does not need to know the transfer function to perform equalization matching. The network learns by itself how to process the audio signal to bring its frequency response to a given target. The authors presented the effectiveness of the network during the equalization matching operation for shelving, peaking, lowpass and highpass IIR, and FIR filters. Therefore, they trained four models (one for each EQ task) via supervised and unsupervised learning procedures. The trained models achieved loss values < 0.333 for all EQ tasks. The proposed models consisted of adaptive front-end, latent-space Deep Neural Network, and synthesis back-end.

Moffat et al. in [73] proposed a pure machine-learning approach to level (gain) mixing of audio drum tracks. The authors noted that there were many approaches to automatically mixing the levels of individual tracks, however, the machine learning approach was missing. The authors' hypothesis was that the lack of such research was caused by the lack of available data that are needed to train the models. The authors used the random forest approach to conduct multiple outputs predictions. These outputs are the levels of values that need to be applied to a given individual track to achieve the desired mix. Finally, the authors compared their results with pre-existing algorithms and "man-made" mixes. Objective and subjective tests have shown that with the proper database, it is possible to train a model that could produce mixes comparable to those made by humans.

The previous solutions were characterized by approaching one problem related to mixing (matchEQ in [98], level in [73]). Steinmetz et al. [125] proposed for the first time to perform automatic multitrack mixing, which involved more than setting the levels. The authors trained a model that produces human-readable mixing parameters, which enable the user to adjust them later manually. The proposed Differentiable Mixing Console (DMC) was trained on a limited and unstructured dataset, and the entire solution was capable of implementing real-world mixes. During the mixing process, the network had at its disposal such parameters as gain, polarity inversion (if needed), 5-band equalizer, compressor, reverb, level fader, and panning knob. The entire system was made of pre-trained subnetworks, weight sharing, and sum/difference stereo loss function. During the evaluation, it was found that the solution worked very well when performing tasks of mixing drums to a stem, while when mixing whole songs, DMC performed better than the considered baselines.

Ramirez et al. in [94] proposed an end-to-end Deep Neural Network based on the Wave-U-Net autoencoder to perform automatic mixing of drums. The authors used a network that was originally intended for audio separation. The ENST drum dataset [33], which is divided

into two groups (“dry” and “wet”), was used to train the models. The first of the groups includes only the level and panorama changes. In the second group, additional effects such as equalization, compression, reverb, and dynamic range control (during the mastering process) were used. Therefore, the authors trained two models (“dry” and “wet” respectively), which were then tested subjectively. They concluded the work by proving that the mixes generated by their model are indistinguishable from mixes prepared professionally by a human, at the same time being much better than the previous state-of-the-art methods. Using this architecture appeared to be the most beneficial in the task of automatic audio mixing.

The current state-of-the-art methods enabled the author of this dissertation to choose the system architecture for fully automatic mixing of audio files, i.e., Wave-U-Net one-dimensional autoencoder. An additional advantage of the chosen architecture is the provided wave-input and wave-output approach, making it a more user-friendly solution. Based on this architecture, a system was designed, which is described in Chapter 4. Moreover, experiments were conducted using the system, which were further described in Chapters 5 and 6.

4. AUTOMATIC AUDIO MIXING BASED ON WAVE-U-NET AUTOENCODER

4.1. System assumptions

Experiments are structured in a three-fold setup (see Fig. 4.1). It consists of designing and building the system (this will be described in Chapter 4.1). Then, preparation of separate tracks of recordings to be mixed and processed automatically is to be described in Chapter 4.2 (based on a custom database). Then, all models are trained and validated based on a deep learning algorithm with details contained in Chapter 4.3.

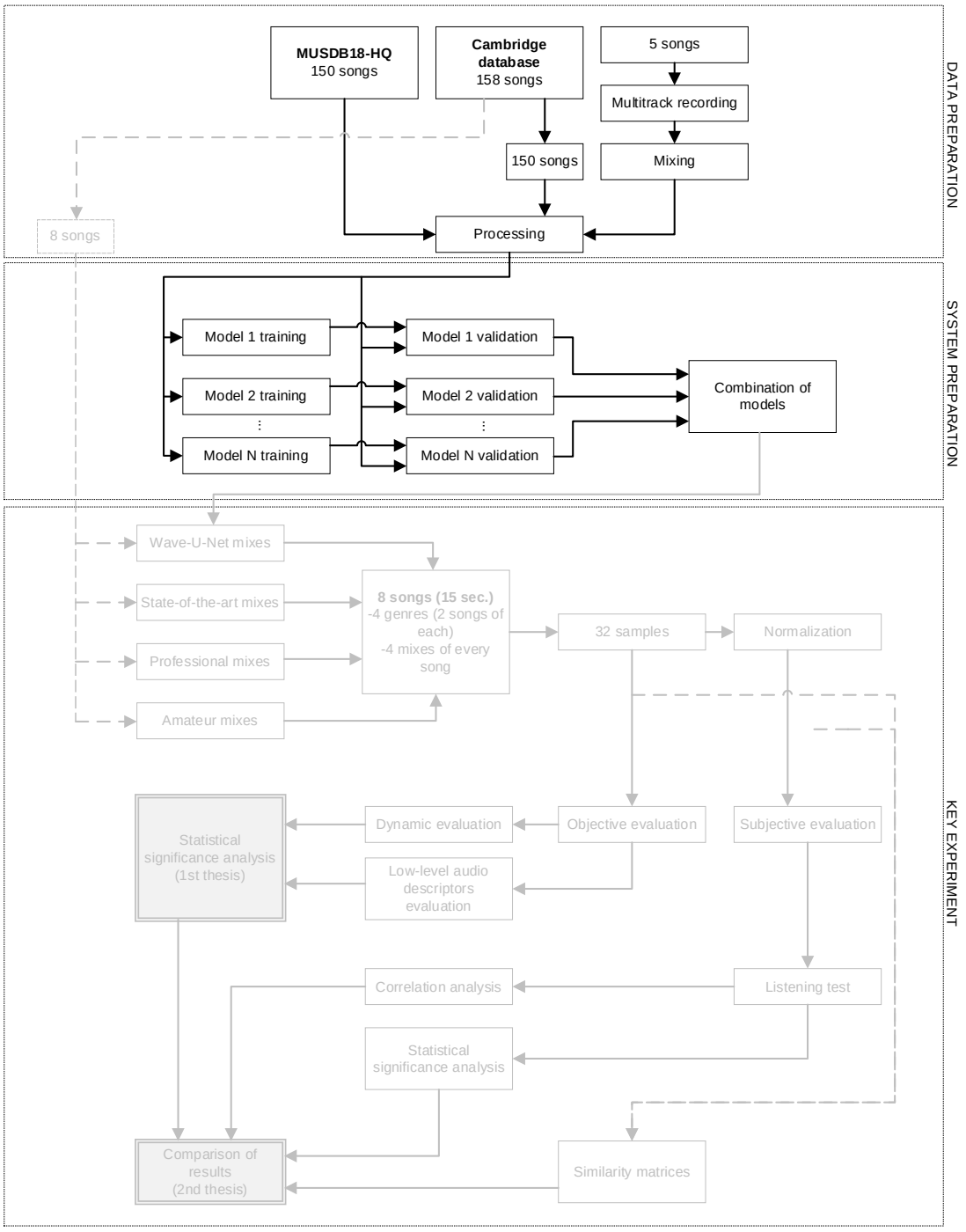


Fig. 4.1. Stages of analysis executed in Chapter 4

The design assumptions of the system were defined with particular consideration for three aspects. First, it was assumed that, based on the created system, the mixing of a song without any human interaction could be done. The system shall mix the song automatically, and the user inputs only the recorded signals in an appropriate format. The system shall not be a mixing assistant nor have parameters that could be accessed or adjustable by the user (it should act as a black box). Second, it was assumed that the system would be independent of the music genre. The user can input tracks from any genre of music and be given a finished mix as the output. Third, it was assumed that, inside the system, a bottom-up approach to mixing would be used – tracks from a given group of instruments will be mixed together into stems, and then the stems will be mixed into the final mix. Bottom-up mixing involves building a mix from the ground up, from single channels to buses, and then the final mix. Such an approach is considered more traditional [49]. After carefully studying various structures and utilizing the knowledge acquired from state-of-the-art articles (details contained in Chapter 3), it was decided to use a one-dimensional Wave-U-Net autoencoder architecture. The same architecture was used by Ramirez et al. [94] to mix drum recordings. Followed by subjective testing, it was proven that the mixes generated by their trained model were undistinguishable from human-made mixes.

4.1.1. System requirements

Listed below are the most critical requirements in the context of creating a system in a way that allows mixing a song by the user, who is not necessarily a professional, i.e., a person who has no previous experience in mixing. It was assumed that in the scope of performing functions directly related to the mixing of audio material, the user's involvement should be minimal, and the user shall have no influence on the operations of the system. Allowed is the possibility for interactions not directly related to mixing, for example, uploading audio signals to the system, to be performed using a keyboard and a mouse. It was assumed that the system should automatically export the finished mix to an output directory preselected by the user.

In the current state of the system, there is no possibility for it to perform on every computer or in a cloud/website. Listed below are the system requirements for performing the models training and the mix predictions (mixing). Although, in theory, the user does not have to train the models from scratch, a possibility to do so was assumed. It was also assumed that during the training process, the system would be supported by a CUDA®-enabled graphics card. In Table 4.1, the software system requirements are presented, and in Table 4.2, the hardware system requirements are shown.

Table 4.1. Software system requirements

Python	3.7 or later
pip	19 or later
Ubuntu	16.04 or later
macOS	10.12.6 (Sierra) or later (64-bit)
Windows	7 or later (64-bit)

Microsoft Visual C++ Redistributable for Visual Studio	2015, 2017 and 2019
NVIDIA® GPU drivers	450.80.02 or later
CUDA® Toolkit	N/A
cuDNN SDK	8.1.0 or later

Table 4.2. Hardware system requirements

NVIDIA® GPU card with CUDA® architectures	3.5, 5.0, 6.0, 7.0, 7.5, 8.0 or later
CPU with AVX Intel*	Sandy Bridge processors, Q1 2011 Sandy Bridge E processors, Q4 2011 Ivy Bridge processors, Q1 2012 Ivy Bridge E processors, Q3 2013 Haswell processors, Q2 2013 Haswell E processors, Q3 2014 Broadwell processors, Q4 2014 Skylake processors, Q3 2015 Broadwell E processors, Q2 2016 Kaby Lake processors, Q3 2016(ULV mobile)/Q1 2017(desktop/mobile) Skylake-X processors, Q2 2017 Coffee Lake processors, Q4 2017 Cannon Lake processors, Q2 2018 Whiskey Lake processors, Q3 2018 Cascade Lake processors, Q4 2018 Ice Lake processors, Q3 2019 Comet Lake processors (only Core and Xeon branded), Q3 2019 Tiger Lake (Core, Pentium and Celeron branded) processors, Q3 2020 Rocket Lake processors, Q1 2021 Alder Lake processors, 2021 Gracemont processors, 2021
CPU with AVX AMD	Jaguar-based processors and later Puma-based processors and later Bulldozer-based processors, Q4 2011 Piledriver-based processors, Q4 2012 Steamroller-based processors, Q1 2014 Excavator-based processors and later, 2015 Zen-based processors, Q1 2017 Zen+-based processors, Q2 2018 Zen 2-based processors, Q3 2019

	Zen 3 processors, Q4 2020

**Not all CPUs from the listed families support AVX. Generally, CPUs with the commercial denomination Core i3/i5/i7/i9 support them, whereas Pentium and Celeron CPUs do not.*

4.1.2. Components and architecture of the system

The entire system consists of five models. All models used in this dissertation were trained separately and connected to one system. The models differ by the number of inputs and outputs (mono/stereo). The system was created from variants of Wave-U-Net networks, suggested in [94][126]. Each individual model utilizes raw (unprocessed) audio input and output with connection to a series of downsampling and upsampling blocks that contain 1D convolution layers. The models also include resampling operations which allow the calculation of features used in the prediction process. A block diagram of the system is presented in Fig. 4.2.

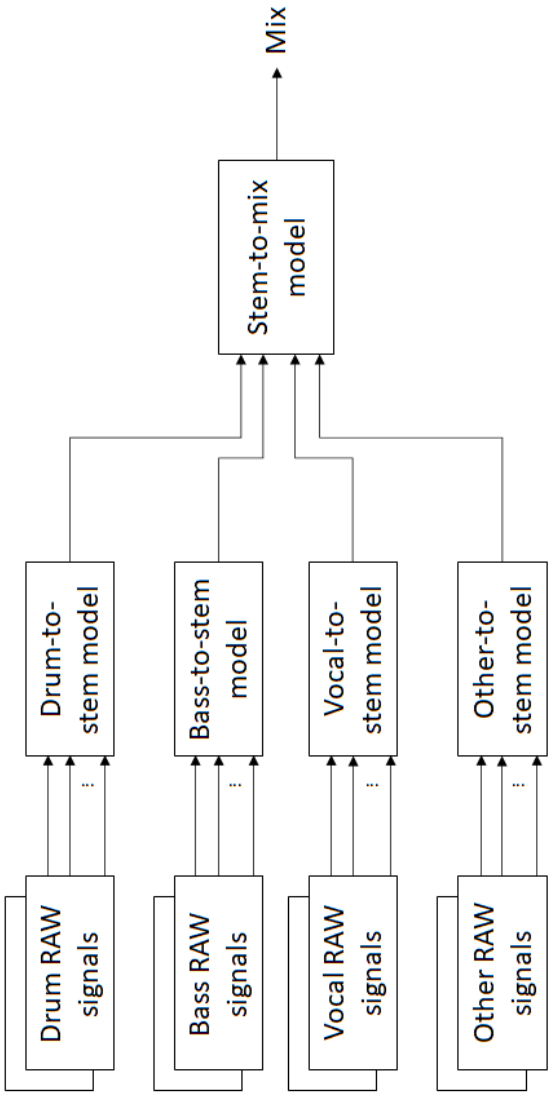


Fig. 4.2. Block diagram of an automatic audio mixing system

The Drum-to-stem, Bass-to-stem, Vocal-to-stem, and Other-to-stem can accept, respectively, up to 10, 4, 4, and 8 mono inputs (provided by the user) and one stereo output. The Stem-to-mix model has a stereo input and output. To achieve this effect, the code of the models was rewritten in a way that adjusts the number of inputs and outputs for each given task. In other words, each model accepts N inputs, and the adjustment of inputs is implemented as follows: if the user provides fewer inputs (than N), then the missing inputs will be automatically generated and they will contain no signal (only zeros), which will not affect the mix in any way. The described process was also a part of the training to ensure that each model will be able to create a proper mix in case of silence in inputs. This approach also gives extra robustness to the system – even if the user passes empty inputs, the model will not be destabilized. As already mentioned, the system works on a principle of a „black box" for the user. The user first provides recorded and synchronized tracks (that have the same length) as the input and receives a finished mix as the output. Therefore, there are no additional preprocessing blocks in the architecture. Also, it is assumed that the user provides signals with a sampling frequency of 44.1 kHz for the input material.

A single model is constructed from 10 layers for the training of the model to be more effective. A block diagram of a singular model is presented in Fig 4.3 (every model has an identical structure.)

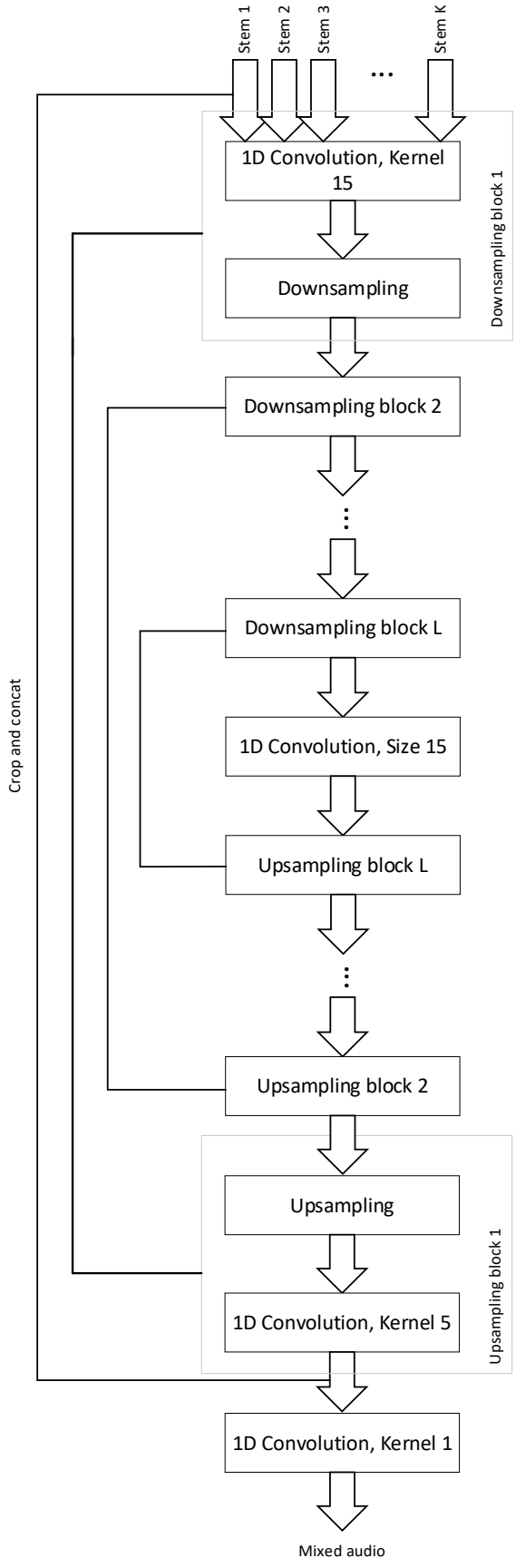


Fig. 4.3. Block diagram of the adapted Wave-U-Net network for automatic mixing K stems using L layers

The downsampling blocks perform one-dimensional convolutions of $Fc \cdot i$ filters (kernel) of $fd = 15$ size with the i layer is in the range of $[1; L]$, where Fc corresponds to the number of initial filters (number of convolutions in topology) and is equal to 24. The convolutional filter decreases the size of data, and samples become convoluted so effectively there are fewer samples. Convolution uses information from each sample. The goal of the decimate operation is to halve the time resolution. Before decimation, the feature map is concentrated with the cropped output of the respective downsampling block by the upsampling blocks. The upsampling blocks are performed with a factor of two by utilizing linear interpolation. These processes are followed by a one-dimensional convolution of $Fc \cdot i$ filters of $fu = 5$ size, where i is within the $[L; 1]$ range. All strides are of unit value and all of the convolutions are along the time dimension with no implicit padding. To ensure the outputs were between +1 and −1 at the time of the test, the outputs were clipped accordingly. LeakyReLU activation function was used, which accelerates the convergence of the training process in the classical framework of deep learning.

4.2. Data preparation for models training

To properly train an individual model, an adequate database is needed. The data should be structured, appropriately differing, and large enough. Databases for tasks in the speech domain, such as speech denoising or speech arrival direction detection, are commonly used. There are, however, very few databases that can be used for mixing purposes. Thus, based on MUSDB18-HQ [89], the best database available, a new database was built by the author, supplemented with individual tracks (from the Cambridge database), and expanded by additional songs prepared by the author. The database had to be prepared in a particular way to be helpful in model training and validation. The preparation process is described in this Chapter.

In the case of using machine learning or neural networks (NN), the pre-processing step may play a crucial role because, depending on the utilized network's topology and the type of problem that needs to be solved, the input may vary. The architecture may require input in the form of, for example, files no longer than 30 seconds, one or multichannel, a specific sampling frequency, or a spectrogram. Pre-processing may be performed in several ways: manually (by trimming audio files in any DAW and exporting them in specific sampling frequency and bit depth) or automatically (by scripting a series of tasks by using tools like, e.g., SoX [124], Matlab [66], or Python [88]). In this case, the preprocessing involved bringing all the input signals into the same sampling frequency of 44.1 kHz and a bit depth of 16 bits.

The MUSDB18-HQ database [89] and five songs recorded by the author were used to train the network. This database consists of 150 songs (approximately 10 hours in total) from various genres. 100 songs were used as a training set, 50 as a testing set. Drum, bass, vocals, and other instrument stems and finished mixes can be found in the database (Fig. 4.4). The database consists of songs from the Cambridge database [72], which means that, in order to acquire individual tracks, they had to be taken from the Cambridge database to be matched appropriately.

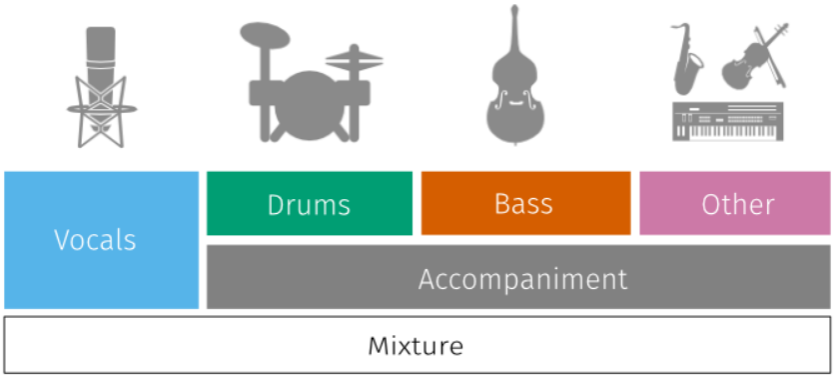


Fig. 4.4. MUSDB18-HQ database structure [89]

As already mentioned, five songs recorded and mixed exclusively by the author were added to the training database. All five songs were recorded in the Auditorium of the Electronics, Telecommunication and Informatics Faculty at the Gdansk University of Technology and in a home studio. The songs consist of drums, bass, guitars, and vocals, and their genre can be classified as rock. The drums were recorded using the multitrack technique listed in Table 4.3. Eleven microphones were used: two for the bass drum (Audix d6 and AKG d112 mkII), two for the snare drum (both being Shure sm57), two for toms (Audix d4 and Audix d6), one for hi-hat (AKG C414), one for the ride (AKG C414), two overheads (Audix SCX-25A) and a room mic (Cascade Fat Head II). The configuration is presented in Fig. 4.5.

Table 4.3. Drum set recording session input list. Particular parts of the set are listed along with used microphones

Instrument	**Instrument notes**	**Mic**	**Stereo configuration**	**Notes (e.g., polar pattern, placing)**
Kick in	Pearl Export kit, Joey Jordison Signature Snare, Zildjian custom cymbals	Audix d6	mono	*Cardioid; Apogee interface*
Kick out		AKG d112 mkII	mono	*Cardioid; Apogee interface*
Snare top		Shure sm57	mono	*Cardioid; Apogee interface*
Snare bottom		Shure sm57	mono	*Cardioid; Apogee interface, inverted phase*
Tom high		Audix d4	mono	*Cardioid; Apogee interface*
Tom low		Audix d6	mono	*Cardioid; Apogee interface*
Hi-hat		AKG C414	mono	*Cardioid; Apogee interface*
Ride		AKG C414	mono	*Cardioid; Apogee interface*
OH		Audix SCX25-A	XY	*Cardioid; Apogee interface, Looptrooter monster compressor*
Room		Cascade Fat Head II	mono	*Ribbon; Apogee interface, Placed in the middle of the room to catch the sound of the room*

Fig. 4.5. Drum set recording setup

The auditorium used for recording was built for lectures only, so the RT60 is approximately 0.6-0.8 s, which is very low in such a vast space (approximately 1150 m^3). The optimal area of the room was found for the drum set, and portable absorbers were used to cancel out the ringing caused by the auditorium piano strings from a piano placed there. All mic placements were standard, but the overhead microphones were placed high.

The guitars (Fender Telecaster as the electric guitar and Schecter Diamond Series as the bass guitar) were connected to the Native Instruments Komplete Audio 6 interface through the DI box, and the tone was set by using VST amplifiers. All recordings were made in a home studio.

All backing vocals, as well as the main vocal, were recorded in the home studio through the abovementioned interface. Due to a suboptimal acoustic situation, it was decided that a Shure SM7b microphone should be used. To compensate for the lack of preamplifiers, which caused problems of setting the appropriate level of the input signal while recording the backing vocals (more subtle than the main vocal), virtual preamplifiers in the DAW were used. All vocals were edited from multiple takes so that before mixing, there were only two tracks of the main vocal (the bridge of the song was on a separate track) and two backing vocals tracks (mostly doubles).

After recording, editing, and mixing, all channels were exported as separate tracks. This also concerned drums, bass, vocal, and other stems, as well as the full non-mastered mix.

Due to the nature of the system's architecture, it was decided to use a fixed number of inputs and outputs for each model. The number of inputs and outputs for the models is presented in Table 4.4. In cases where the number of signals was bigger than the assumed number of inputs, a premix was conducted. The premixing process consisted only of adding the signals together – there was no change applied to their loudness level and loudness in relation to each other, and no effects (such as eq, compression or reverb) were added. In cases where there were too few original signals (for example, there were only two signals for bass), empty tracks were created to meet the set requirement of the input number.

Table 4.4. Models and number of inputs and outputs

Model	Inputs	Outputs
Drum-to-stem	10 (mono)	1 (stereo)
Bass-to-stem	4 (mono)	1 (stereo)
Vocal-to-stem	4 (mono)	1 (stereo)
Other-to-stem	8 (mono)	1 (stereo)
Stem-to-mix	4 (stereo)	1 (stereo)

4.3. Models training and validation

As was mentioned above, the system consists of five models. Each model was trained separately and then connected to create the system. The training was performed using the L2 distance as training loss, as previous observations of neural models have shown that using this distance helps achieve better results [90][93]. The optimizer used was Adam, with a learning rate of 0.0001, decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Also, early stopping patience of 20 epochs was used, and a finetuning step followed. The initial learning rate was 10^{-4} and the batch size was 16. A model with the lowest loss for the validation subset was selected. The test loss function of Stem-to-mix model training is presented in Fig. 4.6.

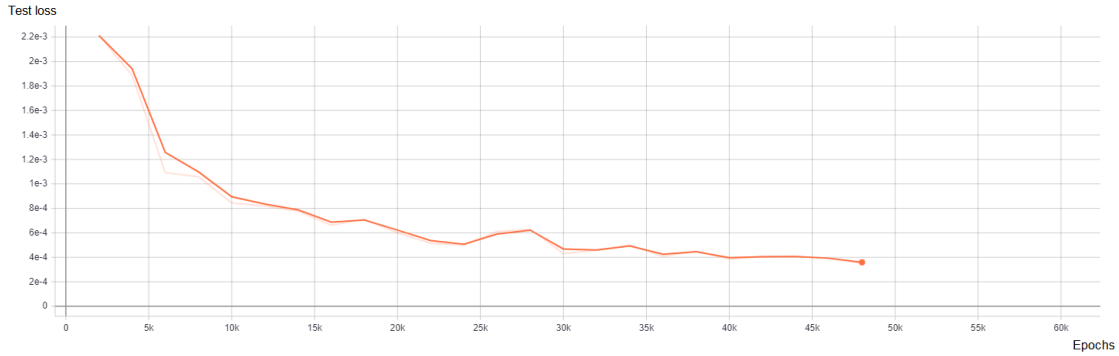


Fig. 4.6. Test loss function of Stem-to-mix model training

The models were trained on a computer, supported by an NVidia GeForce 1080 graphics card. Training an individual model took approximately two days.

5. PREPARATION OF AUDIO MIXES FOR EVALUATION

The primary purpose of training the models was to acquire finished mixes from input files. After creating the network and training the models (detailed in Chapter 4), the main experiment was designed and performed according to Fig. 5.1.

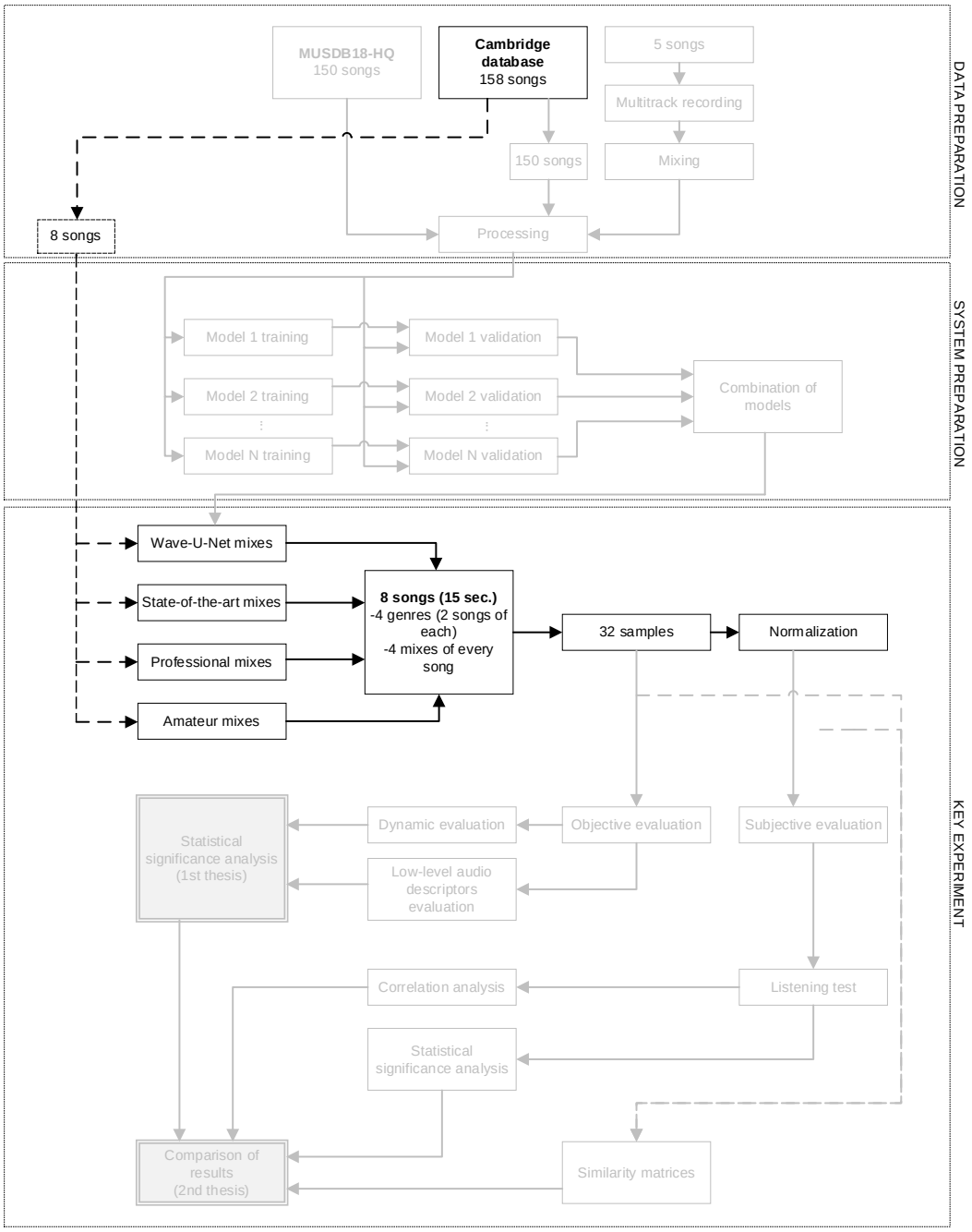


Fig. 5.1. Stages of analysis executed in Chapter 5

In order to perform the experiment, first, the finished mixes were acquired, and the experiment was divided into two parts: objective and subjective. All stages of the testing process are shown in Fig. 5.2. The methods of creating all testing samples are detailed in this Chapter, whereas in Chapter 6, the results of subjective and objective tests are presented.

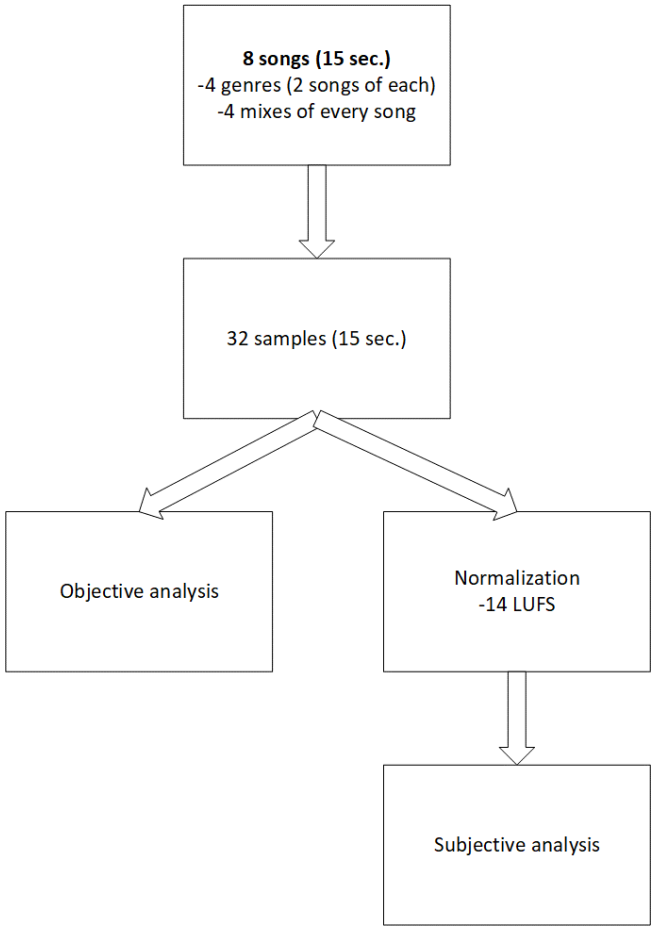


Fig. 5.2. Main test arrangement

For testing purposes, it was decided to create four different mixes of the same song:

- A professional mix.
- An amateur mix.
- A mix created using state-of-the-art software.
- A mix created by the trained models of the Wave-U-Net network.

Clean tracks for eight songs in four music genres were chosen and acquired from the Cambridge database [72]. The list of the selected songs, including their genres and the number of tracks to be mixed, is presented in Table 5.1.

Table 5.1. List of selected songs

No.	Artist name	Name of the song	Genre	No. of tracks
1	Angels in Amplifiers	I'm alright	Pop	13
2	Ben Carrigan	We'll talk about it all tonight	Alternative	51
3	Georgia Wonder	Siren	Electronica	59
4	Secretariat	Over the top	Rock	11
5	Side Effects Project	Sing with me	Electronica	46
6	Speak Softly	Broken man	Pop	17
7	The Doppler Shift	Atrophy	Rock	22
8	Tom McKenzie	Directions	Alternative	31

In Table 5.1, several high-level music genres are presented. In reality, each genre consists of various subgenres to which the songs could belong to [72]. Nevertheless, categorizing song genres is not the focus of the dissertation. Four different music genres were chosen for experiments. Moreover, the songs include varying ensemble of instruments, instrumental compositions, and the number of tracks to be mixed, which introduces an additional degree of freedom. Detailed structures of all chosen songs are presented in Appendix A.

Due to the fact that the songs belong to such vastly different genres of music and the models were trained (more on that subject in Chapter 4.3.) on data from various genres, the evaluation and testing may show interesting results. For example, all 11 tracks from a selected song (Secretariat – Over the top) are shown in the form of a mel spectrogram in Fig. 5.3. All tracks in each song differ from each other in their spectral content.

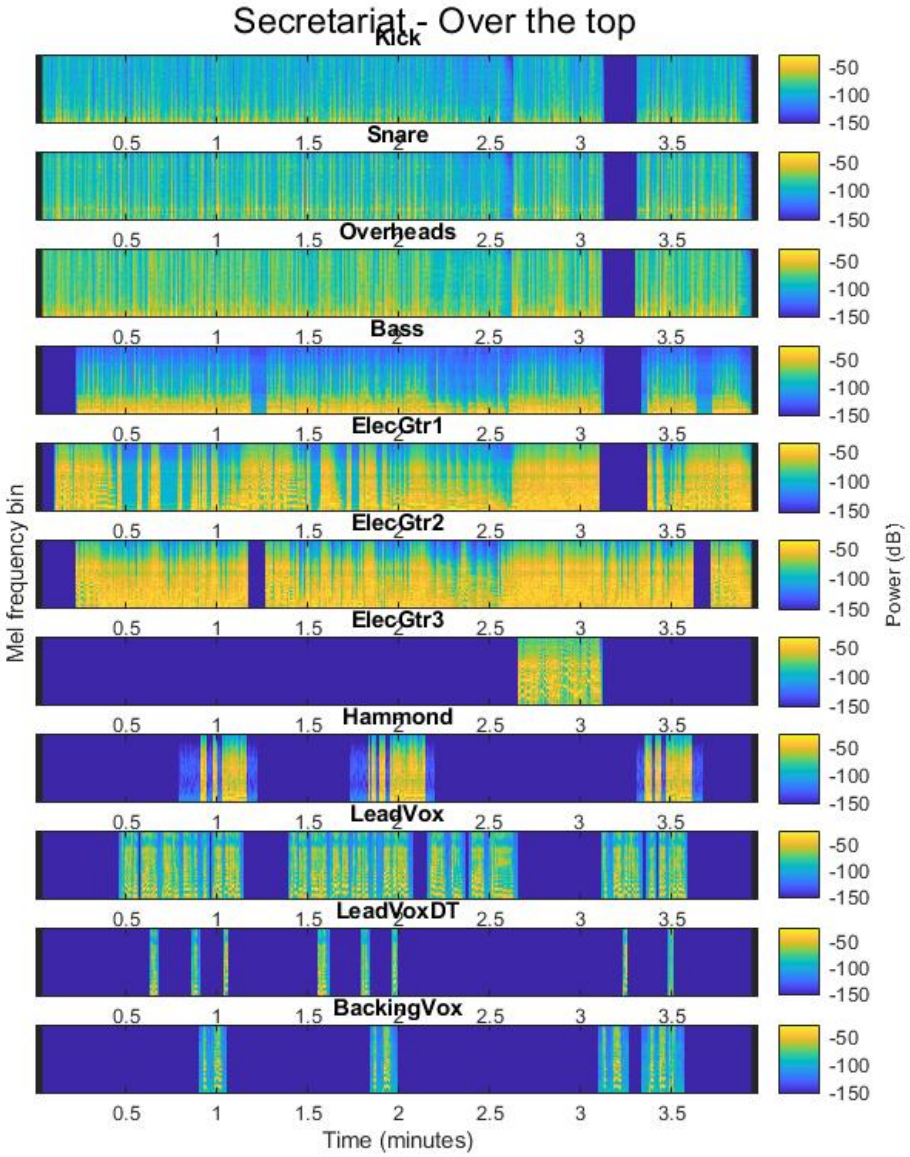


Fig. 5.3. All 11 tracks from Secretariat – Over the top song in the form of a mel spectrogram

5.1. Professional mixes

Known experienced audio engineers created professional mixes. Mixes of the following songs: Angels in Amplifiers – I'm alright, Georgia Wonder – Siren, Side Effects Project – Sing with me, Speak Softly – Broken man, The Doppler Shift – Atrophy, and Tom McKenzie – Directions were created by Mike Senior [9][10][11][12][121][122]. Mike Senior earned a Music Degree at Cambridge University and worked as an assistant engineer in many noted recording studios, such as RG Jones, West Side, Angell Sound, or By Design. Mike Senior is also the creator of the open Cambridge database. He collaborated with many famous artists and is the creator of books such as "Recording Secrets For The Small Studio" and "Mixing Secrets For The Small Studio."

The mix for the song Secretariat – Over the top was created by Brian Garten [68]. Garten is a known recording and mixing engineer. He collaborated with artists like Mariah Carey, Justin Bieber, Britney Spears, and Whitney Houston. He is a four-time nominee for a Grammy award and won one Grammy award for the Best Contemporary R&B Album with Emancipation of Mimi in 2005.

The song Ben Carrigan – We'll talk about it tonight was mixed by Ben Carrigan [22]. Carrigan is a songwriter, composer, and music producer from Dublin, Ireland. He graduated from a music school specializing in jazz, classical, and pop music traditions.

The author of this dissertation did not find detailed information about how each mix was created (e.g., which tools were used). However, he concluded that the mixes sound highly professional and can be used as the "**reference**" (later referred to as the **"Pro" mix**).

5.2. Amateur mixes

The "**Amateur**" mixes were created by a person with experience in both music theory through their education and in practice as a musician. The person, however, did not have any previous experience in audio mixing, neither professional nor recreational. The mixes were created in a home studio using the Cubase 10.5 PRO software. The room in which the mixes were made was treated acoustically. The monitors used during the process were APS Klasik 2020. The digital-to-analog converter used was Apollo Twin.

The length of the mixing process varied for each of the songs, depending on the number of tracks in a given song and the song's genre. The quickest preparation of a mix took approximately 2 hours, the most prolonged – 6 hours. Additionally, in general, the more familiar the genre was to the amateur mixer, the quicker was the process of mixing. The lack of experience in mixing led to a rather intuitive usage of available tools and relying on subjective assumptions about what a mix should sound like. The amateur was, however, free of any habits and mannerisms that a mixer with more experience would have and performed the process with no external guidance. In the "Amateur" mixes author did not exclude any tracks from the final mix. The process of preparing the "Amateur" mixes involved the following steps: first, the loudness levels of the tracks were established; afterward, the panning was set. In the final step, appropriate effects – equalization, reverb, and compression – were used. In Fig. 5.4, the setting

of levels and panning for the song Secretariat – Over the top is shown. In Table 5.2, all effects used on all channels are presented.

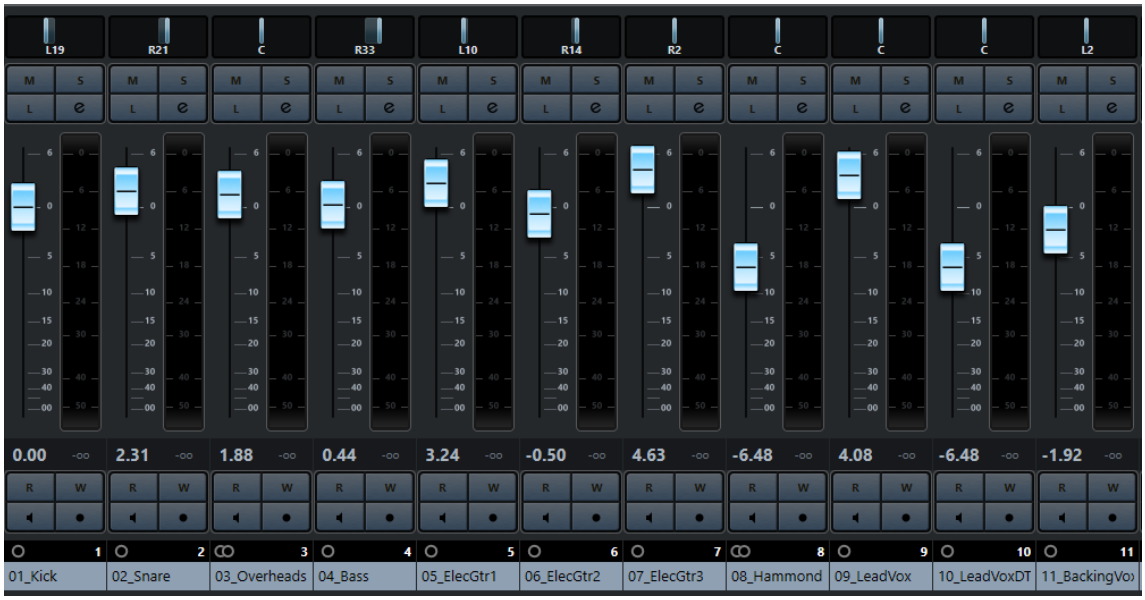


Fig. 5.4. Levels and panning setting in the Amateur mix of Secretariat – Over the top song

As presented in Fig. 5.4., some level values exceed digital 0 (03_Snare: by 2.31, 03_Overheads: by 1.88, 04_Bass: by 0.44, 05_ElecGtr1: by 3.24, 07_ElecGtr3: by 4.63, and finally, 09_LeadVox: by 4.08). It is the result of a common beginner's mistake. Amateur mixers tend to raise the volume of the most important elements in the mix (such as solo guitar or lead vocals). It is widely known that, for human beings, the louder the signal, the better it sounds subjectively [101][142]; thus, before the subjective testing process, the samples needed to be normalized in the context of loudness. Additionally, in Fig. 5.5, the waveforms of each track are presented to compare the values of level changes with respect to the initial level values in the tracks.

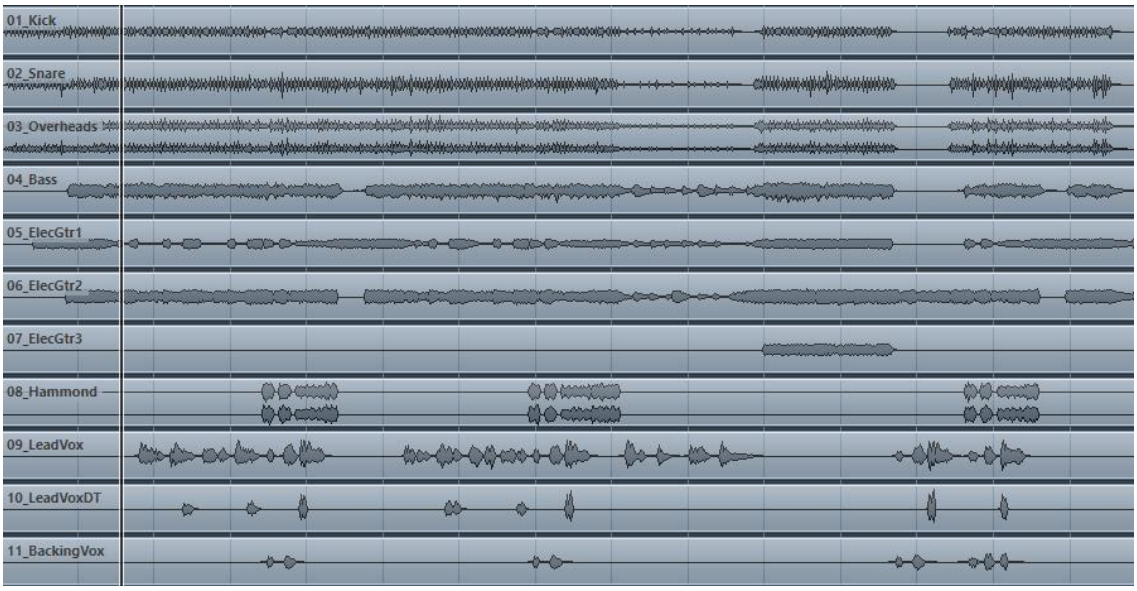


Fig. 5.5. Waveforms of each track in the Secretariat – Over the top song

Table 5.2. Used effects in “Amateur” mix of Secretariat – Over the top song

Channel name	Used effects
01_Kick	Eq: FabFilter Pro-Q3
02_Snare	None
03_Overheads	Eq: FabFilter Pro-Q3 Reverb: FabFilter Pro-R
04_Bass	None
05_ElecGtr1	Eq: FabFilter Pro-Q3 Reverb: FabFilter Pro-R
06_ElecGtr2	Eq: FabFilter Pro-Q3 Reverb: FabFilter Pro-R
07_ElecGtr3	Eq: FabFilter Pro-Q3 Reverb: FabFilter Pro-R
08_Hammond	None
09_LeadVox	Reverb: FabFilter Pro-R
10_LeadVoxDT	Eq: FabFilter Pro-Q3
11_BackingVox	Eq: FabFilter Pro-Q3 Compressor: Stock Cubase Vintage Compressor

5.3. State-of-the-art technology mixes

There are many methods of automatic level balance setting, equalization, compression, or even appropriate reverb matching (to recall: the particular steps were presented in Chapters 2 and 3). Unfortunately, most of them are single solutions to a single problem. As mentioned before, to make a mix sound appealing, one should apply a combination of the aforementioned operations. To create state-of-the-art mixes, a set of Izotope plugins (details on that included in Chapter 3.3) from the music production bundle [44] was used. The plugins included Neutron Pro and Nectar Pro. Their automatic balance and automatic mix features make it possible to mix a song in a semi-automatic way.

First, all recordings were imported into the Cubase 10.5 PRO software. Each track was imported into a separate channel. The semi-automatic processing method with the use of Izotope plugins can be divided into two stages:

- Setting overall balance.
- Creating custom presets for every channel.

In the first stage (setting overall balance), the Relay plugin was applied to each channel. The plugin enables tracks containing Neutron Pro and Nectar Pro to interact with one another for automatic mixing processes [44].

Next, for the master channel, the Neutron Pro plugin was applied. The plugin has a Mix Assistant -> Balance feature. This feature automatically sets the loudness level of every channel in the song while classifying them.

One or more channels can be chosen as the “focus” point with the help of the plugin’s Assistant feature. The focus point is the most critical element in the song, usually being the main vocal. When creating the mixes, the lead vocal was chosen as the element the plugin

should treat as the focus point (Fig. 5.6). Next, the song was played from start to finish following the program instructions.

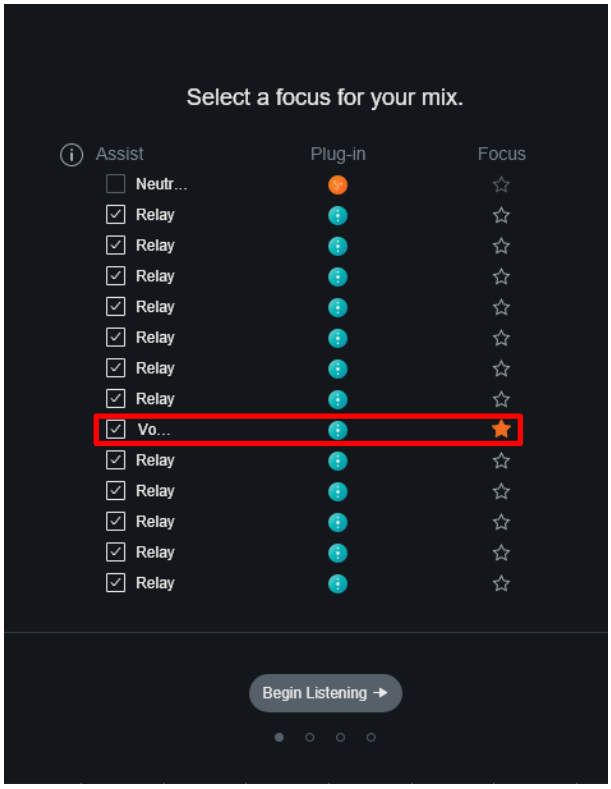


Fig. 5.6. Selection of the focus point of the song (vocal in this case)

After the „listening" stage was completed, the plugin showed the list of tracks and, as mentioned before, automatically classified channels into groups of instruments (Fig. 5.7). At this point, the classification was checked for errors and corrected manually, followed by accepting the Assistant's suggested relative instrument balance.

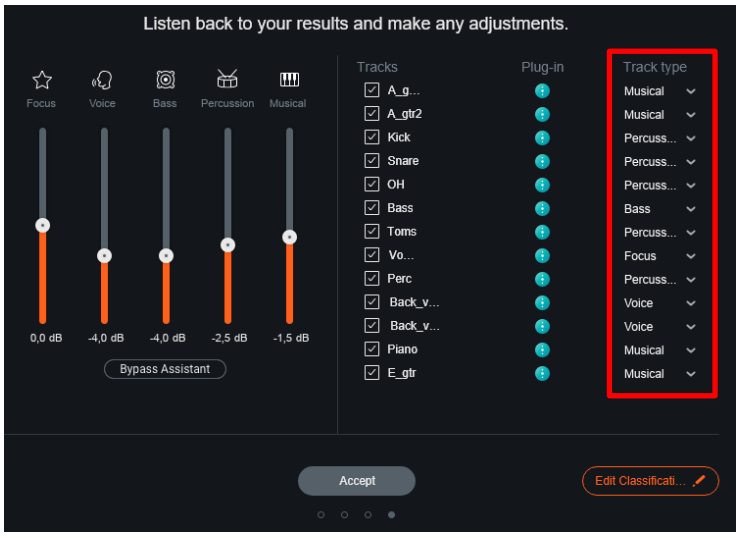


Fig. 5.7. Results of automatic balance settings and instrument classification made by the Neutron Pro plugin (corrected track types marked)

In the second stage (creating a custom preset for every channel), the Neutron Pro plugin was applied to each instrumental channel. After selecting the Mix Assistant option, this time, the Track Enhance feature was chosen.

The plugin gives the user the option to choose an instrument manually or to recognize it automatically. In this case, to avoid incorrect classification, correct instrument labels were assigned manually in GUI. The plugin allows selecting the style in which the preset should be created (warm, balanced, upfront) and the intensity (low, medium, high) with which the instrument should be treated. The balanced style and medium intensity were chosen for every track (Fig. 5.8).

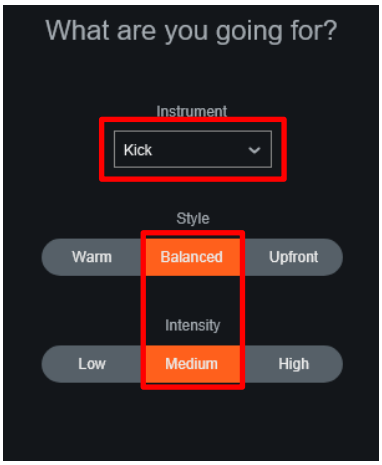


Fig. 5.8. Instrument, Style and Intensity selection

It is crucial to monitor the output level after creating an automatic preset for a selected channel. As mentioned in Chapter 2.6, each operation may change the final loudness level. Therefore, meters available in the plugin were used, which show the levels at the input and output of the plugin. To avoid interfering with the previously set balance, an appropriate output level was set.

Analogically, to channels containing vocals, the Nectar Pro plugin was applied. The plugin works in the same way as Neutron Pro. It is, however, suited for vocal processing. Similarly, as for the channels containing instruments, the Mix Assistant tool was used for the vocal tracks. Next, the type of vocals (singing), the intensity (medium), and the tone (balanced) were selected, as shown in Fig. 5.9.

Fig. 5.9. Selected settings of the Izotope Nectar Pro plugin on a vocal track

Apart from the aforementioned operations, there were no additional operations performed. The mixes were exported in a 44.1 kHz sampling frequency and a 16-bit depth (the same as the source files).

5.4. Wave-U-Net mixes

The "**Unet**" mixes were created using the system presented in Chapter 4. Although in the final version, the system allows for mixing a song without any user intervention, the mixes were created manually. This means that, in the first step, the drum tracks were mixed into a drum stem, the bass tracks into the bass stem, the vocal tracks into the vocal stem, and the remaining tracks into the other stem, using appropriate models. Then, the stems were mixed together using an appropriate stem-to-mix model according to the assumed system architecture.

First, individual tracks for each song were prepared and edited. An appropriate number of tracks that were supposed to be mixed depending on the model used was prepared manually. As mentioned in Chapter 4.2., the drum-to-stem model was constructed to receive 10 mono inputs, the bass-to-stem model four mono inputs, the vocal-to-stem four mono inputs, and the other-to-stem eight mono inputs. If the song contained stereo files, they were separated into two mono files. In cases where the song contained too few tracks of a given instrument group, additional empty tracks were prepared. Moreover, in cases where the number of tracks of a given instrument group was higher than required, it was decided to pre-mix chosen elements manually. There was no level adjustment done between tracks, nor any effects were used (a regular addition of audio signal was used). The tracks which were pre-mixed were chosen so that the signal from individual tracks did not overlap the signal in other tracks (if possible). It is presented in the form of a mel spectrogram in Fig 5.10, where the first three signals were mixed, and a summed signal was then acquired, which is presented at the bottom.

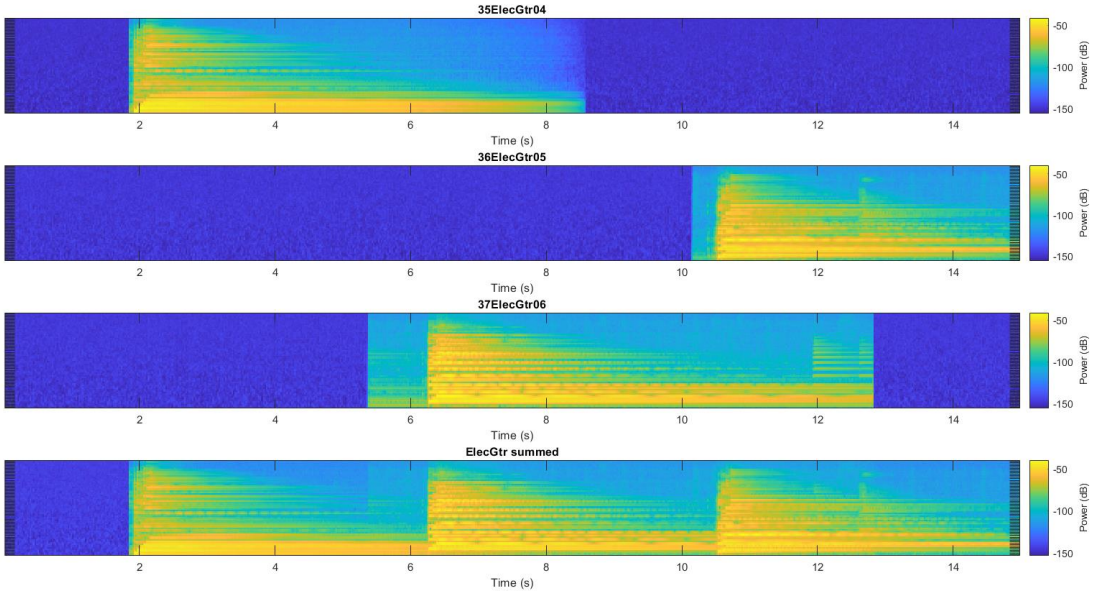


Fig. 5.10. An example of pre-mixing tracks to fit the input of the other-to-stem model in the form of a mel spectrogram

After acquiring the stems of a given song, a stem-to-mix model, which receives four stereo signals, was used, and in effect, a finished stereo mix was obtained. This procedure was repeated for each of the eight chosen songs.

5.5. Postprocessing of mixes

After obtaining all 32 mixes, the postprocessing of the acquired songs was performed. First, from each song, a 15-second clip was selected (duration of an excerpt according to [147]), which best represents the chorus or other loudest part of the song. In other words, a fragment of the song with the most instruments was chosen. An example of a music piece, i.e., “Secretariat – Over the top,” is presented in Fig. 5.11, where all tracks are displayed in the top part, whereas at the bottom, the finished mixes are visible.

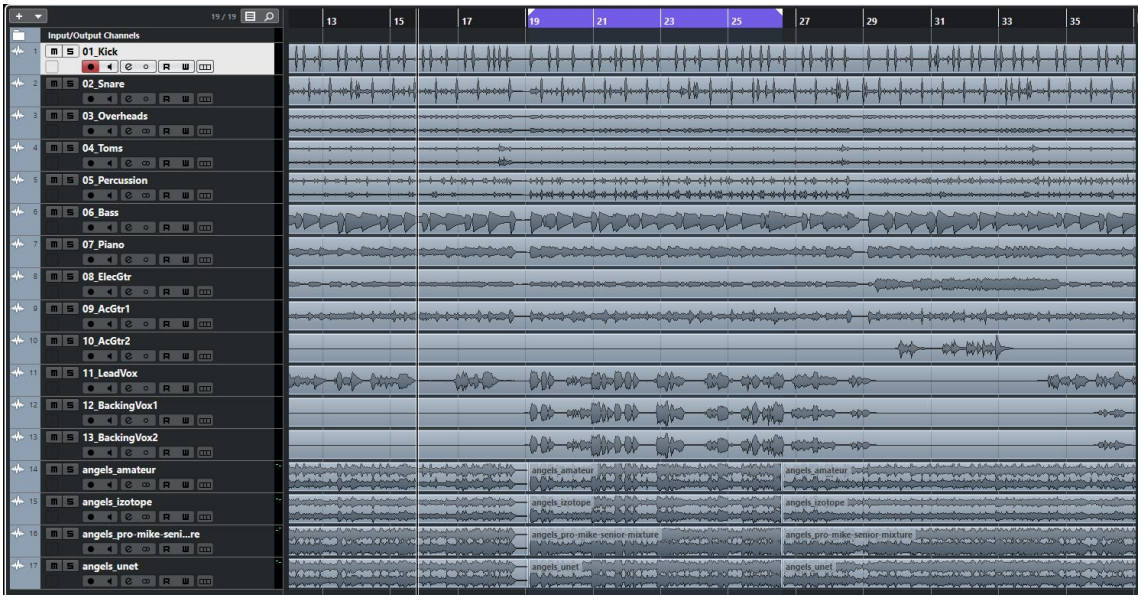


Fig. 5.11. Selected part (15 s) of the Secretariat – Over the top song for the listening evaluation

The completed mixes varied significantly in their loudness level, which is an unwanted characteristic in subjective testing. Therefore, in accordance with AES Technical Committee recommendations, the songs were normalized to a target level of -14 LUFS [96]. The normalization was performed in the Cubase software by setting the level of consecutive samples and observing the level using the Izotope Insight Pro plugin, which shows the Integrated LUFS level.

Subjective tests were performed on normalized samples, whereas objective tests were performed on non-normalized samples. Both of the testing processes (objective and subjective) along with their results are presented in the following chapter. In Fig. 5.12, the spectral content of each of the four mixes of a chosen song (Secretariat – Over the top) is shown. This figure confirms the subjective assumption made during listening that the prepared mixes vary in spectral content.

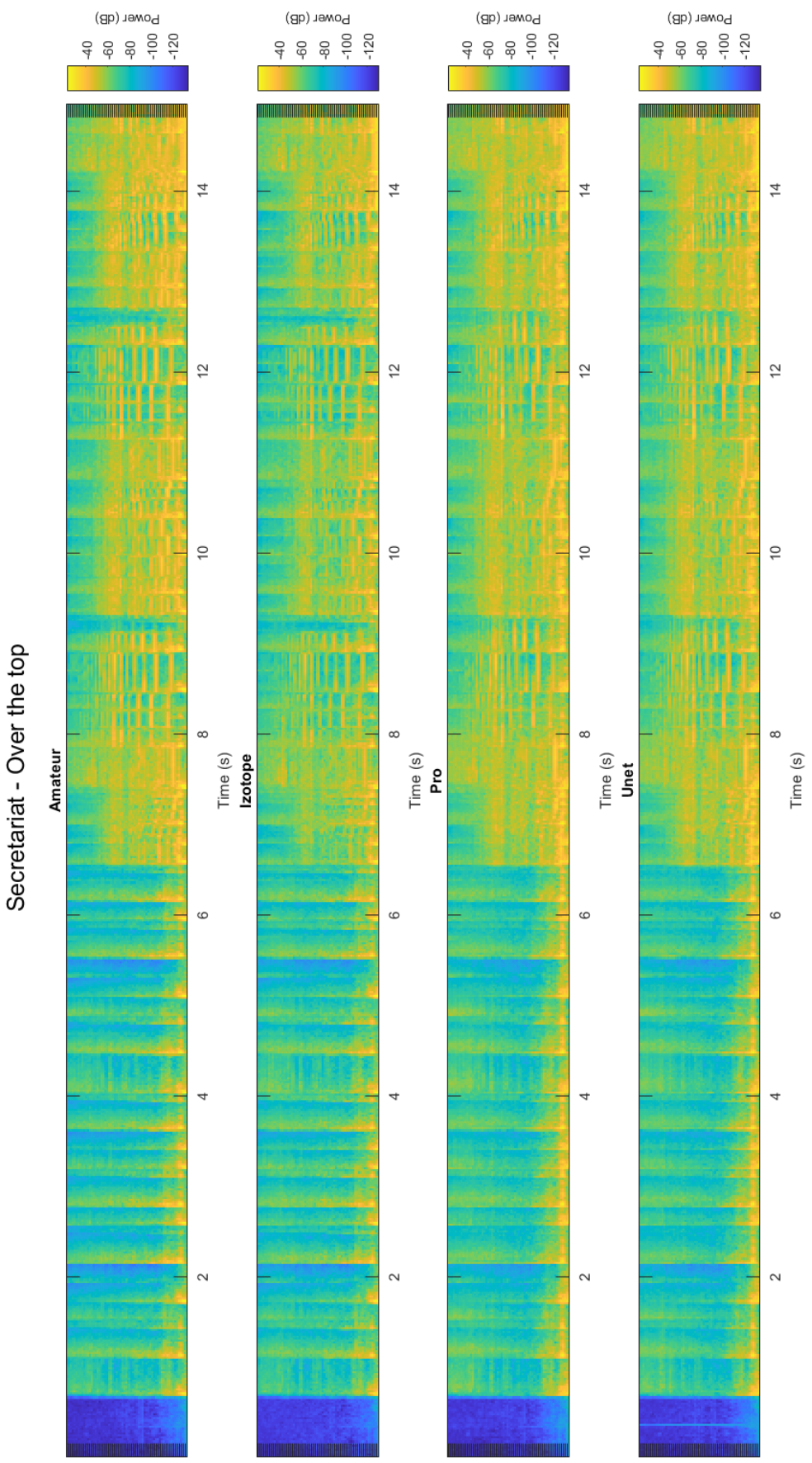


Fig. 5.12. All four mixes of Secretariat – Over the top song in the form of a mel spectrogram

6. EVALUATION OF AUDIO MIXES

In this Chapter, both objective evaluation and subjective test methodology are described (see Fig. 6.1). They constitute an overall quality assessment methodology of mixes obtained.

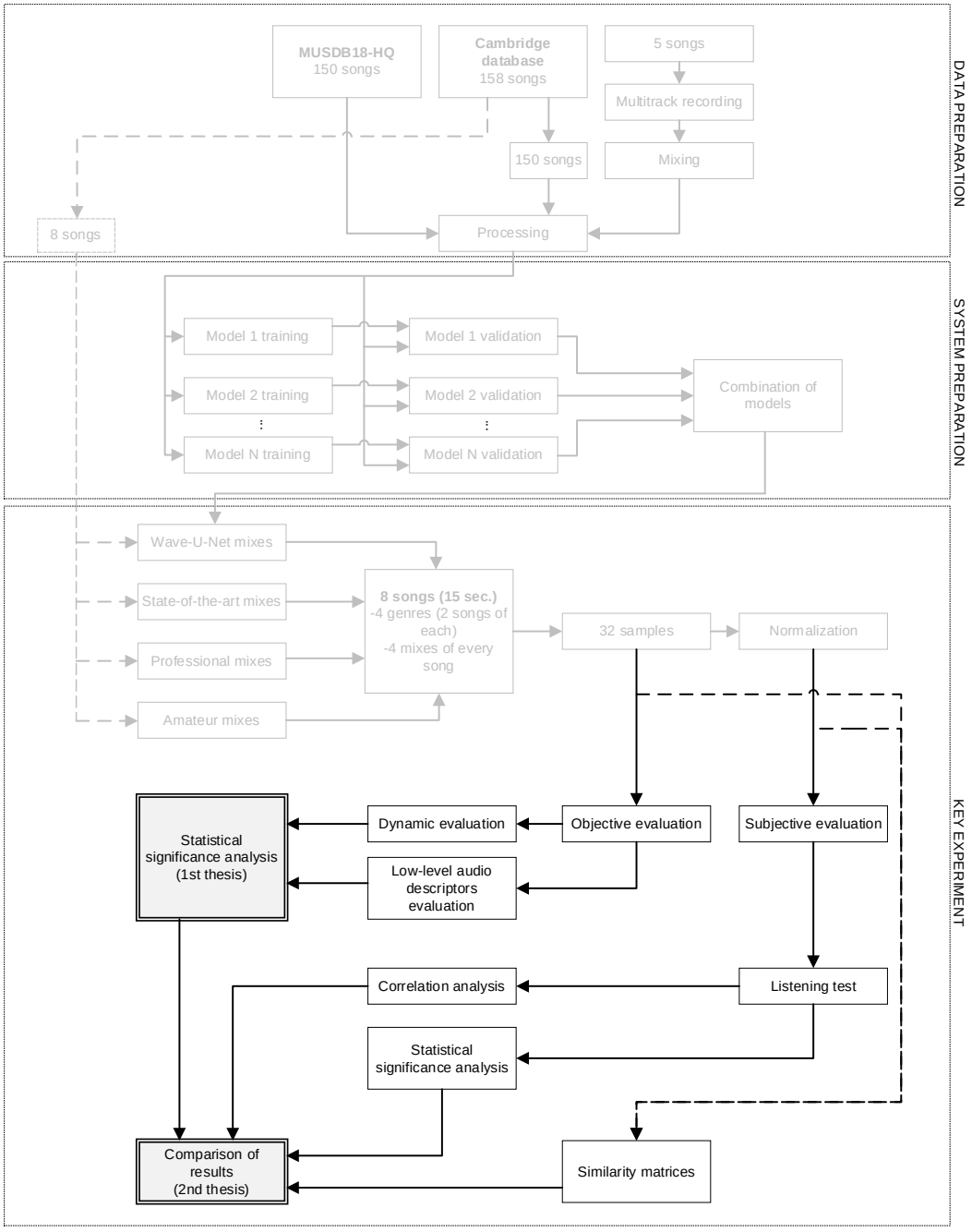


Fig. 6.1. Stages of analysis executed in Chapter 6

First, the several descriptor values related to perceptual characteristics for each mix are calculated. Then, the evaluation methodology and the results of a subjective test conducted on a group of experts are shown. The statistical analysis is performed, and the statistical significance of the achieved results is presented. The main focus is the subjective test analysis

because of the higher priority of this type of test over the objective test results [4][64][137]. This is followed by similarity matrix-based [76][113][118] analyses and discussion.

6.1 Evaluation methodology

6.1.1. Low-level descriptors

There are a variety of low-level descriptors that are used in MIR (Music Information Retrieval) [50][82][103][141]. Many of them are contained in the MPEG-7 standard (Multimedia Content Description Interface). This standard is a document that describes dealing with multimedia data [48][59][133]. In the context of audio signals, the MPEG-7 standard recommends a way of saving sound files and interpreting their parameters. The basic information that the MPEG-7 standard recommends for the description of audio files includes [133]:

- General file information (e.g., copyright, author, year);
- File storing information (e.g., save format, coding);
- Structural information about the spatial, time, or space-time elements of the file;
- Low-level parameter information in the file (e.g., timbre, melody description, decibel level, speed/tempo);
- Additional information – high-level functions (e.g., mood, genre);
- Information about the user interaction with the content (e.g., user preferences – how long the song was listened to, how many times it was skipped);

  Described below are the parameters used in the conducted experiments.

Below, several parameters − employed to compare a given and the reference mixes − are listed. They were chosen due to their correlation with perceptual meaning.

RMS-Energy Envelope

Parameters based on *Root Mean Square* (RMS) are used for calculating the average value of samples in the signal's time-domain (more on that in Chapter 2.1).

Three groups of RMS parameters based on the analysis of the value distribution of sound samples in relation to square means of the signal can be distinguished. The groups are the same as the RMS r_1, r_2, r_3 levels for the analyzed frame of the signal [56]. Based on the exceeding of the RMS value threshold, one can specify the parameters that contain a number of samples that exceed specific RMS thresholds. The following parameters from the RMS group are developed based on smaller fragments of the signal.

Odd-to-Even Harmonic Ratio and Harmonic Envelope

Frequency parameters constitute an essential part of the vector of parameters (i.e., feature vector) built for perceptual analysis. Parameters from the frequency domain describe the content of an audio file due to a breakdown of individual elements of the file into frequency components. The primary way to do frequency analysis of audio signals is by calculating the spectrum of the signal. Then, spectral parameters are calculated based on the signal's

spectrum estimation. Most commonly, the signal's spectrum is obtained using Discrete Fourier Transform (DFT). DFT processes the real sequence of the signal, which is N-samples long, into an M-samples complex representation in the frequency field. Based on the calculated partials, the number of odd and even harmonic components is determined, which can be used to calculate the *Odd-to-Even Harmonic Ratio*. The *Audio Spectrum Envelope* (ASE) is used to calculate the *Harmonic Envelope*. ASE is a short-term power spectrum of P_x for frequencies in logarithmic intervals [133]. These parameters are highly related to the perceptual evaluation of a music signal.

6.1.2. Statistical analysis

There are several statistical tests employed for checking whether differences between parameter values are significant. The Shapiro-Wilk test is used to test the similarity of the distribution of a given variable to the normal distribution. The test tests the null hypothesis, which states that the distribution of the given variable is close to normal. Testing the normality of distribution is necessary when using parametric tests, e.g., variance analysis.

The Shapiro-Wilk test is calculated using the following formula (6.1) [115]:

$$W = \frac{(\sum_i a_i(n)(X_{n-i+1}-X_i))^2}{\sum_{j=1}^{n}(X_j-\bar{X})^2} \tag{6.1}$$

where W is the result of the Shapiro-Wilk test, $a_i(n)$ is a constant, j are the subsequent data points in the sample, and i are the subsequent differences between extreme data points.

A single-factor variance analysis (one-way ANOVA) is performed to test if any individual factor influences the measured dependent variable. It is assumed that the factor has a form of groups varying only by the value of the factor. The single-factor variance analysis is treated as an expansion of the t-Student test. The tests are limited to testing only two groups. Variance analysis does not have these limitations. To compare more than two groups, instead of a triple use of the t-Student test, a single-factor variance analysis is performed. However, a significant result of an F test (variance analysis) does not provide information on which groups of the tested ones vary. The result informs about the existence of variation among the groups (or that the impact of a given factor on the dependent variable was observed.) To confirm which groups are characterized by statistically significant variance, "post-hoc" multiple comparisons need to be performed [83].

The Tukey-Kramer Test (Tukey's Honest Significant Difference test) is a single-step multiple comparison procedure and statistical test [23]. It is a post-hoc ad based on a studentized distribution [23]. As mentioned above, the ANOVA test provides information on the overall statistical significance of the results without confirming the placement of the differences. The Turkey-Kramer test can be performed (after the ANOVA test results proved to be significant) to specify which groups' means are different when compared with each other. The Tukey's Honest Significant Difference test compares each pairwise combination of means [136].

When simultaneously testing multiple hypotheses (a family of hypotheses), one could risk the increase of the α error value, which is the main problem in the field of multiple comparisons. An increase of the α error suggests that the null hypothesis is rejected too often

while being true (the existence of differences is indicated when in reality, there are none.) To prevent the increase of α, a correction (decrease) of the α value or a correction (increase) of the p value of the tests can be done (Sidak correction). The Sidak correction is described by the two Eqs. (6.2) and (6.3) [1]:

$$p_{(Sidak,i)} = 1 - (1 - p_i)^c \tag{6.2}$$

$$\alpha_{(Sidak,i)} = 1 - (1 - \alpha_i)^{1/c} \tag{6.3}$$

The Pearson correlation coefficient (Pearson's r) is used to calculate the relationship between quantitative variables [80]. It informs about the strength and the regression slope between variables. The correlation coefficients can assume values from a $[-1; 1]$ range. The values indicate the strength of the relationship – the closer the value is to "0", the weaker the correlation; the closer the value is to "1" or "-1", the stronger the correlation. A value of „1" implies a perfect linear correlation (all data points lie on a line).

Table 6.1. Interpretation of the correlation coefficient values

r values (absolute)	Interpretation
0 – 0.3	No correlation or very weak correlation
0.3 – 0.5	Moderately strong correlation
0.5 – 0.7	Strong correlation
0.7 – 1	Very strong correlation

6.1.3. Self-similarity matrices

A self-similarity matrix (SSM) converts the sequence of features into 2D feature space by comparing its elements. An idea of comparing each element of the feature sequence with all other elements of the sequence and visualization by a matrix of similarity scores was borrowed from the music information retrieval (MIR) domain [28]. Currently, SSMs are widely used for the analysis and generation of music signals [58][84], as well as for performing other tasks related to audio signals, such as highlighting interlanguage phoneme differences [53] or motion data analysis in manufacturing scenarios [102].

The author focused on music-based features called a chromagram. The chromagram construction method takes into account the fact that pitch consists of two components: tone height and chroma [3][116]. The features represent the distribution of signal energy over chroma and time. The relationship between components can be defined by the following formula:

$$f = 2^{c+h} \tag{6.4}$$

where c is chroma ($c \in [0,1]$), f is frequency, and h denotes the frequency.

The process of feature calculation was organized in a pipelined manner, i.e., the signal was divided into overlapping frames, and for each frame of the chroma, a vector was obtained. The chroma vector sums up the spectral energy into 12 bins corresponding to the 12 semitones within an octave. An example of the chromagram is given in Fig. 6.2, where frames are shown along the *x*-axis, chroma bins are presented along the *y*-axis, and the color saturation indicates the intensity of the sound signal.

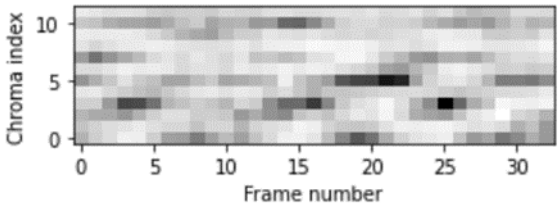


Fig. 6.2. Chromagram of the Secretariat – Over the top "Unet" mix

For the chromagram example (see Fig. 6.2), the following settings were selected: frame size – 2048 samples, overlap – 1025 samples.

The starting point of an SSM construction is the feature normalization procedure. The feature sequence is realized by the $M \times N$ matrix X, where M is the number of acoustic parameters and N is the number of short-time segments. The feature normalization was performed by normalization of each column of the feature matrix.

The normalized values are calculated using the following formula:

$$\hat{x}_n = \frac{x_n - \bar{x}_n}{SD} \tag{6.5}$$

where $\bar{x}_n$ and SD are the mean and standard deviation of non-normalized features, respectively, and $x_n = (x_{1n}, \ldots, x_{Nn})$ is the n-th matrix column $(n = 1, \ldots, N)$. The mean and standard deviation are calculated as follows:

$$\bar{x}_n = \frac{1}{M}\sum_{m=1}^{M} x_{mn} \tag{6.6}$$

$$SD = \sqrt{\frac{\sum_{n=1}^{M}(x_{mn} - \bar{x}_n)}{M-1}} \tag{6.7}$$

To calculate the values of SSM, each column of the normalized feature matrix $\hat{X}$ is compared with each other. For this purpose, the dot product between the feature matrix and its transpose is calculated:

$$S = \hat{X}^T \hat{X} \tag{6.8}$$

The entries of the matrix imply the similarity scores. Each pixel in the matrix obtains a greyscale value corresponding to the given similarity score. The darkest color refers to the smallest similarity. By comparing the given scores, each short-time segment is compared with each other.

6.2 Objective evaluation

Unprocessed samples were used for the objective evaluation. This is because subjecting the recordings to normalization may prevent the correct identification of objective values for the acquired samples.

First, the waveform statistics were calculated, such as RMS level (Fig. 6.3), *Integrated Loudness* (Fig. 6.4), *Loudness Range* (Fig. 6.5), and *True peak* level (Fig. 6.6) for all music excerpts. Further on, selected low-level descriptors MPEG-7 were calculated. For this purpose, the timbre toolbox [134] in the MATLAB environment was used. *Odd-to-even Harmonic Ratio*, RMS-*Energy Envelope*, *Harmonic Energy*, and *Noisiness* were calculated for each music sample. As already mentioned, these descriptors were chosen because of their perceptual

interpretation. An example of all computed data for the “Secretariat – Over the top” sample resulting from the Izotope-based mixing is presented in Fig. 6.7. All calculated descriptor values with corresponding graphs for other songs are included in Appendix B.

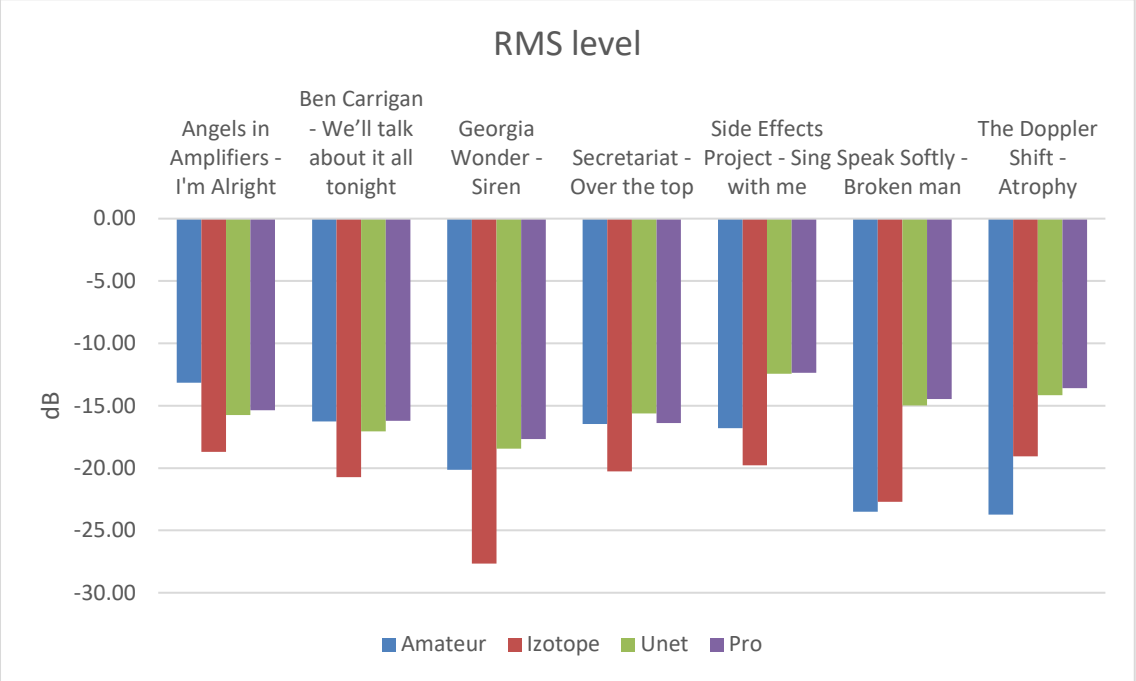


Fig. 6.3. RMS level calculated for all music pieces evaluated

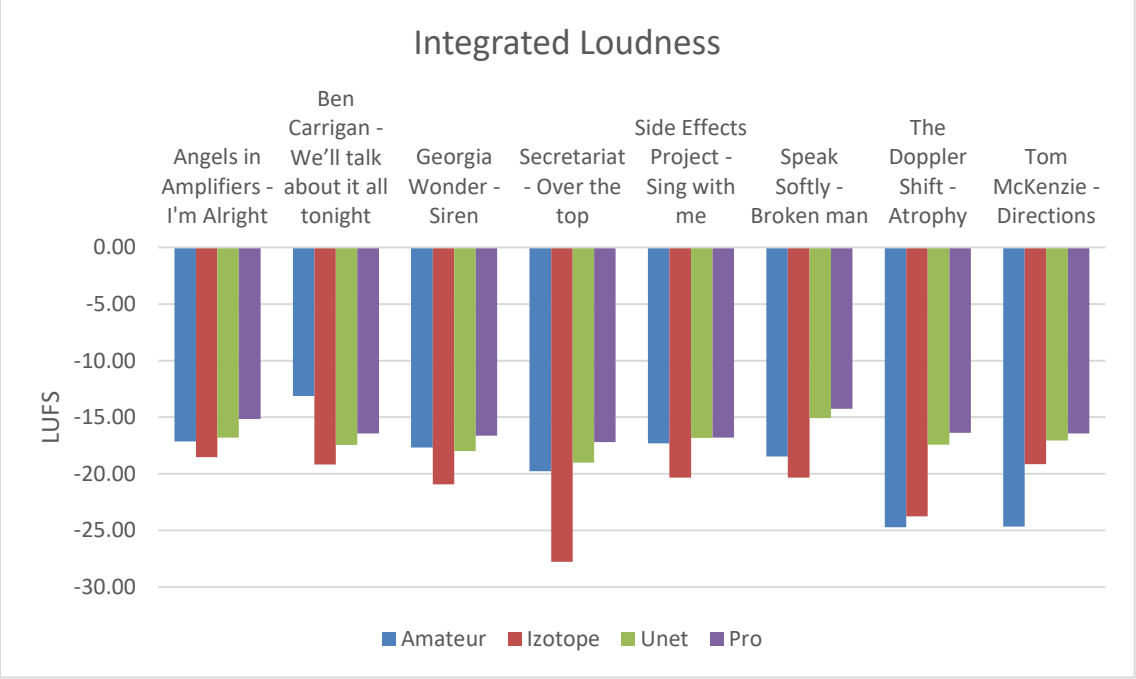


Fig. 6.4. *Integrated Loudness* calculated for all music samples

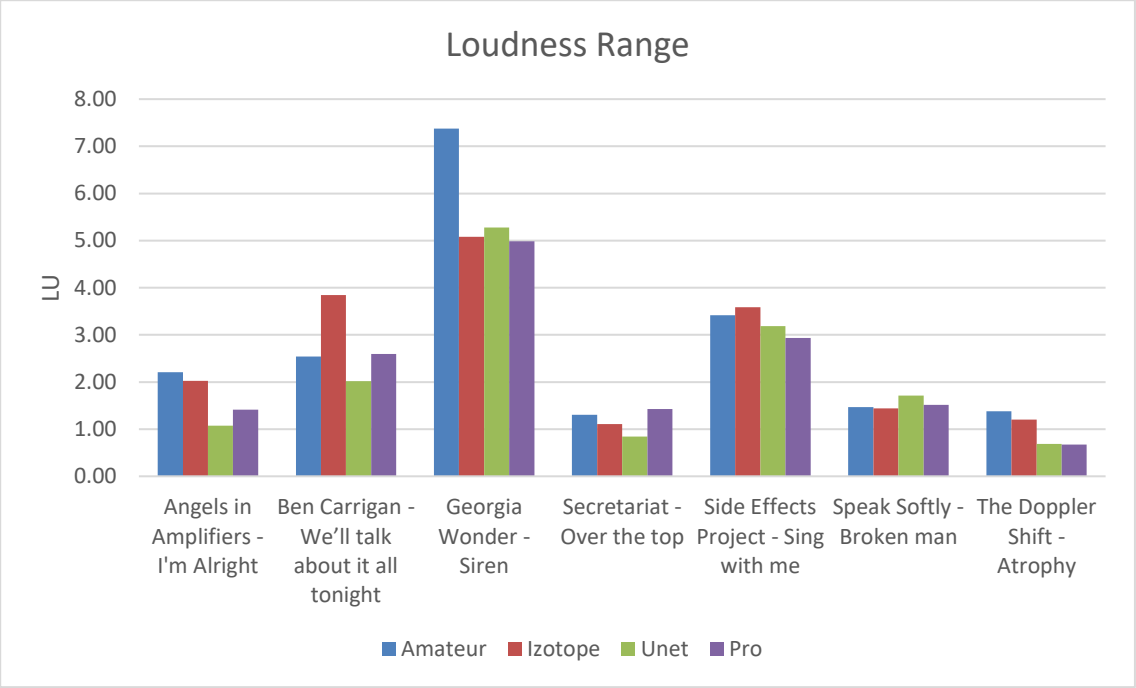


Fig. 6.5. *Loudness Range* calculated for all music samples

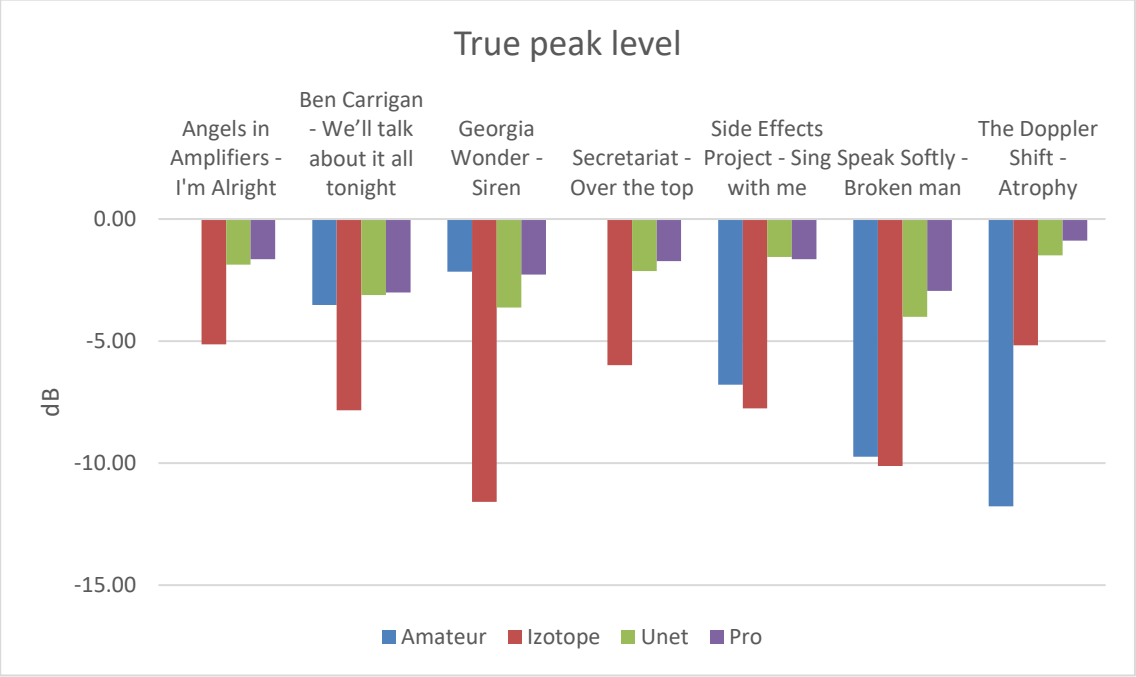


Fig. 6.6. *True peak* level calculated for all music samples

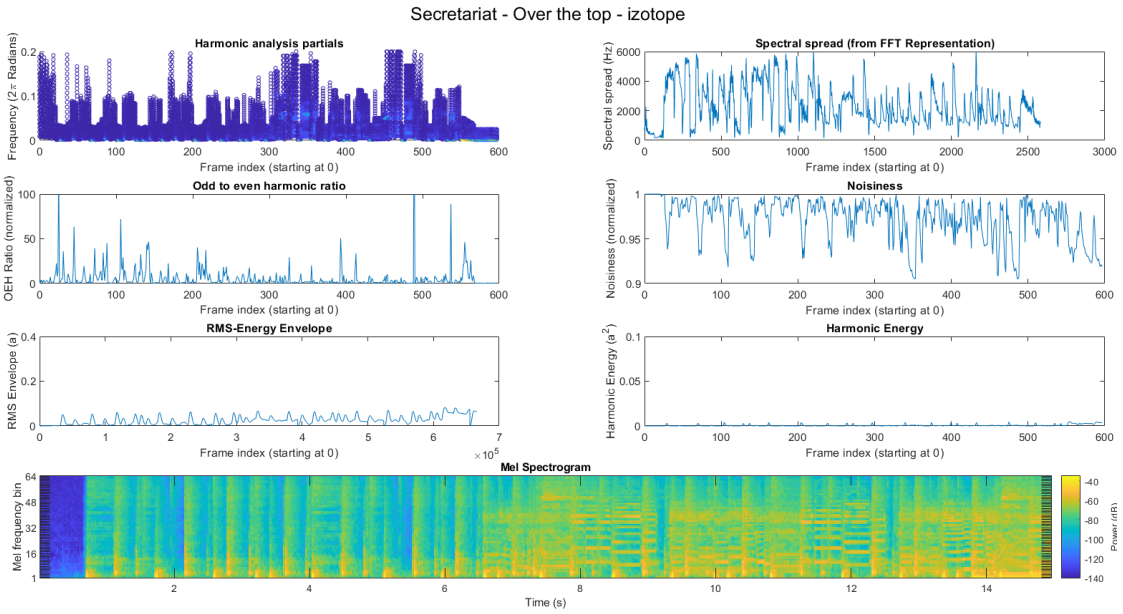


Fig. 6.7. Descriptors calculated for Secretariat – Over the top "Izotope" sample

The RMS level-based results from Fig. 6.3 are presented in Table 6.2. The RMS level is a critical element that every mixing engineer should pay attention to, as it approximates the human perception of the loudness of sound. As seen in Table 6.2, level values differ depending on the recording genre and the number of instrumental tracks mixed together. Assuming the "Pro" mix as the reference, it can be noted that the "Unet" mix is the closest to the said reference. The "Amateur" and "Izotope" mixes differ noticeably from the commonly accepted norm of -16 to -14 dB RMS.

Table 6.2. RMS level calculated for all objective samples

song	**RMS level [dB]**			
	Amateur	*Izotope*	*Unet*	*Pro*
Angels in Amplifiers - I'm Alright	-16.13	-18.12	-15.80	-15.12
Ben Carrigan - We'll talk about it all tonight	-13.16	-18.69	-15.75	-15.37
Georgia Wonder - Siren	-16.26	-20.74	-17.07	-16.22
Secretariat - Over the top	-20.15	-27.66	-18.45	-17.68
Side Effects Project - Sing with me	-16.47	-20.27	-15.62	-16.39
Speak Softly - Broken man	-16.80	-19.77	-12.44	-12.37
The Doppler Shift - Atrophy	-23.51	-22.70	-14.98	-14.46
Tom McKenzie - Directions	-23.73	-19.05	-14.17	-13.60
Standard deviation	3.80	3.09	1.80	**1.68**
Variance	14.42	9.52	3.25	**2.82**

The *Integrated Loudness*-based results from Fig. 6.4 are presented in Table 6.3. This level was defined for audio signal normalization purposes and matched how human ears perceive sound. As can be concluded from the table, the level varies not only among the types of mixes but also between the songs. It is normal because different music genres are

characterized by different target levels. However, assuming the “Pro” mix as the reference, the “Unet” mix is the closest to the reference. The “Amateur” and “Izotope” mixes are vastly different from the assumed norm of -16 to -14 LUFS.

Table 6.3. *Integrated Loudness* calculated for all objective samples

song	**Integrated Loudness (LUFS)**			
	Amateur	*Izotope*	*Unet*	*Pro*
Angels in Amplifiers - I'm Alright	-17.14	-18.52	-16.80	-15.17
Ben Carrigan - We’ll talk about it all tonight	-13.13	-19.18	-17.46	-16.43
Georgia Wonder - Siren	-17.67	-20.92	-17.98	-16.63
Secretariat - Over the top	-19.77	-27.76	-19.00	-17.20
Side Effects Project - Sing with me	-17.30	-20.32	-16.83	-16.80
Speak Softly - Broken man	-18.46	-20.33	-15.08	-14.26
The Doppler Shift - Atrophy	-24.73	-23.74	-17.43	-16.39
Tom McKenzie - Directions	-24.67	-19.15	-17.04	-16.44
Standard deviation	3.93	3.08	1.12	**0.96**
Variance	15.46	9.51	1.26	**0.93**

The *Loudness Range*-based results from Fig. 6.5 are presented in Table 6.4. *Loudness Range* (measured in Loudness Units) shows loudness variation over the entire song. As can be concluded from the table, the “Pro” mixes are characterized by the smallest deviation and variance. In general, it is assumed that with $LU < 4$, a mix is relatively static in dynamics. The calculations presented in the table show that the “Pro” mixes are the most static in dynamic range, followed by “Izotope” and “Unet” mixes. The “Amateur” mixes have the highest values of standard deviation and variance – their loudness widely varies in different songs and different music genres.

Table 6.4. *Loudness Range* calculated for all objective samples

song	**Loudness Range (LU)**			
	Amateur	*Izotope*	*Unet*	*Pro*
Angels in Amplifiers - I'm Alright	2.45	2.71	1.50	1.51
Ben Carrigan - We’ll talk about it all tonight	2.21	2.03	1.07	1.42
Georgia Wonder - Siren	2.54	3.84	2.02	2.60
Secretariat - Over the top	7.38	5.08	5.28	4.99
Side Effects Project - Sing with me	1.30	1.11	0.84	1.43
Speak Softly - Broken man	3.42	3.59	3.19	2.93
The Doppler Shift - Atrophy	1.46	1.44	1.71	1.52
Tom McKenzie - Directions	1.38	1.20	0.69	0.67
Standard deviation	2.00	1.44	1.53	**1.36**
Variance	3.99	2.08	2.34	**1.84**

The results of True peak level from Fig. 6.6 are presented in Table 6.5. As can be observed, the “Pro” mixes are characterized by the lowest values of variance and standard deviation even though the mixes were created by several audio engineers and are closely followed by the “Unet” mixes. The “Amateur” mixes are not only characterized by the highest

variance. This is understandable as amateurs do not have much experience mixing songs in various music genres. Moreover, two songs, i.e., “Ben Carrigan – We’ll talk about it tonight” and Side “Effects Project – Sing with me,” exceed digital “zero,” which means that in those mixes, an unpleasant digital distortion is present. Exceeding digital “zero” in mixes is a typical mistake made by amateur mixers.

Table 6.5. *True peak* level calculated for all objective samples

song	**True peak level (dB)**			
	Amateur	*Izotope*	*Unet*	*Pro*
Angels in Amplifiers - I'm Alright	-2.74	-4.14	-1.82	-2.04
Ben Carrigan - We’ll talk about it all tonight	0.02	-5.13	-1.87	-1.65
Georgia Wonder - Siren	-3.52	-7.84	-3.11	-3.01
Secretariat - Over the top	-2.16	-11.59	-3.62	-2.27
Side Effects Project - Sing with me	0.01	-5.99	-2.13	-1.73
Speak Softly - Broken man	-6.79	-7.76	-1.56	-1.64
The Doppler Shift - Atrophy	-9.74	-10.12	-4.01	-2.95
Tom McKenzie - Directions	-11.77	-5.18	-1.49	-0.89
Standard deviation	4.40	2.61	0.98	**0.71**
Variance	19.40	6.82	0.97	**0.51**

Descriptors such as *Odd-to-Even Harmonic Ratio*, RMS-*Energy Envelope*, and *Harmonic Energy* are given more consideration as they show both the dynamic and spectral content in the given audio signal. In Fig. 6.8, a variation of the RMS-*Energy Envelope* of the “Secretariat – Over the top song” − depending on the mix type − is shown. Graphs of all three descriptors calculated for each sample are presented in Appendix B.

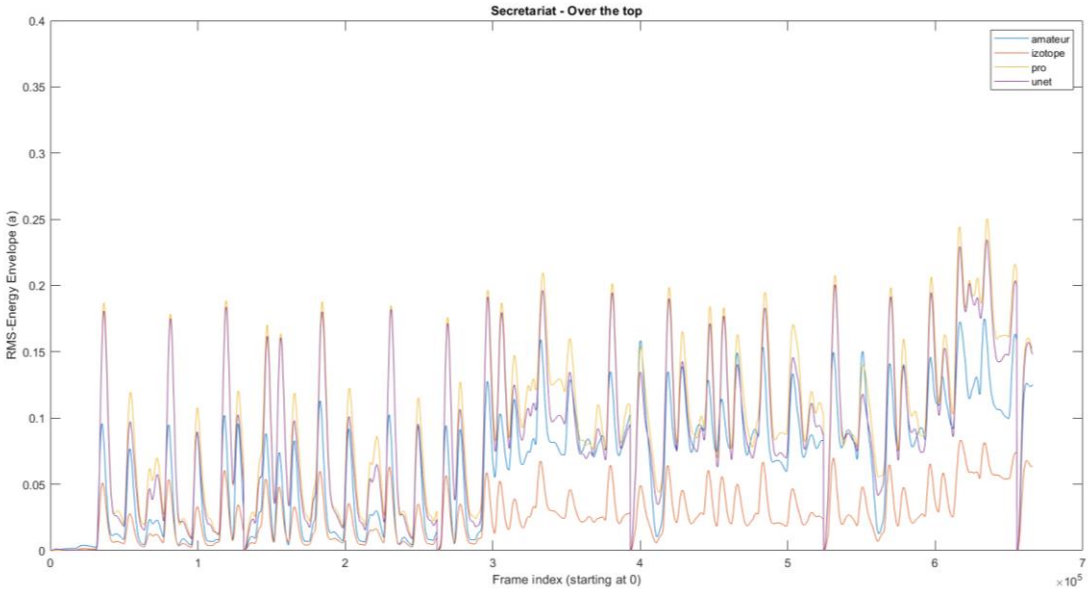


Fig. 6.8. Variation of the RMS-*Energy Envelope* depending on the mix type in the Secretariat – Over the top song

For each mentioned descriptor, an analysis was performed to determine the statistical significance of differences between the mixes. For this purpose, one-way ANOVA series and

the post hoc Tukey Kramer test were executed. The level of significance in this Chapter was assumed to be $\propto = 0{,}05$.

The statistical significance calculation results of the RMS-*Energy Envelope* descriptor value differences for pairs of mixes are presented in Table 6.6. In bold font, the *p*-value for all significant differences between compared mix samples is highlighted. As can be concluded from the table, in the case of this descriptor, almost all pair comparisons are characterized by statistical significance, except the “Amateur” – “Unet” pair in the “Georgia Wonder – Siren” song and the “Unet” – “Pro” in the “The Doppler Shift – Atrophy” song. This means that the differences between the types of mixes in most cases are statistically significant.

Table 6.6. Statistical significance calculation results of the RMS-Energy Envelope descriptor

RMS-*Energy Envelope*					
Samples compared		*Lower confidence*	*Estimate*	*Upper confidence*	*p*
Angels in Amplifiers - I'm Alright					
Amateur	Izotope	0.02	0.02	0.02	**0.00**
Amateur	Unet	-0.02	-0.02	-0.02	**0.00**
Amateur	Pro	-0.01	-0.01	-0.01	**0.00**
Izotope	Unet	-0.04	-0.04	-0.04	**0.00**
Izotope	Pro	-0.03	-0.03	-0.03	**0.00**
Unet	Pro	0.01	0.01	0.01	**0.00**
Ben Carrigan - We'll talk about it tonight					
Amateur	Izotope	0.08	0.08	0.08	**0.00**
Amateur	Unet	0.04	0.04	0.04	**0.00**
Amateur	Pro	0.04	0.04	0.04	**0.00**
Izotope	Unet	-0.04	-0.04	-0.04	**0.00**
Izotope	Pro	-0.03	-0.03	-0.03	**0.00**
Unet	Pro	0.01	0.01	0.01	**0.00**
Georgia Wonder - Siren					
Amateur	Izotope	0.05	0.05	0.05	**0.00**
Amateur	Unet	0.00	0.00	0.00	0.23
Amateur	Pro	0.01	0.01	0.01	**0.00**
Izotope	Unet	-0.05	-0.05	-0.05	**0.00**
Izotope	Pro	-0.03	-0.03	-0.03	**0.00**
Unet	Pro	0.01	0.01	0.01	**0.00**
Secretariat - Over the top					
Amateur	Izotope	0.04	0.04	0.04	**0.00**
Amateur	Unet	-0.03	-0.03	-0.03	**0.00**
Amateur	Pro	-0.02	-0.02	-0.02	**0.00**
Izotope	Unet	-0.07	-0.07	-0.07	**0.00**
Izotope	Pro	-0.06	-0.06	-0.06	**0.00**
Unet	Pro	0.01	0.01	0.01	**0.00**
Side Effects Project - Sing with me					
Amateur	Izotope	0.04	0.04	0.04	**0.00**
Amateur	Unet	0.00	0.00	0.00	**0.00**
Amateur	Pro	-0.01	-0.01	-0.01	**0.00**
Izotope	Unet	-0.04	-0.04	-0.04	**0.00**

Izotope	Pro	-0.05	-0.05	-0.05	**0.00**
Unet	Pro	-0.01	-0.01	-0.01	**0.00**
Speak Softly - Broken man					
Amateur	Izotope	0.03	0.03	0.03	**0.00**
Amateur	Unet	-0.08	-0.08	-0.08	**0.00**
Amateur	Pro	-0.08	-0.08	-0.08	**0.00**
Izotope	Unet	-0.11	-0.11	-0.11	**0.00**
Izotope	Pro	-0.11	-0.11	-0.11	**0.00**
Unet	Pro	0.00	0.00	0.00	**0.00**
The Doppler Shift - Atrophy					
Amateur	Izotope	0.01	0.01	0.02	**0.00**
Amateur	Unet	-0.03	-0.03	-0.02	**0.00**
Amateur	Pro	-0.03	-0.03	-0.02	**0.00**
Izotope	Unet	-0.04	-0.04	-0.03	**0.00**
Izotope	Pro	-0.04	-0.04	-0.03	**0.00**
Unet	Pro	0.00	0.00	0.01	1.00
Tom McKenzie - Directions					
Amateur	Izotope	-0.04	-0.04	-0.04	**0.00**
Amateur	Unet	-0.10	-0.10	-0.10	**0.00**
Amateur	Pro	-0.09	-0.09	-0.09	**0.00**
Izotope	Unet	-0.06	-0.06	-0.06	**0.00**
Izotope	Pro	-0.06	-0.06	-0.06	**0.00**
Unet	Pro	0.01	0.01	0.01	**0.00**

The statistical significance results of the *Harmonic Energy* descriptor value differences for pairs of mixes are presented in Table 6.7. As this table shows, in the case of the *Harmonic Energy* descriptor, most of the pair comparisons are characterized by statistical significance (these values are highlighted in bold) except the “Unet” – “Pro” pair in the “Ben Carrigan – We’ll talk about it tonight” song, “Unet” – “Pro” pair in the “Secretariat – Over the top” song, “Amateur” – “Pro” and “Unet” – “Pro” pair in the “Side Effects Project – Sing with me” song, “Unet” – “Pro” in “Speak Softly – Broken man” and “Amateur” – “Izotope” pair in “The Doppler Shift – Atrophy” song.

Table 6.7. Statistical significance calculation results of the Harmonic Energy descriptor

Harmonic Energy					
Samples being compared		*Lower confidence*	*Estimate*	*Upper confidence*	***p***
Angels in Amplifiers - I'm Alright					
Amateur	Izotope	0.00	0.00	0.00	**0.00**
Amateur	Unet	-0.01	0.00	0.00	**0.00**
Amateur	Pro	0.00	0.00	0.00	**0.00**
Izotope	Unet	-0.01	-0.01	0.00	**0.00**
Izotope	Pro	-0.01	0.00	0.00	**0.00**
Unet	Pro	0.00	0.00	0.00	**0.03**
Ben Carrigan - We'll talk about it tonight					
Amateur	Izotope	0.02	0.02	0.02	**0.00**
Amateur	Unet	0.01	0.01	0.01	**0.00**

Amateur	Pro	0.01	0.01	0.01	**0.00**
Izotope	Unet	-0.01	-0.01	-0.01	**0.00**
Izotope	Pro	-0.01	-0.01	0.00	**0.00**
Unet	Pro	0.00	0.00	0.00	0.88
Georgia Wonder - Siren					
Amateur	Izotope	0.01	0.01	0.01	**0.00**
Amateur	Unet	0.00	0.00	0.00	**0.00**
Amateur	Pro	0.00	0.00	0.00	**0.00**
Izotope	Unet	-0.01	-0.01	0.00	**0.00**
Izotope	Pro	-0.01	0.00	0.00	**0.00**
Unet	Pro	0.00	0.00	0.00	**0.00**
Secretariat - Over the top					
Amateur	Izotope	0.00	0.00	0.00	**0.00**
Amateur	Unet	0.00	0.00	0.00	**0.00**
Amateur	Pro	0.00	0.00	0.00	**0.00**
Izotope	Unet	-0.01	-0.01	-0.01	**0.00**
Izotope	Pro	-0.01	-0.01	0.00	**0.00**
Unet	Pro	0.00	0.00	0.00	0.37
Side Effects Project - Sing with me					
Amateur	Izotope	0.00	0.01	0.01	**0.00**
Amateur	Unet	0.00	0.00	0.00	**0.01**
Amateur	Pro	0.00	0.00	0.00	0.12
Izotope	Unet	-0.01	0.00	0.00	**0.00**
Izotope	Pro	-0.01	0.00	0.00	**0.00**
Unet	Pro	0.00	0.00	0.00	0.80
Speak Softly - Broken man					
Amateur	Izotope	0.01	0.01	0.02	**0.00**
Amateur	Unet	-0.03	-0.03	-0.02	**0.00**
Amateur	Pro	-0.03	-0.03	-0.02	**0.00**
Izotope	Unet	-0.04	-0.04	-0.03	**0.00**
Izotope	Pro	-0.04	-0.04	-0.03	**0.00**
Unet	Pro	0.00	0.00	0.01	1.00
The Doppler Shift - Atrophy					
Amateur	Izotope	0.00	0.00	0.00	0.51
Amateur	Unet	-0.01	-0.01	-0.01	**0.00**
Amateur	Pro	-0.01	-0.01	-0.01	**0.00**
Izotope	Unet	-0.01	-0.01	-0.01	**0.00**
Izotope	Pro	-0.01	-0.01	-0.01	**0.00**
Unet	Pro	0.00	0.00	0.00	**0.00**
Tom McKenzie - Directions					
Amateur	Izotope	-0.04	-0.04	-0.04	**0.00**
Amateur	Unet	-0.10	-0.10	-0.10	**0.00**
Amateur	Pro	-0.09	-0.09	-0.09	**0.00**
Izotope	Unet	-0.06	-0.06	-0.06	**0.00**
Izotope	Pro	-0.06	-0.06	-0.06	**0.00**
Unet	Pro	0.01	0.01	0.01	**0.00**

Results of the statistical significance calculation results of the *Odd-to-Even Harmonic Ratio* descriptor value differences for pairs of mixes are presented in Table 6.8. As shown in this table, in the case of the *Odd-to-Even Harmonic Ratio* descriptor, the differences between mixes are rather statistically insignificant (except for a few examples). Again, all significant differences are highlighted in bold.

Table 6.8. Statistical significance calculation results of the Odd-to-Even Harmonic Ratio

Odd-to-Even Harmonic Ratio					
Samples being compared		*Lower confidence*	*Estimate*	*Upper confidence*	***p***
Angels in Amplifiers - I'm Alright					
Amateur	Izotope	-3.24	-1.10	1.03	0.55
Amateur	Unet	-2.87	-0.73	1.40	0.81
Amateur	Pro	-1.53	0.61	2.74	0.88
Izotope	Unet	-1.77	0.37	2.50	0.97
Izotope	Pro	-0.42	1.71	3.85	0.17
Unet	Pro	-0.79	1.34	3.48	0.37
Ben Carrigan - We'll talk about it tonight					
Amateur	Izotope	-0.41	0.23	0.88	0.78
Amateur	Unet	-0.77	-0.12	0.52	0.96
Amateur	Pro	-0.80	-0.16	0.48	0.92
Izotope	Unet	-1.00	-0.36	0.28	0.48
Izotope	Pro	-1.04	-0.39	0.25	0.39
Unet	Pro	-0.68	-0.03	0.61	1.00
Georgia Wonder - Siren					
Amateur	Izotope	-1.87	0.57	3.01	0.93
Amateur	Unet	-2.06	0.38	2.82	0.98
Amateur	Pro	-2.09	0.35	2.79	0.98
Izotope	Unet	-2.63	-0.19	2.25	1.00
Izotope	Pro	-2.65	-0.22	2.22	1.00
Unet	Pro	-2.46	-0.02	2.41	1.00
Secretariat - Over the top					
Amateur	Izotope	-8.63	-1.49	5.65	0.95
Amateur	Unet	-30.58	-23.44	-16.30	**0.00**
Amateur	Pro	-29.32	-22.18	-15.04	**0.00**
Izotope	Unet	-29.09	-21.95	-14.81	**0.00**
Izotope	Pro	-27.83	-20.69	-13.55	**0.00**
Unet	Pro	-5.88	1.26	8.40	0.97
Side Effects Project - Sing with me					
Amateur	Izotope	-5.50	-0.72	4.05	0.98
Amateur	Unet	-6.10	-1.32	3.45	0.89
Amateur	Pro	-7.78	-3.01	1.77	0.37
Izotope	Unet	-5.37	-0.60	4.17	0.99
Izotope	Pro	-7.05	-2.28	2.49	0.61
Unet	Pro	-6.46	-1.68	3.09	0.80
Speak Softly - Broken man					
Amateur	Izotope	10.00	18.64	27.29	**0.00**

Amateur	Unet	9.56	18.21	26.86	**0.00**
Amateur	Pro	8.49	17.13	25.78	**0.00**
Izotope	Unet	-9.08	-0.44	8.21	1.00
Izotope	Pro	-10.16	-1.51	7.14	0.97
Unet	Pro	-9.72	-1.07	7.57	0.99
The Doppler Shift - Atrophy					
Amateur	Izotope	-4.82	-0.93	2.97	0.93
Amateur	Unet	-7.20	-3.31	0.58	0.13
Amateur	Pro	-11.30	-7.41	-3.51	**0.00**
Izotope	Unet	-6.28	-2.38	1.51	0.40
Izotope	Pro	-10.37	-6.48	-2.59	**0.00**
Unet	Pro	-7.99	-4.10	-0.21	**0.03**
Tom McKenzie - Directions					
Amateur	Izotope	-16.20	-3.09	10.02	0.93
Amateur	Unet	-9.34	3.77	16.87	0.88
Amateur	Pro	-19.85	-6.74	6.36	0.55
Izotope	Unet	-6.25	6.86	19.96	0.54
Izotope	Pro	-16.76	-3.65	9.45	0.89
Unet	Pro	-23.62	-10.51	2.60	0.17

From tables 6.2-6.5, an interesting observation can be made that the professional mixes, although created by various people, are characterized by the smallest standard deviations and variances in RMS, *Integrated Loudness*, *Loudness Range*, and *True peak* level values. For the “Amateur” and “Izotope” mixes, however, the opposite is true. Professional mixing engineers appear to be more consistent in their mixing (in loudness and dynamic range), regardless of the music genre being mixed. The results for the “Unet” mixes were close to those for “Pro” mixes, which indicates that this type of mix resembles professional mixing the most.

As shown in Tables 6.6-6.8, two descriptors (RMS-*Energy Envelope* and *Harmonic Energy*) show statistically significant differences between mixes.

Considering all above results, it can be concluded that the “Unet” mixes are the closest to the “Pro” mixes and the developed system (described in chapter 4) is capable of creating a mix that can be objectively rated as professional or close to professional. Moreover, it can be concluded that the system produces mixes better than amateur mixes and better than mixes created by the well-known state-of-the-art method. The conclusions prove thesis no. 1, i.e., **“It is possible to mix music consisting of separate raw recordings using a one-dimensional adaptation of the Wave-U-Net autoencoder that can objectively be evaluated similarly to a professional mix.”**

In the following Chapter 6.3, the subjective tests are described that were conducted to prove thesis no. 2. The comparison between the subjective and objective tests is presented in Chapter 6.4.

6.3. Subjective evaluation

6.3.1. Listening test

After adequate postprocessing of samples (described in Chapter 5.5), the listeners were asked to fill out a questionnaire and give their subjective rates for each acquired 32 samples. The questionnaire given to the listeners is attached in Appendix C. The rating of samples was conducted in line with the methodology of the rating procedure [147] using a five-point scale (1-5). The subjective test results, including their analysis, are presented in Chapter 6.3.2.

The listeners performed the listening test in the R1 laboratory (mixing room) at the Hamburg University of Applied Sciences. The room is adapted to professional listening and is equipped with multiple pairs of audio monitors. In this case, it was decided to use the "main speakers" pair, i.e., Klein+Hummel 0410. Nuendo 10 software and Audient ASP 8024 mixing console were used for the listening session. All effects on the console were turned off and all faders were set to the unity position. On the same console, the routing of individual channels to subgroups in the middle of the console was performed. All samples were played simultaneously from the DAW and the listeners could freely switch between the different mixes. The system calibration was set to 85db SPL and was performed with the use of the Bruel & Kjaer Precision 732A meter. For the calibration, pink noise correlated to the listening files (i.e., normalized to the -14 LUFS level) was used. The chosen level may seem relatively high for a regular user, but due to the expert character of the testing process and to make the identification of the most minute details possible, the selected level was appropriate. The loudness level is also recommended by the Audio Engineering Society [96].

During the listening sessions, the expert listeners were able to switch between the different mixes in any order and marked their ratings in the questionnaire. The listeners were taking part in the sessions individually. The test was constructed in such a way that each person received samples in a different order – the trial was fully randomized, and there was no possibility for the listener to lean into a specific answer due to the testing samples' order. Every listener was familiar with operating the console and was asked if they understood all questions included in the questionnaire. Due to the fact that the audio jargon used by professional audio engineers may differ in various areas of the world, the author included definitions next to each expression (e.g., balance). Different expressions (mix-defining characteristics to be subjectively rated) are presented in Appendix C. A single listening test session lasted for approximately one hour. The results and their statistical analysis are presented in the following Chapter.

6.3.2. Analysis of the test results

After the subjective tests were completed, a statistical analysis of the results was performed. There were 20 participants in the tests; all of them were students of the Music Production Class and Digital Sound Masters Program at the Hamburg University of Applied Sciences. Among all the subjects, 16 were men, whereas four were women. The average age of participants was 26.9 and the standard deviation of age equaled 4.39. All the participants

confirmed that they listen to music. Music genres that the participants listened to varied, but the most frequent responses were rock, alternative, hip-hop, and jazz. The majority of listeners answered that they were familiar with genres such as rock, pop, alternative, and electronica. 85% of the listeners were musicians, and 60% were also mixing engineers. Presented in Fig. 6.9 are the listeners' years of experience in music mixing.

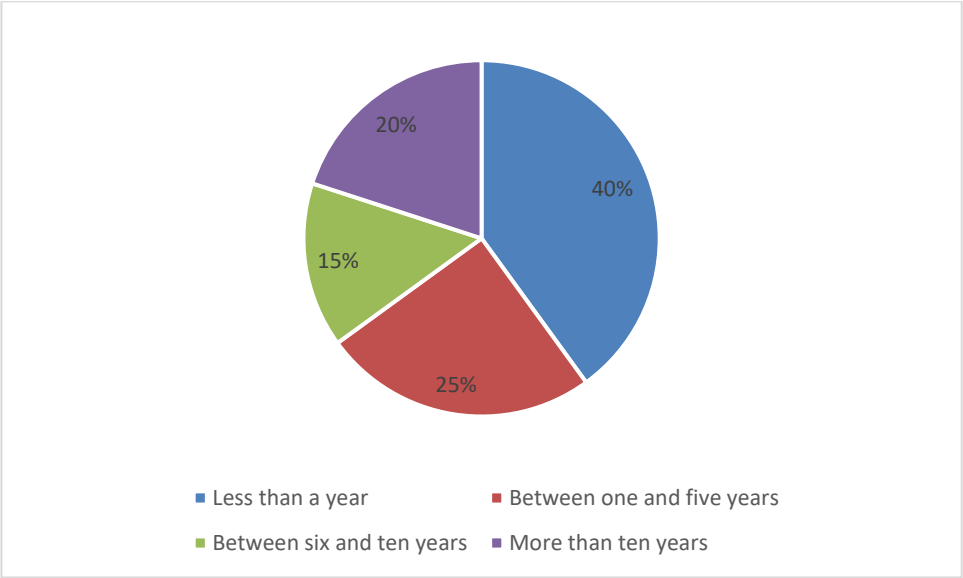


Fig. 6.9. Results of the survey in which the subjects were asked how many years of experience they have in music mixing

To answer the hypotheses, statistical analyses were performed using the IBM SPSS Statistics 25 software. The software was used to calculate the analysis of basic descriptive statistics, the Shapiro-Wilk test of normality, a series of one-way analyses of variance (abbr. one-way ANOVA) for dependent samples, and the linear correlation analysis with the use of the Pearson correlation coefficient (r). The level of significance was assumed to be $\propto = 0{,}05$. Results whose significance was at the level of $0{,}05 < p < 0{,}1$ were assumed to be statistically significant at the level of the statistical trend.

To check whether the assumption about the compliance of the distributions of the measured quantitative variables with the normal distribution has been met, first, the analysis of basic descriptive statistics with the Shapiro-Wilk test was conducted. The test result was statistically significant for a part of the variables (in bold). This means that their distribution deviates from the normal curve with statistical significance. However, the skew value for all variables does not exceed the agreed absolute value of 2, which indicates that the distributions are not extremely asymmetrical to the normal curve even when the normality test result is statistically significant [31]. Due to the above, if the other assumptions are met, parametric tests are to be performed. The basic descriptive statistics, including the Shapiro-Wilk test results, are presented in Tables 6.9-6.18.

Table 6.9. Basic descriptive statistics and Shapiro-Wilk test results for the overall ratings of mixes and the listeners' years of experience in mixing

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Overall rating: Amateur	2.67	2.53	0.47	1.17	1.16	2.05	3.85	0.90	**0.048**
Overall rating: Izotope	2.62	2.63	0.55	0.58	1.36	1.63	4.05	0.97	0.675
Overall rating: Unet	3.58	3.76	0.59	-1.16	0.92	2.05	4.30	0.89	**0.024**
Overall rating: Pro	4.10	4.28	0.54	-0.94	0.87	2.83	5.00	0.91	**0.054**
Years of experience in mixing	4.10	2.00	5.40	1.61	2.65	0.00	20.00	0.79	**<0.001**

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.10. Basic descriptive statistics and Shapiro-Wilk test results for the measured indicators of the Amateur-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Balance Amateur	2.66	2.50	0.54	0.88	0.24	1.88	3.88	0.93	0.122
Clarity Amateur	2.64	2.69	0.54	0.26	-0.75	1.75	3.63	0.95	0.394
Panning Amateur	2.88	2.81	0.50	0.37	-0.20	2.13	4.00	0.97	0.796
Space Amateur	2.64	2.56	0.59	1.58	3.17	1.88	4.38	0.87	**0.010**
Dynamics Amateur	2.51	2.44	0.56	0.35	0.03	1.63	3.75	0.96	0.534

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.11. Basic descriptive statistics and Shapiro-Wilk test results for the measured indicators of the Izotope-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Balance Izotope	2.58	2.56	0.63	0.92	2.24	1.63	4.38	0.92	0.119
Clarity Izotope	2.76	2.81	0.69	0.12	-0.17	1.50	4.25	0.98	0.929
Panning Izotope	2.67	2.63	0.60	0.98	1.45	1.75	4.25	0.92	0.118
Space Izotope	2.54	2.50	0.60	-0.09	-0.07	1.38	3.75	0.98	0.890
Dynamics Izotope	2.56	2.44	0.54	0.22	-0.24	1.50	3.63	0.98	0.927

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.12. Basic descriptive statistics and Shapiro-Wilk test results for the measured indicators of the Unet-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Balance Unet	3.46	3.69	0.79	-1.84	3.28	1.25	4.25	0.79	**0.001**
Clarity Unet	3.49	3.50	0.54	-0.25	-0.76	2.50	4.38	0.97	0.677
Panning Unet	3.71	3.94	0.66	-1.30	1.03	2.00	4.50	0.85	**0.006**
Space Unet	3.58	3.75	0.65	-0.86	0.20	2.13	4.63	0.91	**0.073**
Dynamics_Unet	3.66	3.75	0.62	-0.73	-0.33	2.38	4.50	0.93	0.130

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.13. Basic descriptive statistics and Shapiro-Wilk test results for the measured indicators of the Pro-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Balance Pro	4.08	4.25	0.51	-0.77	1.55	2.75	5.00	0.93	0.132
Clarity Pro	4.04	4.13	0.59	-0.85	0.57	2.63	5.00	0.94	0.227
Panning Pro	4.14	4.13	0.63	-0.76	0.43	2.63	5.00	0.94	0.282
Space Pro	4.11	4.13	0.65	-1.43	2.86	2.25	5.00	0.88	**0.019**
Dynamics Pro	4.14	4.31	0.56	-1.17	1.37	2.75	5.00	0.90	**0.033**

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.14. Basic descriptive statistics and Shapiro-Wilk test results for the overall ratings of mixes in each music genre

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Overall rating: Pop	3.15	3.09	0.35	0.76	1.30	2.55	4.03	0.94	0.243
Overall rating: Alternative	3.18	3.14	0.30	0.62	1.71	2.58	3.95	0.95	0.440
Overall rating: Electronica	3.32	3.38	0.32	-0.65	0.86	2.53	3.93	0.97	0.657
Overall rating: Rock	3.32	3.33	0.39	-0.03	0.56	2.45	4.08	0.98	0.879

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.15. Basic descriptive statistics and Shapiro-Wilk test results for the ratings of music genres in the Amateur-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Pop	2.52	2.20	0.59	0.73	-0.73	1.80	3.80	0.88	**0.014**
Alternative	2.61	2.45	0.56	1.23	1.34	2.00	4.10	0.89	**0.025**
Electronica	2.61	2.65	0.46	-0.31	0.13	1.60	3.50	0.98	0.974
Rock	2.93	2.90	0.61	0.72	1.16	1.80	4.40	0.95	0.342

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.16. Basic descriptive statistics and Shapiro-Wilk test results for the ratings of music genres in the Izotope-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Pop	2.48	2.45	0.74	0.67	0.86	1.30	4.30	0.95	0.425
Alternative	2.45	2.45	0.47	0.11	-0.95	1.70	3.30	0.97	0.654
Electronica	2.87	2.80	0.71	0.42	-0.84	1.90	4.30	0.95	0.350
Rock	2.70	2.60	0.72	0.34	-0.06	1.50	4.30	0.97	0.815

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* –Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.17. Basic descriptive statistics and Shapiro-Wilk test results for the ratings of music genres in the Unet-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*P*
Pop	3.63	3.75	0.61	-0.49	0.00	2.30	4.70	0.97	0.707
Alternative	3.59	3.75	0.58	-0.90	-0.09	2.30	4.30	0.91	**0.052**
Electronica	3.59	3.75	0.71	-0.81	0.38	1.80	4.50	0.93	0.142
Rock	3.50	3.60	0.76	-0.69	0.13	1.80	4.70	0.93	0.174

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

Table 6.18. Basic descriptive statistics and Shapiro-Wilk test results for the ratings of music genres in the Pro-based mix

	M	*Mdn*	*SD*	*Sk.*	*Kurt.*	*Min.*	*Max.*	*W*	*p*
Pop	4.00	4.05	0.58	-0.55	-0.23	2.90	5.00	0.94	0.196
Alternative	4.08	4.20	0.65	-1.02	0.66	2.50	5.00	0.92	0.105

Electronica	4.19	4.20	0.64	-0.77	0.36	2.70	5.00	0.93	0.147
Rock	4.15	4.20	0.65	-0.88	1.24	2.40	5.00	0.93	0.166

M – mean; *Mdn* – median; *SD* – standard deviation; *Sk.* – skew*; Kurt.* – kurtosis; *Min. and Max. – minimum and maximum values in the distribution*; *W* – Shapiro-Wilk test statistic; *p* – statistical significance

As part of the first of the research questions, it was decided to check if the types of mixes ("Amateur," "Izotope," "Unet," and "Pro") differ in how the respondents rated them. For this purpose, conducted was a series of one-way analyses of variance for dependent samples, and individual mixes were compared in the following categories: overall rating, balance, clarity, panning, space, and dynamics.

First, an analysis of the overall rating of mixes was executed. The result is statistically significant, and the effect size coefficient indicates strong differences. The pairwise comparisons with the Šidák correction demonstrated that the "Pro" mixes were rated the highest by the respondents, followed by "Unet". The "Amateur" and "Izotope" mixes were rated the lowest without a significant difference in ratings between them.

A visual representation of the results is shown in Fig. 6.10.

Table 6.19. The overall rating of the mix as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Overall rating	2.67_a	0.47	2.62_a	0.55	3.58_b	0.59	4.10_c	0.54	39.09	<0.001	0.67

Note: The means that do not share the letter index differ from each other at a $p < 0.05$ level—pairwise comparisons with the Šidák correction.

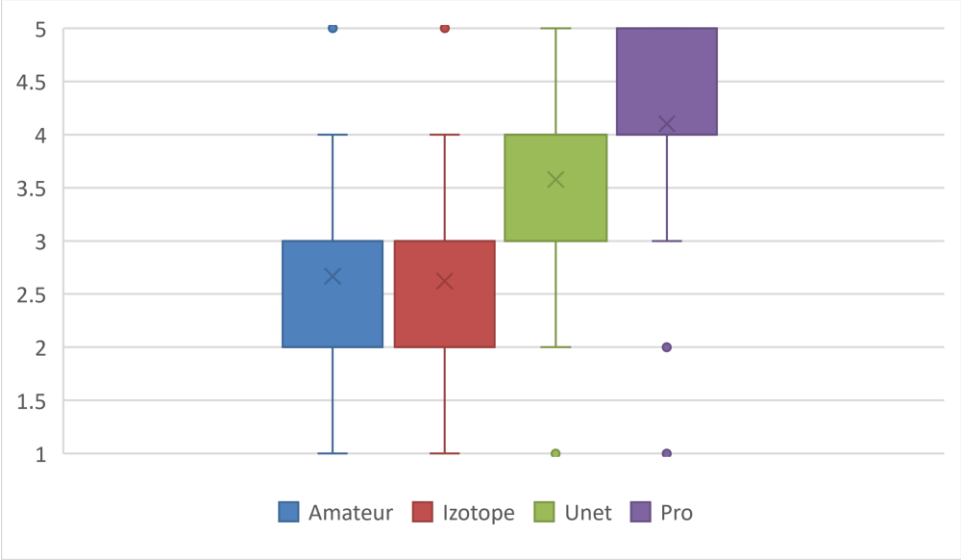


Fig. 6.10. Box plot showing the distribution of the overall ratings for the "Amateur", "Izotope", "Unet" and "Pro" mixes

Next, the mixes were compared within the balance category. The result was statistically significant and the η^2 value signified strong differences. The pairwise comparisons with the Šidák correction demonstrated that the highest-rated mixes in the balance category were the

"Pro" mixes, followed by the "Unet" mixes. The lowest-rated mixes were "Amateur" and "Izotope," without any significant differences in results between them.

A visual representation of the results is shown in Fig. 6.11.

Table 6.20. Balance as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Balance	2.66_a	0.54	2.58_a	0.63	3.46_b	0.79	4.08_c	0.51	27.62	<0.001	0.59

Note: The means that do not share the letter index differ from each other at a $p < 0.05$ level—pairwise comparisons with the Šidák correction.

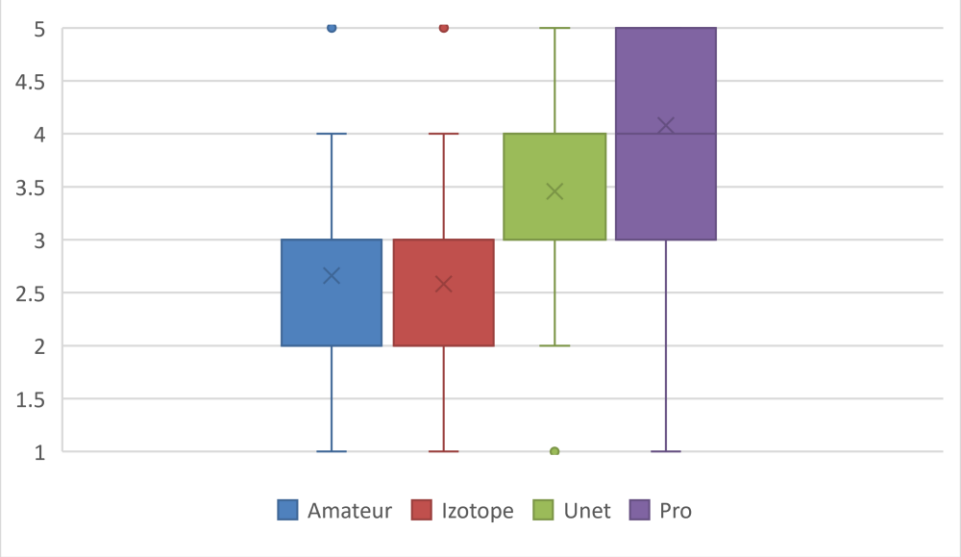


Fig. 6.11. Box plot showing the distribution of the Balance ratings for the "Amateur", "Izotope", "Unet", and "Pro" mixes

An analogous analysis was conducted with the use of the clarity variable. The analysis results show very strong and statistically significant differences, and the pairwise comparisons with the Šidák correction show that the highest-rated mixes in the clarity category were the "Pro" mixes, followed by the "Unet" mixes. The lowest-rated mixes were "Amateur" and "Izotope," without any significant differences in their results.

A visual representation of the results is shown in Fig. 6.12.

Table 6.21. Clarity as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Clarity	2.64_a	0.54	2.76_a	0.69	3.49_b	0.54	4.04_c	0.59	22.71	<0.001	0.54

Note: The means that do not share the letter index differ from each other at a $p < 0.05$ level—pairwise comparisons with the Šidák correction.

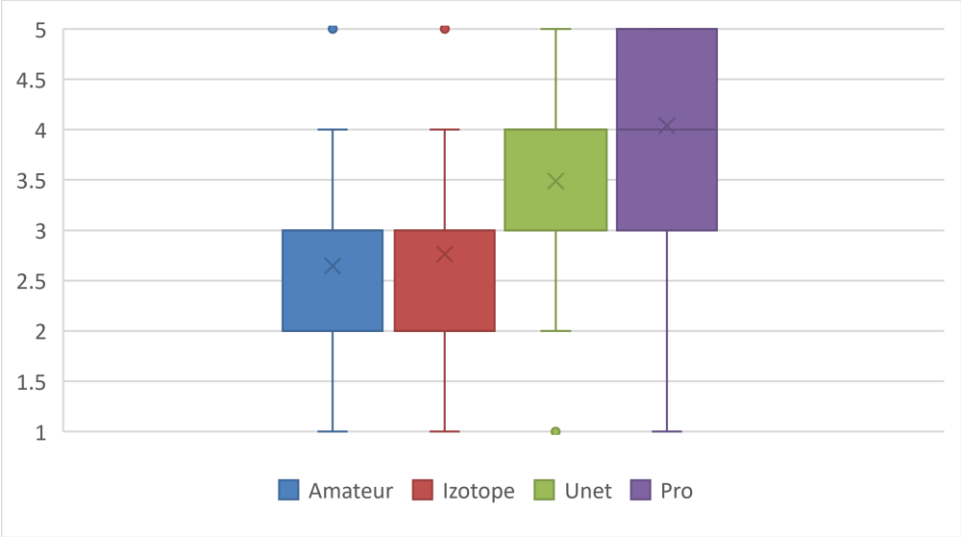


Fig. 6.12. Box plot showing the distribution of the Clarity ratings for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

The next comparison of mixes was conducted within the panning category. The analysis results show very strong and statistically significant differences, and the pairwise comparisons with the Šidák correction show that the highest-rated mixes in the panning category were the “Pro” mixes, followed by the “Unet” mixes. The lowest-rated mixes were “Amateur” and “Izotope”, without any significant differences in their results.

A visual representation of the results is shown in Fig. 6.13.

Table 6.22. Panning as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Panning	$2{,}88_a$	0,50	$2{,}67_a$	0,60	$3{,}71_b$	0,66	$4{,}14_c$	0,63	27,24	<0,001	0,59

Note: The means that do not share the letter index differ from each other at a $p < 0{,}05$ level—pairwise comparisons with the Šidák correction.

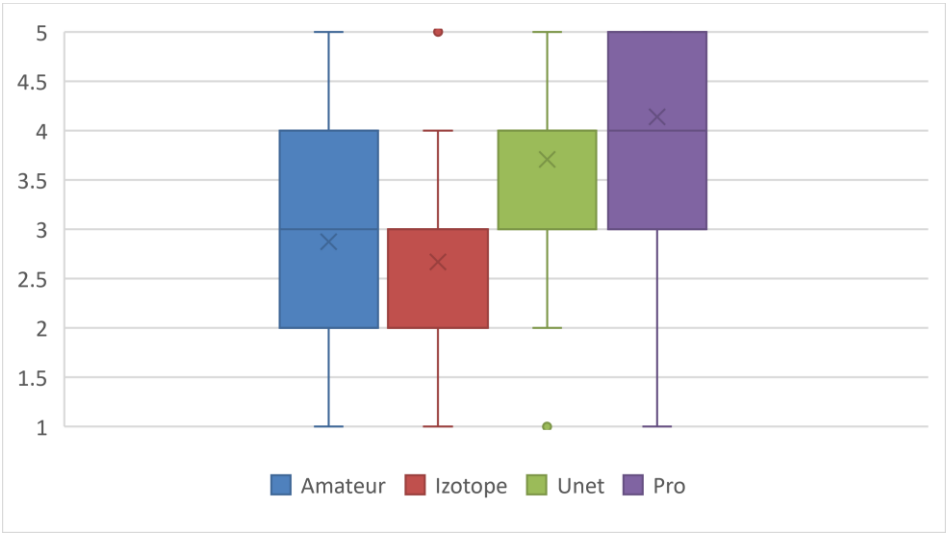


Fig. 6.13. Box plot showing the distribution of the Panning ratings for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

Next, the mixes were compared using the space variable. The results, as in the previous analyses, proved very strong and statistically significant differences between the types of mixes. The pairwise comparisons with the Šidák correction proved the “Pro” mixes to be the highest-rated mixes in the space category, followed by “Unet”. The “Amateur” and “Izotope” mixes were rated the lowest, with no significant difference between them.

A visual representation of the results is shown in Fig. 6.14.

Table 6.23. Space as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Space	2.64_a	0.59	2.54_a	0.60	3.58b	0.65	4.11c	0.65	33.40	<0.001	0.64

Note: The means that do not share the letter index differ from each other at a $p < 0,05$ level—pairwise comparisons with the Šidák correction.

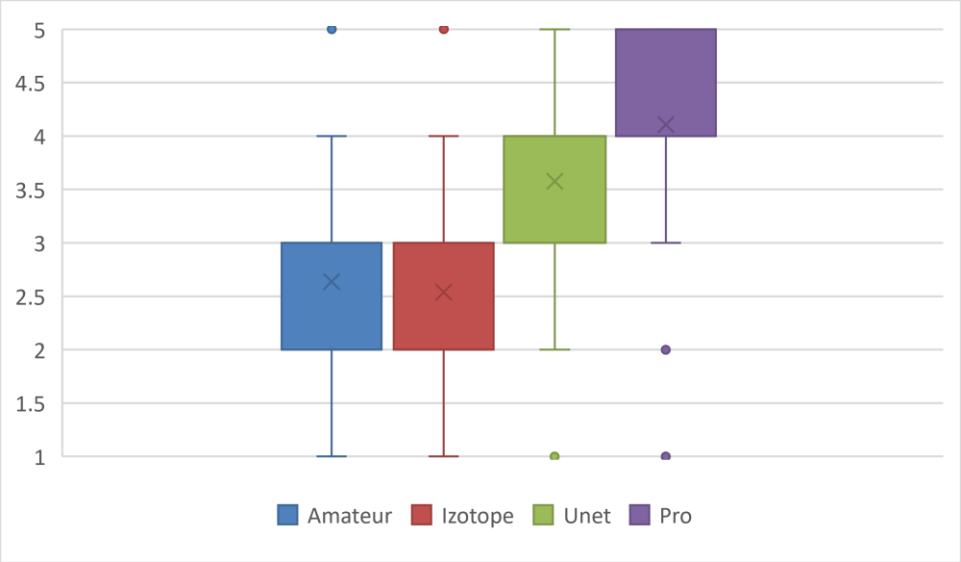


Fig. 6.14. Box plot showing the distribution of the Space ratings for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

The last variable used for the comparison of mix types was dynamics. Analogously to the previous analyses, the results showed very strong and statistically significant differences. The pairwise comparisons with the Šidák correction proved the “Pro” mixes to be the highest-rated mixes in terms of dynamics, followed by “Unet”. The “Amateur” and “Izotope” mixes were rated the lowest by respondents, with no significant difference between them.

A visual representation of the results is shown in Fig. 6.15.

Table 6.24. Dynamics as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Dynamics	2.51_a	0.56	2.56_a	0.54	3.66_b	0.62	4.14_c	0.56	45.38	<0.001	0.70

Note: The means that do not share the letter index differ from each other at a $p < 0,05$ level—pairwise comparisons with the Šidák correction.

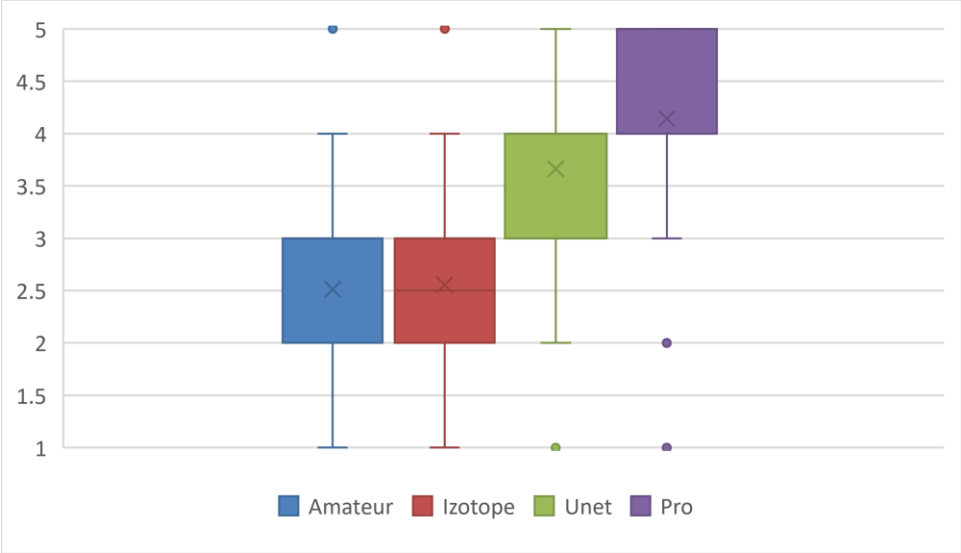


Fig. 6.15. Box plot showing the distribution of the Dynamics ratings for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

In the next step, it was decided to check if the mix types varied in terms of overall ratings within different music genres. For this purpose, a series of one-way analyses of variance for dependent samples was used.

First, the mixes were compared in the Pop category. The result indicates very strong and statistically important differences, and the pairwise comparisons with the Šidák correction show that the respondents rated the “Pro” mixes the highest in this category, followed by the “Unet” mixes. The mixes rated the lowest were the “Amateur” and “Izotope” mixes, with no significant differences between them.

A visual representation of the results is shown in Fig. 6.16.

Table 6.25. Overall rating of the Pop mixes as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Overall rating in Pop	2.52_a	0.59	2.48_a	0.74	3.63_b	0.61	4.00_c	0.58	32.06	<0.001	0.63

Note: The means that do not share the letter index differ from each other at a $p < 0{,}05$ level—pairwise comparisons with the Šidák correction.

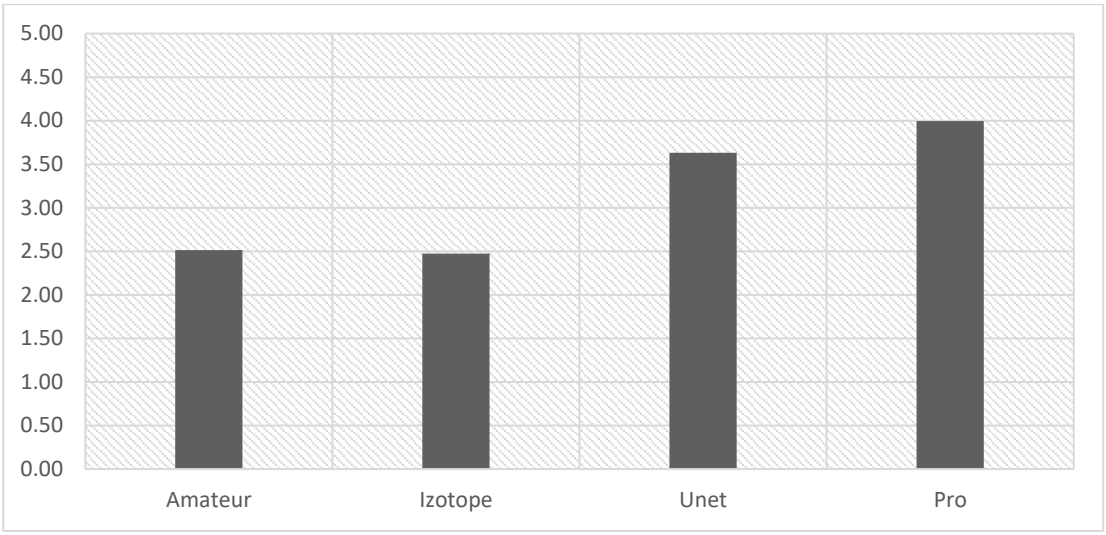


Fig. 6.16. Average overall ratings of mixes in the Pop genre for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

Next, the same analysis was conducted for the Alternative category. Again, the result was statistically significant and the η^2 value indicates very strong differences. The pairwise comparisons with the Šidák correction show that the respondents rated the “Pro” mixes the highest in the Alternative category, followed by the “Unet” mixes. The mixes rated the lowest were the “Amateur” and “Izotope” mixes, with no significant differences between them.

A visual representation of the results is shown in Fig. 6.17.

Table 6.26. Overall rating of the Alternative mixes as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Overall rating in Alternative	2.61_a	0.56	2.45_a	0.47	3.59_b	0.58	4.08_c	0.65	39.07	<0.001	0.67

Note: The means that do not share the letter index differ from each other at a $p < 0{,}05$ level—pairwise comparisons with the Šidák correction.

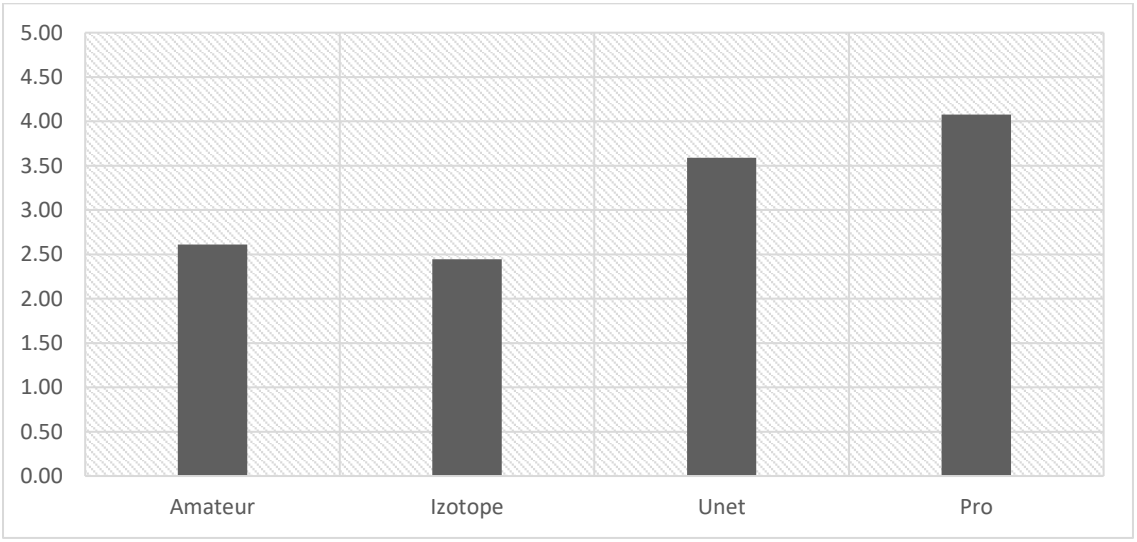


Fig. 6.17. Average overall ratings of mixes in the Alternative genre for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

Further on, the same analysis was repeated for the Electronica genre. The highest rate was achieved by the “Pro” mixes, next the “Unet” mixes, while the “Amateur” and “Izotope” mixes had the lowest rates. The “Amateur” and “Izotope” mixes showed no differences between their ratings, the “Unet” and “Izotope” mixes differ with statistical significance, whereas the “Pro” mixes differ from other types with statistical significance.

A visual representation of the results is shown in Fig. 6.18.

Table 6.27. Overall rating of the Electronica mixes as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Overall rating in Electronica	2.61_a	0.46	2.87_a	0.71	3.59	0.71	4.19_c	0.64	25.57	<0.001	0.57

Note: The means that do not share the letter index differ from each other at a $p < 0,05$ or $p < 0,1$ level—pairwise comparisons with the Šidák correction.

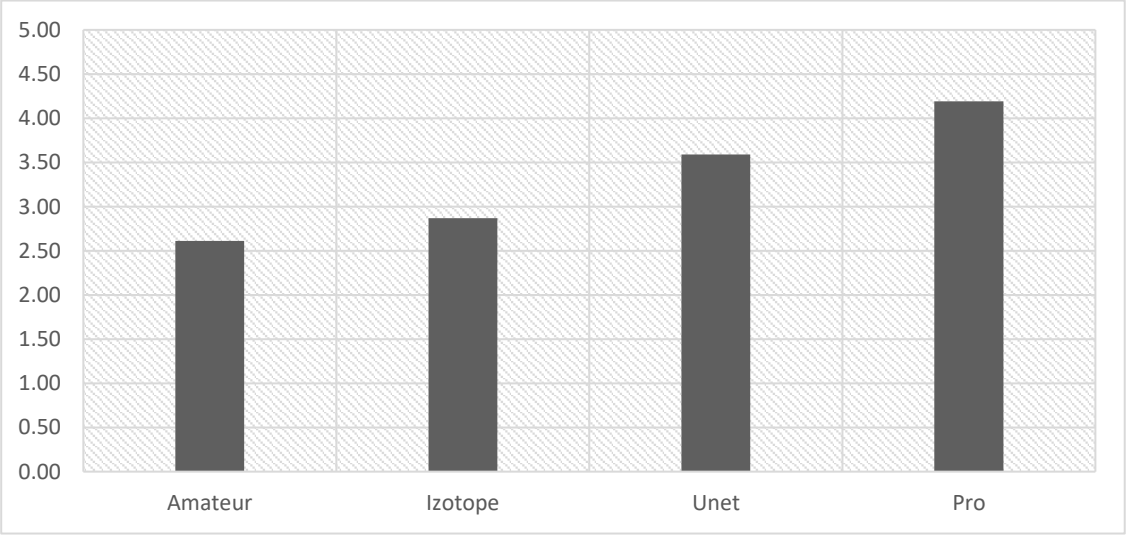


Fig. 6.18. Average overall ratings of mixes in the Electronica genre for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

As part of the last analysis of variance, differences in ratings of mixes in the Rock genre were examined. The pairwise comparisons with the Šidák correction revealed that the “Pro” mixes were rated the highest, followed by “Unet”. The “Amateur” and “Izotope” mixes were rated the lowest with no differences between them. The “Unet” and “Amateur” mixes differ on a statistical significance level.

A visual representation of the results is shown in Fig. 6.19.

Table 6.28. Overall rating of Rock mixes as a function of the mix type

	Amateur		Izotope		Unet		Pro		*F*	*p*	η^2
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*			
Overall rating in Rock	2.93_a	0.61	2.70_a	0.72	3.50_b	0.76	4.15_c	0.65	19.51	<0.001	0.51

Note: The means that do not share the letter index differ from each other at a $p < 0,05$ or $p < 0,1$ level—pairwise comparisons with the Šidák correction.

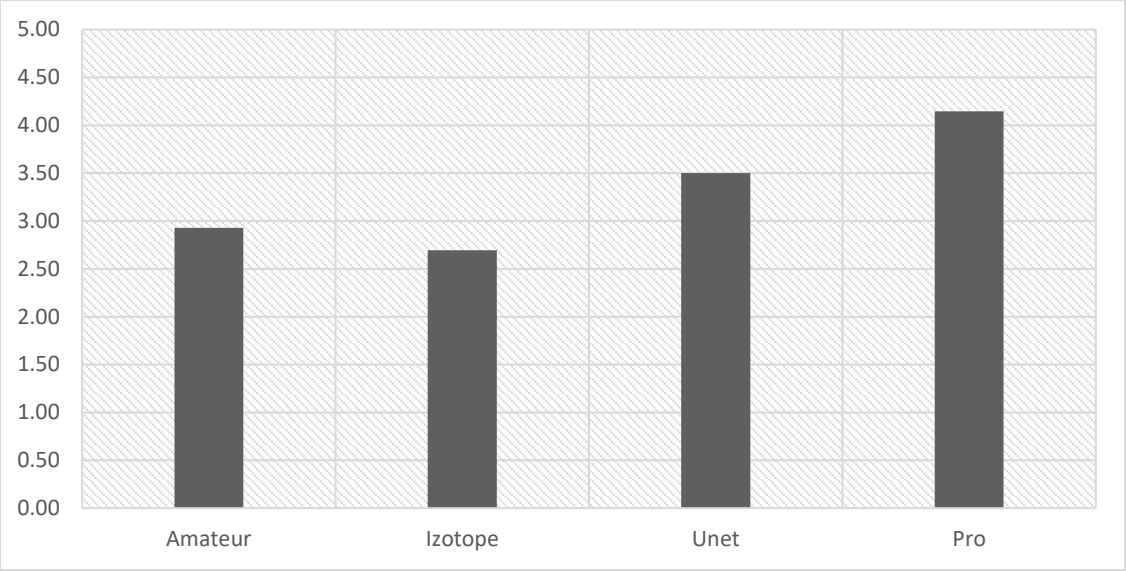


Fig. 6.19. Average overall ratings of mixes in the Rock genre for the “Amateur”, “Izotope”, “Unet”, and “Pro” mixes

The last step of the analysis encompassed examining the correlation between respondents’ experience in mixing and their overall ratings of each mix type. For this purpose, correlation analysis with the use of the Pearson correlation coefficient *r* was conducted.

The analysis proved a statistically significant correlation between the number of years of experience in mixing with the rating of “Amateur” and “Pro” mixes and a correlation at a level of statistical significance for the “Unet” mixes.

The negative value of the *r* coefficient for the correlation of experience and ratings of the “Izotope” and “Amateur” mixes means that the more years of experience the listeners have, the lower they rate the mixes. In the case of the “Unet” and “Pro” mixes, the correlation is positive, and it is either moderately strong or strong, which means that when the number of years of experience in mixing grows, the overall rating of those mixes increases.

Table 6.29. Correlation between the experience in mixing and the overall ratings of mixes

		Experience in mixing
Amateur	Pearson’s *r*	-0.31
	Significance	0.186
Izotope	Pearson’s *r*	-0.52
	Significance	0.018
Unet	Pearson’s *r*	0.38
	Significance	0.098
Pro	Pearson’s *r*	0.69
	Significance	0.001

6.4. Discussion

After testing and analyzing the objective and subjective samples from each mix, self-similarity matrices (SSM) based on chromagrams were constructed. The graphical representation of the SSM of “Secretariat – Over the top” objective and subjective samples is given in Fig. 6.20. Figures that correspond to the remaining songs are presented in Appendix D.

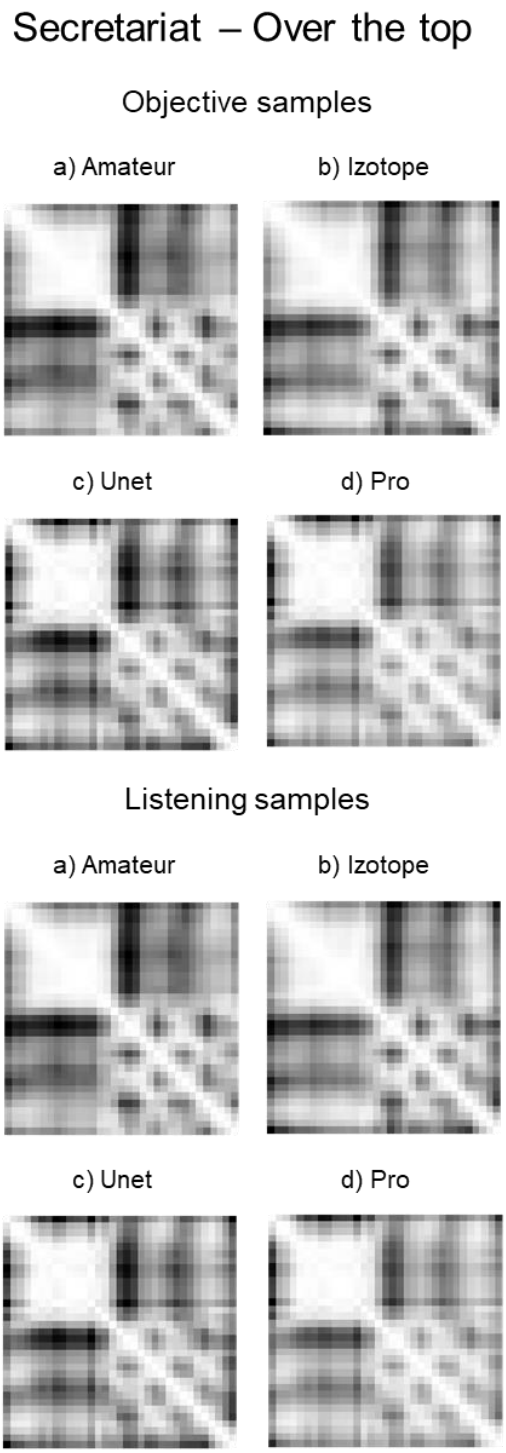


Fig. 6.20. Graphical representation of the SSM of “Secretariat – Over the top” objective and subjective samples

As already mentioned, each pixel in the matrix obtains a greyscale value corresponding to the given similarity score. The darkest color refers to the smallest similarity.

Next, all matrices were compared to each other using the Root Mean Square Error (RMSE), Structural Similarity Index (SSIM), used for measuring similarity between images, and Visual Information Fidelity (VIF), treated as a full-reference image quality related to image information extracted by the human visual system.

The results obtained are presented in Tables 6.30-6.32. Values of means in bold in Table 6.30 signify the smallest differences between mixes. The values obtained indicate that the "Pro" and "Unet" mixes are the most similar. In contrast, means highlighted in bold in Tabs. 6.31-6.32 are the biggest.

Table 6.30. Root Mean Square Error (RMSE) calculation for all samples

Song	Objective samples			Listening samples		
	Pro/Unet	Pro/Izotope	Pro/Amator	Pro/Unet	Pro/Izotope	Pro/Amator
Angels in Amplifiers - I'm Alright	2.75	16.54	9.00	2.75	16.53	8.99
Ben Carrigan - We'll talk about it all tonight	9.09	16.40	16.68	9.11	16.40	16.68
Georgia Wonder - Siren	4.05	16.91	16.97	4.05	16.91	16.97
Secretariat - Over the top	11.92	16.94	18.83	11.90	17.03	18.89
Side Effects Project - Sing with me	20.22	27.27	19.88	20.23	27.27	19.87
Speak Softly - Broken man	4.23	27.33	23.44	4.23	27.33	23.44
The Doppler Shift - Atrophy	3.57	9.58	12.19	3.57	9.58	12.17
Tom McKenzie - Directions	1.71	18.40	13.30	1.71	18.40	13.31
Mean	**7.19**	18.67	16.29	**7.19**	18.68	16.29

Table 6.31. Structural similarity index (SSIM) calculation for all samples

Song	Objective samples			Listening samples		
	Pro/Unet	Pro/Izotope	Pro/Amator	Pro/Unet	Pro/Izotope	Pro/Amator
Angels in Amplifiers - I'm Alright	0.9977	0.9433	0.9799	0.9977	0.9433	0.9800
Ben Carrigan - We'll talk about it all tonight	0.9895	0.9441	0.9337	0.9895	0.9441	0.9338
Georgia Wonder - Siren	0.9967	0.9521	0.9343	0.9967	0.9521	0.9343
Secretariat - Over	0.9898	0.9360	0.9386	0.9898	0.9359	0.9386

the top						
Side Effects Project - Sing with me	0.8638	0.8470	0.8548	0.8638	0.8471	0.8548
Speak Softly - Broken man	0.9945	0.8777	0.8878	0.9945	0.8777	0.8878
The Doppler Shift - Atrophy	0.9947	0.9437	0.9452	0.9947	0.9437	0.9454
Tom McKenzie - Directions	0.9985	0.9264	0.9700	0.9984	0.9265	0.9700
Mean	**0.9782**	0.9213	0.9305	**0.9781**	0.9213	0.9306

Table 6.32. Visual Information Fidelity (VIF) calculation for all samples

Song	Objective samples			Listening samples		
	Pro/Unet	Pro/Izotope	Pro/Amator	Pro/Unet	Pro/Izotope	Pro/Amator
Angels in Amplifiers - I'm Alright	0.89	0.56	0.70	0.89	0.56	0.70
Ben Carrigan - We'll talk about it all tonight	0.82	0.61	0.58	0.82	0.61	0.58
Georgia Wonder - Siren	0.93	0.65	0.61	0.93	0.65	0.61
Secretariat - Over the top	0.92	0.60	0.59	0.92	0.60	0.60
Side Effects Project - Sing with me	0.46	0.44	0.45	0.46	0.44	0.45
Speak Softly - Broken man	0.86	0.50	0.50	0.86	0.50	0.50
The Doppler Shift - Atrophy	0.87	0.63	0.64	0.87	0.63	0.64
Tom McKenzie - Directions	0.95	0.57	0.70	0.95	0.57	0.70
Mean	**0.84**	0.57	0.60	**0.84**	0.57	0.60

As the tables above demonstrate, the "Unet" mixes are the closest to the "Pro" mixes. Both the objective and subjective samples achieve similar results. An interesting example is the "Side Effects Project – Sing with me" song, which achieves similar scores regardless of the mix type ("Amateur"/"Izotope"/"Unet"). This is possibly due to the fact that the "Pro" mix, apart from the mixing itself, was edited manually, i.e., the vocal track was tuned and moved manually to achieve a better synchronization [121].

In this chapter, the results of objective and subjective tests of the obtained mixes were presented. The goal of these tests was to confirm if the system presented in Chapter 4 is capable to mix songs in different music genres in such a way that the obtained mixes could be evaluated as very good/better than amateur mixes and mixes made using known state-of-the-art methods are similar to professional mixes. The conclusions from the abovementioned tests are

unequivocally favorable for the proposed solution. The conducted analyzes allowed for answering the previously posed questions:

1. Is it possible to create a system that can automatically mix a song in a given music genre?

The very creation of the mixes used in the tests allows for a positive answer to this question. The obtained mixes are free from distortion and the models were trained to be independent of music genres.

2. Can the proposed system objectively produce satisfactory mixes?

The results unequivocally prove that the developed system, in an objective way, produces mixes that are very similar to mixes created professionally. This means that the developed system can successfully be used to mix songs in a given music genre.

3. Are the mixes produced by the proposed system rated as high in subjective tests?

The participants of the subjective tests were people with experience in mixing. The results prove that the mixes produced by the developed system are rated only slightly worse than professional mixes. Additionally, the correlation analysis allows for stating that the more experienced a person is in mixing, the more likely they are to choose the professional mix and the mixes produced by the proposed system. The listeners rated the samples in multiple subjective evaluation categories (i.e., Balance, Clarity, Panning, Space, and Dynamics). In each of these categories, the “Pro” mix ranked best, followed closely by the “Unet” mix. Additionally, it was decided to check whether the system performs better in any specific music genre than other mixes obtained, but the results prove that, regardless of the genre, the best mix was always the “Pro” mix, again − followed closely by the “Unet” mix.

Most test results are characterized by statistical significance.

Overall, the results achieved prove thesis no. 2, i.e., **“The prepared mixes may subjectively be evaluated as better ones than recordings created by an amateur engineer or mixes produced using state-of-the-art methods and can be comparable to mixes produced by a professional mixer.”**

7. SUMMARY

The last chapter summarizes the experiments performed while preparing the dissertation and the results of the proposed research theses. Additionally, presented are the directions of work development which deserve to be pointed out, briefly discussed, and explored.

The main goal of the research was to develop and test an audio file mixing system that allows creating mixes from raw audio signals in a given music genre automatically, without user intervention, which would match professionally made mixes in quality. As part of the system concept, an architecture based on a one-dimensional Wave-U-Net encoder was designed. The implemented system consists of five models that have been trained. A specially prepared MUSDB18-HQ database, which was enriched by individual tracks from the Cambridge database and five original compositions from the author, was used for training purposes.

In order to test the validity of the theses posed, multiple experiments were conducted. The first of them concerned the objective features of the obtained mixes. The developed system should automatically mix the input tracks in such a way that the mix obtained as the output will be objectively better than the state-of-the-art method and comparable to (or indistinguishable from) a mix created by a professional mixing engineer. In the dissertation, it was proven that it is possible to automatically mix input tracks provided by the user, using previously trained models, in a way that the final effect would be objectively very close to mixes prepared by a professional mixing engineer.

The research conducted for the system was divided into two parts. In the first phase of the research study, analyzed were samples that were not normalized and were subjected to objective analysis, i.e., waveform statistics (RMS level, Integrated Loudness, Loudness Range, and True peak level) and low-level MPEG-7 descriptors (Odd-to-Even Harmonic Ratio, RMS-Energy Envelope, and Harmonic Energy). Statistical analysis was conducted based on the results. The results obtained from the analysis of the developed automatic mixing system prove **thesis no. 1,** i.e., **"It is possible to mix music consisting of separate raw recordings using a one-dimensional adaptation of the Wave-U-Net autoencoder that can objectively be evaluated similarly to a professional mix."**

In the second phase, listening tests were conducted on normalized samples, where the listeners rated each sample in multiple evaluation categories (*Balance*, *Clarity*, *Panning*, *Space*, and *Dynamics*). Based on the results, a statistical analysis was performed. The results achieved prove **thesis no. 2,** i.e., **"The prepared mixes may subjectively be evaluated as better ones than recordings created by an amateur engineer or mixes produced using state-of-the-art methods and can be comparable to mixes produced by a professional mixer."**

In this dissertation, several original achievements accomplished by the author can be distinguished:

- An automatic audio file mixing method using Wave-U-Net autoencoders was proposed.
- A custom database for the model training purposes, including songs recorded by the author, was prepared.

- A methodology that includes objective and subjective evaluation and a comparison between the results obtained in the assessment process was introduced.
- A series of tests that enabled to rate the quality of the obtained mixes objectively was proposed.
- A listening test allowing for testing multiple characteristics of the obtained mixes was prepared.
- A series of experiments in the form of subjective tests where listeners rated the quality of the obtained mixes were performed.
- The correlation between objective and subjective results was derived, showing that it should be a part of the evaluation methodology.
- A method of comparing differences between the prepared mixes using the similarity matrices was proposed, which allowed for additional verification of the results of the objective and subjective quality assessment.

Overall, the methodology proposed shows the possibility of mixing audio signals of good quality automatically. This is especially important in applications designed for the game development industry, where audio quality is important; still, the whole effort is on visual effects or custom music branding. In the latter case, it may concern combining songs focused on matching the end and the beginning of tracks to be mixed. These areas are open to such findings as automatization of the audio mixing process.

Further research directions

In the extended plans of the proposed method, it is anticipated to include an additional module in the proposed system, i.e., the integration of an automatic instrument classification module at the system's input. This way, the user would not need to introduce appropriate tracks to respective inputs in the system manually. In the current form, for the system to work correctly, the user needs to assign bass tracks to the bass model, drum tracks to the drums model, etc. Automatic instrument classification is possible [7][39][57][60][100] and would improve the performance of the system in the context of the length of the process. It would also enhance the user's experience and the ease of use for beginner users who are not trained sound engineers.

The system proposed by the author could additionally be expanded by implementing an automatic detector of the song's music genre. It could be implemented at the system's input (during prediction) as well as at the output (during training). This solution would bring additional benefits during the training of models that are later used for prediction (mix). The system could acquire information about the input genre during training directly from the database; thus, the models would learn how to mix (react) songs in a given genre. During prediction, automatic genre classification would allow for better mixing of songs in different genres because the models would be trained in an analogical way. Moreover, it would be possible to mix the same sets of tracks in different music genres. The possibility of mixing a song simultaneously in multiple varying genres (where the user could choose which mix they prefer most) or combining

genres (e.g., 60% Rock and 40% Electronica) might be an interesting solution. Currently, the models are trained on a database consisting of four music genres – Pop, Alternative, Rock, Electronica. Most mixing engineers work with a few preferred music genres, so this addition to the system could be used for genres with which the engineers are not familiar.

Another proposed direction of further research and development is an additional module that could edit individual tracks. Such a module would allow synchronizing tracks with each other automatically (for example, in multitrack drum recordings) and automatically deleting (or scaling down the volume of) unwanted sounds (such as the vocalist's breathing or accidental microphone hits in between the desired signal). The module should be implemented at the system's input so that all tracks can be edited before mixing. Currently, the user needs to synchronize all tracks and edit unwanted or accidental sounds manually.

A particularly interesting direction, which is a scope separate from the research conducted within this dissertation, is an automatic recognition of rhythm and tempo. The knowledge of a given song's tempo in Beats Per Minutes (BPM) units may be crucial for using effects such as reverb or delay. If the reverb applied to a track is too long, a post-masking effect can occur. Knowing the rhythm and tempo of a song could also allow for setting a perfect delay tempo. Synchronizing the delay effect to the song's tempo is a universal procedure done by mixing engineers. For example, knowing that a song's tempo is 120 BPM, a mixer is able to set the tempo of subsequent delay's playback as quarter notes. Most popular plugins that offer a delay effect enable the user to select such an option as "1/4" or if it is necessary to provide the value in milliseconds, the user can calculate it (in the example given above, it would be 500 ms).

The proposed development directions will be the subject of research plans in the near future. It will allow for the conceptual development of the methods proposed and tested for the dissertation purposes.

REFERENCES

[1] Abdi H., “The Conferonni and Sidak Corrections for Multiple Comparisons”, Encyclopedia of measurements and statistics, 3, 2007.

[2] Audio Unity Group, https://www.audio-unity-group.com/andrew-scheps-on-mixing-100-in-the-box/, access: 09.11.2021.

[3] Bachem A., “Tone height and tone chroma as two different pitch qualities”, Acta Psychologica, 1950.

[4] Barbedo J. G. A., Lopes A., “A New Cognitive Model for Objective Assessment of Audio Quality”, JAES, 53, ½, pp. 22-31, 2005.

[5] Bendiksen R., “Digitale Lydeffekter”, Norwegian University of Science and Technology, 1997.

[6] Benito A. L., Reiss J. D., “Intelligent multitrack reverberation based on hinge-loss markov random fields”, Audio Engineering Society Int. Conf. (Semantic Audio), June 2017.

[7] Blaszke M., Koszewski D., “Determination of Low-Level Audio Descriptors of a Musical Instrument Sound Using Neural Network”, Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) Proceedings, 2020.

[8] Cai D., Wang W., Li M., "Incorporating Visual Information in Audio Based Self-Supervised Speaker Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1422-1435, doi: 10.1109/TASLP.2022.3162078, 2022.

[9] Cambridge-MT Patrons Podcast, https://www.patreon.com/posts/cambridge-mt-13924190, access: 19.04.2022.

[10] Cambridge-MT Patrons Podcast, https://www.patreon.com/posts/cambridge-mt-12239764, access: 19.04.2022.

[11] Cambridge-MT Patrons Podcast, https://www.patreon.com/posts/cambridge-mt-15135897, access: 19.04.2022.

[12] Cambridge-MT Patrons Podcast, https://www.patreon.com/posts/cambridge-mt-25725471, access: 19.04.2022.

[13] Bevelle M., “Compressors and limiters”, Studio Sound, pp. 32, 1977.

[14] Bromham G., “How can academic practice inform mix-craft?”, Mixing Music, Routledge, 2017.

[15] Chourdakis E.T., Reiss J. D., “Automatic control of a digital reverberation effect using hybrid models”, Audio Engineering Society 60th Int. Conf. (DREAMS), February 2016.

[16] Chourdakis E. T., Reiss J. D., “A machine learning approach to application of intelligent artificial reverberation”, J. Audio Eng. Soc., vol. 65, January/February 2017.

[17] De Man B., “Towards a better understanding of mix engineering”, PhD thesis, Queen Mary University of London, May 2017.

[18] De Man B., Mora M., Fazekas G., Reiss J. D., “The Open Multitrack Testbed”, Audio Engineering Society Convention e-Brief, Los Angeles, USA, October 2014.

[19] De Man B., Reiss J. D., Stables R., “Ten years of automatic mixing”, Proceedings of the 3rd Workshop on Intelligent Music Production, Salford, UK, September 2017.

[20] Deruty E., “Goal-oriented mixing”, in 2nd AES Workshop on Intelligent Music Production, vol. 13, 2016.

[21] Diener L., Sootla S., Branets S., Saabas A., Aichner R., Cutler R., “INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge”, arXiv:2204.05222, 2022.

[22] Discogs, https://www.discogs.com/artist/447075-Ben-Carrigan, access: 19.04.2022.

[23] Driscoll W. C., “Robustness of the ANOVA and Tukey-Kramer statistical tests”, Computers & Industrial Engineering, vol. 31, 1–2, pp. 265-268, October 1996.

[24] Dutilleux P., “Vers la machine a sculpter le son, modification en temps reel des carasteristiques frequentielles et temporelles des sons”, PhD thesis, University of Aix-Marseille II, 1991.

[25] Engel J., Resnick C., Roberts A., Dieleman S., Eck D., Simonyan K., Norouzi M., “Neural audio synthesis of musical notes with wavenet autoencoders”, 34th International Conference on Machine Learning, 2017.

[26] Everardo F., “Towards an Automated Multitrack Mixing Tool using Answer Set Programming”, Proceedings of the 14th Sound and Music Computing Conference, July 5-8, Espoo, Finland, 2017.

[27] FabFilter ProQ 3, https://www.fabfilter.com/downloads/pdf/help/ffproq3-manual.pdf, access: 15.04.2022.

[28] Foote J., “Visualizing music and audio using self-similarity”, In Proceedings of the seventh ACM international conference on Multimedia, Part 1, pp. 77-80, October 1999.

[29] FxDSP API Reference, https://fxdsp.readthedocs.io/en/latest/api/pan.html, access: 10.03.2022.

[30] Gardner W. G., “Reverberation algorithms”, In M. Kahrs and K. Brandenburg (eds), Applications of digital signal processing to audio and acoustics, Kluwer Academic Publishers, pp. 85-131, 1997.

[31] George D., Mallery P., “IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference”, Routledge Taylor & Francis Group, New York 2020.

[32] Gerzon M. A., “Unitary (energy preserving) multichannel networks with feedback”, Electron. Lett. V, 12, pp. 278-279, 1976.

[33] Gillet O., Richard G., “ENST-Drums: An Extensive Audio-Visual Database for Drum Signals Processing”, in Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR), October 2006.

[34] Gonzalez E. P., Reiss J. D., “An automatic maximum gain normalization technique with applications to audio mixing”, Audio Engineering Society Conv. 124, May 2008.

[35] Gonzalez E. P., Reiss J.D., “Automatic equalization of multichannel audio using cross-adaptive methods”, Audio Engineering Society Conv. 127, October 2009.

[36] Gonzalez E. P., Reiss J. D., “Automatic gain and fader control for live mixing”, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2009.

[37] Hafezi S., Reiss J. D., “Autonomous multitrack equalization based on masking reduction”, J. Audio Eng. Soc., vol. 63, May 2015.

[38] Haykin S. S., “Neural networks and learning machines”, 3rd ed. New York, Prentice Hall, 2009.

[39] Herrera P., Peeters G., Dubnov S., “Automatic Classification of Musical Instrument Sounds”, Journal of New Music Research 32(1), August 2010.

[40] Higham C. F., Higham D. J., “Deep Learning: and Introduction for Applied Mathematicians”, http://arxiv.org/abs/1801.05894, 2018.

[41] Hulse R., “A different way of looking at compression”, Studio Sound, 1997.

[42] International Telecommunication Union, “Algorithms to measure audio programme loudness and true-peak audio level," ITU-R BS.1770-2, 2011.

[43] Izhaki R., “Mixing Audio: Concepts, Practices and Tools”, Focal Press, 2010.

[44] Izotope Neutron 3 documentation, https://support.izotope.com/hc/en-us/articles/360051377974-Neutron-3-Help-Documentation, access: 10.02.2022.

[45] Izotope Ozone 9 documentation, https://support.izotope.com/hc/en-us/articles/360046031314-Ozone-9-Help-Documentation, access: 18.04.2022.

[46] Katayose H., Yatsui A., Goto M., “A mix-down assistant interface with reuse of examples”, Int. Conf. on Automated Production of Cross Media Content for Multi-Channel Distribution, November 2005.

[47] Katz B., “Mastering Audio: The art and the science”, Focal Press, 2007.

[48] Kim H. G., Moreau N., Sikora T., “MPEG-7 audio and beyond: audio content indexing and retrieval”, Chichester, West Sussex, England, Hoboken, NJ, USA, J. Wiley, 2005.

[49] Kleczkowski P., “Perception of Mixture of Musical Instruments with Spectral Overlap Removed”, Archives of Acoustics, 37, 3, pp. 355-363, 2012.

[50] Knees P., Schedl M., “Music Retrieval and Recommendation: A Tutorial Overview”, 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1133---1136, 2015.

[51] Kolasinski B., “A framework for automatic mixing using timbral similarity measures and genetic optimization”, Audio Engineering Society Conv. 124, May 2008.

[52] Korbicz J., Obuchowicz A., Uciński D., „Sztuczne sieci neuronowe: podstawy i zastosowania”, Warsaw, Akademicka Oficyna Wydawnicza PLJ, 1994.

[53] Korvel G., Treigys P., Kostek B., “Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network”, The Journal of the Acoustical Society of America, 149(1), 508-523, 2021.

[54] Kostek B., “Musical Data Musical Instrument Classification and Duet Analysis Employing Music Information Retrieval Techniques”, Proceedings of the IEEE 92, 4, 712-729, 2004.

[55] Kostek B., Czyzewski A., Krolikowski R., “Neural Networks Applied to Sound Localization Detection”, 110th AES Convention, Paper no. 5375, 2011.

[56] Kostek B., Kupryjanow A., Zwan P., Jiang W., Ras Z., Wojnarski M., Swietlicka J., Report of the ISMIS 2011 Contest: Music Information Retrieval, Foundations of Intelligent Systems, Springer Verlag, Berlin, Heidelberg, 715–724., 2011, pp. 715–724, doi: 10.1007/978-3-642-21916-0_75.

[57] Koszewski D., Kostek B., “Musical instrument tagging using data augmentation and effective noisy data processing”, Journal of Audio Engineering Society, JAES vol. 68 Issue 1/2 pp. 57-65, January 2020.

[58] Lattner S., Grachten M., Widmer G., “Imposing higher-level structure in polyphonic music generation using convolutional restricted boltzmann machines and constraints”, Journal of Creative Music Systems, 2, pp. 1-31, 2018.

[59] Lindsay A.T., Herre J., “MPEG-7 and MPEG-7 Audio’ An Overview”, Journal of Audio Engineering Society, vol. 49, no. 7/8, pp. 589–594, 2001.

[60] Liu J., Xie L., „SVM-Based Automatic Classification of Musical Instruments”, Intelligent Computation Technology and Automation (ICICTA), vol. 3, June 2010.

[61] Lukin A., Todd J., “Adaptive time-frequency resolution for analysis and processing of audio”, AES Convention paper no. 6717, May 2006.

[62] Lund T., “Level and distortion in digital broadcasting”, EBU Technical Review, April 2007.

[63] Ma Z., Reiss J. D., Black D. A. A., “Implementation of an intelligent equalization tool using YuleWalker for music mixing and mastering”, Audio Engineering Society Conv. 134, May 2013.

[64] Malecki P., “Evaluation of objective and subjective factors of highly reverberant acoustic field”, PhD Thesis, AGH University of Science and Technology, Krakow, 2013.

[65] Mansbridge S., Finn S., Reiss J. D., “Implementation and evaluation of autonomous multi-track fader control”, Audio Engineering Society Conv. 132, April 2012.

[66] Mathworks Matlab manual, https://www.mathworks.com/help/matlab/, access: 28.10.2021.

[67] Merriam-Webster, https://www.merriam-webster.com/dictionary/automatic, access: 29.03.2022.

[68] Midside, http://www.midside.com/music/acremaker/, access: 19.04.2022.

[69] Mimilakis S. I. Cano E., Abfer J., Schuller G., “New sonorities for jazz recordings: Separation and mixing using deep neural networks”, 2nd Workshop on Intelligent Music Production, September 2016.

[70] Mimilakis S. I., Drossos K., Floros A., Katerelos D., “Automated tonal balance enhancement for audio mastering applications”, Audio Engineering Society Conv. 134, May 2013.

[71] Mimilakis S.I., Drossos K., Virtanen T., Schuller G., “Deep neural networks for dynamic range compression in mastering applications”, Audio Engineering Society Conv. 140, May 2016.

[72] “Mixing Secrets” free multitrack library, https://www.cambridge-mt.com/ms/mtk/, access: 25.04.2019.

[73] Moffat D., Sandler M., “Machine Learning Multitrack Gain Mixing of Drums”, Audio Engineering Society 147th Convention, New York, October 16-19, 2019.

[74] Moffat D., Thalmann F., Sandler M., “Towards a semantic web representation and application of audio mixing rules”, Proceedings of the 4th Workshop on Intelligent Music Production, Huddersfield, UK, 14 September 2018.

[75] Moorer J. A., “About this reverberation business”, Comp. Music J., 3, pp. 13-28, 1979.

[76] Müller M., Kurth, F., “Enhancing Similarity Matrices for Music Audio Analysis” 5. V - V. 10.1109/ICASSP.2006.1661199, 2006.

[77] Najafabadi M. M., Villanustre F., Khoshgoftaar T. M., Seliya N., Wald R., Muharemagic E., “Deep learning applications and challenges in big data analytics”, J. Big data, vol. 2, no. 1, 2015.

[78] Nielsen S., “Personal Communication”, TC Electronics A/S, 2000.

[79] Oeksound Soothe 2 manual, https://oeksound.com/manuals/soothe2/, access: 18.04.2022.

[80] Okwonu F. Z., Asaju B. L., Arunaye F. I., “Breakdown Analysis of Pearson Correlation Coefficient and Robust Correlation Methods”, International Conference on Technology, Engineering and Sciences (ICTES), 2020.

[81] Orfanidis S. J., “Introduction to Signal Processing”, Prentice Hall, 1996.

[82] Osmalskyj J., Van Droogenbroeck M., Embrechts J. J., “Performances of low-level audio classifiers for large scale music similarity”, International Conference on Systems, Signals and Image Processing, 2014.

[83] Ostertagova E., Ostertag O., “Methodology and Application of One-way ANOVA”, American Journal of Mechanical Engineering, vol. 1, no. 7, pp. 256-261, 2013.

[84] Paulus J., Müller M., Klapuri A., “State of the Art Report: Audio-Based Music Structure Analysis”, Ismir, pp. 625-636, August 2020.

[85] Pras A., Guastavino C., Lavoie M., “The impact of technological advances on recording studio practices”, Journal of the American Society for Information Science and Technology, January 2013.

[86] Pure Mix, https://www.puremix.net/video/andrew-scheps-mixing-ziggy-marley-in-the-box.html, access: 15.10.2021.

[87] Puri T., Soni M., Dhiman G., Khalaf O. I., Alazzam M., Khan I. R., "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network", Journal of Healthcare Engineering, vol. 2022, Article ID 8472947, 9 pages, https://doi.org/10.1155/2022/8472947, 2022.

[88] Python documentation, https://docs.python.org/3/, access: 18.03.2022.

[89] Rafii Z., Liutkus A., Stoter F. R., Mimilakis S. I., Bittner R., “MUSDB18-HQ – an uncompressed version of MUSDB18”, https://doi.org/10.5281/zenodo.3338373, August 2019.

[90] Ramirez M. A., Benetos E., Reiss J. D., „Deep Learning for Black-Box Modeling of Audio Effects”, Applied Sciences, vol. 10, no. 2, p. 638, 2020.

[91] Ramirez M. A., Reiss J. D., „Analysis and Prediction of the Audio Feature Space when Mixing Raw Recordings into Individual Stems”, Audio Engineering Society Convention, 143, Conv. Paper no. 9848, October 2017.

[92] Ramirez M. A., Reiss J. D., “Deep learning and intelligent audio mixing”, Proceedings of the 3rd Workshop on Intelligent Music Production, Salford, UK, September 2017.

[93] Ramirez M. A., Reiss J. D., „Modeling nonlinear audio effects with end-to-end deep

neural networks", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019.

[94] Ramirez M. A., Stoller M. A., Moffat D., "A Deep Learning Approach to Intelligent Drum Mixing With the Wave-U-Net", J. Audio Eng. Soc., vol. 69, no. 3, pp. 142–151, March, https://doi.org/10.17743/jaes.2020.0031, 2021.

[95] Ravichandran N. K., "Tamil Natural Language Voice Classification using Recurrent Neural Networks", IJRESM, vol. 5, no. 1, pp. 79–82, January 2022.

[96] "Recommendations for Loudness of Internet Audio Streaming and On-Demand Distribution", Technical Document AESTD1008.1.21-9, September 2021.

[97] Reed D., "A perceptual assistant to do sound equalization", in Proceedings of the 5th International Conference on Intelligent User Interfaces, pp. 212/218, January 2000.

[98] Reiss J. D., Ramirez M., "End-to-end equalization with Convolutional Neural Networks", Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18), Aveiro, Portugal, September 4-8, 2018.

[99] Ronan D., Ma Z., Namara P., Reiss D. J., „Automatic Minimisation of Masking in Multitrack Audio using Subgroups", https://doi.org/10.48550/arXiv.1803.09960, 2018.

[100] Rosner A., Kostek B., „Automatic music genre classification based on musical instrument track separation", Journal of Intelligent Information Systems, 50(2), pp. 363-384, 2018.

[101] Rumsey F., "Loudness revisited", JAES, 62, 12, pp. 906-910, 2014.

[102] Santos A., Rodrigues J., Folgado D., Santos S., Fujao C., Gamboa H., "Self-Similarity Matrix of Morphological Features for Motion Data Analysis in Manufacturing Scenarios", BIOSIGNALS pp. 80-90, 2021.

[103] Sarroff A., Casey M., "Groove Kernels as Rhythmic Acoustic Motif Descriptors", 14th International Society for Music Information Retrieval Conference (ISMIR), 2013.

[104] Schroeder M. R., "Improved quasi-stereophony and colorless artificial reverberation", J. Acoust. Soc. Am., 33(8), pp. 1061-1064, 1961.

[105] Schroeder M. R., "Natural-sounding artificial reverberation", J. Audio Eng. Soc., 10(3), pp. 219-223, 1962.

[106] Schroeder M. R., "Digital simulation of sound transmission in reverberant spaces", J. Acoust. Soc. Am., 47(2 Part 1), pp. 424-431, 1970.

[107] Schroeder M. R., "Computer models for concert hall acoustics", Am. J. Physics, 41, pp. 461-471, 1973.

[108] Schroeder M. R., Logan B., "Colorless artificial reverberation", J. Audio Eng. Soc., 9, pp. 192-197, 1961.

[109] Schulze-Forster K., C., Richard G., Badeau R., "Unsupervised Audio Source Separation Using Differentiable Parametric Source Models", arXiv:2201.09592, January 2022.

[110] Scott J., et al., "Automatic multi-track mixing using linear dynamical systems", 8th Sound and Music Computing Conference, July 2011.

[111] Scott J., Kim Y. E., "Analysis of acoustic features for automated multi-track mixing", 12th International Society for Music Information Retrieval Conference, October 2011.

[112] Sigtia S., Benetos E., Dixon S., “An end-to-end neural network for polyphonic piano music transcription”, IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 24, no. 5, pp. 927– 939, 2016.

[113] Silva D. F., Yeh C. M., Zhu Y., Batista G. E. A. P. A., Keogh E., “Fast Similarity Matrix Profile for Music Analysis and Exploration”, IEEE Transactions on multimedia, Vol. 14, no. 8, August, 2015.

[114] Shah A. P., Hori T., Le Roux J., Hori C., DSTC10-AVSD Submission System with Reasoning using Audio-Visual Transformers with Joint Student-Teacher Learning, The 10th Dialog System Technology Challenge Workshop at AAAI 2022.

[115] Shapiro S. S., Wilk M. B., “An Analysis of Variance Test for Normality (complete samples)”, JSTOR, Oxford University Press, vol. 52, no. 3/4, pp. 591-611, December 1965.

[116] Shepard R. N., “Circularity in judgments of relative pitch”, The journal of the acoustical society of America, 36(12), pp. 2346-2353, 1964.

[117] Shi J., Ma C., “Unsupervised Sounding Object Localization With Bottom-Up and Top-Down Attention”, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1737-1746, 2022.

[118] Shiu Y., Jeong, H., Kuo, C.C. J., “Similarity matrix processing for music structure analysis”, 10.1145/1178723.1178734, 2006.

[119] Sonible smart EQ+ manual, https://www.sonible.com/wp-content/uploads/2016/11/manual-smartEQ.pdf, access: 15.02.2022.

[120] Sonible smart:comp manual, https://www.sonible.com/wp-content/uploads/2019/07/manual-smartComp.pdf, access: 15.02.2022.

[121] Sound on sound, https://www.soundonsound.com/techniques/mix-rescue-preslin-davis, access: 19.04.2022.

[122] Sound on sound, https://www.soundonsound.com/techniques/mix-rescue-tom-mckenzie, access: 19.04.2022.

[123] Soundtheory Gullfoss manual, https://www.soundtheory.com/static/Gullfoss%20Manual.pdf, access: 18.04.2022.

[124] SoX manual, http://sox.sourceforge.net/sox.pdf, access: 18.07.2021.

[125] Steinmetz J. C., Pons J., Pascual S., Serra J., “Automatic multitrack mixing with a differentiable mixing console of neural audio effects”, arXiv:2010.10291 2020.

[126] Stoller D., Ewert S., Dixon S., “Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation”, in Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018), June 2018.

[127 Tadeusiewicz R., „Neural Networks”, Warsaw, Akademicka Oficyna Wydawnicza, 1993.

[128] TechTarget, https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network, access: 10.04.2022.

[129] Ten Myths About Normalization, https://www.hometracked.com/2008/04/20/10-myths-about-normalization/, access: 18.04.2022.

[130] Terrell M. J., Reiss J. D., “Automatic monitor mixing for live musical performance”, J.

Audio Eng. Soc., vol. 57, November 2009.

[131] Terrell M. J., Sandler M., “An offline, automatic mixing method for live music, incorporating multiple sources, loudspeakers, and room effects”, Computer Music Journal, vol. 36, May 2012.

[132] Terrell M. J., Simpson A., Sandler M., “The mathematics of mixing”, J. Audio Eng. Soc., vol. 62, January/February 2014.

[133] The MPEG-7 Standard, http://mpeg7.org/mpeg-7-standard, access: 20.03.2022.

[134] Timbre toolbox, https://github.com/mondaugen/timbretoolbox, access: 06.07.2021.

[135] Toulson R., “Can we fix it? – The consequences of ‘fixing it in the mix’ with common equalisation techniques are scientifically evaluated”, J. Art of Record Production, vol. 3, November 2008.

[136] Tukey J. W., “Exploratory data analysis”, Addison-Wesley, Reading, 1977.

[137] Unehara M., Yamada K., Shimada T., „Subjective evaluation of music with brain wave analysis for interactive music composition by IEC”, Soft Computing and Intelligent Systems (SCIS), pp. 66-70, 2014.

[138] Wakefield J.P., Dewey C., “An investigation into the efficacy of methods commonly employed by mix engineers to reduce frequency masking in the mixing of multitrack musical recordings”, 138th International AES Convention, May 2015.

[139] Wang N., Fang Y., “Music Recognition and Classification Algorithm considering Audio Emotion”, Scientific Programming, vol. 2022, Article ID 3138851, 10 pages, https://doi.org/10.1155/2022/3138851, 2022.

[140] Ward D., Reiss J. D., Athwal C., “Multitrack mixing using a model of loudness and partial loudness”, Audio Engineering Society Convention 133, October 2012.

[141] Weissenberger L., “Toward a Universal, Meta-Theoretical Framework for Music Information Classification and Retrieval”, Journal of Documentation, 71, 5, 2015.

[142] Westhausen N. L., Huber R., Baumgartner H., Sinha R., Rennies J., Meyer B. T., “Reduction of Subjective Listening Effort for TV Broadcast Signals with Recurrent Neural Networks”, https://doi.org/10.48550/arXiv.2111.01914, 2021.

[143] Wichern G., et al., “Comparison of loudness features for automatic level adjustment in mixing”, Audio Engineering Society Conv. 139, October 2015.

[144] Wilson A., Fazenda B., “An evolutionary computation approach to intelligent music production, informed by experimentally gathered domain knowledge”, 2nd Workshop on Intelligent Music Production, September 2016.

[145] Wilson A., Fazenda B.M., “Populating the Mix Space: Parametric Methods for Generating Multitrack Audio Mixtures”. Appl. Sci. 2017, 7, 1329. https://doi.org/10.3390/app7121329.

[146] Yang D., Wang H., Zou Y., Wang W., “A Two-student Learning Framework for Mixed Supervised Target Sound Detection “, arXiv:2204.02088, April 2022.

[147] Zacharov N., Huopaniemi, J., „Results of a Round Robin Subjective Evaluation of Virtual Home Theatre Sound Systems”, Proceedings of the Audio Engineering Society 107th International Convention, January 1998.

[148] Zadeh L. A., „Fuzzy logic, neural networks, and soft computing”, Commun. ACM, t.37,

no. 3, pp. 77-84, 1994.

[149] Zhu L., Rahtu E., “Visually Guided Sound Source Separation and Localization Using Self-Supervised Motion Representations”, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1289-1299, 2022.

[150] Zolzer U., “Digital Audio Signal Processing”, John Wiley & Sons, Ltd. 2nd edition, 2011.

[151] Zubayer I., Abdel-Aty M., “Real-time Emergency Vehicle Event Detection Using Audio Data”, https://doi.org/10.48550/arXiv.2202.01367, February 2022.

[152] Zwicker E., Fastl H., “Psychoacoustics: Facts and models”, Springer-Verlag, 1990.

LIST OF FIGURES

LIST OF TABLES

Appendix A: Detailed structure of selected songs

Table A.1. Structure of the “Angels in Amplifiers” song

Angels in Amplifiers – I’m alright		
No.	Track name	Mono/Stereo
1	Kick	Mono
2	Snare	Mono
3	Overheads	Stereo
4	Toms	Stereo
5	Percussion	Stereo
6	Bass	Mono
7	Piano	Mono
8	Electric guitar	Mono
9	Acoustic guitar 1	Mono
10	Acoustic guitar 2	Mono
11	Lead vocal	Mono
12	Backing vocal 1	Mono
13	Backing vocal 2	Mono

Table A.2. Structure of the “Ben Carrigan” song

Ben Carrigan – We’ll talk about it all tonight		
No.	Track name	Mono/Stereo
1	Drum machine	Mono
2	Kick	Mono
3	Kick sample	Mono
4	Snare	Mono
5	Snare sample 1	Mono
6	Snare sample 2	Mono
7	Hihat	Mono
8	Tom 1	Mono
9	Tom 2	Mono
10	Cymbal	Mono
11	Cymbal sample 1	Mono
12	Cymbal sample 2	Mono
13	Drums room	Mono
14	Shaker	Mono
15	Sleigh bells	Mono
16	Tambourine	Mono
17	Hand claps	Mono
18	Bass	Mono
19	Acoustic guitar 1	Stereo
20	Acoustic guitar 2	Mono
21	Rhythm guitar	Stereo
22	Piano	Stereo
23	Organ	Mono
24	Keyboard	Mono
25	Piano SFX	Stereo

26	Glockenspiel	Stereo
27	Violin 1	Mono
28	Violin 2	Mono
29	Viola	Mono
30	Cello	Mono
31	Strings 1	Stereo
32	Strings 2	Stereo
33	Violin sample	Stereo
34	Viola sample	Stereo
35	Cello sample 1	Stereo
36	Cello sample 2	Stereo
37	Trumpet sample	Stereo
38	French horn sample	Stereo
39	Harmonica	Mono
40	Lead vocal	Mono
41	Backing vocal 1	Mono
42	Backing vocal 2	Mono
43	Backing vocal 3	Mono
44	Backing vocal 4	Mono
45	Backing vocal 5	Mono
46	Backing vocal 6	Mono
47	Backing vocal 7	Mono
48	Backing vocal 8	Mono
49	Backing vocal 9	Mono
50	Backing vocal 10	Mono
51	Backing vocal 11	Mono

Table A.3. Structure of the “Georgia Wonder” song

No.	Track name	Mono/Stereo
1	Loop 1	Stereo
2	Loop 2	Stereo
3	Loop 3	Stereo
4	Loop 4	Stereo
5	Kick	Mono
6	Snare	Mono
7	Hihat	Stereo
8	Cymbal rolls	Stereo
9	Cymbals	Stereo
10	Bass	Mono
11	Synth 1	Stereo
12	Synth 2	Stereo
13	Synth 3	Stereo
14	Synth 4	Stereo
15	Synth 5	Stereo
16	Synth 6	Stereo
17	Synth pad 1	Stereo

18	Synth pad 2	Stereo
19	Synth pad 3	Stereo
20	Strings	Stereo
21	SFX 1	Stereo
22	SFX 2	Stereo
23	SFX 3	Stereo
24	SFX 4	Stereo
25	SFX 5	Stereo
26	SFX 6	Stereo
27	Acoustic guitar 1	Stereo
28	Acoustic guitar 2	Stereo
29	Acoustic guitar 3	Stereo
30	Acoustic guitar 4	Stereo
31	Acoustic guitar 5	Mono
32	Electric guitar 1	Stereo
33	Electric guitar 2	Stereo
34	Electric guitar 3	Mono
35	Electric guitar 4	Mono
36	Electric guitar 5	Mono
37	Electric guitar 6	Mono
38	Electric guitar 7	Mono
39	Electric guitar 8	Mono
40	Electric guitar 9	Mono
41	Electric guitar 10	Mono
42	Electric guitar 11	Mono
43	Electric guitar 12	Mono
44	Electric guitar 13	Mono
45	Electric guitar 14	Mono
46	Electric guitar 15	Mono
47	Sitar	Stereo
48	Lead vocal	Mono
49	Lead vocal doubles	Stereo
50	Backing vocal 1	Mono
51	Backing vocal 2	Mono
52	Backing vocal 3	Mono
53	Backing vocal 4	Mono
54	Backing vocal 5	Mono
55	Backing vocal 6	Mono
56	Backing vocal 7	Mono
57	Backing vocal 8	Mono
58	Backing vocal 9	Mono
59	Backing vocal 10	Mono

Table A.4. Structure of the “Secretariat” song

No.	Track name	Spatiality
1	Kick	Mono

2	Snare	Mono
3	Overheads	Stereo
4	Bass	Mono
5	Electric guitar 1	Mono
6	Electric guitar 2	Mono
7	Electric guitar 3	Mono
8	Hammond	Stereo
9	Lead vocal	Mono
10	Lead vocal doubles	Mono
11	Backing vocal	Mono

Table A.5. Structure of the “Side Effects Project” song

No.	Track name	Spatiality
1	Kick	Stereo
2	Snare 1	Stereo
3	Snare 2	Stereo
4	Clap 1	Mono
5	Clap 2	Mono
6	Clap 3	Mono
7	Clap 4	Stereo
8	Hihat 1	Mono
9	Hihat 2	Stereo
10	Hihat 3	Stereo
11	Hihat 4	Mono
12	Bongo	Stereo
13	Timbale	Stereo
14	Tom 1	Stereo
15	Tom 2	Stereo
16	Reverse Cymbal	Stereo
17	Shaker	Mono
18	Bass	Stereo
19	Harp	Stereo
20	Synth	Mono
21	Electric guitar	Stereo
22	Sample 1	Stereo
23	Sample 2	Stereo
24	Sample 3	Stereo
25	Sample 4	Stereo
26	Sample 5	Stereo
27	Lead vocal	Mono
28	Lead vocal doubles 1	Mono
29	Lead vocal doubles 2	Mono
30	Lead vocal doubles 3	Mono
31	Backing vocal 1	Mono
32	Backing vocal 2	Mono
33	Backing vocal 3	Mono

34	Backing vocal 4	Mono
35	Backing vocal 5	Mono
36	Backing vocal 6	Mono
37	Backing vocal 7	Mono
38	Backing vocal 8	Mono
39	Backing vocal 9	Mono
40	Backing vocal 10	Mono
41	Backing vocal 11	Mono
42	Backing vocal 12	Mono
43	Backing vocal 13	Mono
44	Backing vocal 14	Mono
45	Backing vocal 15	Mono
46	Backing vocal 16	Mono

Table A.6. Structure of the “Speak Softly” song

No.	Track name	Spatiality
1	Drums 1	Stereo
2	Drums 2	Stereo
3	Percussion 1	Mono
4	Percussion 2	Mono
5	Bass synth	Mono
6	Piano 1	Mono
7	Piano 2	Mono
8	Piano 3	Mono
9	Piano 4	Mono
10	Rhodes 1	Stereo
11	Rhodes 2	Mono
12	Rhodes 3	Stereo
13	Synth 1	Stereo
14	Synth 2	Stereo
15	Lead vocal	Mono
16	Backing vocal 1	Mono
17	Backing vocal 2	mono

Table A.7. Structure of the “The Doppler Shift” song

No.	Track name	Spatiality
1	Kick	Mono
2	Snare top	Mono
3	Snare bottom	Mono
4	Hihat	Mono
5	Overheads	Stereo
6	Drums room	Stereo
7	Ride	Mono
8	Tom 1	Mono
9	Tom 2	Mono
10	Tom 3	Mono

11	Bass	Mono
12	Electric guitar 1	Mono
13	Electric guitar 2	Mono
14	Electric guitar 3	Mono
15	Electric guitar 4	Mono
16	Electric guitar 5	Mono
17	Electric guitar 6	Mono
18	Electric piano 1	Stereo
19	Electric piano 2	Stereo
20	Synth 1	Stereo
21	Synth 2	Stereo
22	Lead vocal	Mono

Table A.8. Structure of the “Tom McKenzie” song

No.	Track name	Spatiality
1	Kick	Mono
2	Snare top	Mono
3	Snare bottom	Mono
4	Hihat	Mono
5	Overheads	Stereo
6	Shaker	Stereo
7	Tambourine	Stereo
8	Bongos	Mono
9	Bass 1	Mono
10	Bass 2	Mono
11	Acoustic guitar 1	Mono
12	Acoustic guitar 2	Mono
13	Acoustic guitar 3	Mono
14	Acoustic guitar 4	Mono
15	Acoustic guitar 5	Mono
16	Acoustic guitar 6	Mono
17	Electric guitar 1	Mono
18	Electric guitar 2	Mono
19	Electric guitar 3	Mono
20	Electric guitar 4	Mono
21	Lead vocal	Mono
22	Backing vocal 1	Mono
23	Backing vocal 2	Mono
24	Backing vocal 3	Mono
25	Backing vocal 4	Mono
26	Backing vocal 5	Mono
27	Backing vocal 6	Mono
28	Backing vocal 7	Mono
29	Backing vocal 8	Mono
30	Backing vocal 9	Mono
31	Backing vocal 10	Mono

Appendix B: Objective results, figures and tables

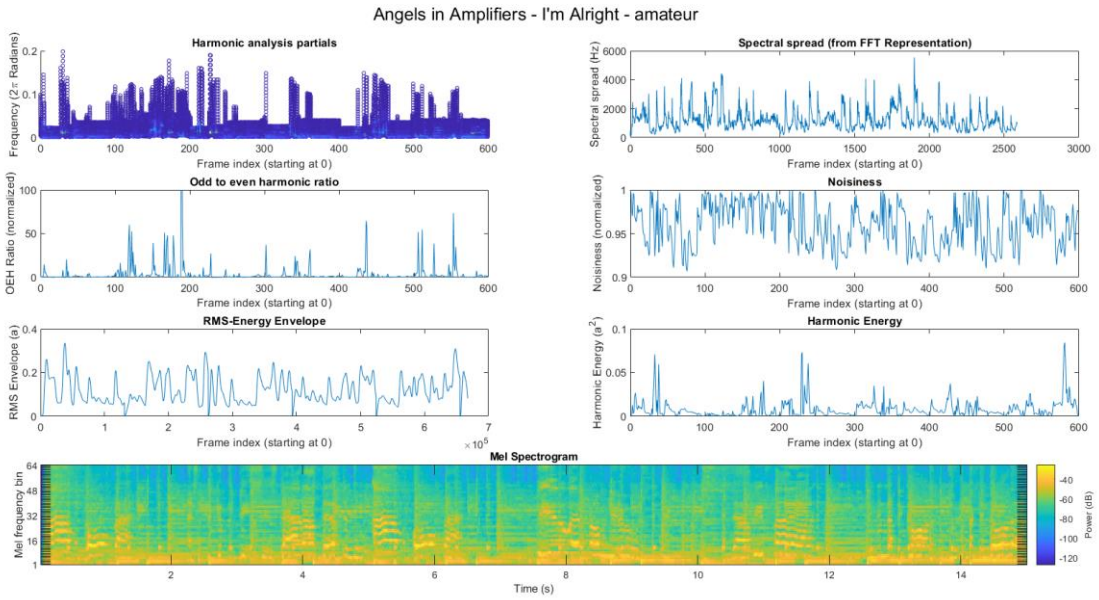


Fig. B.1. Descriptors calculated for Angels in Amplifiers – I'm Alright – "Amateur" sample

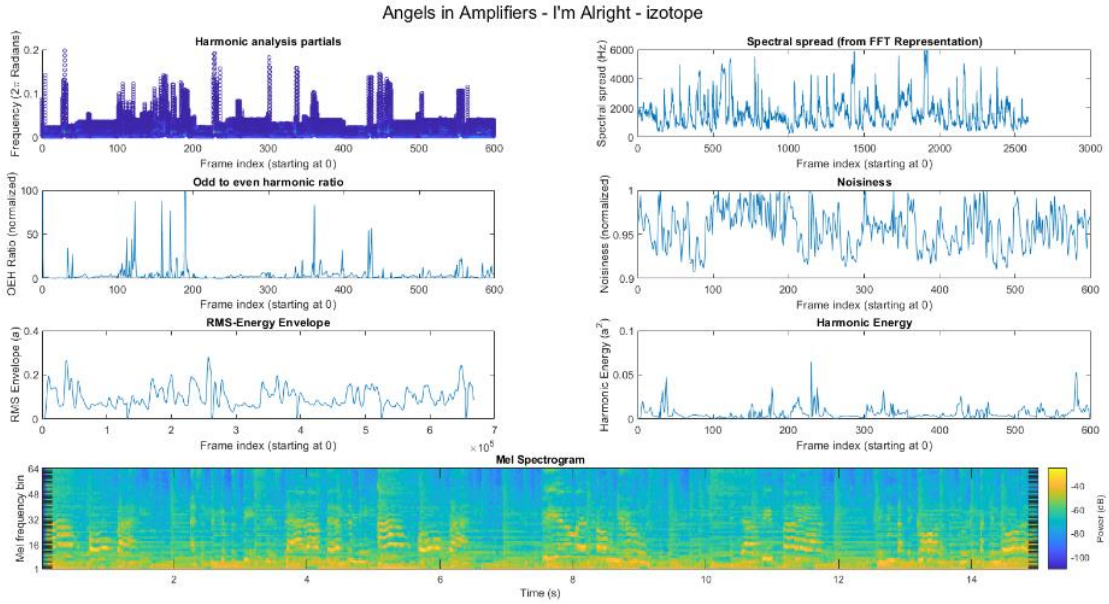


Fig. B.2. Descriptors calculated for Angels in Amplifiers – I'm Alright – "Izotope" sample

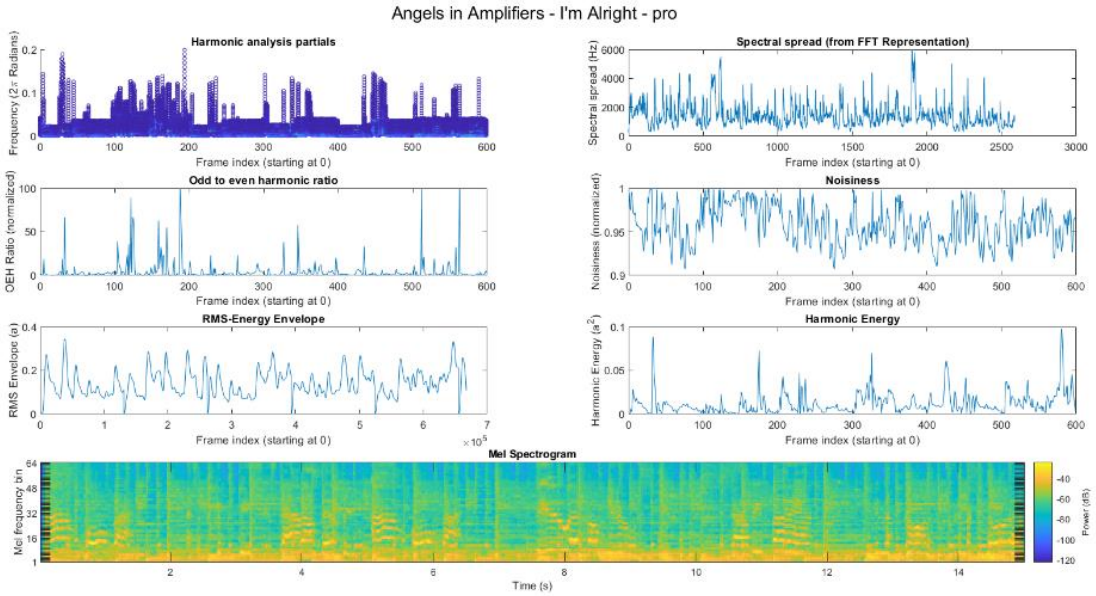


Fig. B.3. Descriptors calculated for Angels in Amplifiers – I'm Alright – "Pro" sample

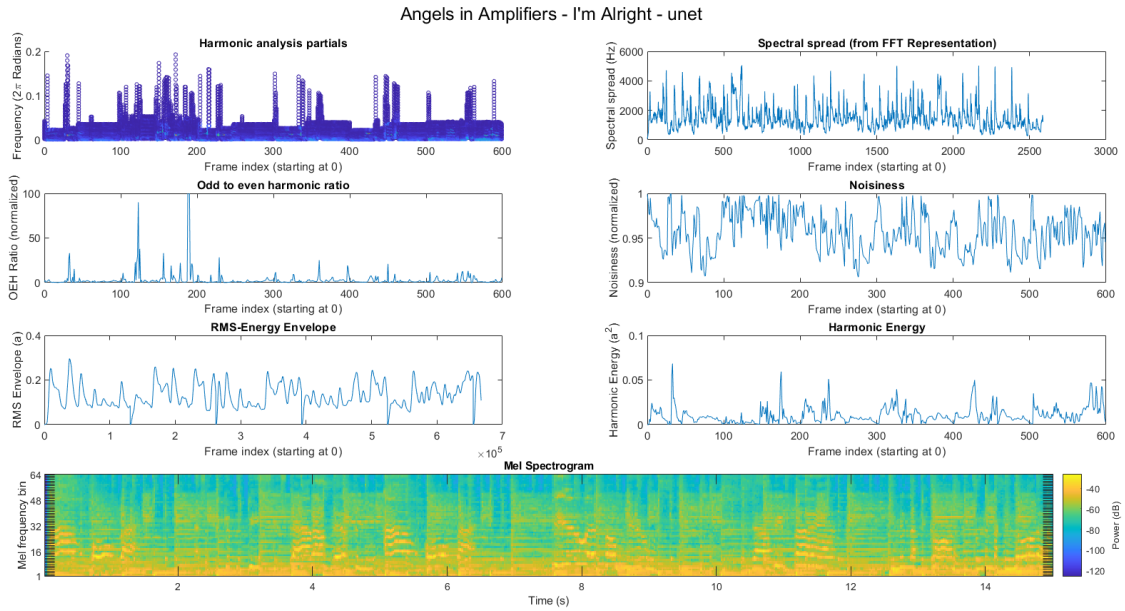


Fig. B.4. Descriptors calculated for Angels in Amplifiers – I'm Alright – "Unet" sample

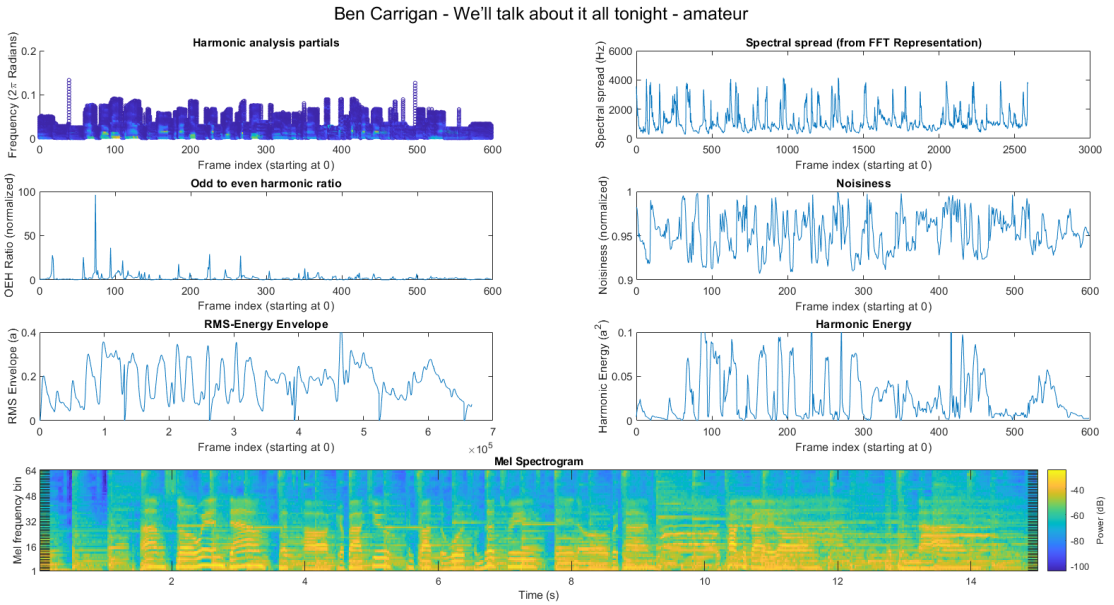


Fig. B.5. Descriptors calculated for Ben Carrigan – We'll talk about it tonight – "Amateur" sample

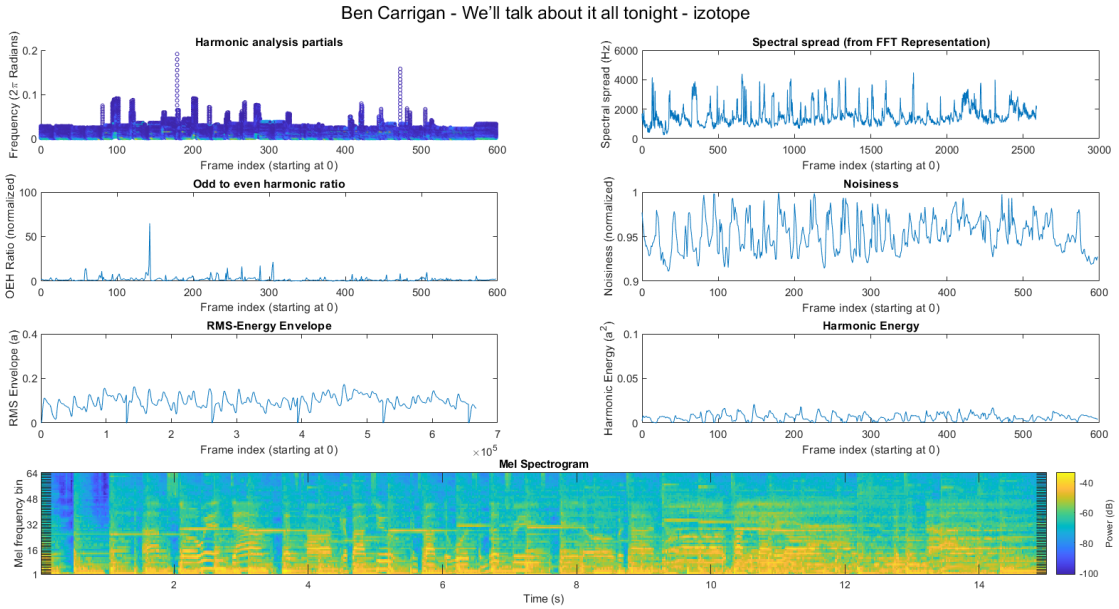


Fig. B.6. Descriptors calculated for Ben Carrigan – We'll talk about it tonight – "Izotope" sample

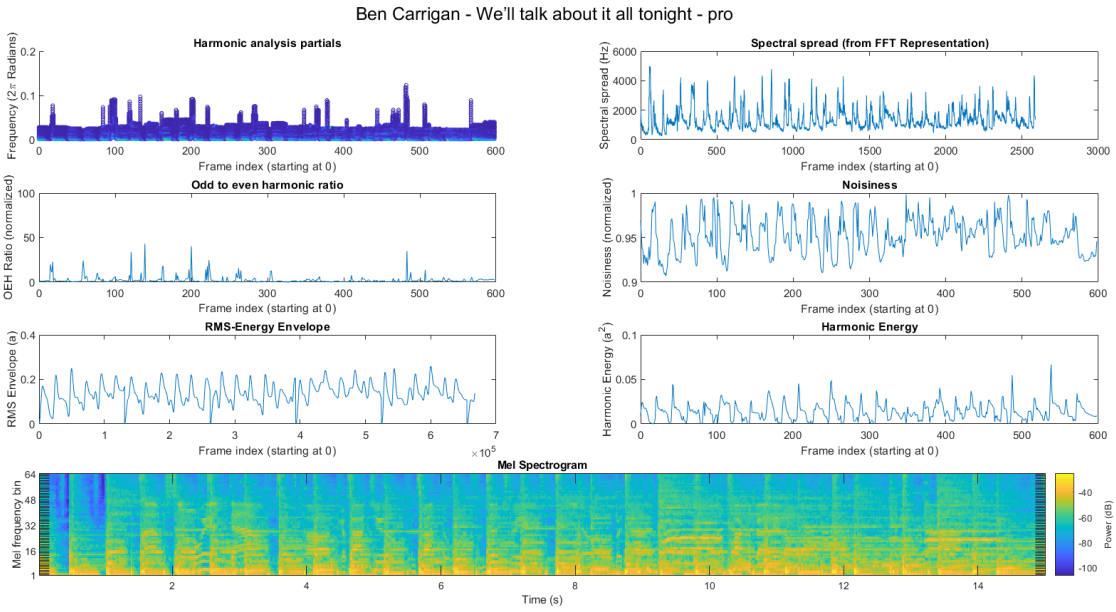


Fig. B.7. Descriptors calculated for Ben Carrigan – We'll talk about it tonight – "Pro" sample

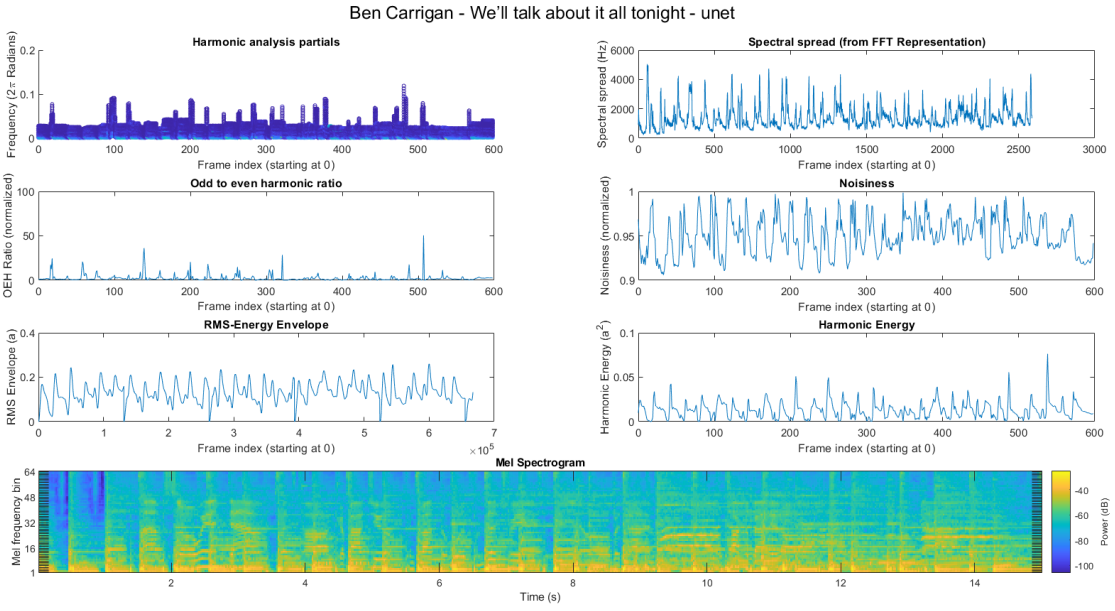


Fig. B.8. Descriptors calculated for Ben Carrigan – We'll talk about it tonight – "Unet" sample

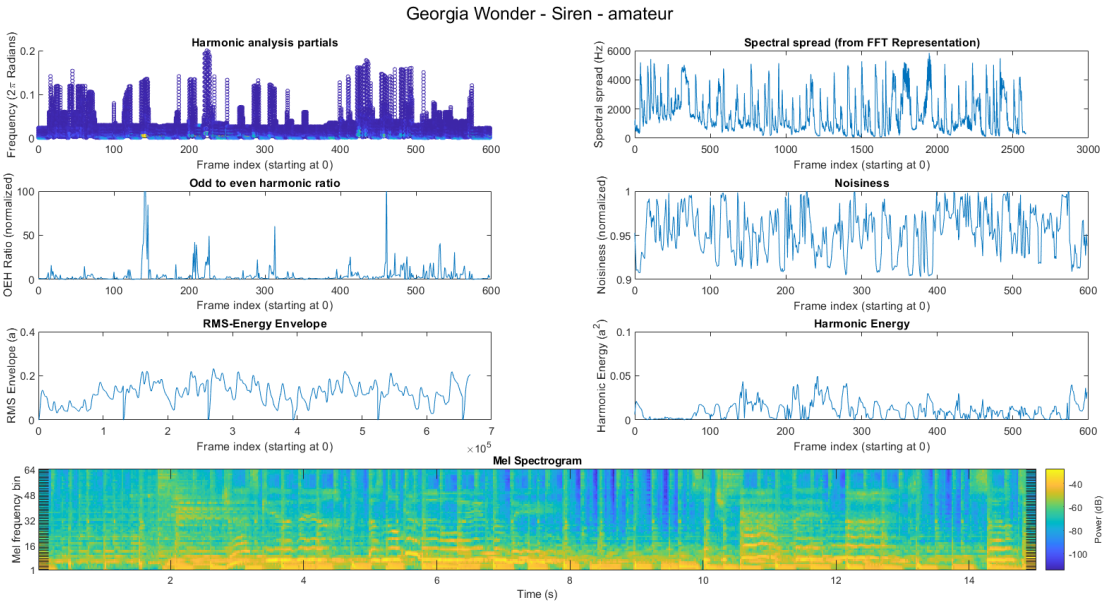


Fig. B.9. Descriptors calculated for Georgia Wonder – Siren – “Amateur” sample

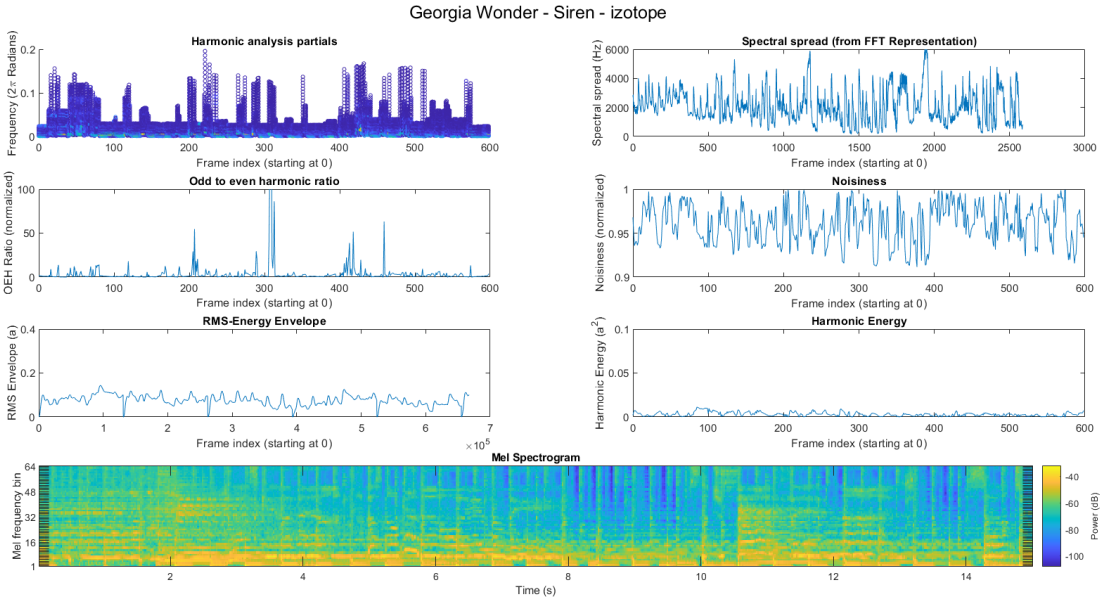


Fig. B.10. Descriptors calculated for Georgia Wonder – Siren – “Izotope” sample

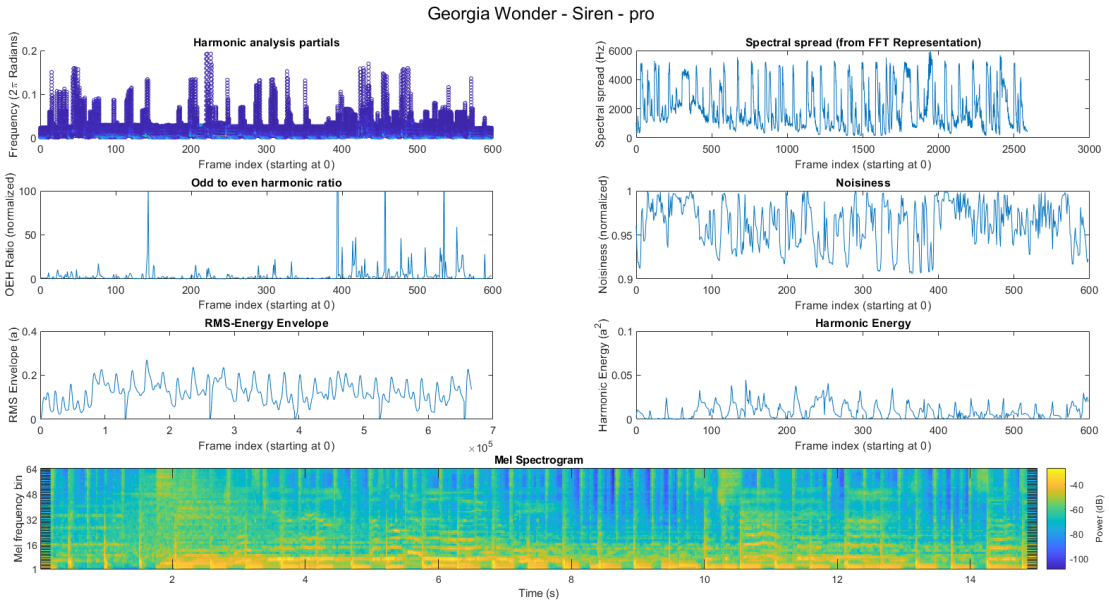


Fig. B.11. Descriptors calculated for Georgia Wonder – Siren – “Pro” sample

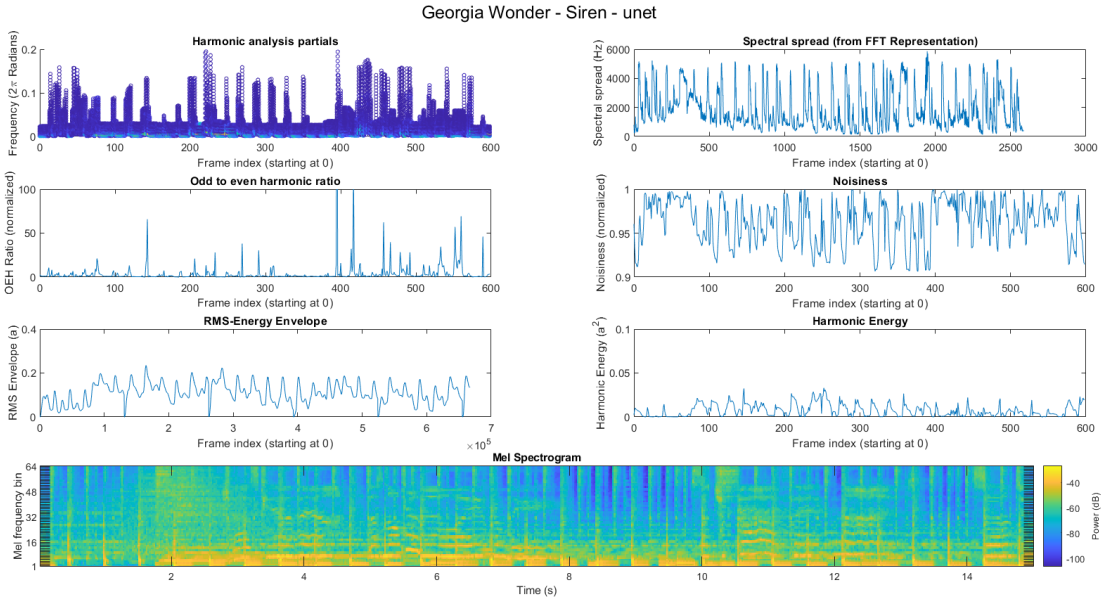


Fig. B.12. Descriptors calculated for Georgia Wonder – Siren – “Unet” sample

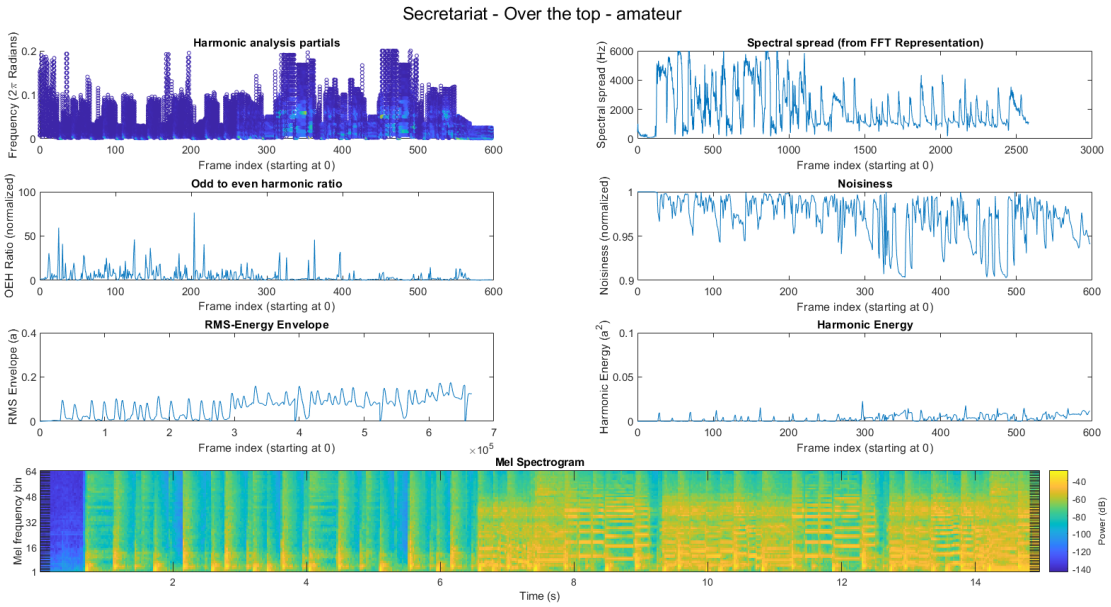


Fig. B.13. Descriptors calculated for Secretariat – Over the top – “Amateur” sample

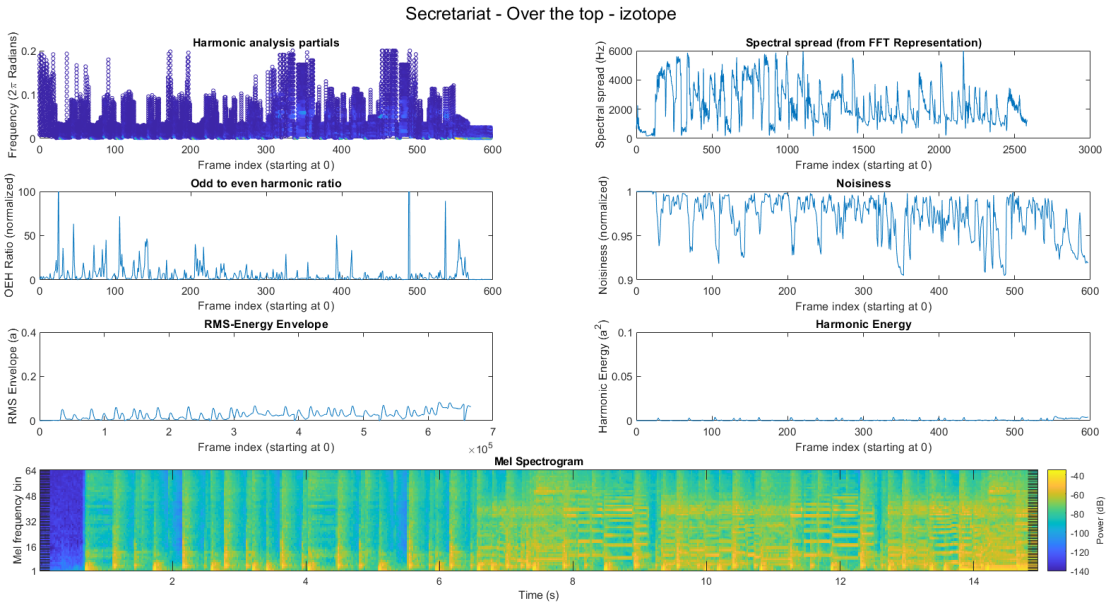


Fig. B.14. Descriptors calculated for Secretariat – Over the top – “Izotope” sample

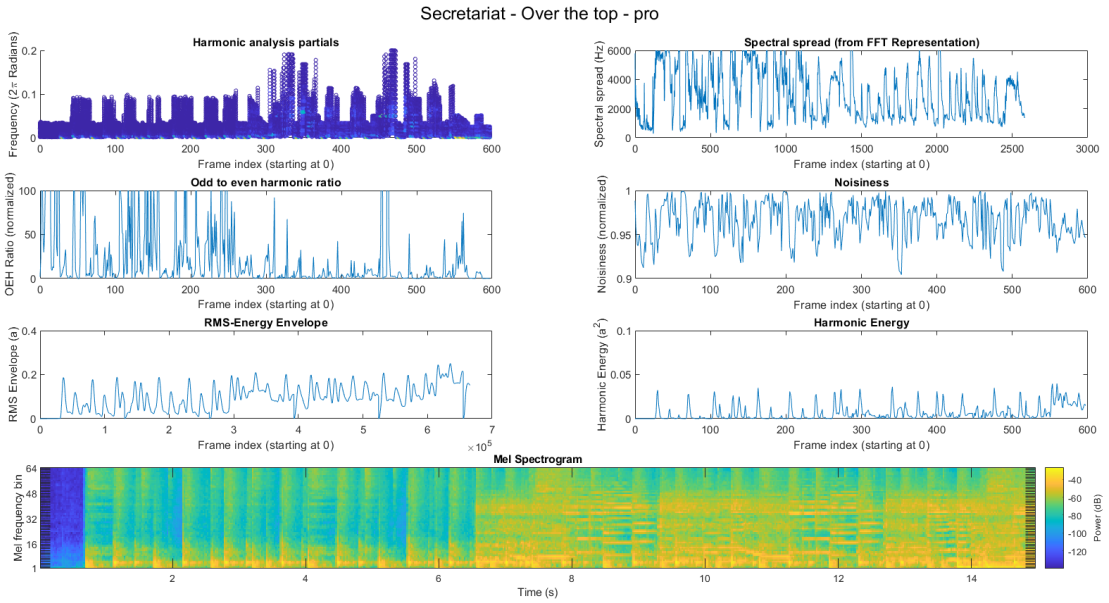


Fig. B.15. Descriptors calculated for Secretariat – Over the top – “Pro” sample

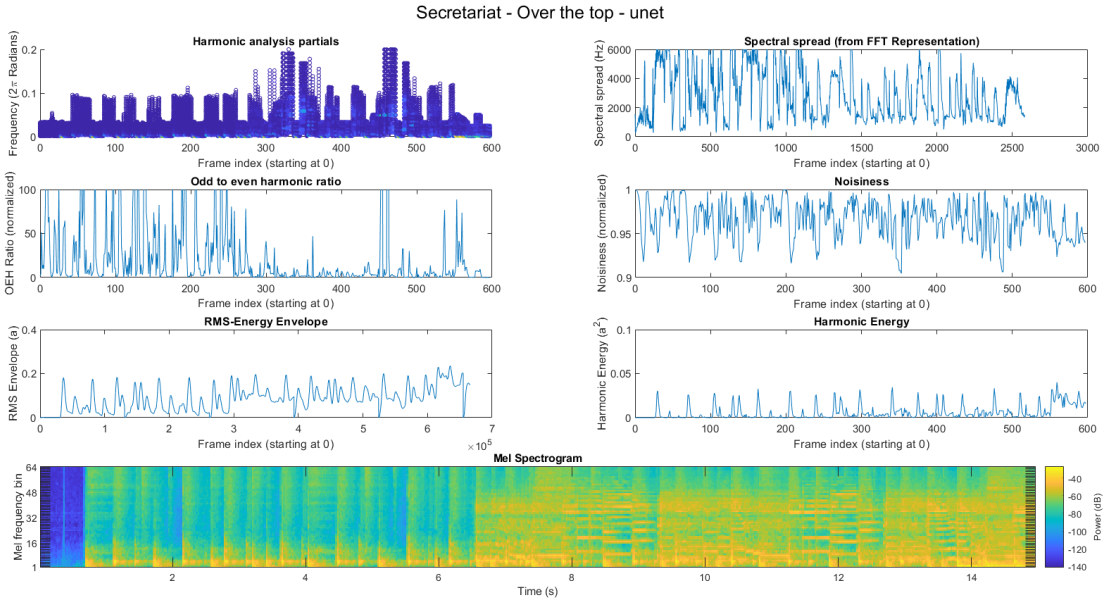


Fig. B.16. Descriptors calculated for Secretariat – Over the top – “Unet” sample

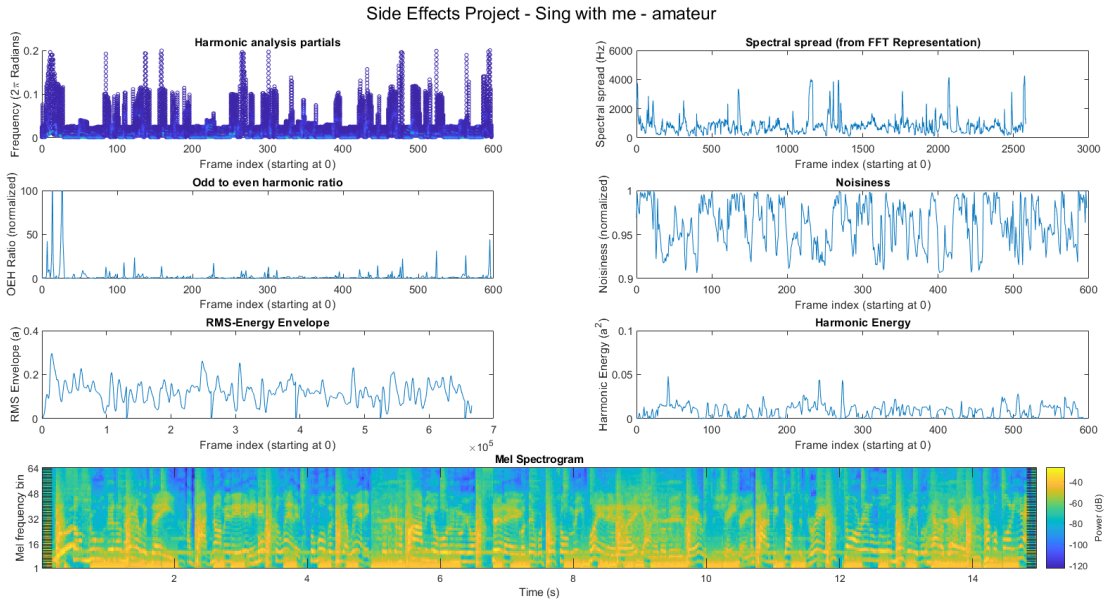


Fig. B.17. Descriptors calculated for Side Effects Project – Sing with me – “Amateur” sample

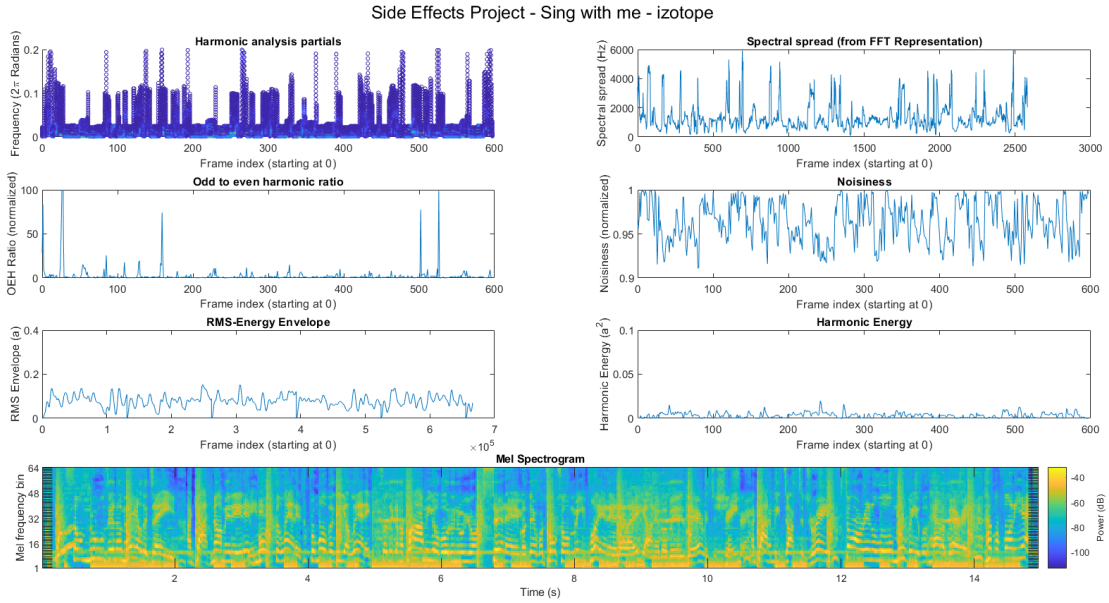


Fig. B.18. Descriptors calculated for Side Effects Project – Sing with me – “Izotope” sample

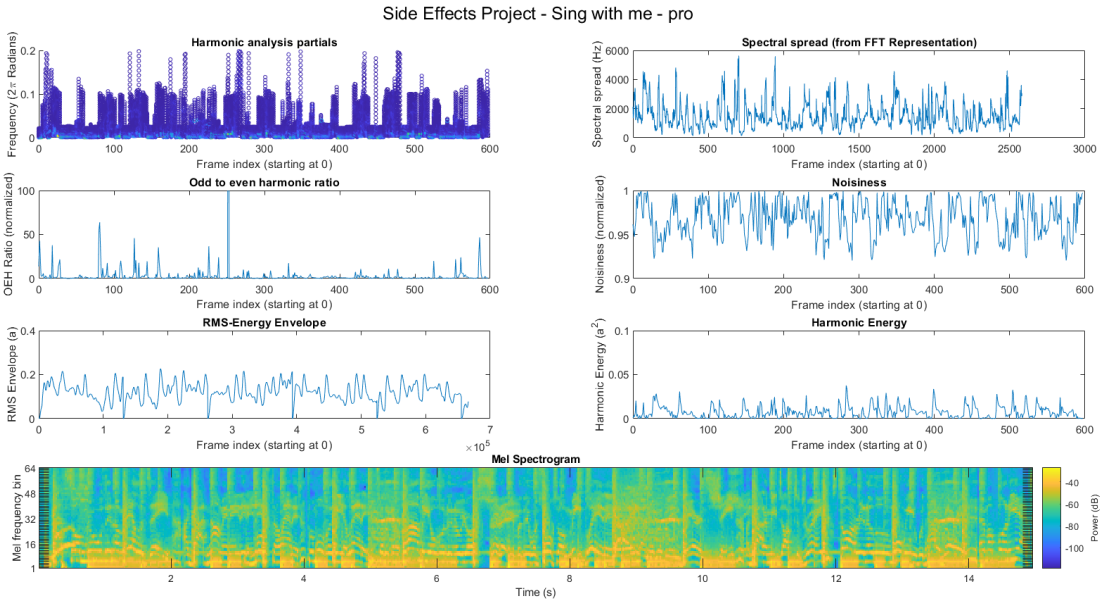


Fig. B.19. Descriptors calculated for Side Effects Project – Sing with me – “Pro” sample

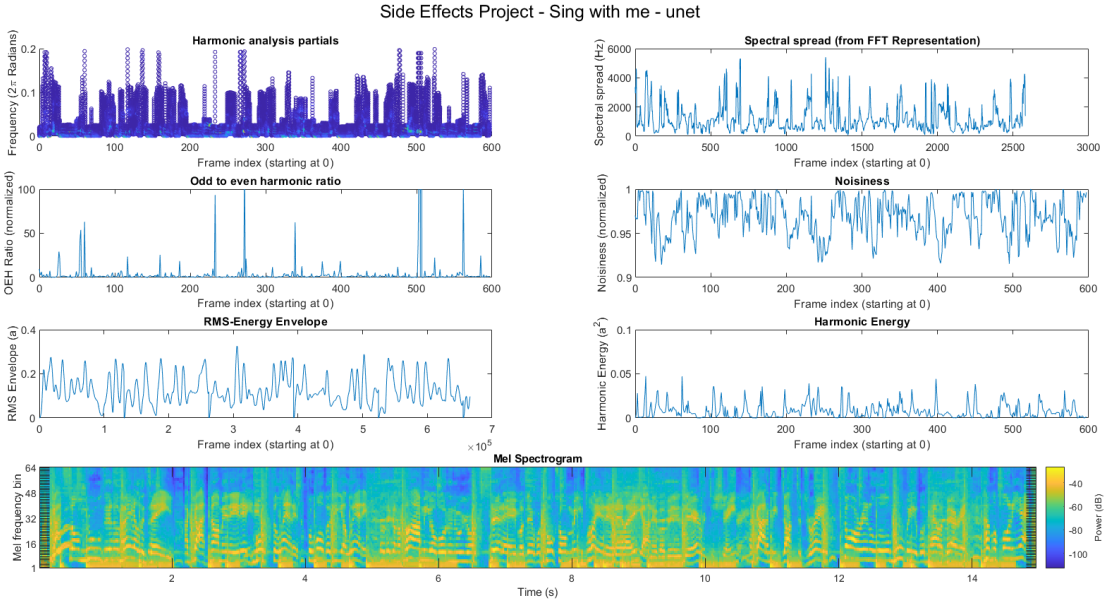


Fig. B.20. Descriptors calculated for Side Effects Project – Sing with me – “Unet” sample

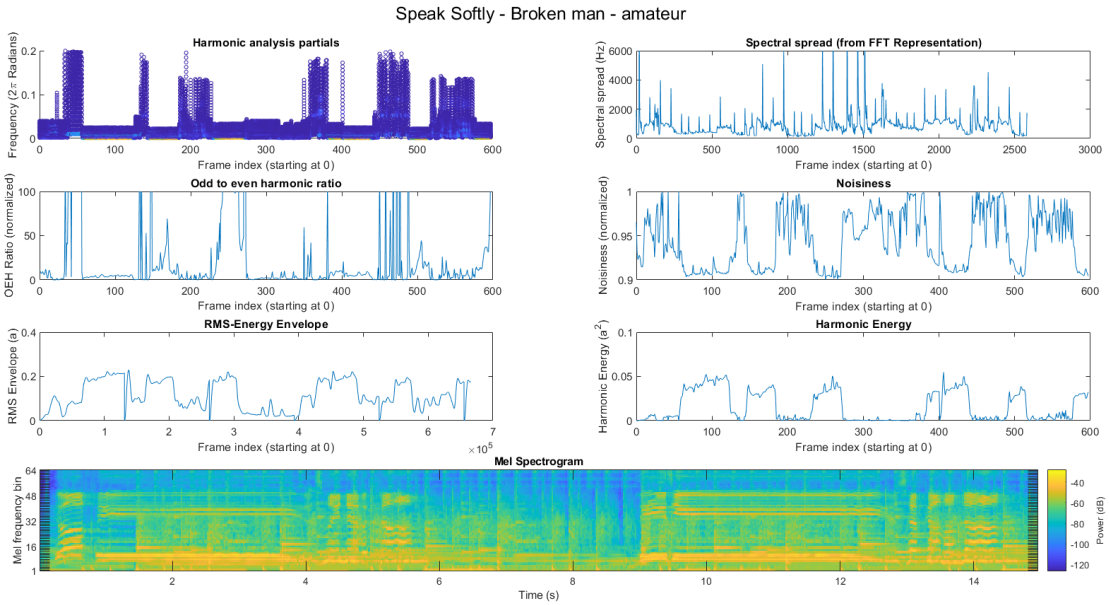


Fig. B.21. Descriptors calculated for Speak Softly – Broken man – “Amateur” sample

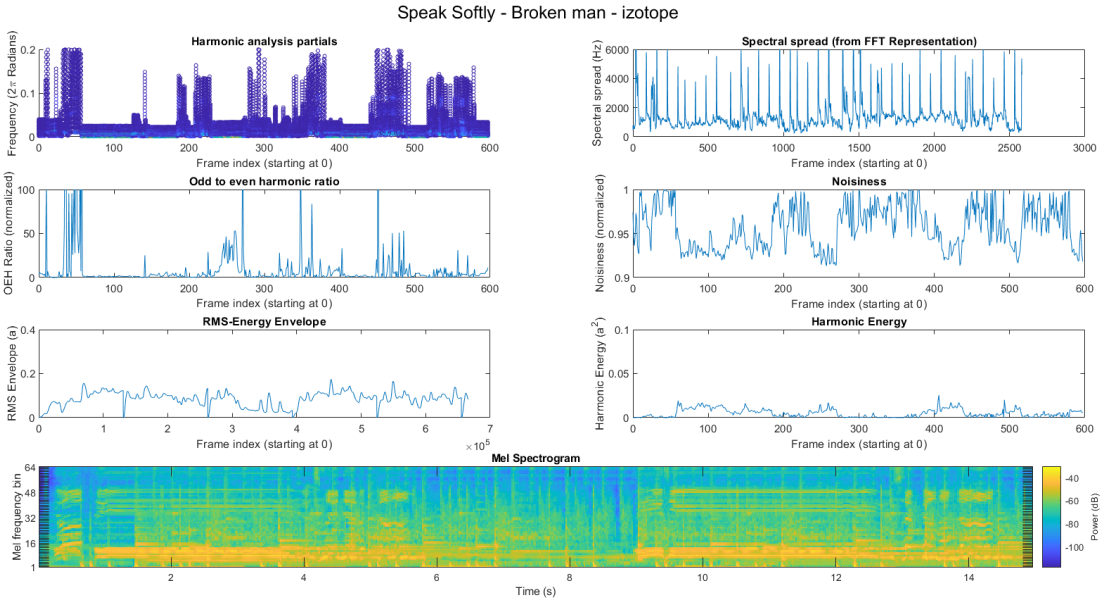


Fig. B.22. Descriptors calculated for Speak Softly – Broken man – “Izotope” sample

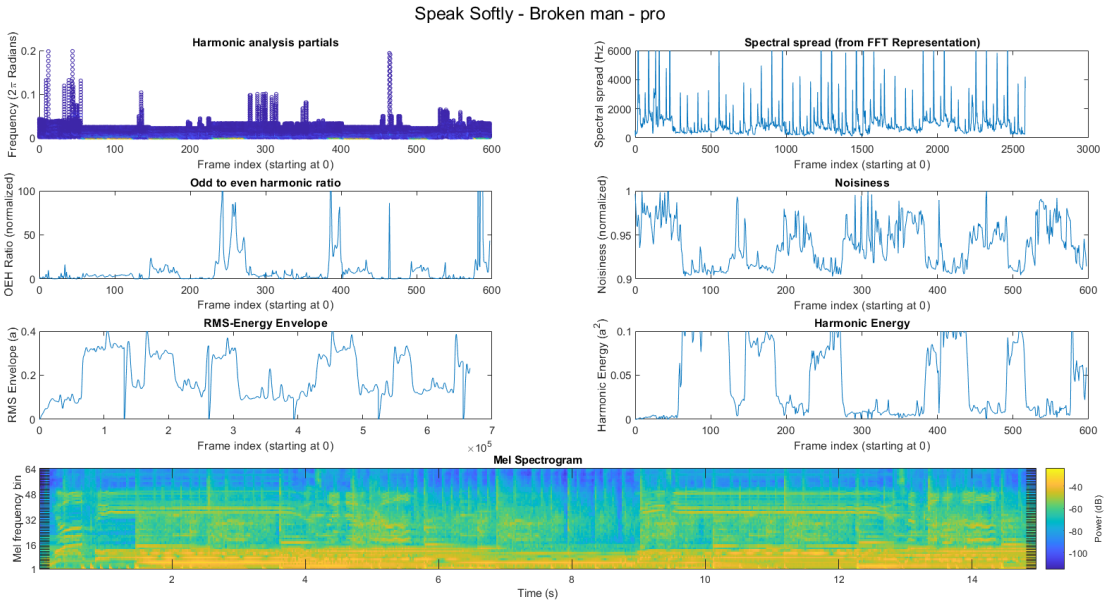


Fig. B.23. Descriptors calculated for Speak Softly – Broken man – “Pro” sample

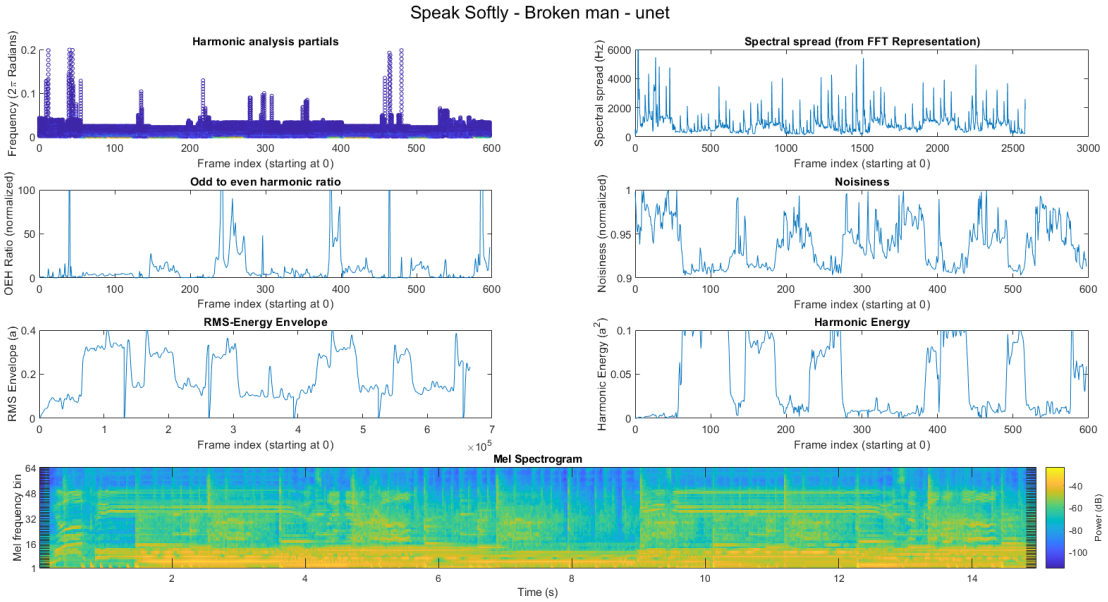


Fig. B.24. Descriptors calculated for Speak Softly – Broken man – “Unet” sample

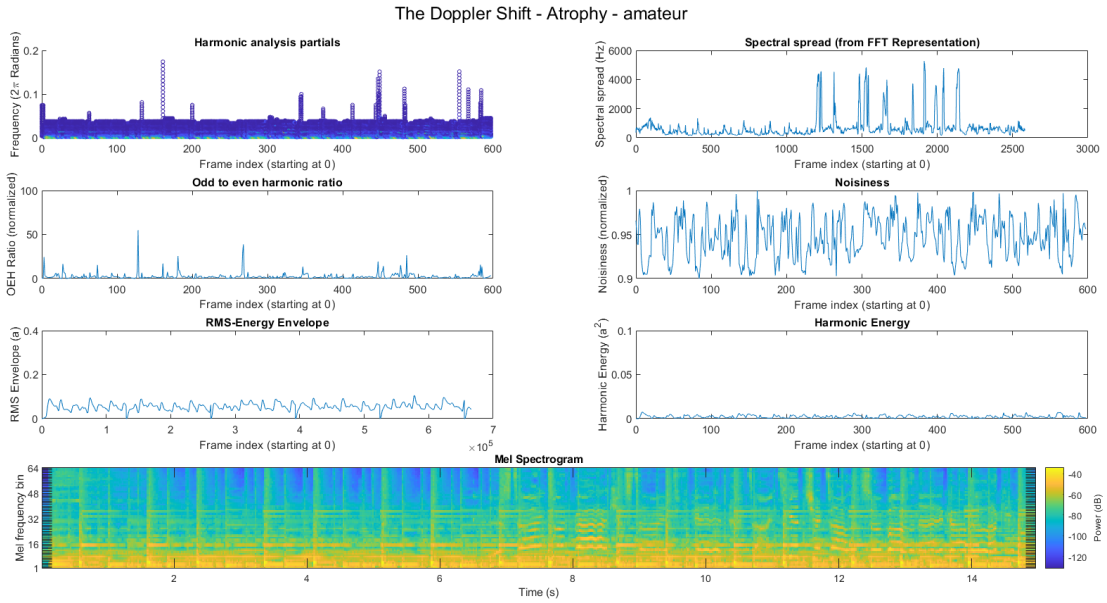


Fig. B.25. Descriptors calculated for The Doppler Shift – Atrophy – “Amateur” sample

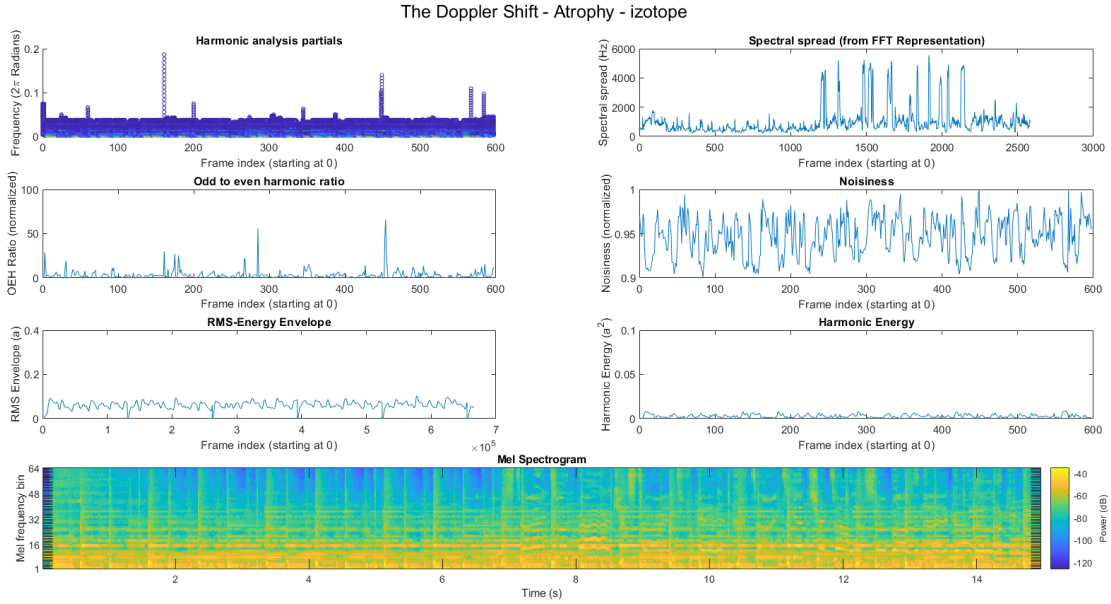


Fig. B.26. Descriptors calculated for The Doppler Shift – Atrophy – “Izotope” sample

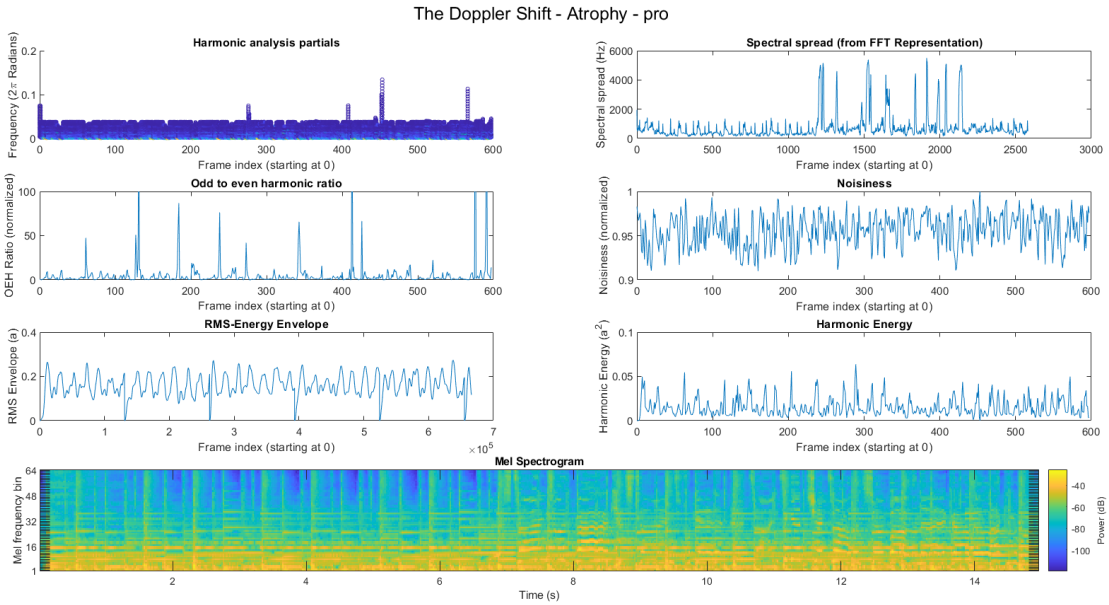


Fig. B.27. Descriptors calculated for The Doppler Shift – Atrophy – “Pro” sample

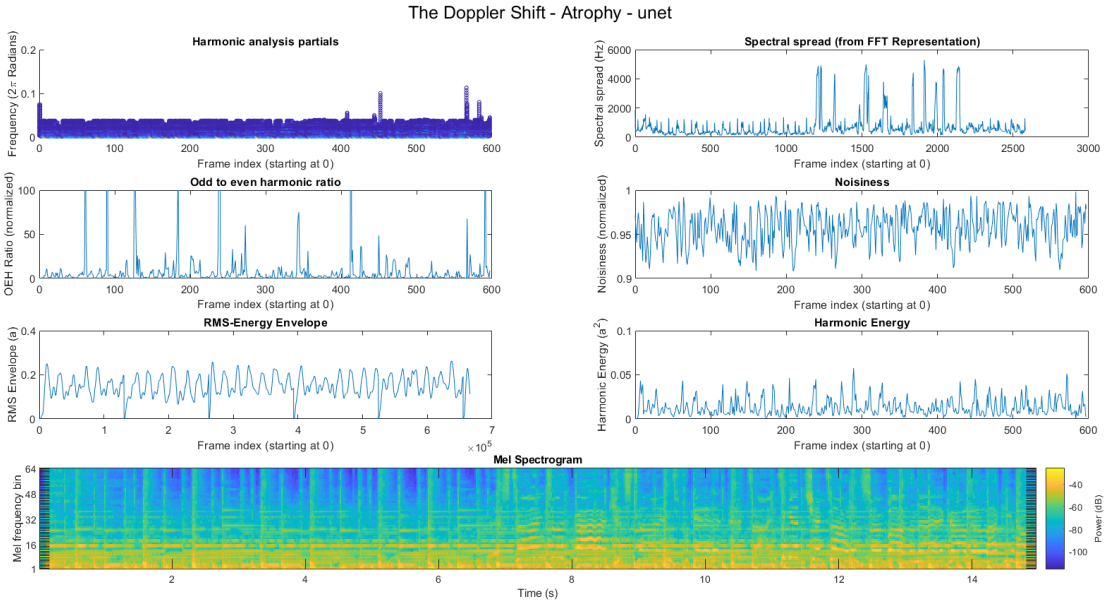


Fig. B.28. Descriptors calculated for The Doppler Shift – Atrophy – “Unet” sample

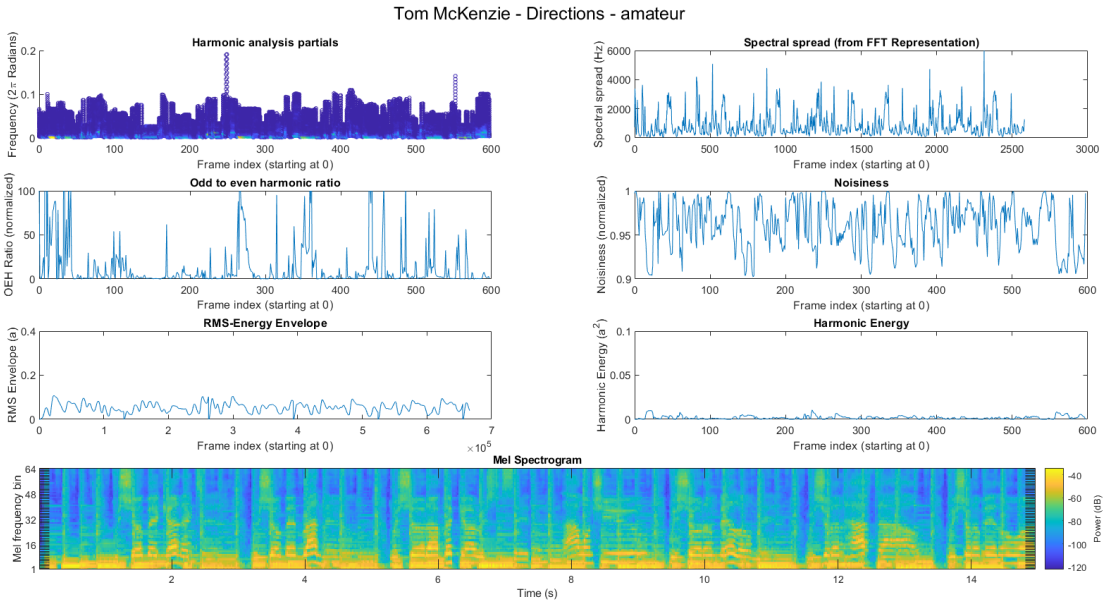


Fig. B.29. Descriptors calculated for Tom McKenzie – Directions – “Amateur” sample

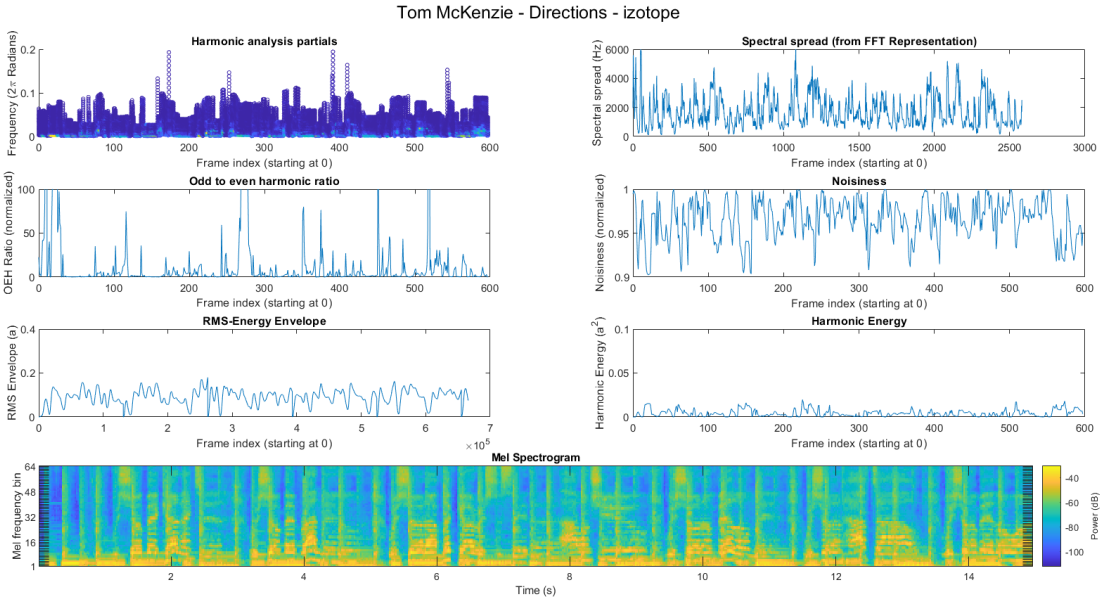


Fig. B.30. Descriptors calculated for Tom McKenzie – Directions – “Izotope” sample

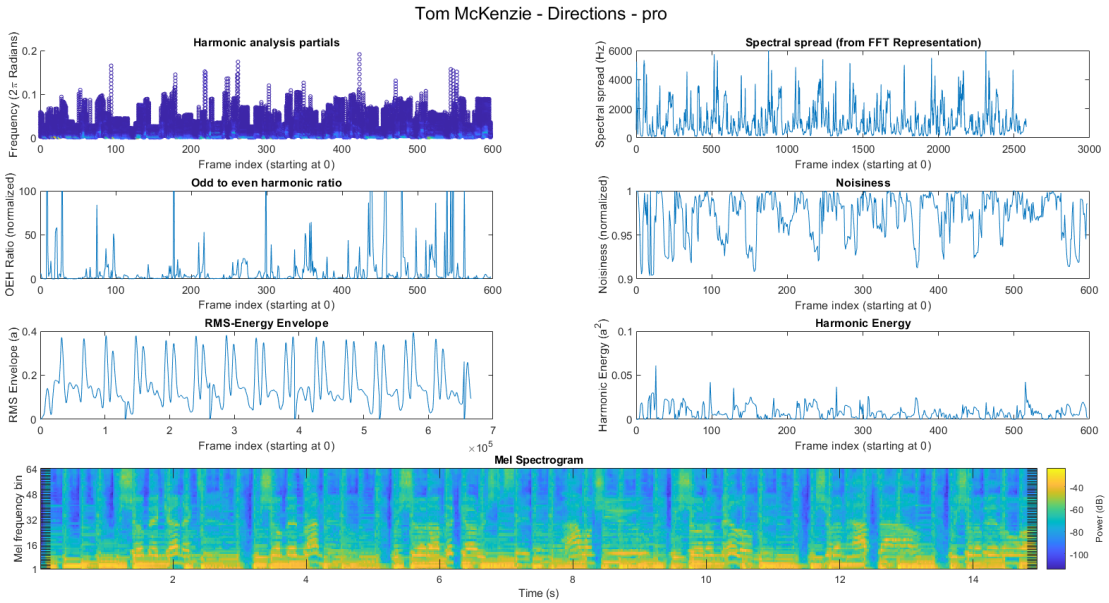


Fig. B.31. Descriptors calculated for Tom McKenzie – Directions – “Pro” sample

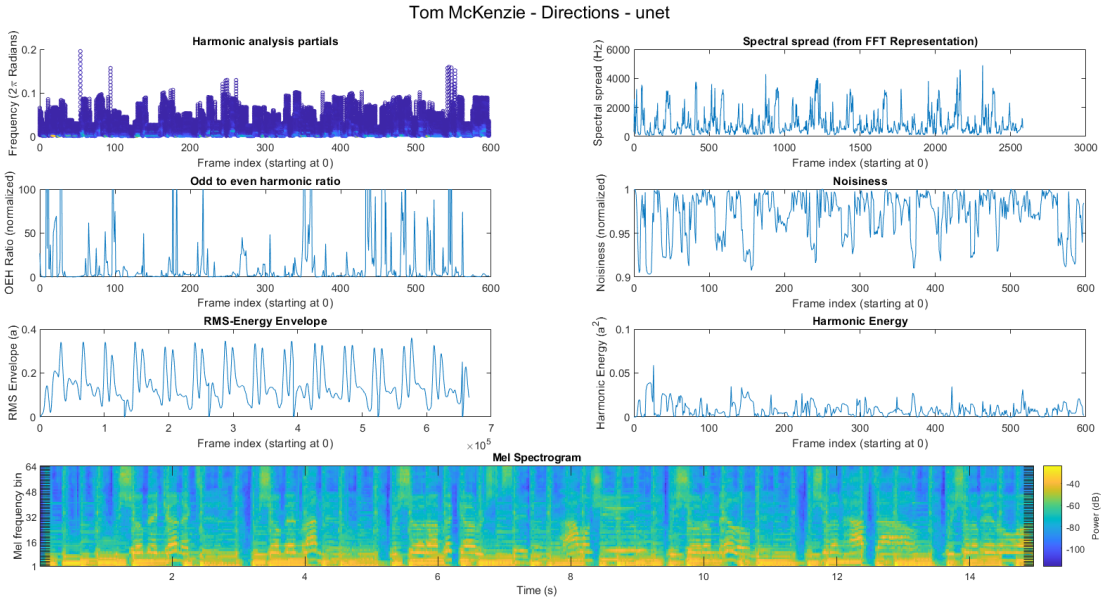
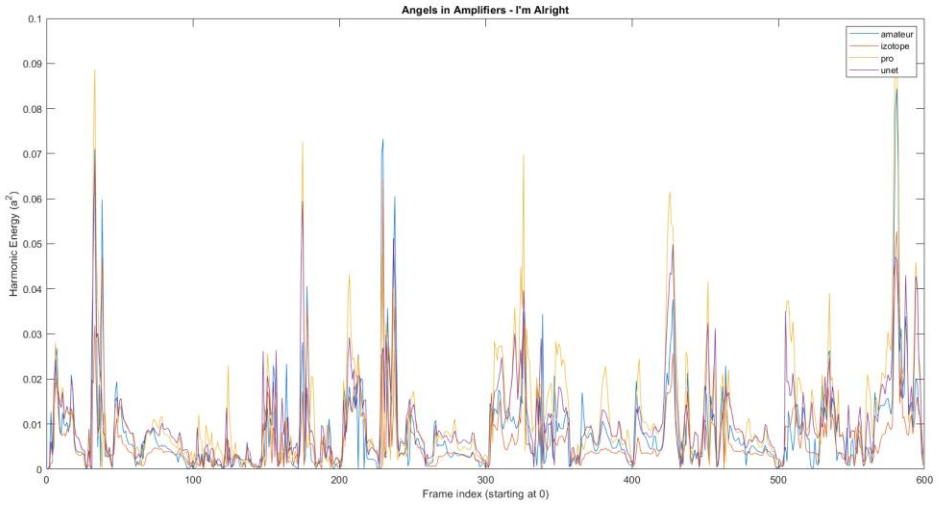
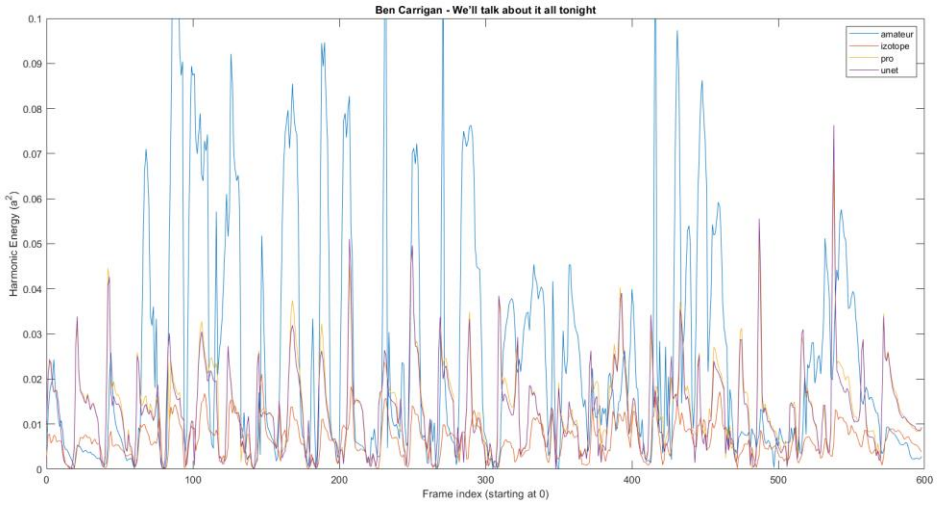
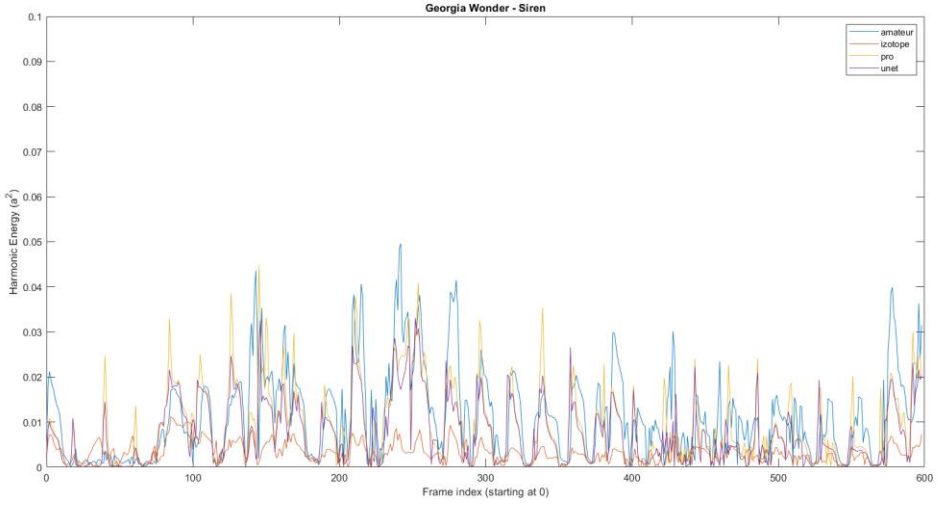
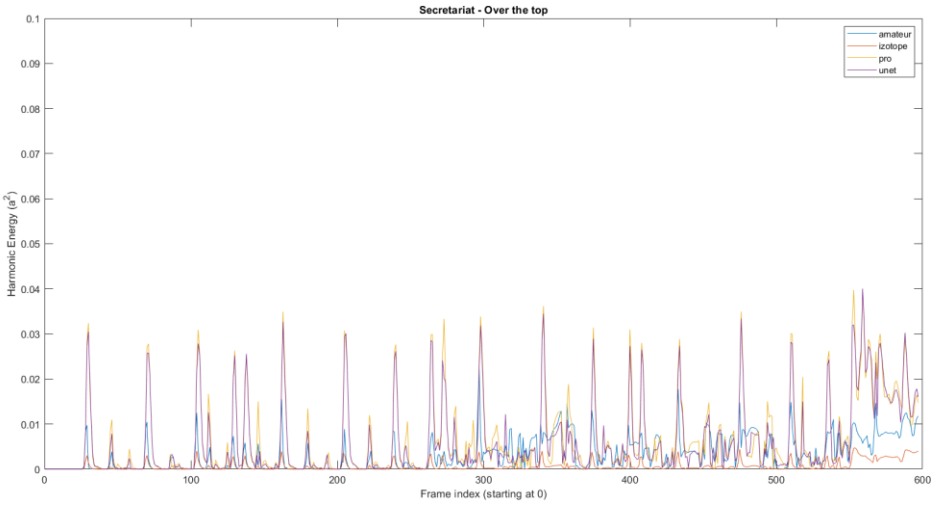


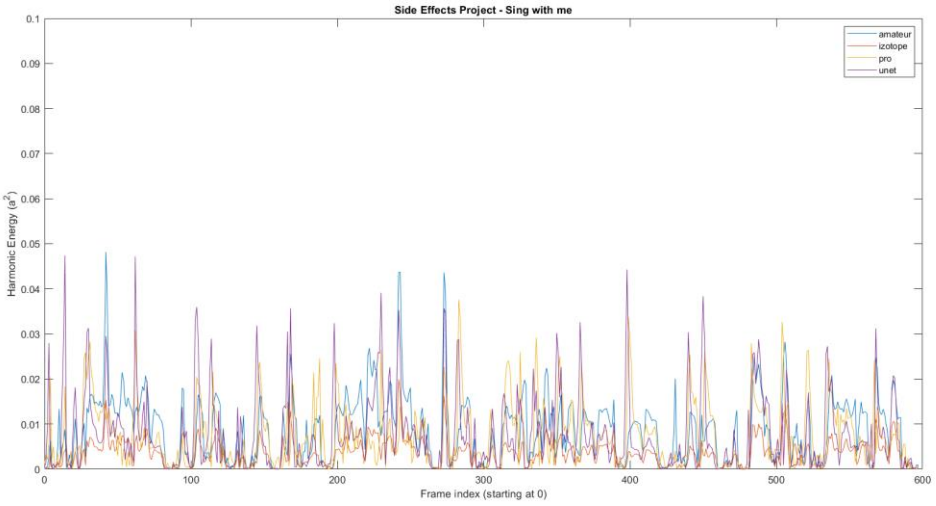
Fig. B.32. Descriptors calculated for Tom McKenzie – Directions – “Unet” sample

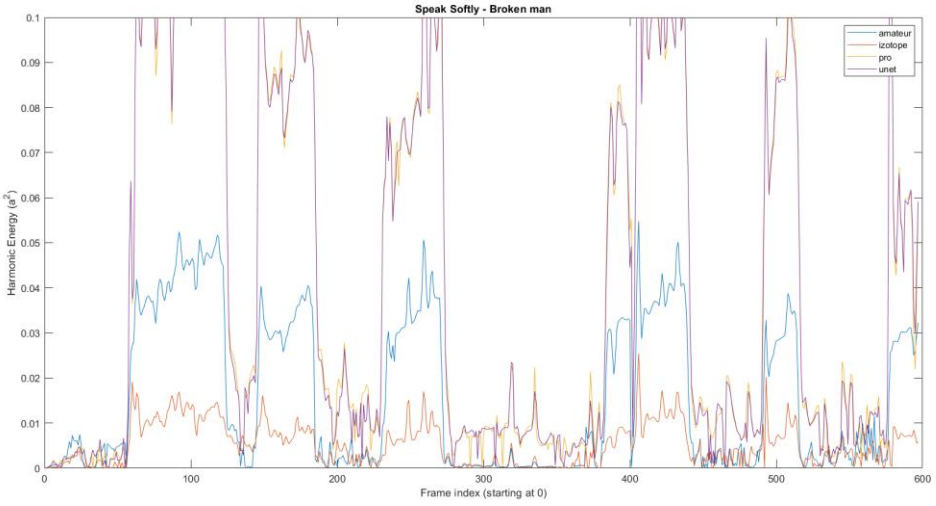












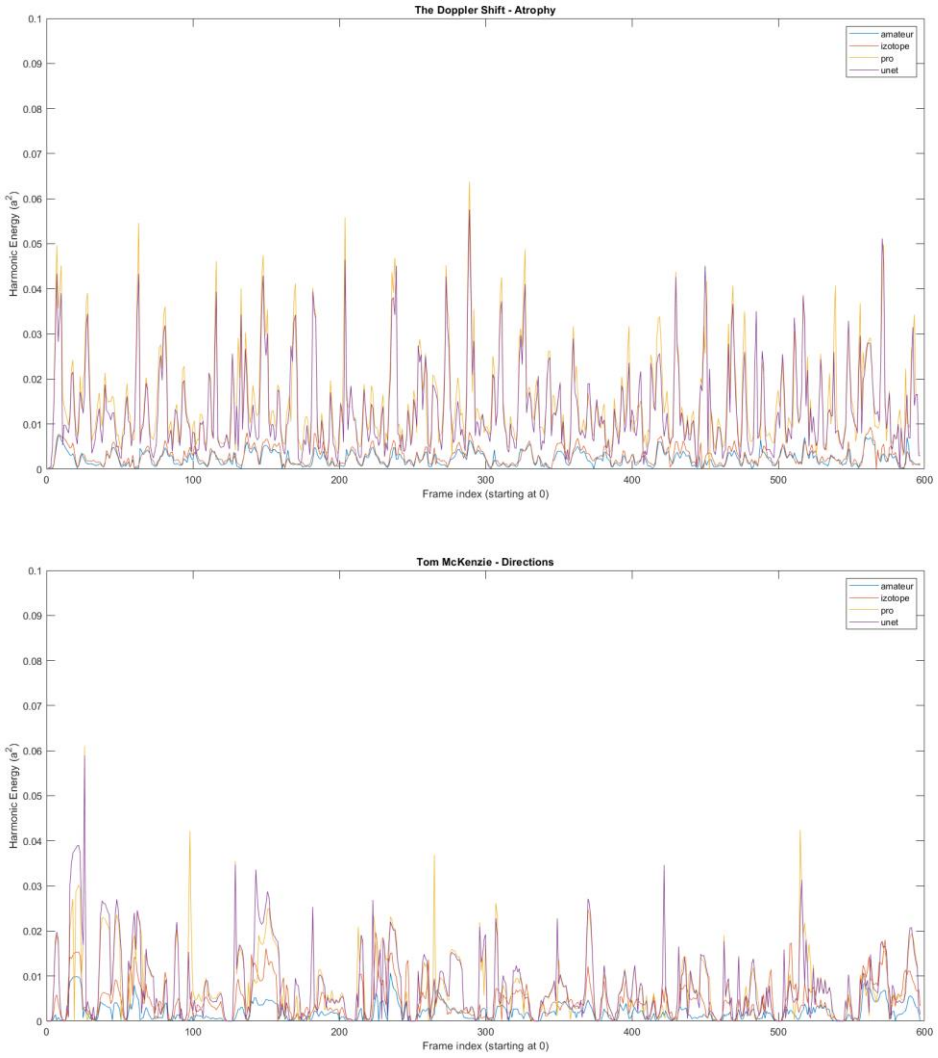
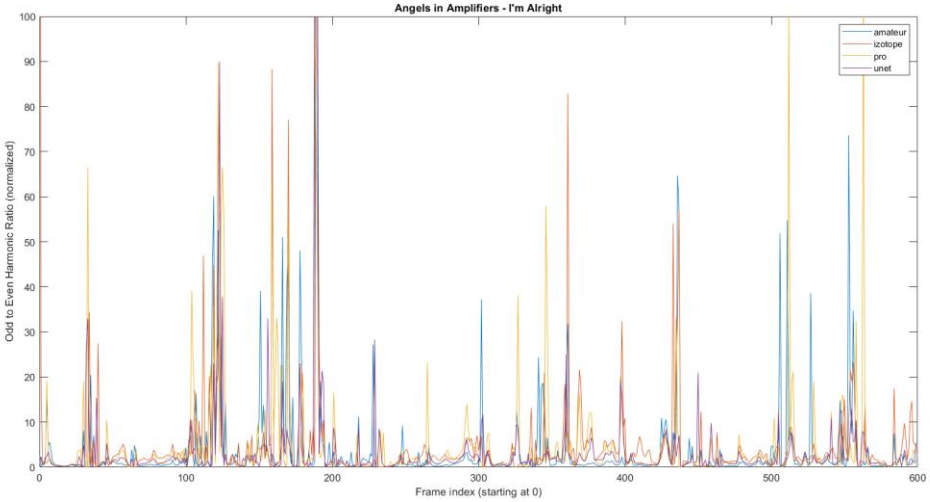
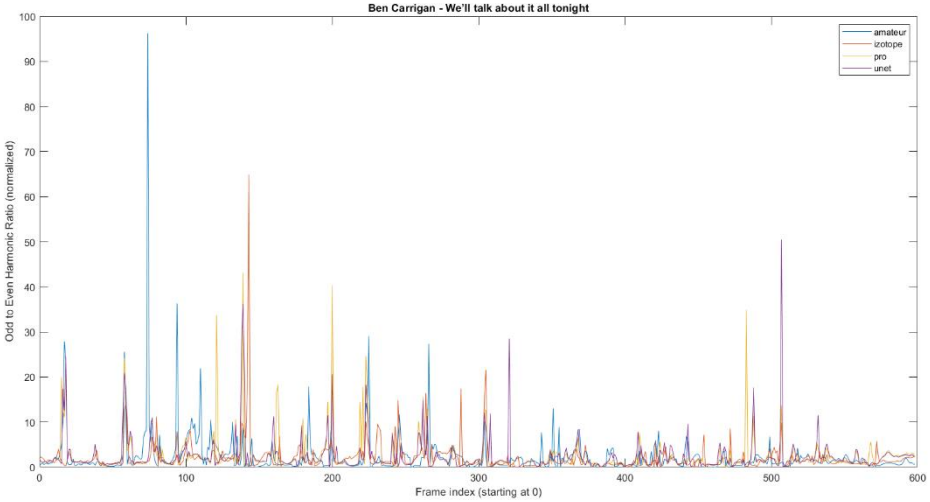
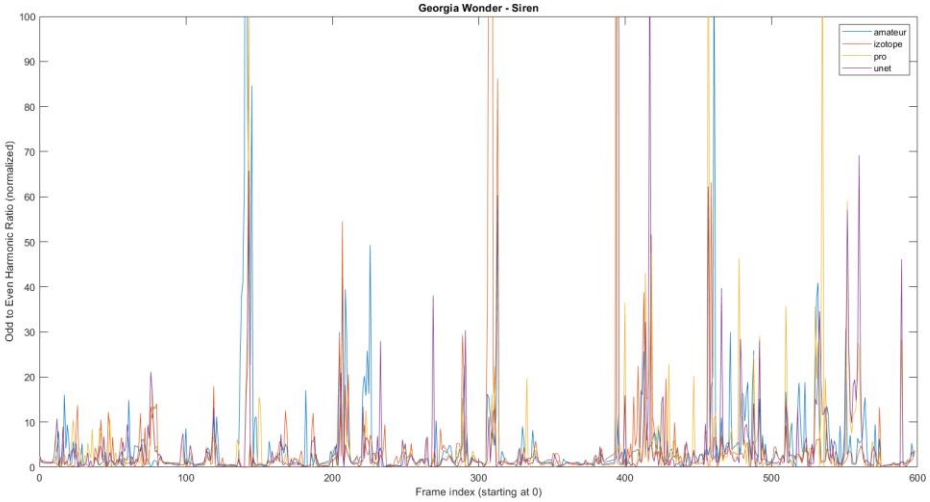
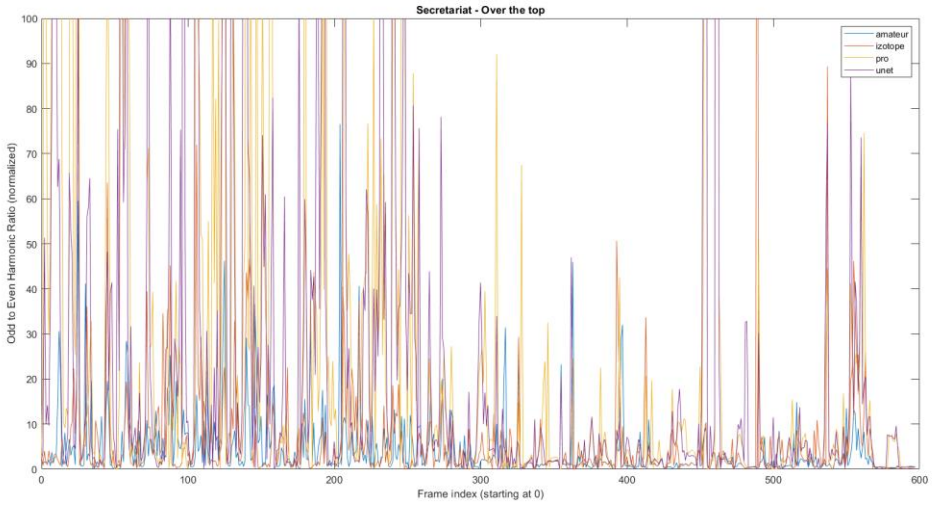


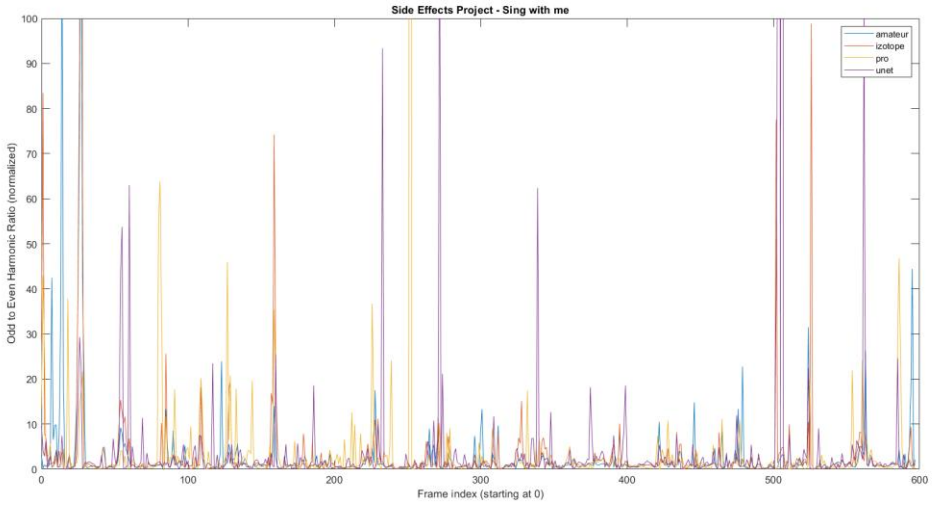
Fig. B.33. Harmonic Energy descriptor calculated for all songs depending on mix type

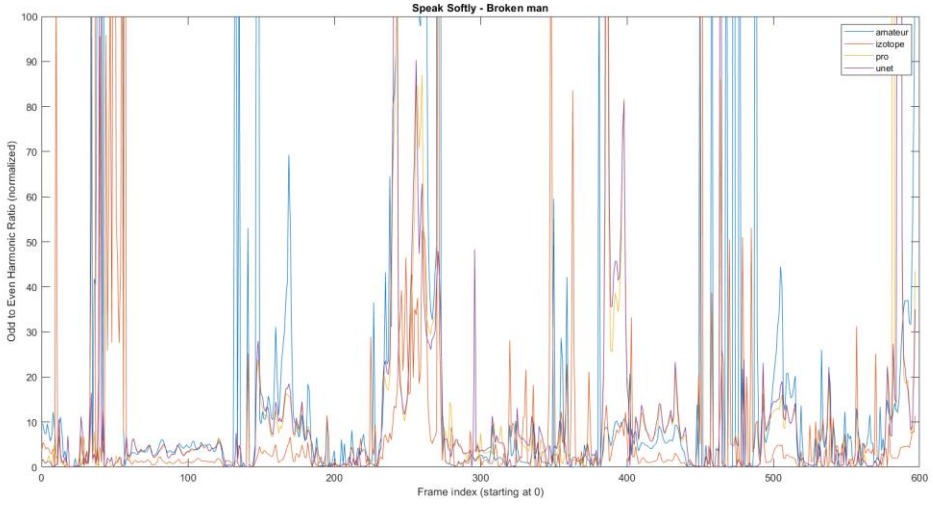












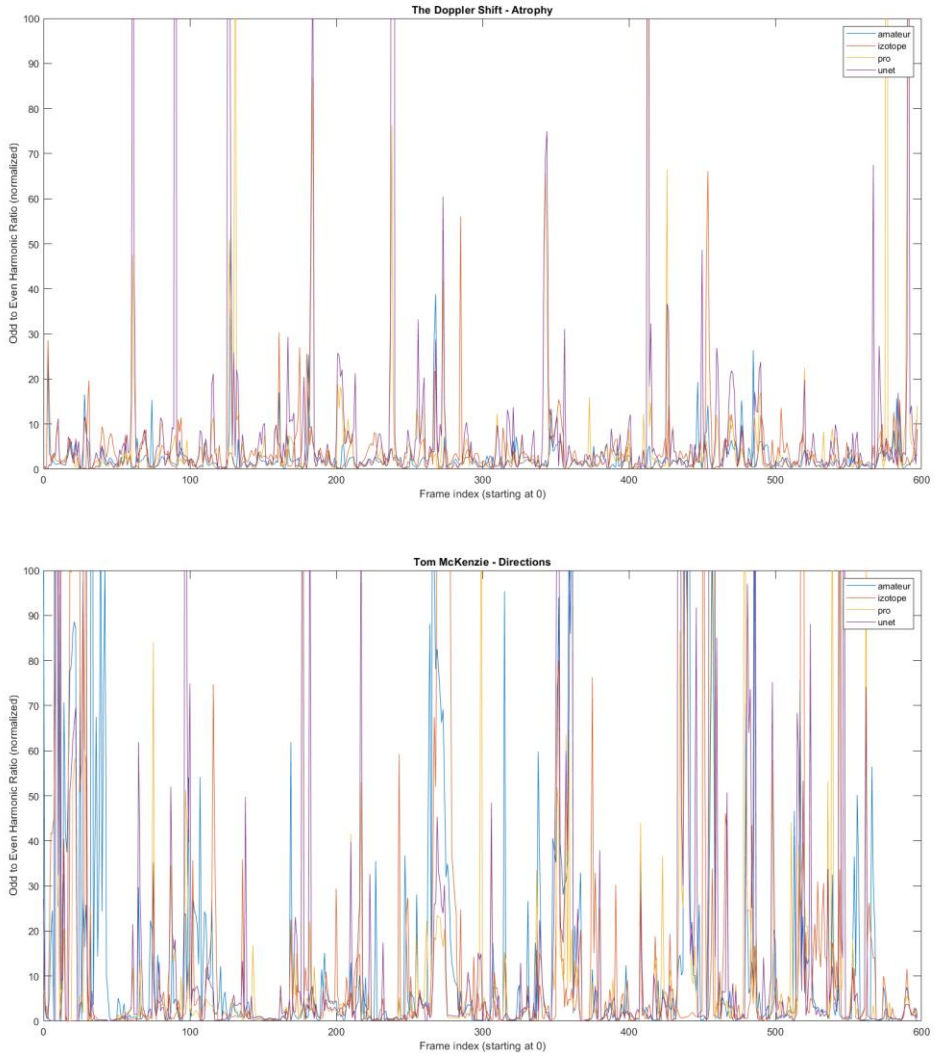
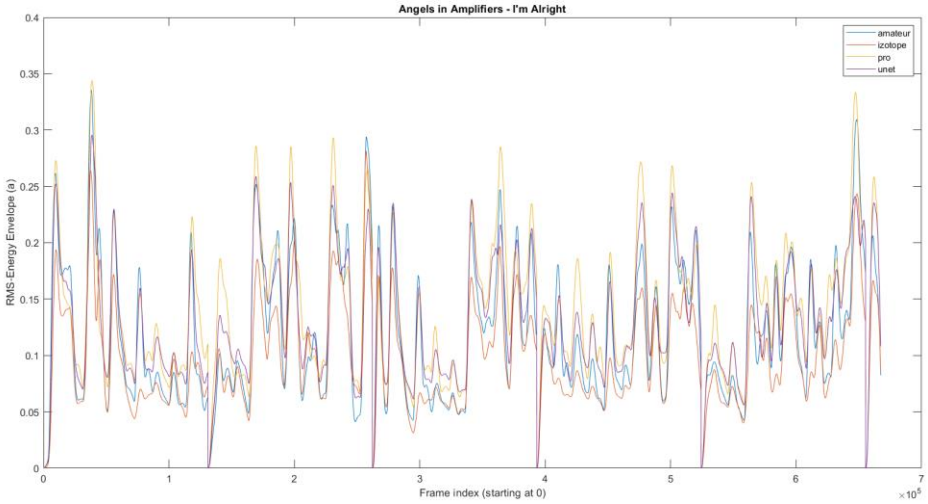
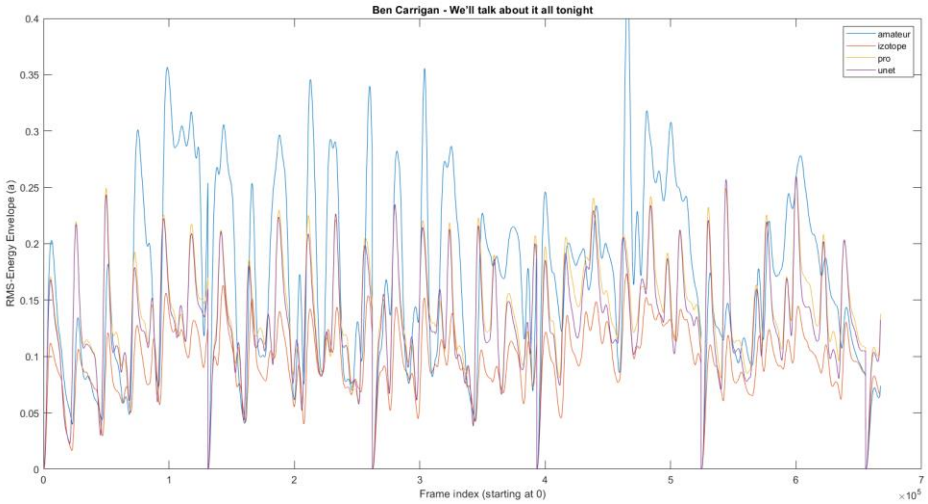
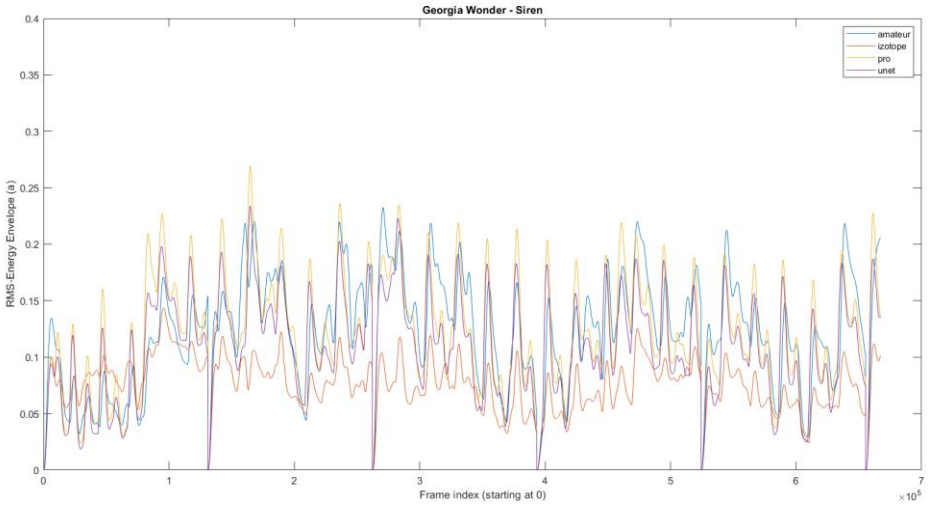
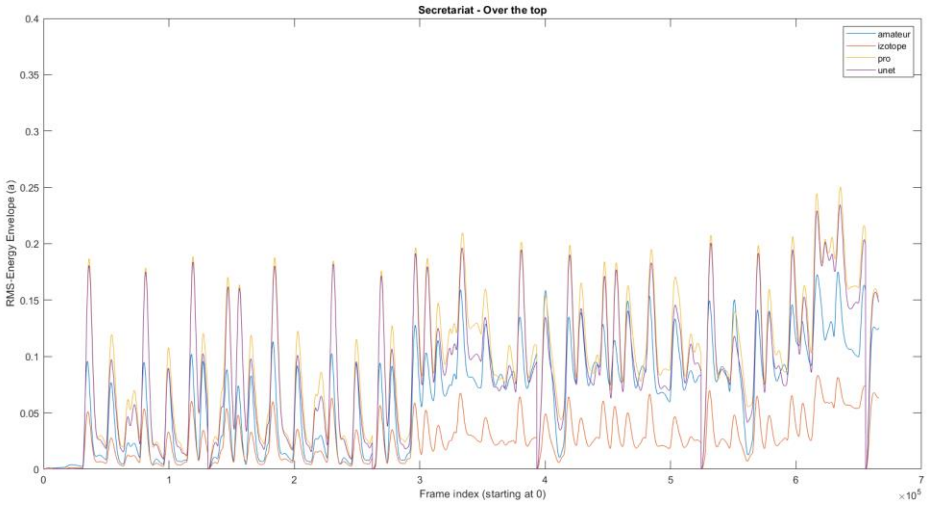


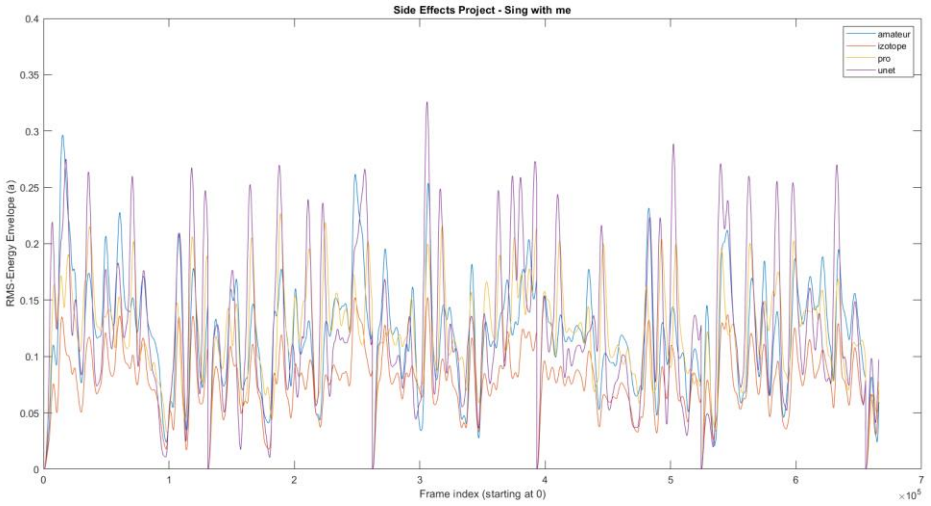
Fig. B.34. Odd-to-Even Harmonic Ratio descriptor calculated for all songs depending on mix type

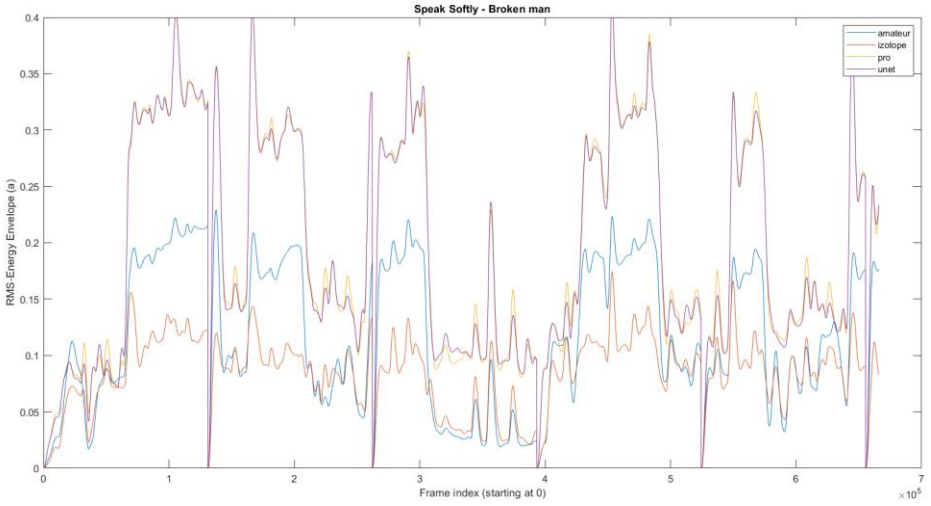












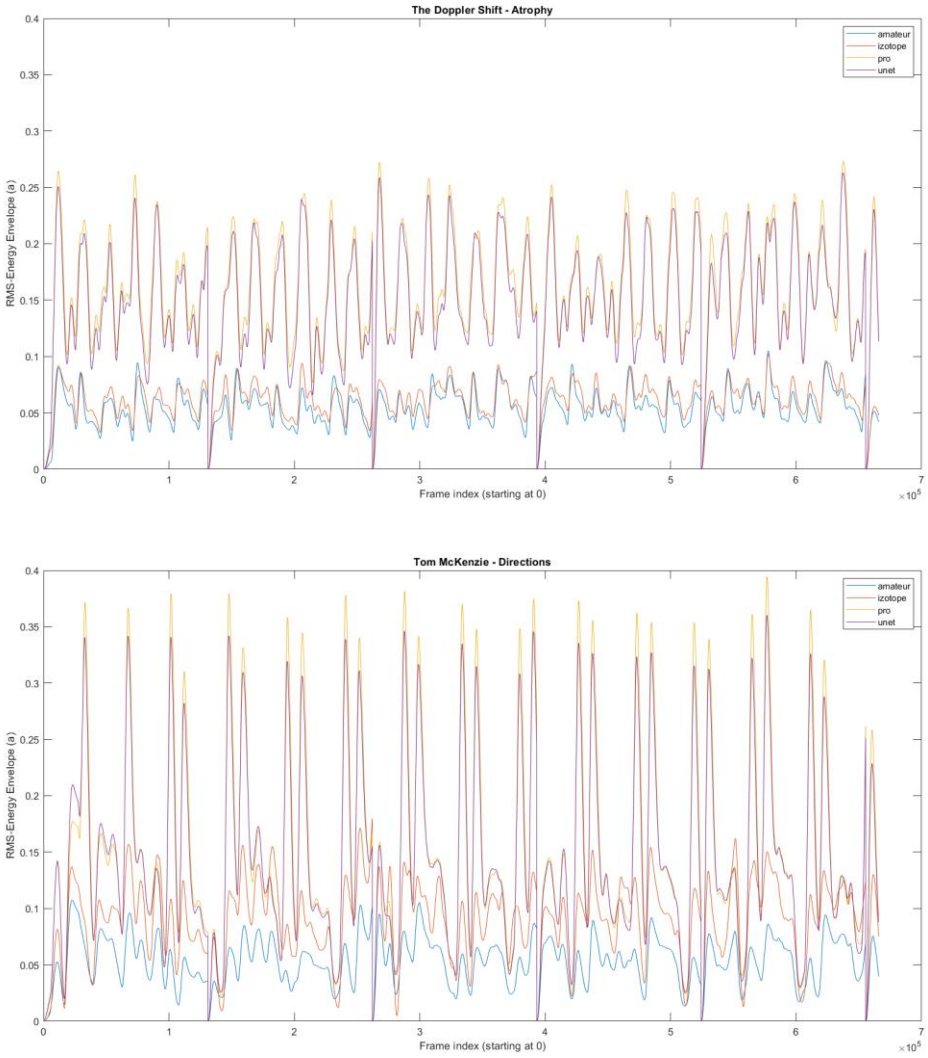


Fig. B.35. RMS-Energy Envelope descriptor calculated for all songs depending on mix type

Appendix C: Questionnaire for listeners

LISTENING TEST

You will hear eight songs mixed in four different ways. The samples are 15-seconds long, played in groups of four. Listen carefully and decide how you perceive each sample in these categories:

*- **Balance:** how well-balanced are the levels of each instrument, vocals, etc.?*

*- **Clarity:** how well does the mix represent the entire frequency range of the instruments?*

*- **Panning:** is every element in the song placed in the panorama in a way that makes sense to the artist and the listener?*

*- **Space:** do various elements in the song (as well as the overall mix) have a proper ambiance?*

*- **Dynamics:** does the mix allow each part of the song to "breathe" and develop? Is the compression of each element (and the overall mix) adequate?*

You can listen to the group of samples and compare them to each other in any order you choose. ***Rate the samples in each of the above categories from 1 (bad) to 5 (good).***

Please fill questionnaire before listening

Age: ……..

Gender: ……

Do you listen to music? ……. If yes, what music genre (list up to three most important to you): ……

What music genres are you familiar with (select all that apply):

- Jazz
- Folk Music
- Hip Hop Music
- K-Pop
- Blues
- Pop Music
- Country Music
- Rapping
- Reggae

- Rock Music
- Rhythm and Blues
- Punk Rock
- Classical
- Disco
- Heavy Metal
- Funk
- Techno
- Opera
- Gospel Music
- Other (please specify)

Are you a musician? ……. and /or sound engineer? …….

Are you a mixing engineer? ..... If yes, how many years of experience in mixing do you have: ……

What music genres are you usually mix (select all that apply):

- Jazz
- Folk Music
- Hip Hop Music
- K-Pop
- Blues
- Pop Music
- Country Music
- Rapping
- Reggae
- Rock Music
- Rhythm and Blues
- Punk Rock
- Classical
- Disco
- Heavy Metal
- Funk
- Techno
- Opera
- Gospel Music
- Other (please specify)

	1 - A	1 - B	1 - C	1 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

	2 - A	2 - B	2 - C	2 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

	3 - A	3 - B	3 - C	3 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

	4 - A	4 - B	4 - C	4 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

	5 - A	5 - B	5 - C	5 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

	6 - A	6 - B	6 - C	6 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

	7 - A	7 - B	7 - C	7 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

	8 - A	8 - B	8 - C	8 - D
Balance				
Clarity				
Panning				
Space				
Dynamics				

Appendix D: Self-similarity matrices

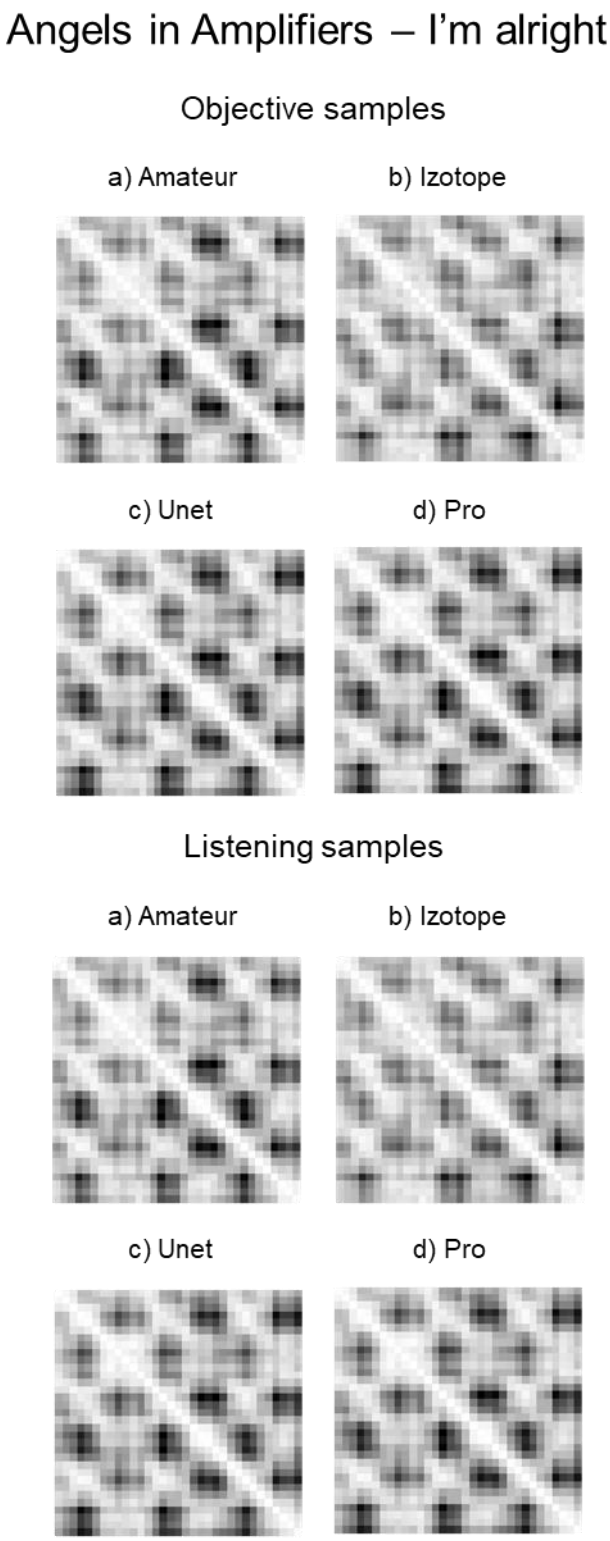


Fig. D.1. Graphical representation of the SSM of the Angels in Amplifiers – I'm alright objective and subjective samples

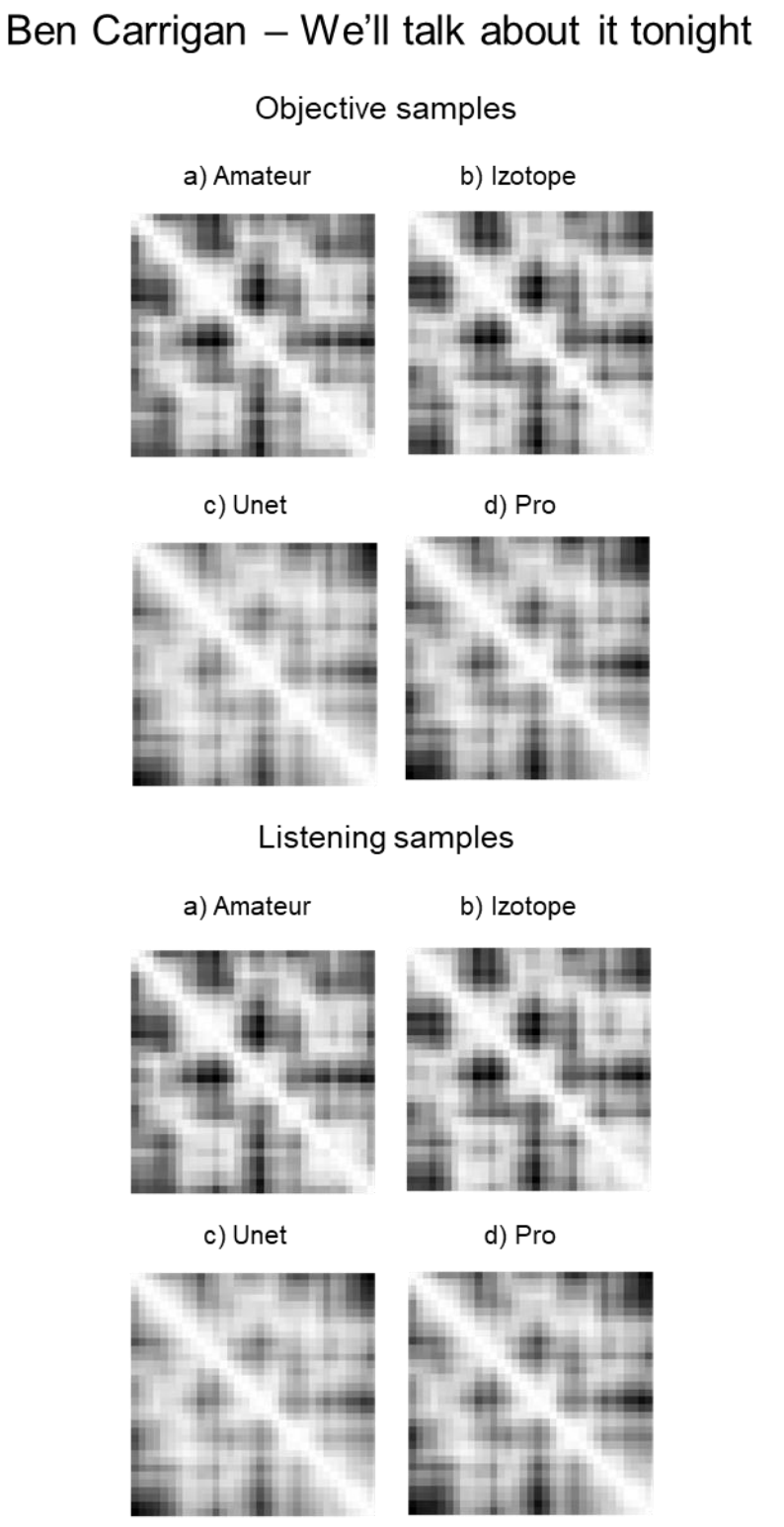


Fig. D.2. Graphical representation of the SSM of the Ben Carrigan – We'll talk about it tonight objective and subjective samples

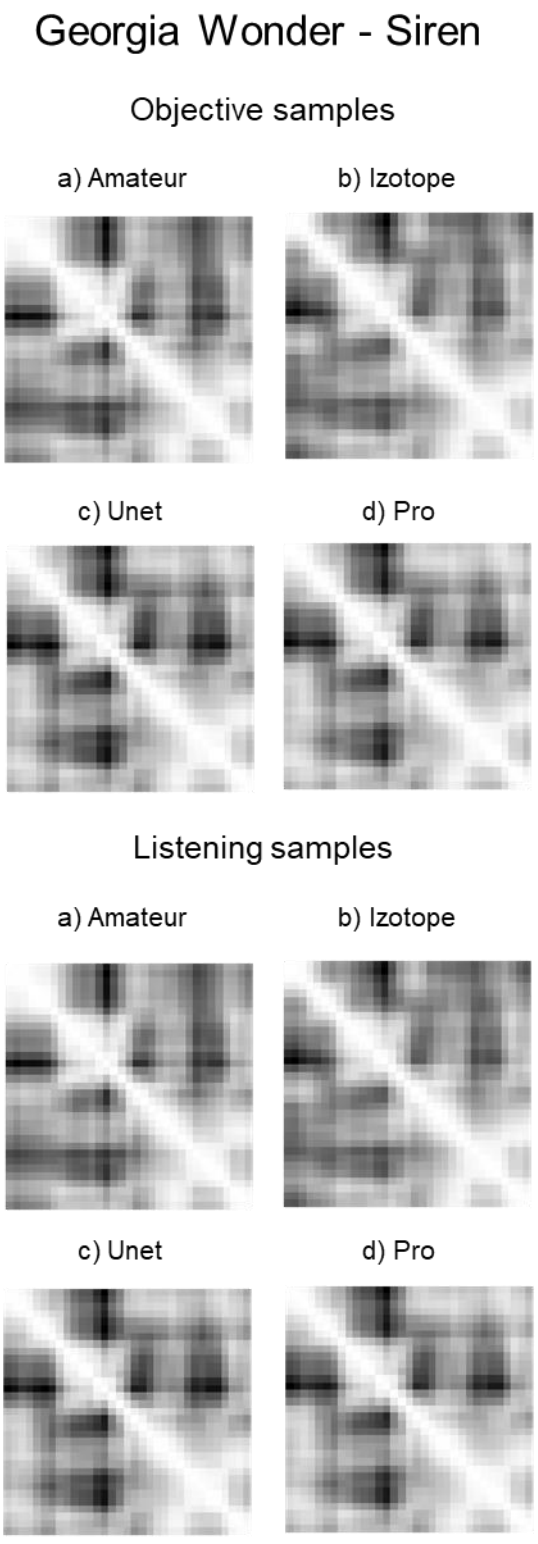


Fig. D.3. Graphical representation of the SSM of the Georgia Wonder – Siren objective and subjective samples

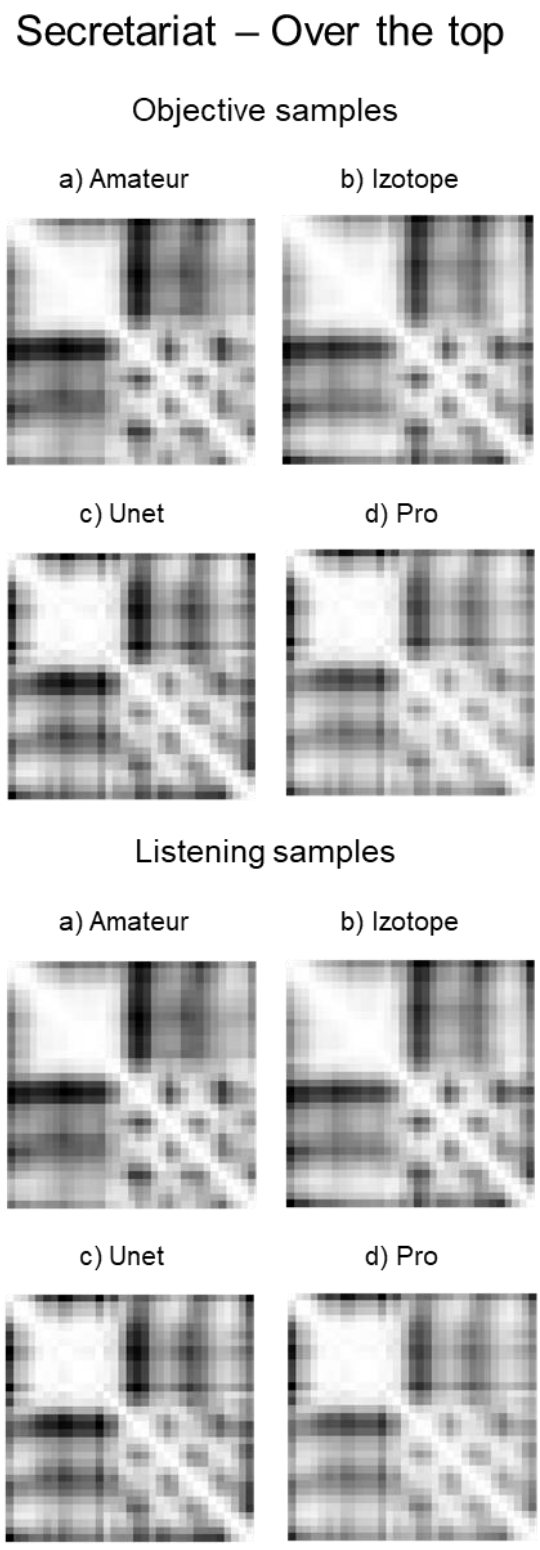


Fig. D.4. Graphical representation of the SSM of the Secretariat – Over the top objective and subjective samples

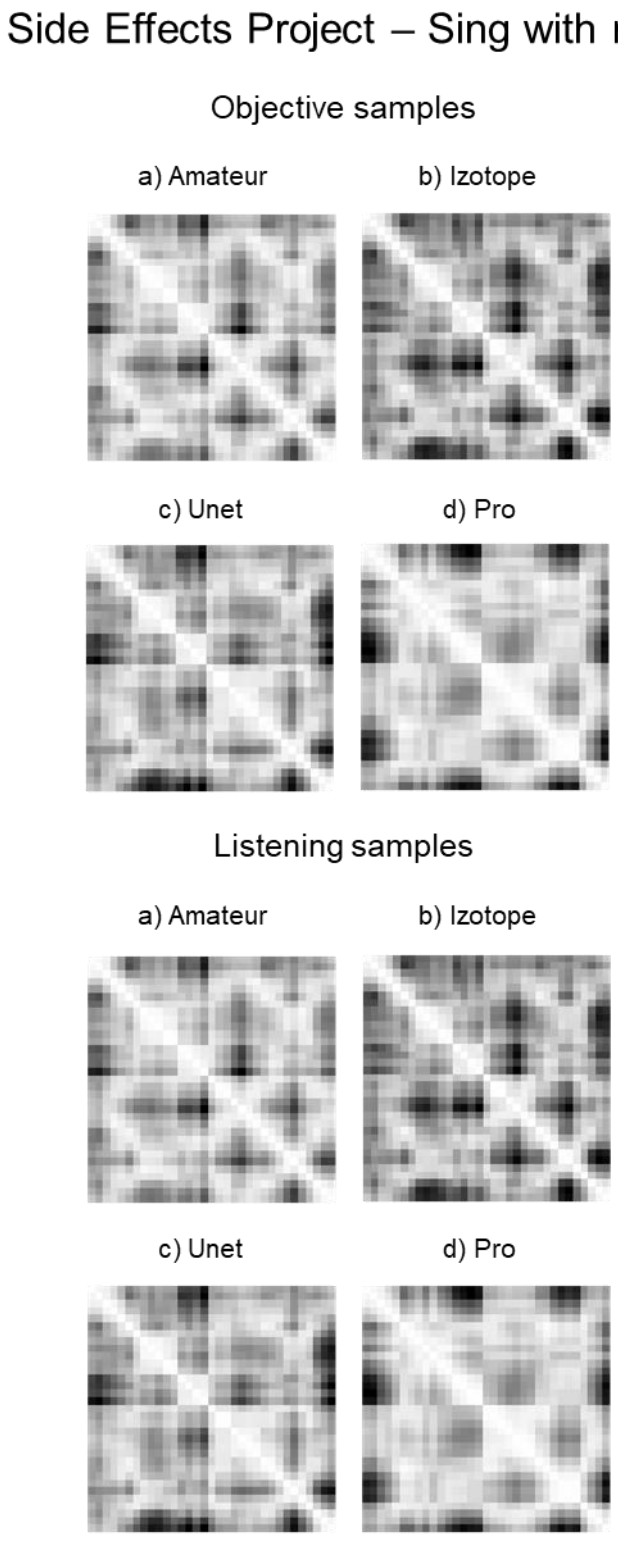


Fig. D.5. Graphical representation of the SSM of the Side Effects Project – Sing with me objective and subjective samples

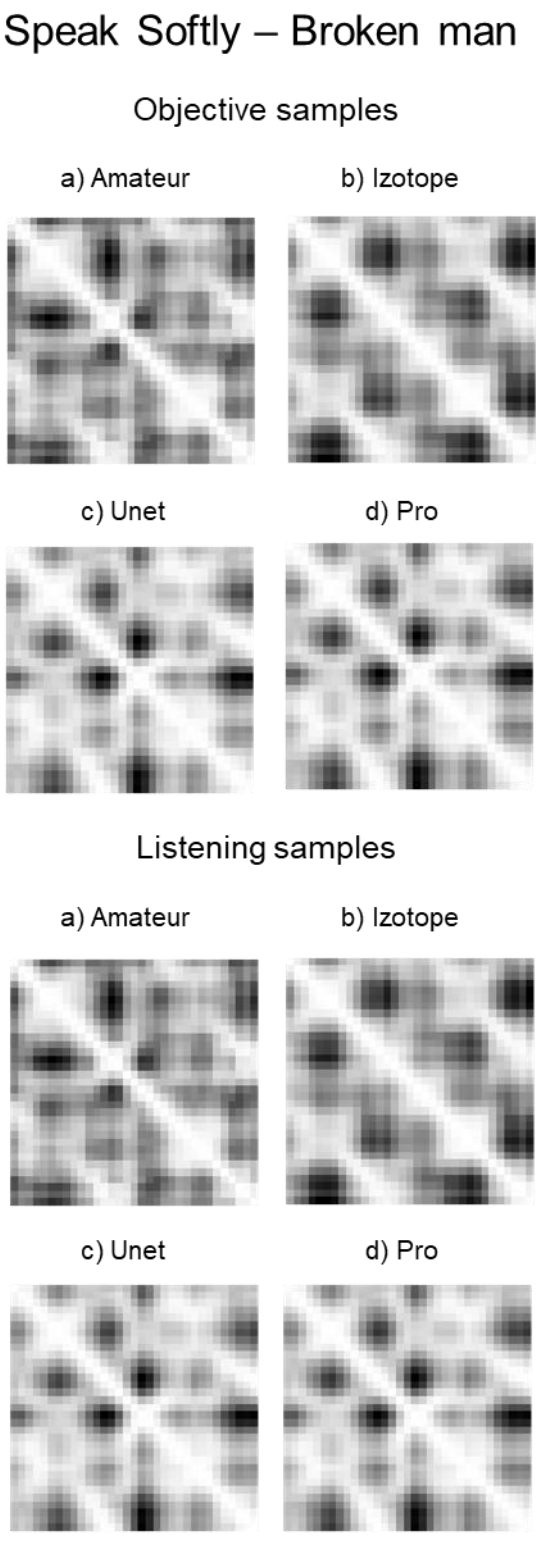


Fig. D.6. Graphical representation of the SSM of the Speak Softly – Broken man objective and subjective samples

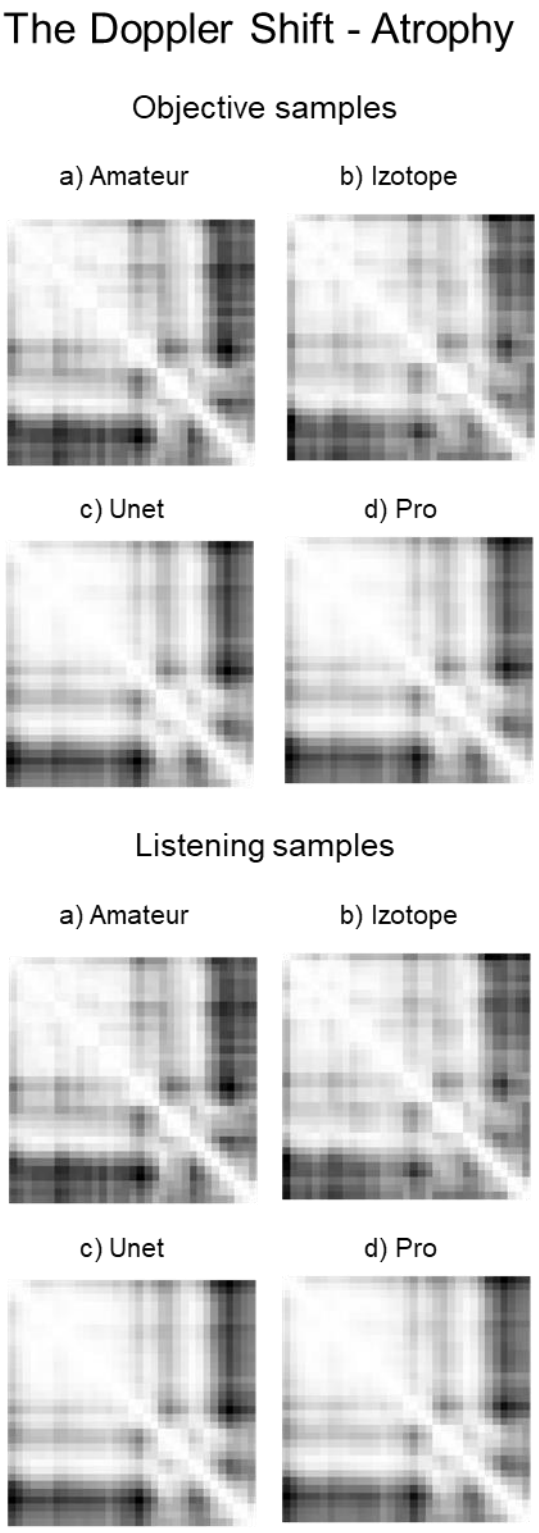


Fig. D.7. Graphical representation of the SSM of the The Doppler Shift – Atrophy objective and subjective samples

Appendix E: List of publications and patents

1. Koszewski, D., Marciniuk, K., Kostek, B., „Badanie wierności brzmienia dźwięku instrumentów wirtualnych VST/TRTAS", Aspekty Komputerowej inżynierii dźwięku. Od Metafory Do Standaryzacji, pp. 95-101, 2017.

2. Koszewski D., Kostek B., "Low-level audio descriptors-based analysis of music mixes from different Digital Audio Workstations - case study", IEEE SPA Proceedings, ISBN 978-83-62065-31-8, 2018 (WoS).

3. Koszewski D., Weber D., „Przykład zastosowania przetworników piezoelektrycznych do stworzenia elektronicznych padów na platformie sprzętowej Arduino", Studium Badawcze Młodych Akustyków ed. Katedra Mechaniki i Wibroakustyki AGH Kraków: Akademia Górniczo-Hutnicza, pp. 87-97, 2018.

4. Kurowski A., Koszewski D., Kotus J., Kostek B., "A Stand for Measurement and Prediction of Scattering Properties of Diffusers", 144 Audio Engineering Society Convention, e-brief, 2018.

5. Blaszke M., Koszewski D., Zaporowski Sz., "Real and Virtual Instruments in Machine Learning – Training and Comparison of Classification Results", IEEE SPA Proceedings, ISBN: 978-83-62065-34-9, 2019 (WoS).

6. Czyżewski A., Cygert S., Szwoch G., Kotus j., Weber D., Szczodrak M., Koszewski D., Jamroz K., Kustra W., Sroczyński A., Śmiałkowski T., Hoffmann P., "Comparative study on the effectiveness of various types of road traffic intensity detectors", 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2019.

7. Koszewski D., Kostek B., "Musical instrument tagging using data augmentation and effective noisy data processing", Journal of Audio Engineering Society, JAES Volume 68 Issue ½, pp. 57-65; January, 2020.

8. Blaszke M., Koszewski D., "Determination of Low-Level Audio Descriptors of a Musical Instrument Sound Using Neural Network", Signal Processing: Algorithms, Architectures, Arrangements, and Applications Proceedings, 2020.

9. Kalman M., Koszewski D., Mróz B., "Comparison of sound of organ pipes in contemporary and historical instruments", 148 Audio Engineering Society Convention, e-brief, 2020.

10. Maziewski P., Banas J., Koszewski D., Stanczak D., Trella P., "Analysis of Nonlinear Distortions in a Digital MEMS Microphone", 148 Audio Engineering Society Convention, e-brief, 2020.

11. Duzinkiewicz K., Koszewski D., Pietrusinska K., Trella P., "Overview of speech quality metrics in terms of automated evaluation of signal denoising in a presence of non-stationary noise", 149 Audio Engineering Society Conventio, e-brief, 2020.

12. Maziewski P., Banas J., Klinke P., Koszewski D., Pach P., Stanczak D., Trella P., "Environment classifier for detection of laser-based audio injection attacks", filed with the US Patent Office, patent pending publication number 20200243067, July 30, 2020.

13. Klinke P., Banas J., Koszewski D., Maziewski P., Pach P., Trella P., "Open-loop multichannel audio impulse response measurement and capture path evaluation", filed

with the US Patent Office, patent pending publication number 20200359146, November 12, 2020.

14. Klinke P., Koszewski D., Maziewski P., Banas J., Lopatka K., Kupryjanow A., Trella P., Pach P., “Acoustic signal processing adaptive to user-to-microphone distances”, filed with the US Patent Office, patent pending publication number 20210120353, April 22, 2021.

15. Klinke P., Trella P., Koszewski D., Pach P., Maziewski P., Banas J., “Method and system of audio device performance testing”, filed with the US Patent Office, patent pending publication number 20210306782, September 30, 2021.

16. Blaszke M., Koszewski D., Kostek B., „Skuteczność klasyfikacji gatunków muzycznych za pomocą sieci neuronowej w zależności od typu danych wejściowych”, Wydawnictwo Politechniki Wrocławskiej, (rozdział w książce) 207-224, 2021.

17. Koszewski D., Banas J., Maziewski P., Trella P., Pach P., Klinke P., Stanczak D., Kuklinowski M., “Recreating complex soundscapes for audio quality evaluation”, e-brief, 151 Audio Engineering Society Convention, 2021.

18. Banas J., Koszewski D., Trella P., Klinke P., Pach P., Grzywa M., Jezierski R., Maziewski P., Stanczak D., Kuklinowski M., “Overview of Evaluation Methods of Sound Field Reproduction Systems”, 152 Audio Engineering Society Convention, e-brief, 2022.

19. Koszewski D., Görne T., Korvel G., Kostek B., “Automatic music mixing system based on one-dimensional Wave-U-Net autoencoders”, EURASIP Journal on Audio, Speech, and Music Processing (in review), 2022.