NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY POLITEHNICA BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

DOCTORAL THESIS SUMMARY

Bogdan MOROŞANU

CONTRIBUŢII ÎN DOMENIUL PRODUCŢIEI MUZICALE FOLOSIND REŢELE NEURONALE

CONTRIBUTIONS IN THE FIELD OF MUSIC PRODUCTION USING NEURAL NETWORKS

DOCTORAL COMMITTEE

Prof. Univ. Politehnica din Bucureşti	Committee Chair
Prof. dr. ing. Constantin PALEOLOGU Univ. Politehnica din Bucureşti	Doctoral Supervisor
Prof. dr. ing. Cristian NEGRESCU Univ. Politehnica din Bucureşti	Reviewer
Prof. dr. ing. Radu-Mihnea UDREA Univ. Politehnica din Bucureşti	Reviewer
Dr. ing. Daiana-Amelia SARU Univ. Politehnica din Bucureşti	Reviewer

BUCHAREST 2025

Abstract

This thesis, entitled "Contributions in the Field of Music Production Using Neural Networks," proposes innovative solutions to the contemporary challenges of audio production, in a context characterized by an exponential increase in musical content and limited technical resources, especially in Romania. The research addresses the significant discrepancy between the growing number of artists and musical productions and the available professional technical infrastructure, exploring the potential of neural networks to optimize critical processes in the audio production chain.

The work develops a solid theoretical foundation of the principles of auditory perception and music production techniques, before proposing four major original contributions: an automatic system for recognizing musical instruments in complex recordings; a personalized level control algorithm for multi-channel recordings, based on perceptual loudness optimization; a multi-band compressor with machine learning parameter tuning, capable of dynamically adapting settings according to the audio material; and a spatial expansion technology from stereo to multi-channel, using a hybrid model of source separation and ambient extraction.

Each of these solutions has been implemented, tested, and evaluated both through objective metrics and subjective assessments carried out by professionals from the music industry. The results demonstrate that integrating neural networks into audio workflows can significantly improve the efficiency and quality of music productions, providing sound engineers and producers with advanced, accessible, and intuitive tools to meet the challenges of the current music market.

The contributions brought by this research have the potential to democratize access to professional sound quality and to enhance the competitiveness of Romanian music productions on the international market, in the context of an industry where the volume of content and qualitative expectations are constantly increasing.

Table of contents

Chapter 1

Introduction

The field of music production is undergoing an unprecedented period of transformation, characterized by the democratization of production tools and a digital revolution in content distribution. In 2022, over 34.1 million new tracks were uploaded to streaming platforms [50], and globally, approximately 120,000 new tracks are added daily [7], with an estimated annual total of over 43 million tracks.

The global market includes over 3 million artists on Spotify alone [41], and according to UNESCO, nearly 4 million jobs are directly related to the music industry [3]. In Romania, data indicates over 7,000 releases and 5,000 unique artists in 2022 [39], with the specific characteristic that the local market operates with limited technical resources and a deficit of specialists in advanced sound engineering.

1.1 Presentation of the thesis field

The thesis is positioned at the intersection of three fields: audio engineering, artificial intelligence, and contemporary music production. Audio engineering has migrated from the analog paradigm to predominantly digital ecosystems, while music production operates in an extremely competitive environment with massive volumes of content.

In Romania, there are approximately 10,000 active musicians, of which 88% are emerging or niche performers [38], in contrast to only 600-1,200 professionals in the field of sound engineering [25]. This discrepancy generates significant pressure on the existing technical infrastructure, with approximately 5,000-6,000 new compositions recorded annually [5].

The digital context of music consumption in Romania, a country with over 91% internet penetration [51] and ranked among the top 5 most developed digital markets in the region, creates a favorable environment for the implementation of advanced technologies based on artificial intelligence.

1.2 Objectives of the doctoral thesis

The main objective is the development, implementation, and evaluation of solutions based on neural networks for optimizing critical processes in modern music production. From a scientific perspective, the thesis aims to advance knowledge in the application of artificial intelligence in audio signal processing, contributing with original models adapted to the specific challenges of music production.

From a practical perspective, the research seeks to provide the professional community with advanced and accessible technological tools, in a global context where over 3 million artists compete for listeners' attention [41].

1.3 Content of the doctoral thesis

The thesis is structured into five main chapters:

- **Chapter 2 - Fundamentals of Perception and Sound Analysis in Audio Production** establishes the theoretical foundation regarding sound perception, psychoacoustics, and monitoring environments.
- **Chapter 3 - Music Production** details the fundamental processes of the audio production chain, from capture to mastering.
- **Chapter 4 - Neural Networks in Audio Production** presents the original contributions, including a personalized level control system for multi-channel recordings, a multi-band compressor with machine learning, and a technology for spatial expansion from stereo to multi-channel.
- **Chapter 5 - Conclusions** synthesizes the obtained results and outlines future development perspectives.

Overall, the work proposes an integrated approach to contemporary music production, in which neural network-based technologies are strategically used to overcome traditional limitations and democratize access to professional sound quality, with particular applicability in the context of the Romanian music industry.

Chapter 2

Fundamentals of perception and sound analysis in audio production

2.1 Critical listening

Critical listening is the foundation of the entire audio production process, being an essential skill that distinguishes professional sound engineers from regular listeners. Transcending the mere pleasure of musical listening, it involves a methodical analysis of sound components to achieve the desired artistic and technical result.

At its core, critical listening is based on developing an extraordinary awareness of sound details. Izhaki *et al.* [28] refer to this ability as "*the golden ear*" - the capacity to discriminate the subtlest details in a production that would go unnoticed by untrained listeners.

Critical listening operates simultaneously on two complementary planes:

- **Objective** - focused on the technical and measurable aspects of sound (frequency, amplitude, phase, dynamics) using spectrum analyzers, level meters, and phase indicators;
- **Subjective** - evaluates the emotional impact and musicality of the audio material.

Studies show that experienced professionals can identify frequencies with a margin of error of less than a quarter of an octave [33], and in the field of dynamic processing, they can perceive changes in compression ratio as small as 0.5:1 and variations in attack and release times on the order of milliseconds [18].

2.2 The ear and psychoacoustics in the context of audio production

2.2.1 Anatomy and functioning of the auditory system

The human auditory system is structured into three main sections:

External ear The auricle directs sound waves, achieving a natural amplification of 5-10 dB in the 2-7 kHz range [2]. One of the most remarkable functions of the external ear is its ability to modify the spectrum of incoming sound depending on its direction of propagation. This phenomenon, known in the literature as *"Head Related Transfer Function"* (HRTF), is fundamental to our ability to localize sound sources in three-dimensional space [35].

Middle ear Functions as an acoustic-mechanical transformer, amplifying sound vibrations through the principle of leverage and the difference between the areas of the tympanic membrane and the oval window, producing an amplification of approximately 26 dB. The middle ear also features a sophisticated protection mechanism - the stapedius reflex. This involuntary reflex is triggered by loud sounds (over approximately 85 dB SPL) and involves the contraction of two tiny muscles: the stapedius muscle and the tensor tympani muscle.

Inner ear The cochlea converts mechanical vibrations into neural signals. The basilar membrane varies in stiffness and mass, allowing spectral analysis according to Greenwood's relation [20]. The outer hair cells act as molecular amplifiers, enhancing sensitivity and sonic selectivity.

The auditory nervous system, from the cochlea to the primary auditory cortex, represents a complex network of hierarchical sound information processing [46]. This neural pathway is not merely a relay of information but a sophisticated system for analyzing and extracting relevant characteristics of the sound stimulus.

2.2.2 Fundamental principles of psychoacoustics

Psychoacoustics studies the relationship between the physical properties of sound and their perception by the human auditory system, essential for high-quality audio production.

Perception of frequency At low frequencies (below 500 Hz), the auditory system exhibits remarkably fine spectral differentiation, which progressively degrades as frequency increases, following an approximately logarithmic scale [23]. This characteristic

explains the consistent perception of musical intervals when the frequency ratio remains unchanged.

Perception of intensity Weber-Fechner's law describes how the auditory system compresses the wide range of physical intensities into a narrower perceptual range, allowing efficient operation both in quiet and noisy environments [23].

Spectral and temporal masking The presence of one sound affects the ability to perceive other sounds, with a characteristic asymmetry: low-frequency sounds can effectively mask high-frequency sounds ("*upward spread of masking*"). Temporal masking extends up to 20 ms before (pre-masking) and 200 ms after (post-masking) the masking sound.

2.2.3 Head related transfer function (HRTF)

HRTF is a complex function of frequency and direction that characterizes the acoustic modifications introduced by the listener's anatomy on incoming sounds:

$$\text{HRTF}(f,\theta,\varphi) = \frac{P(f,\theta,\varphi)}{P_0(f)} \tag{2.1}$$

where P represents the sound pressure at the eardrum for a sound with frequency f coming from the direction defined by azimuth θ and elevation φ, and P_0 is the sound pressure in free field.

Recent research [43] demonstrates that measuring HRTFs is possible even in reverberant rooms using advanced signal processing techniques. The total measured acoustic response can be expressed as:

$$h_{total,L/R}(t) = h_{amp,L/R} * h_{spk,L/R} * h_{room,L/R} * h_{L/R} * h_{mic,L/R} \tag{2.2}$$

where the components are:

- $h_{amp,L/R}(t)$ - impulse response of the audio amplifier
- $h_{spk,L/R}(t)$ - impulse response of the loudspeaker
- $h_{room,L/R}(t)$ - impulse response of the room
- $h_{L/R}(t)$ - desired HRIR for the left/right ear
- $h_{mic,L/R}(t)$ - impulse response of the microphones

In audio production, HRTFs allow the creation of a three-dimensional sound image using only two playback channels, being essential for spatial mixing and the production of content for virtual and augmented reality.

2.3 Control room

The control room, situated at the intersection of art and technology, is the essential space where audio material is evaluated and processed. The international standards ITU-R BS.1116 and EBU Tech 3276 establish rigorous parameters for these spaces.

Room geometry Dimensional ratios are governed by the relation:

$$1.1\frac{w}{h} \leq \frac{l}{h} \leq 4.5\frac{w}{h} - 4 \tag{2.3}$$

where l, w, and h represent the length, width, and height of the room, ensuring optimal distribution of room modes.

Reverberation time Calibrated according to the room's volume using the equation:

$$T_m = 0.25\left(\frac{V}{V_0}\right)^{1/3} \tag{2.4}$$

where V is the actual volume of the room, and V_0 is the reference volume of 100 m³.

Distribution of room modes Can be analyzed using:

$$F = \frac{c}{2}\sqrt{\frac{x^2}{l^2} + \frac{y^2}{w^2} + \frac{z^2}{h^2}} \tag{2.5}$$

where c is the speed of sound, and x, y, z are integers defining the type of mode.

Acoustic treatment The LEDE (Live-End Dead-End) concept involves predominant absorption in the front part and controlled diffusion in the rear area. For low-frequency control, bass traps and Helmholtz resonators are used, calibrated according to:

$$f = \frac{100R}{\sqrt{V(l + 1.6R)}} \tag{2.6}$$

where f is the desired resonance frequency, R is the radius of the opening, V is the cavity volume, and l represents the neck length of the resonator.

The research methodology adopted combines theoretical foundation with practical validation, using as a case study the development process of the specialized laboratory within the Faculty of Electronics, Telecommunications, and Information Technology, National University of Science and Technology POLITEHNICA Bucharest.

Experience shows that, although challenging, creating a space that meets international standards is achievable through the systematic application of acoustic principles, combined with precise measurements and progressive adjustments [43].

Chapter 3

Music production

Music production represents the complex process through which musical ideas are transformed into finished recordings, ready for distribution and consumption [24]. It is the point of convergence between art and technology, where creative vision meets technical expertise. In the digital era, the democratization of technology has made it possible to create high-quality music in personal studios, while also increasing the complexity of the process [53].

The process consists of two major interconnected stages: mixing and mastering. Mixing is the blending of recorded elements into a coherent and unified whole [28], manipulating relative levels, stereo field positioning, frequency bands, and dynamics. Mastering, the final stage, is the process of finishing and optimizing for distribution [30], ensuring that the material sounds consistent on any playback system.

3.1 The mixing process

Mixing represents the crucial point where art meets technology, transforming multiple audio sources into a coherent and expressive final product [28]. The process builds upon the foundation of the capture stage, as the quality of the recorded material defines the limits and possibilities of the mix.

Mathematically, for a stereo system with n sources, mixing can be modeled as a transfer function:

$$M_{L,R}(t) = \sum_{i=1}^{n} [S_i(t) \cdot A_i(t) \cdot P_i(t) \cdot F_i(\omega,t) \cdot D_i(t)] \tag{3.1}$$

where $M_{L,R}(t)$ represents the mixed signals for the left and right channels, $S_i(t)$ is the source audio signal, $A_i(t)$ is the amplitude, $P_i(t)$ is the panning function, $F_i(\omega,t)$ represents spectral processing, and $D_i(t)$ is the dynamic processing applied.

3.1.1 Establishing fundamental sound proportions

Establishing the fundamental sound proportions is one of the first and most critical steps in building a successful mix [48]. This starts with the elements that define the character and energy of the track, often the rhythm section (drums and bass) and the lead vocal in modern music.

The engineer must identify the center of gravity of the track — the frequency or frequency zone on which the entire sound construction relies [19]. For fast-tempo music, this center tends to be positioned higher in the frequency spectrum, while slower music can benefit from a lower center of gravity.

Optimal tonal balance cannot be defined in absolute terms, only in relation to the specific musical context and artistic intent [42], being influenced by the music genre, production era, and audience expectations.

3.1.2 Panning

Panning configures the distribution of sound energy between the left and right channels, playing an important role in creating the illusion of space and clarifying the sound architecture. This process is governed by psychoacoustic mechanisms: interaural intensity differences (IID) and interaural time differences (ITD) [6].

In current practice, panning fulfills two essential functions:

- Spectral separation and reduction of masking between instruments, especially those occupying similar spectral regions
- Creation of a perceptual hierarchy, with centrally placed sounds receiving increased attention, which is why key elements (lead vocal, kick drum) are positioned centrally

3.1.3 Spectral processing

Spectral processing forms the foundation of sound shaping in contemporary audio production. Equalization transcends simple technical correction, becoming an essential artistic tool for defining the character of a mix.

Mathematically, the transfer function of a parametric equalizer can be represented by:

$$H(f) = A_0 \cdot \prod_{i=1}^{n} \frac{1 + \frac{Q_i}{g_i} \cdot \left(\frac{f}{f_i}\right) + \left(\frac{f}{f_i}\right)^2}{1 + \frac{Q_i}{g_i \cdot A_i} \cdot \left(\frac{f}{f_i}\right) + \left(\frac{f}{f_i}\right)^2} \tag{3.2}$$

where A_0 represents the overall amplification, f_i is the center frequency of band i, Q_i is the quality factor determining the bandwidth, and g_i represents the gain or attenuation applied to that band.

Beyond the mathematical formulation, equalization involves understanding the spectral behavior of musical instruments, each occupying a specific spectral space that can be strategically shaped [28]. In this sense, equalization becomes an expressive language through which the sound engineer can:

- Define the hierarchy and roles of elements in a mix, highlighting certain instruments over others
- Create the sensation of spatiality and depth by manipulating spectral content
- Shape the timbral character of each element to fit the overall aesthetic vision
- Emphasize musical tension and release by strategically boosting or attenuating certain spectral regions

3.1.4 Dynamic processing

Dynamic processing directly influences the perception of intensity, impact, and cohesion of the audio material. The transfer function of a compressor can be represented by:

$$y(t) = \begin{cases} x(t), & \text{for } |x(t)| \leq T \\ T + \frac{|x(t)|-T}{R}, & \text{for } |x(t)| > T \end{cases} \tag{3.3}$$

where $x(t)$ represents the input signal, $y(t)$ the output signal, T the threshold, and R the compression ratio.

The effect of dynamic processing is not merely a modification of the signal level; it directly influences the three-dimensional perception and hierarchy of elements in a mix. The impact of dynamic processing is closely linked to fundamental psychoacoustic mechanisms [56], including:

- Perception of sound intensity, which is not linear and depends on spectral content
- Temporal masking, where loud sounds can mask softer sounds that precede or follow them
- Temporal integration of acoustic energy in the auditory system
- Adaptation to constant levels and sensitivity to changes

3.2 The mastering process

Mastering represents the final creative step in the music production chain, transforming the mix into the finished product intended for the listener [30]. In the contemporary era, optimization for streaming platforms has revolutionized the approach by introducing loudness normalization according to the EBU R128 and ITU BS.1770 standards [47].

3.2.1 Tonal balance

Tonal balance gives the music production clarity, depth, and spectral coherence [30], referring to the optimal distribution of sound energy across the entire frequency spectrum.

Technically, the process begins with a careful analysis of the mix using tools such as spectrum analyzers, whose results are compared with reference materials from the same music genre. This allows the identification of any spectral anomalies, such as excessive energy build-ups in certain frequency bands or deficiencies in others [10]. For example, a mix with too much energy in the 200-300 Hz range will sound muddy or "boxy," while lacking energy in the 3-5 kHz range may cause vocal details to be insufficiently present.

An important aspect to mention is that tonal balance in mastering involves complex interaction with other processes, especially dynamic compression [14]. Multi-band compressors, for example, significantly influence tonal balance by altering the ratio between various frequency ranges depending on the dynamic content of the material. This interconnection requires a holistic approach, where tonal and dynamic adjustments complement each other.

3.2.2 Saturation

Saturation contributes to shaping the sonic character and enriching the harmonic palette [30], being a controlled distortion process that emulates the behavior of analog circuits. The main types include:

- **Tape saturation** - generates predominantly even harmonics and gently compresses transients
- **Tube saturation** - adds predominantly even harmonics in a progressive manner
- **Transformer saturation** - introduces subtle waveform modifications
- **Digital waveshaping saturation** - uses mathematical algorithms for controlled distortion

3.2.3 Loudness optimization

Loudness optimization aims to maximize perceptual impact without compromising artistic quality and musical dynamics [30]. The ITU-R BS.1770 standard [26] provided the mathematical basis for measuring perceived loudness, being adopted and extended by EBU R128 [16].

Multi-stage compression Provides nuanced control over dynamics without sacrificing natural character [14], based on the principle that multiple moderate stages result in a more musical outcome than aggressive compression.

Limiting Limiters have a dual role: preventing digital distortion and maximizing perceptual level [30]. In the era of perceived loudness normalization, the maximization strategy has evolved towards maintaining a moderate level of limiting, preserving dynamics and impact [54].

3.3 Conclusions

Music production is a complex symbiosis between art and technology. Mixing and mastering rely on rigorous technical principles, but transcend the purely mathematical dimension by incorporating indispensable artistic elements.

In the contemporary era, dominated by streaming platforms with loudness normalization algorithms, the traditional approach to mastering has undergone fundamental transformations. The focus has shifted from simply maximizing loudness to maintaining natural dynamics and timbral characteristics.

The artistic value of a music production does not derive from its perceived loudness alone, but from its ability to convey emotion and expressiveness through optimal sound quality.

Chapter 4

Neural networks in audio production

The revolution of artificial intelligence has fundamentally transformed audio production in recent years, extending to almost all aspects of the process: from synthesis and processing to mixing, mastering, and restoration [45]. At the heart of this transformation are neural networks, which offer unprecedented capabilities to model, generate, and manipulate complex audio data [15].

However, most proposed approaches aim to develop universal solutions applicable to a wide range of audio problems [9]. While this generalist orientation is attractive from the perspective of applicability and commercialization, it often leads to suboptimal results when faced with the specific challenges of various audio production subfields [22].

In this context, the thesis proposes a reevaluation of the current paradigm, advocating for the development of specialized neural architectures optimized for specific challenges in audio production. This shift does not represent a step back in the universality of AI-based solutions but rather a necessary maturation of the field, similar to the evolution of other technologies that have progressed from generalist solutions to high-performance specialized tools [55].

4.1 Musical instrument classification

Recognition and classification of instruments represent a fundamental component in modern audio processing, with multiple applications in automatic mixing, sound library organization, and music recording analysis [21]. This section presents a hierarchical classification system that uses both spectral characteristics and textual information for precise identification of musical instruments in multi-channel recordings [36].

Implementing a robust instrument recognition system enables the automation of laborious steps, facilitating the organization of tracks into functional groups, the application of appropriate templates, and the setting of optimized initial levels for each type of instrument. This automation not only accelerates the workflow but also standardizes

the mixing practice, reducing inconsistencies between projects and minimizing human errors.

A significant limitation of current systems is that most models are trained on isolated high-quality recordings, which do not reflect the characteristics of signals encountered in real recording sessions. These often contain channel bleed, varying microphone positions, and room acoustics that significantly alter the spectral fingerprints, complicating the classification process [30].

The proposed architecture is structured into two classification levels: initially, a primary classifier separates the audio material into four fundamental categories (drums, bass, vocals, and others), followed by specialized secondary classifiers providing more detailed identifications for highly diverse categories such as drums and "others."

To improve classification performance, especially in recordings with bleed or ambiguous spectral characteristics, a multi-modal approach was developed combining spectral analysis with textual metadata analysis (track names). This method is based on the observation that, in sound engineering practice, track names often contain valuable clues about the type of instrument recorded.

The system's inference process selects the three most representative windows with the highest RMS energy from each audio track and processes both the audio content and the track name in parallel. The primary classifier categorizes the audio material into the four main classes, and for the Drums or Others categories, a secondary classifier provides detailed identification. The multi-modal model combines audio spectrograms with the track name text. The final class is determined by comparing the confidence scores from valid predictions, favoring the result with the highest confidence.

The performance of the instrument classification system, detailed in Table 4.1, was evaluated using multiple datasets, including MUSDB18HQ [11], the Cambridge Multitrack collection [12], and custom datasets for specialized instrument categories. The primary classification model achieved an accuracy of 83.89% on the MUSDB18HQ validation set, demonstrating its ability to efficiently distinguish between the four main instrument categories. The specialized classifier for drums recorded excellent performance with an accuracy of 95.67%, while the classifier for the "others" category achieved 89.10% accuracy.

Tabel 4.1 Acuratețea de validare pentru fiecare model de clasificare.

Model	**Set de date**	**Tip intrare**	**Clase**	**Acuratețe [%]**
ResNet-18	MUSDB18HQ [11]	STFT	4	83,89
ResNet-18	Tobe personalizat	STFT	5	95,67
ResNet-18	Altele personalizat	STFT	3	89,10
Multi-modal	Cambridge [12]	Text + STFT	27	82,19

Applied in real multi-channel projects, the system demonstrated a variable adaptability depending on the complexity of the music production scenario. For small-scale projects (10-22 tracks), the results were remarkable, with classification accuracy ranging between 90% and 100%. However, in medium and large-scale projects (34-60 tracks), the system encountered significant difficulties, with accuracy dropping to values between 47.72% and 70.58%.

Despite the identified limitations, the instrument recognition system demonstrates the potential of machine learning applications in professional audio production. By combining spectral analysis with textual information processing in a hierarchical decision framework, we obtain a robust classification system that can operate efficiently in various music production contexts, significantly reducing the time required for organizing and manually labeling complex audio sessions.

4.2 Custom level control for multi-channel recordings

The field of mixing is undergoing continuous transformation due to the introduction of automated solutions aimed at imitating or assisting the complex and artistic process that traditionally depends on the expertise of a professional. This expertise is essential for adjusting dynamics, spatial characteristics, timbre, and tonality in multi-channel recordings. Modifying these components involves a series of procedural steps using linear and non-linear processes to achieve the desired auditory result [28].

For this project, a dataset was created comprising 20 music tracks, combining pieces from the *Cambridge Music Technology* database [48] and recordings of local Romanian bands provided by the engineers participating in the experiment. The selected tracks cover a wide variety of genres and styles, providing a solid base for analyzing the specifics of the audio mixing process.

Genetic algorithms [29, 32] were chosen to efficiently manage this complex problem with multiple solutions. The method uses a genetic algorithm to iteratively refine the mixing coefficients, aiming to obtain an optimal mix. The customized genetic algorithm ensures that only the best solutions survive and pass on their characteristics.

To evaluate the feasibility of the proposed method, in addition to the genetic algorithm, a neural network was implemented to solve the problem of automatic perceptual loudness control in an adaptive manner. A convolutional neural network (CNN) architecture was chosen for audio signal analysis as they excel at identifying patterns in sequential data such as sound.

Both proposed methods (the genetic algorithm and the neural network) were applied for each of the 10 sound engineers, aiming to capture the preferences of each. Thus, 10 customized genetic algorithm solutions and 10 trained neural network models resulted.

Following the analysis of the subjective test results, detailed in Table 4.2, on the 15 tracks used for training, it is observed that the genetic algorithm obtained scores equal to

or higher than the mixes created by the engineers in half of the cases, specifically for 5 of the 10 engineers. The CNN neural network also performed well, matching or exceeding the MIX in 4 of the 10 cases. In situations where GA and CNN did not surpass the MIX scores, the differences were minor.

Tabel 4.2 Scorurile medii pentru setul de date de antrenare format din 15 melodii. Cele mai bune rezultate sunt evidențiate cu albastru.

Versiune	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**Medie**
RAW	3,5	3,9	3,5	3,1	3,5	2,9	3,7	1,2	3,5	1,5	3,0
MIX	4,3	4,1	4,0	3,7	3,3	3,3	4,3	2,2	4,0	4,0	3,7
GA	4,6	3,9	4,2	3,6	3,1	3,4	3,9	2,9	4,2	4,5	3,9
CNN)	4,5	4,2	4,5	3,7	3,1	3,7	4,1	3,3	4,5	3,7	3,9

The results obtained from the implementation and evaluation of the algorithm for custom level control in multi-channel recordings, also presented in the previously published paper in the *Algorithms* journal [37], demonstrate the considerable potential of automated methods in the field of audio mixing.

4.3 Multi-band compressor with machine learning parameter tuning

Multi-band compression is one of the most versatile and complex audio processing techniques used in modern music production. Unlike conventional compressors that process the entire frequency spectrum uniformly, multi-band compressors split the audio signal into separate frequency bands, thus allowing individualized dynamic processing for each spectral region [17].

However, optimal parameter configuration of a multi-band compressor represents a considerable challenge even for experienced engineers. The complex interactions between the parameters of multiple bands, along with the dynamic behavior of the audio material, create a vast and non-linear decision space. This context has stimulated intensive research towards developing automated assistance or dynamic processing configuration systems.

The software implementation of the multi-band compressor with machine learning was carried out using JUCE (*Jules' Utility Class Extensions*), a strategic choice based on multiple technical and practical advantages. JUCE currently represents the industry standard for developing professional audio applications, offering a complete and mature ecosystem for implementing audio effect processors [52].

The implementation of the multi-band processing component was one of the central challenges of developing the compressor, requiring complex technical solutions to ensure

precise spectral separation without introducing noticeable artifacts. The division of the audio spectrum into four distinct bands (low, low-mid, mid, and high frequencies) was achieved using fourth-order Linkwitz-Riley (LR4) filters [31], a choice determined by their superior characteristics in audio band crossover applications.

The architecture of the machine learning system for the multi-band compressor parameters is based on a hybrid solution combining siamese convolutional neural networks (CNN) with regression using decision trees (*Random Forest*). This structure was designed to extract relevant features from audio signals and model the complex relationships between compressor parameters and the spectral changes introduced by it.

The training of the model for predicting the parameters of the multi-band compressor involved a complex approach based on siamese networks and style transfer learning. The process used audio recordings from 3 sound engineers, having both the unprocessed and the processed versions for five songs.

The results obtained, detailed in Table 4.3, present the prediction errors and reveal significant aspects regarding the model's behavior. With a total average prediction error of 10.2% for all parameters, the system demonstrates a promising capability to approximate the optimal settings of the multi-band compressor. This overall precision, combined with the excellent performance for critical parameters such as attack time, validates the approach based on siamese neural networks for partial automation of the dynamic compression process in music production.

Tabel 4.3 Eroare medie de a predicției per parametru și eroare totală

Parametru	**Eroare Medie**
Atac	3,17%
Revenire	7,44%
Câștig	3,76%
Rație	20,36%
Prag	16,34%
Frecvență	10,26%
Media erorilor	**10,2%**

The perceptual evaluation of the results, conducted through critical listening tests with three sound engineers (two professionals and one amateur), provides valuable insights into the sonic quality of the processing performed by the implemented model. The audio material processed through the neural network (NN) obtained an average score of 3.7 on a scale from 1 to 5, significantly surpassing the unprocessed material (RAW) and approaching the professionally processed material (MIX), which scored 4.1.

4.4 Spatial expansion from stereo to multi-channel

The development of multi-channel sound systems has evolved significantly over time, starting with simple mono systems and progressing to stereo configurations. This evolution later continued by adding more channels and implementing advanced audio processing technologies such as 5.1 formats (standardized by ITU 775) [49] or 7.1, which are frequently used in cinemas and personal *home theater* systems.

Despite the emergence of these modern technologies, the stereo format, developed by Blumlein in the 1930s [1], remains the most widespread audio system, consisting of two speakers (left and right) designed to simulate the way human hearing perceives sounds. The multi-channel expansion process refers to transforming audio content from n to m channels, where $n < m$.

Various multi-channel expansion methods have been previously developed, based on separating the content into primary and ambient components [4, 44]. The software implementation of the hybrid model for source separation and ambient extraction combines some of the most efficient strategies for extracting multiple channels from a stereo mix. A source separation algorithm was used to extract 4 separate musical sources: drums, bass, vocals, and other sounds.

This module is based on a deep neural network (DNN) developed by Défossez *et al.* [13], a model that receives as input a 44-second stereo audio sample at 44.1 kHz and extracts 4 audio sources (drums, bass, others, vocals), using a hybrid approach based on source separation in both the spectral and waveform domains. This algorithm ranked first at the *Music Demixing Challenge* (MDX) in 2021, organized by Sony [34].

In addition to the four sources obtained through the source separation module, a primary-ambient extraction algorithm was also used, but only the ambient sounds from the stereo input signal were used, as these significantly contribute to creating an immersive effect. For this purpose, the *Adaptive Panning* (ADP) algorithm [8, 27] was used, based on the LMS adaptive filter, which estimates the primary and ambient sounds from the stereo channels.

The audio signal resulting from the multi-channel expansion process must combine all sources harmoniously to create an engaging multi-channel sound experience. The separated audio sources must be carefully distributed in the 5.1 system to ensure a faithful representation of the original signal and maintain the complex relationships between the sound sources.

The experimental results, published at the *SpeD* 2023 conference [40], convincingly demonstrate the superiority of the proposed method. Compared to classical approaches, the proposed method reflects the advantages of using the combination of source separation and ambient extraction techniques to generate a natural immersive sound.

The SSPAE spatial expansion algorithm demonstrated its superiority over traditional methods, offering significant improvements in the immersive playback of stereo materials

through multi-channel formats. The integration of the LFE channel, combined with advanced source separation techniques, led to superior results in objective tests based on the SDR metric.

Chapter 5

Conclusions

This thesis presented contributions in the field of music production using neural networks, focusing on the development of automated and adaptive solutions to enhance audio mixing processes and the listener's experience. The proposed investigations and solutions address essential aspects of modern audio production, from sound perception and analysis to instrument recognition, audio level control, dynamic processing, and spatial expansion.

The research was oriented towards creating specialized solutions for specific challenges in audio production, focusing on machine learning techniques capable of emulating expert human decisions while improving existing workflows. This scientific endeavor brought together knowledge from diverse domains such as psychoacoustics, audio engineering, signal processing, and artificial intelligence.

Particular attention was given to the context of critical listening, understanding auditory perception mechanisms, and the physical listening environment, considered essential foundations for developing efficient algorithms. Throughout the research, the proposed methods were evaluated both through objective metrics and subjective tests, thus ensuring the validation of solutions in real usage contexts.

5.1 Results

Chapter two provided a detailed analysis of the fundamentals of sound perception and analysis, essential for developing algorithms in music production:

- The principles of critical listening were investigated and documented, highlighting both the objective and subjective aspects involved in evaluating audio quality.
- The anatomy and functioning of the human auditory system were analyzed, with emphasis on spatial perception mechanisms and neural sound processing.
- The fundamental psychoacoustic principles were presented, including frequency perception, intensity perception, and spectral and temporal masking phenomena.

- A detailed study of the head-related transfer function (HRTF) was elaborated, discussing its implications in spatial sound reproduction.
- A case study was conducted regarding the design and implementation of a control room for critical listening, including aspects of acoustic isolation, acoustic treatment, and compliance with international standards.

Chapter three offered a comprehensive analysis of the mixing and mastering processes, integrating both technical and artistic perspectives:

- The fundamental principles governing the mixing process were systematized, with emphasis on setting sound proportions, strategic panning, and spectral perspective manipulation.
- A conceptual framework for spectral processing in mixing was developed and documented, including additive and subtractive equalization techniques, as well as methods for building an intuitive library of spectral associations.
- Advanced dynamic processing techniques were analyzed, including serial, parallel, and side-chain compression, evaluating their impact on spatial perception and sound cohesion.
- A mathematical model for loudness optimization was presented, considering modern normalization standards (EBU R128, ITU BS.1770), offering an integrative perspective on the evolution of standards in the streaming platform ecosystem.
- Harmonic saturation techniques in mastering were investigated, analyzing the role of controlled distortions (tape, tubes, transformers) in shaping sonic character and optimizing spectral density.

Chapter four presented original contributions in applying neural networks to automate and enhance essential processes in audio production:

- A hierarchical system for automatic musical instrument recognition was designed and implemented, aimed at optimizing session organization in digital audio workstations (DAWs). The system combines multiple classification models to precisely identify 27 distinct types of instruments, achieving up to 100% accuracy for small-scale projects and an average accuracy of 68.33% for complex projects recorded in multiple studios with varying track naming conventions.
- A personalized level control system for multi-channel recordings was developed, using both genetic algorithms (GA) and neural networks (CNN), achieving performances competitive with those of professional sound engineers. The system demonstrated its ability to reproduce the sonic preferences of sound engineers, with an average error below 5% for the attack parameter.

- A multi-band compressor with automatic parameter learning was implemented, using a siamese neural network architecture capable of determining optimal dynamic settings for different types of musical instruments. Perceptual tests showed an average difference of only 0.4 points between manual and automated processing on a 5-point scale.
- A hybrid algorithm for spatial expansion of stereo audio signals to 5.1 multichannel format was developed, combining source separation and ambient extraction techniques. The proposed method (SSPAE) demonstrated superior performance compared to existing methods, with improvements of up to 6 dB in the SDR (Signal to Distortion Ratio) metric.

5.2 Original contributions

The main original contributions of this thesis are categorized into four major directions:

1. **Design and implementation of a control room for critical listening** [**1**, **2**]
    - An acoustic design methodology based on international standards and experimental research was developed.
    - Modern measurement and acoustic treatment techniques were implemented and validated to obtain a neutral and precise listening environment.
    - A method for measuring HRTFs in reverberant rooms was elaborated, eliminating the need for costly anechoic chambers.
2. **Classification and automatic recognition of musical instruments in multichannel recordings** [**3**]
    - A hierarchical system for musical instrument classification was developed, based on specialized CNN models for different instrument categories.
    - A multi-modal approach was implemented, integrating both spectral analysis and textual metadata processing from track names.
    - A decision algorithm based on adaptive confidence thresholds was designed to maximize classification accuracy in real production scenarios.
    - The system's performance was evaluated on complex recordings from multiple studios, demonstrating the solution's robustness against naming inconsistencies and acoustic bleed between tracks.
3. **Development of automatic audio level control algorithms in multi-channel production** [**4**, **5**]

- A genetic algorithm for optimizing perceptual loudness was designed and implemented, capable of approximating the subjective preferences of sound engineers.
- An adaptive neural network (CNN) for level control was developed, offering increased flexibility depending on the processed audio material.
- Both methods were experimentally validated, demonstrating performances comparable or superior to manually created mixes by professionals.

4. **Adaptive systems for dynamic processing and spatial audio expansion** [**6**, **7**]

- A multi-band compressor with automatic parameter learning was implemented, using a siamese neural network architecture for stylistic transfer between sound engineers.
- A hybrid stereo-to-5.1 spatial expansion system was developed, integrating neural network-based source separation techniques with ambient extraction algorithms.
- Acoustic quality evaluation methods were proposed and validated for spatial expansion systems, demonstrating the superiority of the proposed solution compared to existing methods.

5.3 List of original papers

[**1**] **Moroşanu, Bogdan**, Victor Popa, Cristian Negrescu, and Ionuț Ficîu. "*Control room design for subjective audio critical listening.*" In Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies XI, vol. 12493, pp. 414-421. SPIE, 2023.

[**2**] Popa, Victor, **Bogdan Moroșanu**, and Cristian Negrescu. "*Head related transfer function measurement in reverberant rooms.*" In Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies XI, vol. 12493, pp. 617-622. SPIE, 2023.

[**3**] **Moroşanu, Bogdan**, Marian Negru, Georgian Nicolae, Horia Ioniţă, and Constantin Paleologu. "*A Machine Learning-Assisted Automation System for Optimizing Session Preparation Time in Digital Audio Workstations.*" Information 16, under review (2025). Journal Rank: CiteScore - **Q2** (Information Systems).

[**4**] **Moroşanu, Bogdan**, Marian Negru, and Constantin Paleologu. " *Automated Personalized Loudness Control for Multi-Track Recordings.*" Algorithms 17, no. 6 (2024): 228. (Journal Rank: JCR - **Q2** (Computer Science, Theory and Methods) / CiteScore - **Q1** (Numerical Analysis)), WOS:001254517300001.

[**5**] **Moroşanu, Bogdan**, Marian Negru, Ana Neacşu, Cristian Negrescu, and Constantin Paleologu. "*Personalized Multi-Track Leveling Algorithm.*" In 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 24-29. IEEE, 2023.

[**6**] Negru, Marian, **Bogdan Moroşanu**, Ana Neacşu, Dragoş Drăghicescu, and Cristian Negrescu. "*Automatic Audio Upmixing Based on Source Separation and Ambient Extraction Algorithms.*" In 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 12-17. IEEE, 2023.

[**7**] Ioniţă, Horia Sebastian, Victor Popa, and **Bogdan Moroşanu**. "*Multiband Dynamics Compressor with Automatic Parameter Learning.*" In 2024 15th International Conference on Communications (COMM), pp. 1-4. IEEE, 2024.

5.4 Future development perspectives

The research presented in this thesis opens multiple directions for future development, including:

1. **Extending AI algorithms for music production**
    - Development of complete automatic mixing systems integrating instrument recognition, level control, dynamic processing, and spatial positioning in a unified framework.
    - Designing deep learning models capable of capturing and reproducing the specific styles of producers or musical genres.
    - Implementation of optimization algorithms for processor parameter settings based on contextual analysis of the audio material.
    - Refinement of multi-modal instrument recognition models by integrating temporal and spectral information at the project level, not just individual tracks.
2. **Adaptive systems for immersive audio experiences**
    - Development of spatial expansion algorithms adapted for advanced formats such as Dolby Atmos or object-based audio systems.

- Integration of personalized HRTF techniques to enhance spatial sound reproduction on headphones.
- Development of interactive music production systems that adapt in real time to user preferences.
- Extension of the instrument recognition system to automatically identify optimal spatial positions in surround or object-based mixes.

3. **Advanced perceptual evaluation**

   - Development of standardized protocols for the subjective evaluation of automated music production solutions.
   - Creation of improved objective metrics for quantifying the quality of dynamic processing and spatial expansion.
   - Integration of user feedback into the learning cycle of models for continuous performance refinement.
   - Development of methods to evaluate the impact of automatic instrument classification on workflow efficiency and mixing processes.

4. **Building a comprehensive audio research database in the Romanian context**

   - Systematic collection of audio materials by involving students in practical recording, mixing, and mastering projects, thus leveraging the educational and research potential of university laboratories.
   - Establishment of strategic partnerships with recording studios, production houses, and artists from the Romanian music industry to gain access to professional multi-track recordings from various musical genres.
   - Documentation and standardization of metadata associated with recordings, including information on capture techniques, equipment used, and processing parameters, to facilitate comparative research and the development of algorithms specialized in the specific characteristics of Romanian productions.
   - Integration of this database into a collaborative platform open to the academic community, enabling the evaluation and validation of new audio processing algorithms, thus contributing to the evolution of the Romanian audio research ecosystem.
   - Development of a nomenclature standard for musical instruments in the Romanian context, facilitating automatic instrument recognition in local recordings and optimizing workflows in Romanian studios.

These research directions will contribute to advancing the field of AI-assisted music production, facilitating the democratization of high-quality audio production expertise, and opening new creative possibilities for artists and producers.

References

[1] Alexander, R. (2013). *The inventor of stereo: The life and works of Alan Dower Blumlein*. CRC Press.

[2] Alvord, L. S. and Farmer, B. L. (1997). Anatomy and orientation of the human external ear. *Journal of the American Academy of Audiology*, 8(6).

[3] Askwonder (2017). Innovators choose Wonder. https://askwonder.com/research/total-addressable-market-music-makers-instrumentalists-rappers-producers-ebtqmsw4k. [Accessed 13-04-2025].

[4] Bai, M. R. and Shih, G.-Y. (2007). Upmixing and downmixing two-channel stereo audio for consumer electronics. *IEEE Transactions on Consumer Electronics*, 53(3):1011–1019.

[5] Biem (2025). BIEM. https://www.biem.org/members/directorySociety.do?method=detail&societyId=52&rMethod=membersDirectoryList&by=society&domain=&alpha=. [Accessed 13-04-2025].

[6] Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, Cambridge.

[7] Buckle, B. (2023). 120,000 new tracks released on streaming services every day, report finds — mixmag.net. https://mixmag.net/read/120-000-new-tracks-are-released-on-streaming-services-every-day-report-finds-tech. [Accessed 13-04-2025].

[8] Chun, C. J., Kim, Y. G., Yang, J. Y., and Kim, H. K. (2009). Real-time conversion of stereo audio to 5.1 channel audio for providing realistic sounds. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2(4):85–94.

[9] Colonel, J., Javed, S., and Valero-Fernandez, L. (2021). A general approach to transfer learning for audio classification tasks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

[10] Cousins, M. and Hepworth-Sawyer, R. (2013). *Practical mastering: A guide to mastering in the modern studio*. Routledge.

[11] De Man, B., Mora-Mcginity, M., Fazekas, G., and Reiss, J. D. (2014a). The open multitrack testbed. In *Audio Engineering Society Convention 137*. Audio Engineering Society.

[12] De Man, B., Mora-Mcginity, M., Fazekas, G., and Reiss, J. D. (2014b). The open multitrack testbed: A new approach to the study of multitrack mixing. In *Audio Engineering Society Convention 137*, Los Angeles, USA. Audio Engineering Society.

[13] Défossez, A. (2021). Hybrid spectrogram and waveform source separation. *arXiv preprint arXiv:2111.03600*.

[14] Deruty, E. and Tardieu, D. (2011). About dynamic processing in mainstream music. *Journal of the Audio Engineering Society*, 59(5):300–311.

[15] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. (2017). Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR.

[16] European Broadcasting Union (2014). Loudness normalisation and permitted maximum level of audio signals. Recommendation R128, EBU, Geneva, Switzerland. Retrieved from https://tech.ebu.ch/publications/r128.

[17] Giannoulis, D., Massberg, M., and Reiss, J. D. (2012). Digital dynamic range compressor design—a tutorial and analysis. *Journal of the Audio Engineering Society*, 60(6):399–408.

[18] Giannoulis, D., Massberg, M., and Reiss, J. D. (2013). Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 61(10):716–726.

[19] Gibson, D. (2019). *The Art of Mixing: A Visual Guide to Recording, Engineering, and Production*. Routledge, New York, 3 edition.

[20] Greenwood, D. D. (1991). Critical bandwidth and consonance in relation to cochlear frequency-position coordinates. *Hearing research*, 54(2):164–208.

[21] Han, Y. and Lee, K. (2016). Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation. *IEEE Signal Processing Letters*, 23(12):1649–1653.

[22] Hawley, S. H. and Scheirer, W. J. (2020). Speech enhancement using deep learning and its application to the detection, classification and segmentation of sound events. *Frontiers in Computer Science*, 2:5.

[23] Howard, D. and Angus, J. (2013). *Acoustics and psychoacoustics*. Routledge.

[24] Huber, D. M., Caballero, E., and Runstein, R. (2023). *Modern Recording Techniques: A Practical Guide to Modern Music Production*. CRC Press.

[25] IBISWorld, I. (2024). Sound Recording & Music Publishing in Romania - Market Research Report (2014-2029). https://www.ibisworld.com/romania/industry/sound-recording-music-publishing/200261/#IndustryStatisticsAndTrends. [Accessed 13-04-2025].

[26] International Telecommunication Union (2015). Algorithms to measure audio programme loudness and true-peak audio level. Recommendation BS.1770-4, International Electrotechnical Commission, Geneva, Switzerland. Retrieved from https://www.itu.int/rec/R-REC-BS.1770.

[27] Irwan, R. and Aarts, R. M. (2002). Two-to-five channel sound processing. *Journal of the Audio Engineering Society*, 50(11):914–926.

[28] Izhaki, R. (2017). *Mixing Audio: Concepts, Practices, and Tools*. Focal Press, London, 3 edition.

[29] Katoch, S., Chauhan, S. S., and Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 80(5):8091–8126.

[30] Katz, B. and Katz, R. A. (2003). *Mastering audio: the art and the science*. Butterworth-Heinemann.

[31] Linkwitz, S. H. (1978). Passive crossover networks for noncoincident drivers. *Journal of the Audio Engineering Society*, 26(3):149–150.

[32] Mirjalili, S. (2019). Genetic algorithm: Theory, literature review, and application in image reconstruction. *Nature-Inspired Optimizers*, pages 69–85.

[33] Miskiewicz, A. (1992). Timbre solfege: A course in technical listening for sound engineers. *Journal of the Audio Engineering Society*, 40(7/8):621–625.

[34] Mitsufuji, Y., Fabbro, G., Uhlich, S., Stöter, F.-R., Défossez, A., Kim, M., Choi, W., Yu, C.-Y., and Cheuk, K.-W. (2022). Music demixing challenge 2021. *Frontiers in Signal Processing*, 1:18.

[35] Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995). Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society*, 43(5):300–321.

[36] Moroșanu, B., Negru, M., Nicolae, G., Ioniță, H., and Paleologu, C. (2025). Automated personalized loudness control for multi-track recordings. *Information*, 16(5).

[37] Moroșanu, B., Negru, M., and Paleologu, C. (2024). Automated personalized loudness control for multi-track recordings. *Algorithms*, 17(6):228.

[38] Motoc, G. (2024). Sizing up the romanian music industry. https://georgemotoc.com/2024/01/01/sizing-up-the-romanian-music-industry/. [Accessed 13-04-2025].

[39] MusicBrainz (2023). MusicBrainz - the open music encyclopedia. https://musicbrainz.org/. [Accessed 13-04-2025].

[40] Negru, M., Moroşanu, B., Neacşu, A., Drăghicescu, D., and Negrescu, C. (2023). Automatic audio upmixing based on source separation and ambient extraction algorithms. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 12–17. IEEE.

[41] Niall (2023). How Many Official Musicians Are There in the World? https://bigtimemusicians.com/how-many-official-musicians-are-there-in-the-world/. [Accessed 13-04-2025].

[42] Owsinski, B. (2006). *The mixing engineer's handbook*. Boston: Thomson Course Technology.

[43] Popa, V., Moroșanu, B., and Negrescu, C. (2023). Head related transfer function measurement in reverberant rooms. In *Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies XI*, volume 12493, pages 617–622. SPIE.

[44] Pulkki, V. (2006). Directional audio coding in spatial sound reproduction and stereo upmixing. In *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology–Surround and Beyond*. Audio Engineering Society.

[45] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

[46] Robles, L. and Ruggero, M. A. (2001). Mechanics of the mammalian cochlea. *Physiological reviews*, 81(3):1305–1352.

[47] Rumsey, F. (2013). Mastering for today's media. *Journal of the Audio Engineering Society*, 61(1/2):79–83.

[48] Senior, M. (2018). Mixing secrets for the small studio. *Taylor & Francis*.

[49] Series, B. (2010). Multichannel stereophonic sound system with and without accompanying picture. *International Telecommunication Union Radiocommunication Assembly*.

[50] Stassen, M. (2023). There are now 120,000 new tracks hitting music streaming services each day. https://www.musicbusinessworldwide.com/there-are-now-120000-new-tracks-hitting-music-streaming-services-each-day/. [Accessed 13-04-2025].

[51] Stassen, M. (2024). 'Romania is an emerging music market that's growing at a fast pace. It's catching up with more developed markets.' - Music Business Worldwide — musicbusinessworldwide.com. http://disq.us/t/4omwokx. [Accessed 13-04-2025].

[52] Storer, J. (2004). Home - JUCE. https://juce.com/. [Accessed 01-04-2025].

[53] Théberge, P. (2012). The end of the world as we know it: The changing role of the studio in the age of the internet. *The art of record production: An introductory reader for a new academic field*, pages 77–90.

[54] Vickers, E. (2010). The loudness war: Background, speculation, and recommendations. In *Audio Engineering Society Convention 129*. Audio Engineering Society.

[55] Wang, S., Germain, F. G., Simionato, R., and Bello, J. P. (2023). Neural parametric equalizer matching using differentiable biquads. *Applied Sciences*, 13(2):800.

[56] Zwicker, E. and Fastl, H. (2013). *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media.