Koszewski *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2023) 2023:1
https://doi.org/10.1186/s13636-022-00266-3

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

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

Damian Koszewski[1], Thomas Görne[2], Grazina Korvel[3] and Bozena Kostek[4*]

Abstract

The purpose of this paper is to show a music mixing system that is capable of automatically mixing separate raw recordings with good quality regardless of the music genre. This work recalls selected methods for automatic audio mixing first. Then, a novel deep model based on one-dimensional Wave-U-Net autoencoders is proposed for automatic music mixing. The model is trained on a custom-prepared database. Mixes created using the proposed system are compared with amateur, state-of-the-art software, and professional mixes prepared by audio engineers. The results obtained prove that mixes created automatically by Wave-U-Net can objectively be evaluated as highly as mixes prepared professionally. This is also confirmed by the statistical analysis of the results of the conducted listening tests. Moreover, 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-based and professional mixes than the amateur ones or the mix prepared using state-of-the-art software. These results are also confirmed by the outcome of the similarity matrix-based analysis.

Keywords: Automatic music mixing, Wave-U-Net autoencoder, Music signal parameterization, Listening tests, Similarity matrix

1 Introduction

The aim of this paper is to build a music mixing system that is capable of automatically mixing separate raw recordings with good quality regardless of the music genre. The realm of modern music and sound production is complex and diverse. To achieve a single end product—music that reaches the world—takes effort, immense commitment, and the combined creative talents of all kinds of experts. The music world comprises 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, graphics, and sales. Collaboration within the music production process enables to transform creativity into a product that can be enjoyable for the end-user [1–3]. The underlying drive of the teams of people throughout the recorded sound practices concerns the cultural tastes, the art of music, and the ever-present changes and challenges in production technology and industry [4].

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. 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. 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 produced, the result is a recorded song where each instrument is given a separate mono track and, in some cases, multichannel. After

*Correspondence: bokostek@audioakustyka.org

[4] Audio Acoustics Laboratory, Telecommunications and Informatics, Faculty of Electronics, Gdansk University of Technology, 80-233 Gdansk, Poland
Full list of author information is available at the end of the article

an artist decides to record a musical piece or song, the sound engineer uses appropriate microphones, records the multitrack material, and edits it. The mixer's role is to set proper proportions between the signal elements and adapt their time and frequency-based properties [5]. A more than adequate mix can emphasize the artistic character of a song or even define the music genre [6–8]. 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.

Regardless of the approach chosen, mixing is used to shape the character, tone, and intention of the production in relation to [4, 9]:

- Relative level between tracks (how loud are tracks relative to the others [2, 10]).
- 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—the ratio of the softest to the loudest peak, expressed in decibels [11])—of a single track, a group, or an output bus to optimize the levels or allow it to not stand out within a mix for the duration of the entire song).
- 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 [12–14]). It should be noted that audio effects, sound effects, and sound transformation, as these terms are used interchangeably [12], are understood as signal processing functions that change, modify, or augment an audio signal [10, 11].

Nowadays, fully analog studios are very rare. Analog equipment is expensive and requires special care and effort to upkeep. 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 their approach to mixing [15]. The second mixing approach is called hybrid mixing, where songs from the DAW are routed to an analog mixing console or single tracks are channeled down to outboard hardware, e.g., compressor, equalizer, or reverb. This approach is precisely in-between in the cost/effect category [16]. The least costly and the easiest method of mixing a song is the fully digital approach, called in-the-box [17]. Many renowned engineers changed their approach to mixing from analog to digital entirely [18, 19]. The in-the-box way of mixing has various advantages, i.e., there is a possibility of going back to a project with one click of the computer mouse, and free software programs that emulate the analog equipment are more and more faithfully to the original hardware.

Amateur and professional mixing may differ in talent, skills, experience, music background, artistry, and knowledge. So, the motivation behind our study is to see where an automatic mix may be positioned relative to them, i.e., whether it is closer to the amateur or between these two approaches. Moreover, we should stress that our intention is not to build an automatic mixing system for music mixing per se; this should stay with a professional sound engineer. In contrast, such an approach may help in gaming or branding areas where the focus is not on music quality but on effective ways of mixing audio [20]. Also, we decided to test "automatic mixing" versus human-made mixes. Moreover, when referring to "automatic mixing," we differentiate between the use of Wave-U-Net network and mixes prepared with one of the popular plugins. On the "human" side, we decided to test "amateur" and "professional mixes."

Therefore, in the paper, two hypotheses are posed. The first one considers whether 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 the human-based mix. The second one is related to subjective evaluation and tries to answer the question of whether the prepared mixes may subjectively be assessed as better ones than recordings created by an amateur engineer or mixes produced using state-of-the-art technology and can be comparable to mixes produced by a professional mixer.

The paper is structured as follows. First, literature background is shortly reviewed. This is followed by methodology, focusing on data preparation, deep model training and validation, and preparation of audio mixes. The consecutive section is devoted to the quality evaluation of audio mixes employing objective and subjective approaches, qualitative analysis as well as self-similarity matrices-based analysis. Besides, this section contains statistical analyses and discussion. Finally, a summary is given, along with the proposed direction of further research and development of automated mixing.

2 Literature background

When searching the term "automatic mixing," Google delivers 249,000,000 documents/links in 0.35 s, so the relevance of this area is easily seen. It should be noted that automatic mixing is a part of intelligent music production (IMP) [21], as the latter encompasses the application of artificial intelligence to mixing and mastering. De Man and his colleagues regard automating music production as introducing intelligence in audio effects,

sound engineering, and music production interfaces [10]. Classification of IMP research may differ [17] as it may refer to the audio effect that was automated [22, 23]. Some other aspects researched concern live downmixing stereo [24], selective minimization of masking [25], automatic mixing method for live music, incorporating multiple sources, loudspeakers, and room effects [26], multi-track mixing [27, 28].

Overall, IMP deals with data collection, perceptual evaluation, systems, processes, and interfaces [10]. According to Reiss [28], de Man et al. [10], and a review paper published by Moffat and Sandler [21], IMP is still emerging and under development, even if it is not a new field as it dates back to the 1975 Dugan's paper on a fully deterministic adaptive gain mixing system [29, 30]. It is also important to note that several on IMP may be observed throughout the years as they depend on machine-learning methods and music resources (e.g., [31, 32]), starting with baseline algorithms, knowledge-based approaches [21, 30, 33, 34], and ending with deep models [35–39].

Undoubtedly, the references included do not exhaust literature sources related to automatic music mixing; thus, the list of pioneers in automatic mixing provided by de Man and collaborators [17] should be referred to. Also, Moffat and Sandler [21] and de Man et al. [10] gave insight into intelligent music production and its history, in general.

Up-to-date, automatic mixing mainly regards more manageable tasks, such as setting the maximum level of the microphone in a live situation in a way that does not allow for the system's feedback or distortion of the speakers. The more manageable tasks that can be performed are also automatic mixing of audio elements in cases where artistic quality is not the most crucial aspect, e.g., in video games [40] or audio/music branding (for instance, in stores) where the songs are automatically mixed together one after the other [20, 41–46]. In the latter case, the mixing happens not in the context of multiple tracks in one piece but in the entire music project, where the previous song is smoothly mixed with the following. Examples of such work are described in several papers [47–50].

At the same time, the productization of technology and user-friendly interfaces influence the growth of technology and allow for more advanced automatic sound manipulation. Martinez-Ramirez and his co-authors [36] provided a very useful definition of audio effect units that refer to analog or digital signal processing systems that transform specific sound source characteristics. These transformations can be linear or nonlinear, time-invariant or time-varying, and with short-term and long-term memory. Most typical audio effect transformations are based on dynamics, such as compression; tone, such as distortion; frequency, such as equalization; and time, such as artificial reverberation or modulation-based audio effects [36].

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 dozen of analog devices. Moreover, modern plugins actively help the user to execute tasks that would be unachievable otherwise, e.g., treating one signal with 28 different filters. Some plugins offer genre or instrument detection and either an entire automatic mixing routine or a part of it (i.e., balance, equalization, or compression-only) [51].

In contrast, knowledge-based audio mixing can be described as a departure from the standard automation methods [21, 30, 33, 34, 52]. Still, many of these methods, except for specific ones, e.g., involving certain data augmentation procedures [37, 38], use large databases to train machine-learning algorithms and models. In the process, multiple parameters, e.g., level, panning, and equalization, are changed at the same time. The most commonly used databases are Open Multitrack Testbed [53] and MUSDB-18 [54]. The methods found in the literature use expert-based knowledge during training or creating a specific model or application. Examples of such work are described in several projects [55–59].

As de Man stated in his work [17], mixing is a multidimensional process. 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. With this said, the various types of processing cannot be done separately; instead, this challenge should be set as an all-in-one task. Isolating one problem will lead to another unresolved issue. As shown by state-of-the-art research in music production, there are a lot of tasks in mixing mastering, and beyond that are approached based on machine learning [49, 60–63]. Also, deep learning has gained much acceptance in recent years [63, 64].

3 Methodology

3.1 Data preparation

To properly train a neural network, 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. Based on MUSDB18-HQ [54], the most rewarding database available, a new dataset was built by the authors, supplemented with individual tracks from

the Cambridge database [65], and expanded by additional songs recorded by one of the authors. The dataset had to be prepared in a particular way to be helpful in model training and validation. The stems contained in MUSDB-18-HQ are wet, and the mixture is the summation of the stems. However, since the song-mixing process is more about altering individual tracks rather than stems, it was decided that this database would be sufficient for this study's purpose. Moreover, in the Cambridge database, one can find individual tracks and finished mixes. The instrument-to-stem models were trained on a combination of Cambridge and MUSDB18-HQ as well as the stem-to-mix model.

The MUSDB18-HQ database [54] and five songs recorded by the authors were used to train the deep model. This database consists of 150 songs (approximately 10 h in total) belonging to various genres. One hundred songs were used for training and 50 as a validation set. Drum, bass, vocals, and other instrument stems and finished mixes (summation of the stems) can be found in the database. The database consists of songs from the Cambridge database [65], which means that to acquire individual tracks, they had to be taken from the Cambridge database to be appropriately matched. Since the song-mixing process is more about altering individual tracks rather than stems, it was decided that this database would be sufficient for this paper's purpose.

As already mentioned, five songs recorded and mixed by the authors 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 music genre can be classified as rock.

Due to the nature of the system's architecture, based on Wave-U-Net autoencoders, 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 1. In cases where the number of signals was larger 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 equalization, 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. There was not any other processing done to the individual signals (tracks).

3.2 Deep model training and validation

The system consists of five deep models. The models were trained separately and connected to one system.

Table 1 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)

The models differ by the number of inputs and outputs (mono/stereo). The system was created from variants of Wave-U-Net networks, as suggested by other authors [64, 66]. So, the depth of the network models was the same as the one used by Martinez-Ramirez et al. [64]. The change introduced to the original Wave-U-Net enables the network to work on stereo signals. A single model was trained on a network with the following parameters:

- U-net layers: 10
- Filter size of convolution in downsampling block (max number of inputs): 15
- Upsampling: linear
- Type of output layer: linear without activation
- Learning rate: 1e−4
- Augmentation: false
- Batch size: 16
- Number of update steps per epoch: 200
- Optimizer: Adam

Each model utilizes raw (unprocessed) audio input and output with a connection to a series of downsampling and upsampling blocks that contain 1D convolution layers and is used separately. 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. 1.

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 [36, 67]. 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 batch size was 16. A model with the lowest loss for the validation subset was selected. The validation loss function of the stem-to-mix model training is presented in Fig. 2.

The models were trained on a computer supported by a NVidia GeForce 1080 graphics card. Training an individual model took approximately 2 days.

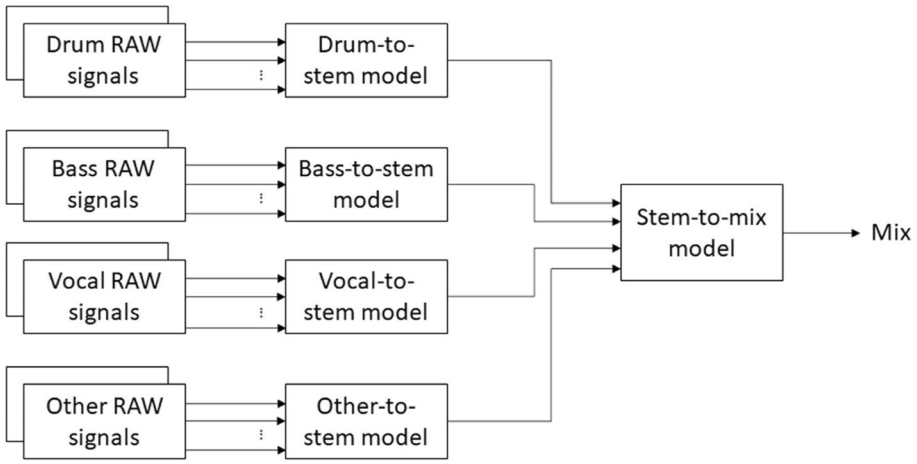

Fig. 1 Block diagram of an automatic audio mixing system

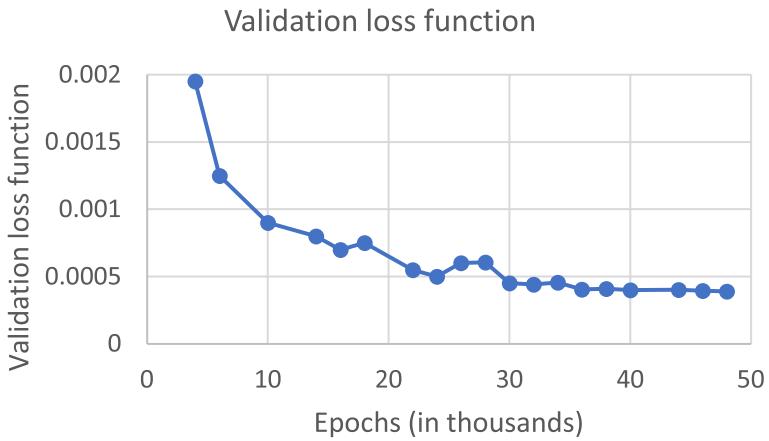

Fig. 2 The validation loss function of stem-to-mix model training

3.3 Preparation of audio mixes

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

- A professional mix (called "Pro")
- An amateur mix (called "Amateur")
- A mix created using state-of-the-art software (called "Izotope" [68])
- A mix created by the trained models of the Wave-U-Net network (called "Unet")

Clean tracks for eight songs in four music genres were chosen and acquired from the Cambridge database [65]. The list of the selected songs, including their genres and the number of tracks to be mixed, is presented in Table 2.

Due to the fact that the songs belong to different music genres and the models were trained on data from various genres, the evaluation and testing may show interesting results. For example, all 11 tracks from a selected song (i.e., "Secretariat – Over the top") are shown in the form of mel spectrograms in Fig. 3. All tracks in each song differ from each other in their spectral content. Moreover, all selected songs differ in the number of individual tracks, and even within the particular genre, they are dissimilar both sonically and emotionally. Also, the songs were specifically chosen to have different tempos and instrumentation.

The professional mixes "Pro" were made by well-known experienced audio engineers. Mixes of the following songs: "Angels in Amplifiers—I'm alright," "Georgia Wonder—Siren," "Side Effects Project—Sing

Table 2 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

with me," "Speak Softly—Broken man," "The Doppler Shift—Atrophy," and "Tom McKenzie—Directions" were created by Mike Senior who 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. He 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, who 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, who 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 "Amateur" mixes were prepared by a person with experience in music theory through education and 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 song, depending on the number of tracks in a given piece and its music genre. The quickest preparation of a mix took approximately 2 h, and the most prolonged—6 h. In general, the more familiar the genre was to the amateur mixer, the quicker the mixing process was. The lack of experience in mixing led to a relatively 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, the mixer did not exclude any raw tracks from the final mix.

To create state-of-the-art mixes, a set of Izotope plugins from the music production packet was used. The plugins included Neutron Pro and Nectar Pro. Their automatic balance and automatic mix features make it possible to mix a song semi-automatic. 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

So, the "Unet" mixes were created using the system presented above. Although in the final version, the system enables to mix a song without any user intervention, the mixes were created manually—each submodel was used separately. 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.

After obtaining all 32 mixes, the postprocessing of the acquired songs was performed. First, from each song, a 15-s clip was selected (duration of an excerpt according to [69]), 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 for the last step of mix preparation.

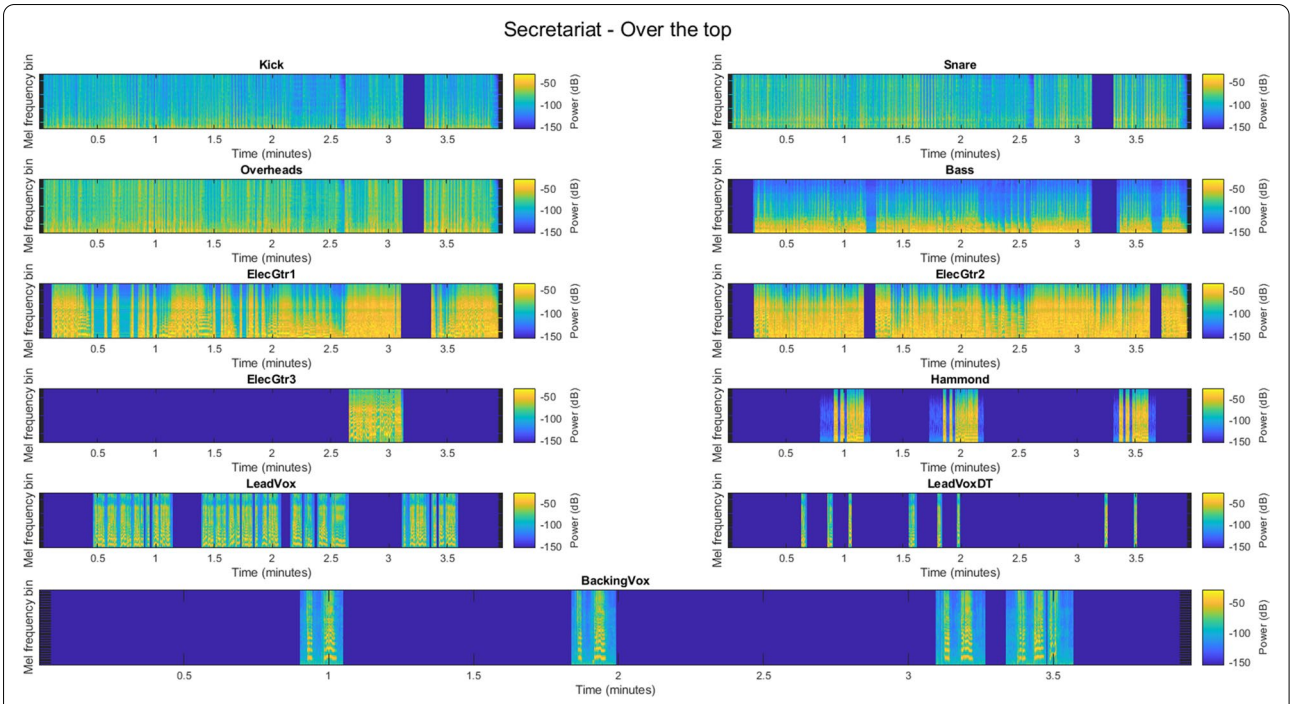

Fig. 3 All 11 tracks from "Secretariat—Over the top" song in the form of mel spectrogram representation

4 Quality evaluation of audio mixes

It should be noted that QoE is related to both subjective evaluation and objective metrics [70–74]. The users' experience, based on several factors, such as fulfilling their expectations, emotions, and preferences while interacting with technology, can be evaluated in subjective tests. In contrast, an objective investigation is both content- and context-related, so—in the absence of such a metric in the music mix quality area—several level-oriented parameters were proposed to be tested on the resulting mixes. Still, there is a need to find a dedicated measure that correlates with subjective evaluation results. That is why an approach based on the self-similarity matrix (SSM) analysis was proposed that may achieve such a goal. This is further examined in Section 4.3.

Sections 4.1 and 4.2 present the evaluation process, carried out in two ways, i.e., objectively and subjectively. First, several descriptor values related to perceptual characteristics for each mix are calculated [75]. The selected parameters are level-oriented as they are easy to calculate and understand. However, we do not compare these parameters between songs but rather between different mixing approaches. From an objective point of view, these parameters can be beneficial for determining the dynamic content of the song, even if it is distorted. This is very important when sending a prepared song to the mastering engineer. Samples that were subjected to objective analysis, i.e., waveform statistics based on RMS level, integrated loudness, loudness range, and true peak level, as well as low-level MPEG-7 descriptors (odd-to-even harmonic ratio, RMS-energy envelope, and harmonic energy) were not normalized.

In addition, a qualitative analysis took place. The test participants filled in a questionnaire form, answering several questions about their listening habits and experience. An example of the answers obtained is presented further on.

Moreover, the evaluation methodology and the results of a subjective test are shown as such evaluation has a higher priority over the objective assessment results [76–78]. It should be noted that listening tests were conducted on normalized samples, where the listeners rated each sample in multiple evaluation categories (balance, clarity, panning, space, and dynamics).

The statistical analysis is then performed, and the statistical significance of the achieved results is commented. This is followed by similarity matrix-based [79–81] analyses and the discussion.

4.1 Objective quality evaluation

Unprocessed samples were used for the objective evaluation. This is because subjecting the recordings to normalization may prevent the correct identification of accurate values for the acquired music signal samples. First, the waveform-based parameters were calculated, such as

RMS (root mean square) level (Fig. 4), integrated loudness (Fig. 5), loudness range (Fig. 6), and true peak level (Fig. 7) [82] for all music excerpts. These parameters were judged to be important in the evaluation process.

Further on, selected low-level descriptors MPEG-7 were calculated [83]. For this purpose, the timbre toolbox [84] in the MATLAB environment was used. Odd-to-even harmonic ratio, RMS-energy envelope, harmonic energy, and noisiness were calculated for each music sample. These descriptors were chosen because of their perceptual interpretation. In Fig. 8, a variation of the harmonic energy of the "Secretariat—Over the top" song—depending on the mix type—is shown.

For each mentioned descriptor, an analysis was performed to determine the statistical significance of differences between the mixes. For this purpose, the one-way ANOVA series [85] and the post hoc Tukey-Kramer test [86] were executed. The level of significance was assumed to be $\alpha = .05$. For most calculated descriptors, i.e., odd-to-even harmonic ratio, RMS-energy envelope, and harmonic energy, the differences between mixes are statistically significant (values are

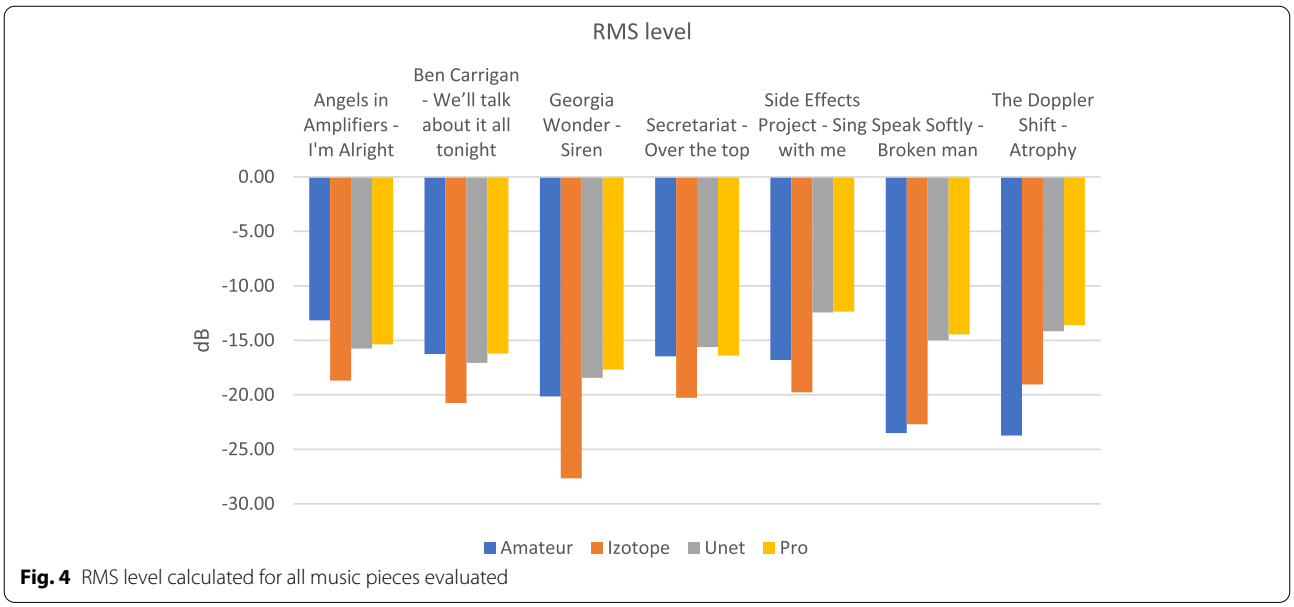

Fig. 4 RMS level calculated for all music pieces evaluated

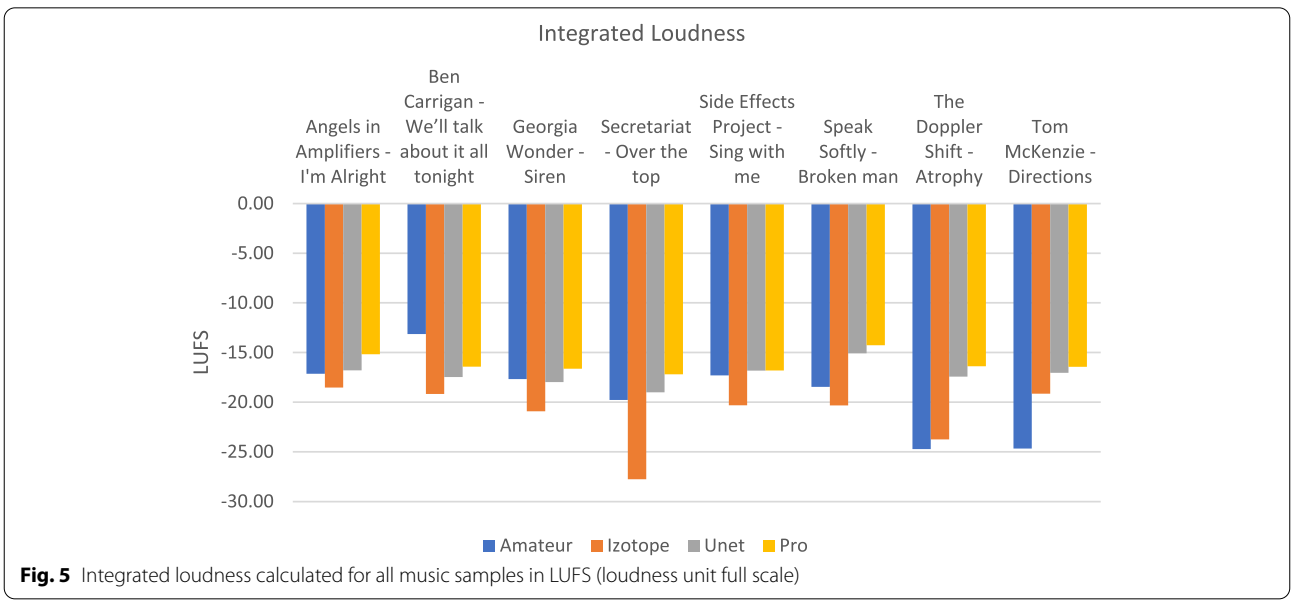

Fig. 5 Integrated loudness calculated for all music samples in LUFS (loudness unit full scale)

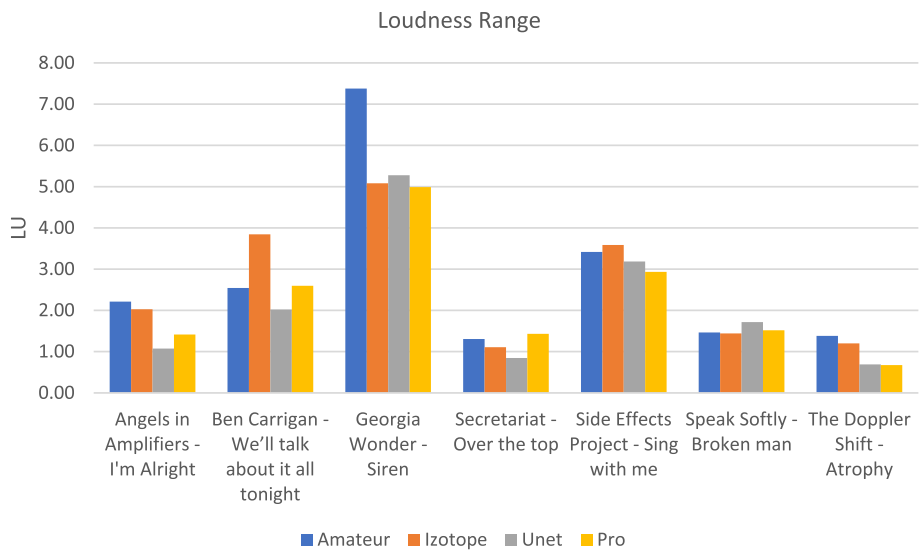

Fig. 6 Loudness range calculated for all music samples in LU (loudness units)

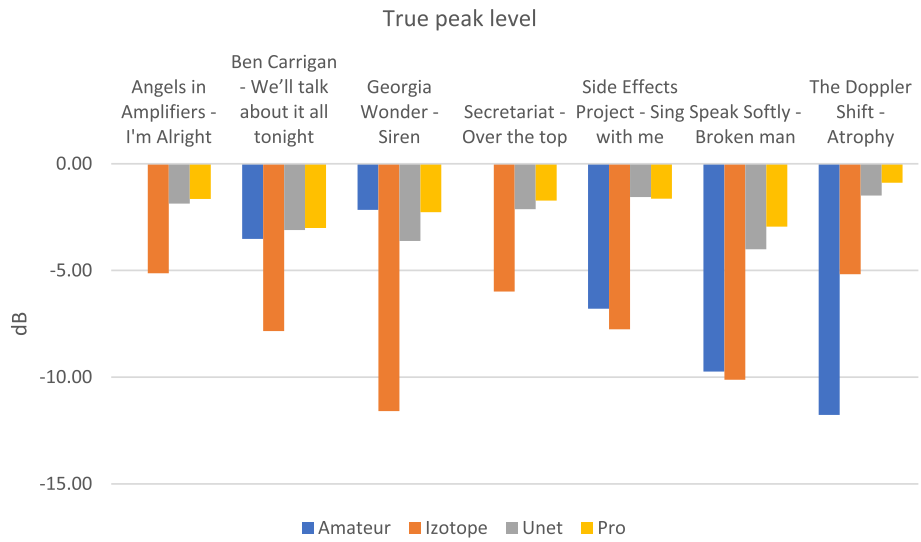

Fig. 7 True peak level calculated for all music samples

highlighted in bold font in Table 3) except for Unet-Pro pairs. In Table 3, the results of the statistical significance analysis for the harmonic energy descriptor for the "Secretariat—Over the top" song are presented.

Considering all the results obtained, it can be concluded that the "Unet" mixes are the closest to the "Pro" mixes, and the developed system 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 well-known state-of-the-art software.

4.2 Subjective quality evaluation

Before the listening test, the participants were asked to fill in a questionnaire form. There were questions concerning what they listen to, whether they are familiar with a particular music genre, and their music and mixing experience. Music genres that the participants listened to varied, but the most frequent responses were rock, alternative, hip-hop, and jazz. Listeners answered

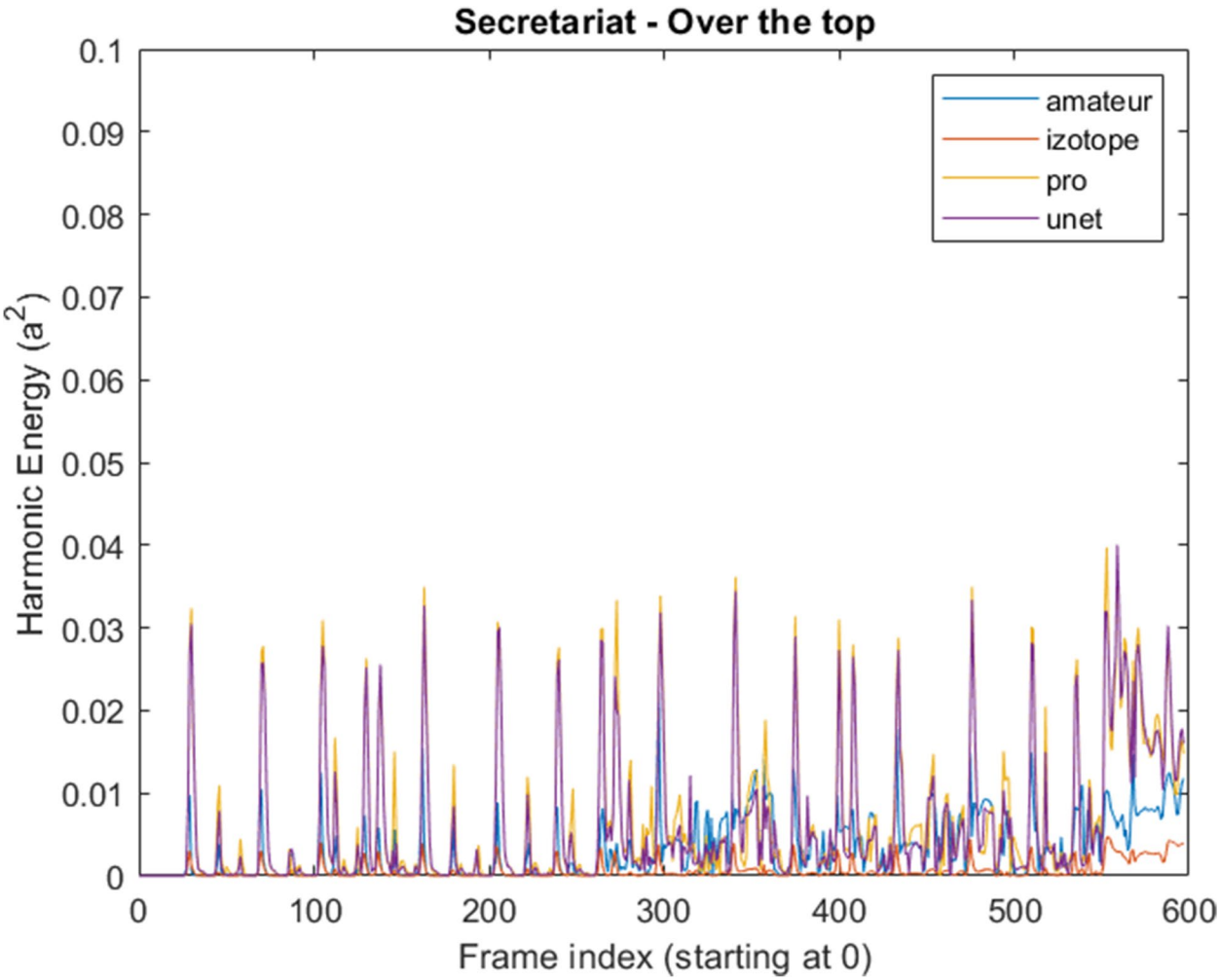

Fig. 8 Variation of the harmonic energy a^2 depending on the mix type in the "Secretariat—Over the top" song

that they were familiar with genres such as rock, pop, alternative, and electronica. Eighty-five percent of the listeners were musicians, and 60% were mixing engineers.

Table 3 Statistical significance analysis results of the harmonic energy descriptor for the "Secretariat—Over the top" song

Secretariat—Over the top				
Samples compared	***Lower confidence***	***Estimate***	***Upper confidence***	***p***
Amateur/Izotope	0.00	0.00	0.00	**.00**
Amateur/Unet	0.00	0.00	0.00	**.00**
Amateur/Pro	0.00	0.00	0.00	**.00**
Izotope/Unet	−0.01	−0.01	−0.01	**.00**
Izotope/Pro	−0.01	−0.01	−0.00	**.00**
Unet/Pro	0.00	0.00	0.00	.37

In Fig. 9, the listeners' years of experience in music mixing are presented.

After adequate postprocessing of samples, the listeners were asked to fill in a questionnaire and give their subjective rates for each acquired 32 samples (the test samples are available under the link provided the "Availability of data and materials" section). The rating of samples was conducted in line with the methodology of the rank-order procedure proposed by Zacharov and Huopaniemi in the round-robin subjective test devoted to evaluating virtual home theater systems [69]; however, using a five-point scale (1 = lowest–5 = highest). Such a subjective test can be considered MOS-like (mean opinion score). It was suitable for this particular listening test since it was easy to conduct and easy for the listeners to follow [69].

The aim of the tests was presented to potential participants before the tests took place. All persons taking part in subjective listening tests gave informed consent to

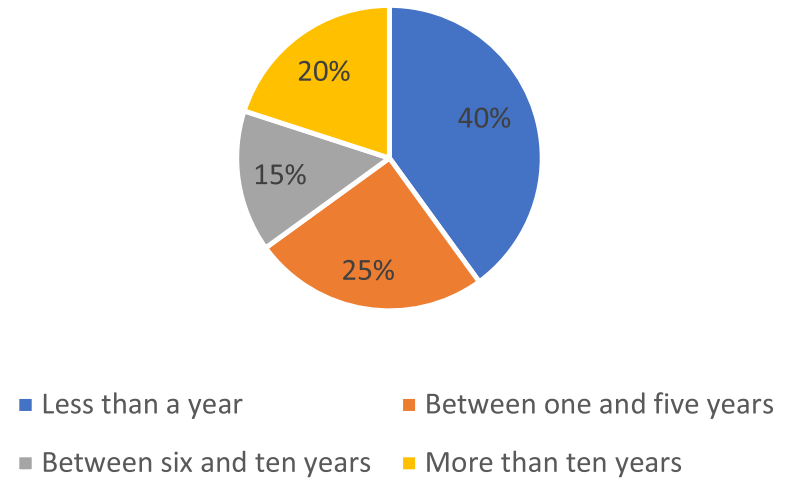

Fig. 9 Results of the survey on how many years of experience listeners have in music mixing

participate. All participants voluntarily decided whether or not to participate in the subjective tests.

The listeners performed the listening test in the R1 laboratory (mixing room) at the Hamburg University of Applied Sciences. The participants of the subjective tests were experts in the audio mixing area; moreover, they were provided with an explanation of the term "good quality" of the mix, understood as "as free of any distortions/artifacts, with properly controlled dynamics with good frequency balance" [87, 88].

The room at the university was 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—this approach was user-friendly since all participants were familiar with the console. Moreover, when listening to different mixes, the listener would not be introduced to any silence in-between and could easily detect all differences between samples.

The system calibration was set to 85 dB SPL and was performed with the use of the Bruel and Kjaer precision 732A m. 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 the identification of the most minute details possible, the selected level was appropriate. The loudness level is also recommended by the Audio Engineering Society [89].

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 authors included definitions next to each expression (e.g., balance).

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. All the participants confirmed that they listened 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. Eighty-five percent of the listeners were musicians, and 60% were mixing engineers.

Statistical analyses of the data resulting from subjective tests were performed using the IBM SPSS Statistics 25 software [90]. The software was used to calculate 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 using the Pearson correlation coefficient (r) [86]. The level of significance was assumed to be $\propto = .05$. Results whose significance was at the level of $.05 < p < .1$ were assumed to be statistically significant at the level of the statistical trend.

As part 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, a series of one-way analyses of variance for dependent samples was conducted, and individual mixes were compared in the following categories: overall rating, balance, clarity, panning, space, and dynamics. The outcome of the analysis is a probability called the p value. To identify homogeneous subsets of means that are not significantly different from each other, a pairwise comparison with Šidák correction was performed. The significance level was set at $p<0.05$. The different homogeneous subsets are denoted by different letter indexes (i.e., a, b, c).

First, an analysis of the overall rating of mixes was executed (see Table 4). The result is statistically significant (highlighted in bold font), and the effect size coefficient indicates large 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.

Next, the mixes were compared within the balance category. The result was statistically significant and the effect size (η^2) value signified large 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. An analogous analysis was conducted with the use of the clarity variable. The analysis results show very big 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.

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 significant differences in their results. 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.

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. All results are presented in Table 4.

Table 4 The overall rating of the mix as a function of the mix type (*M*, mean; *SD*, standard deviation; *p p* value; *F*, *F* ratio; η^2, a measure of the effect size) indicating groups forming separate homogeneous subsets (denoted by a, b, and c)

	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	**< .001**	.67
Balance	2.66*a*	0.54	2.58*a*	0.63	3.46*b*	0.79	4.08*c*	0.51	27.62	**< .001**	.59
Clarity	2.64*a*	0.54	2.76*a*	0.69	3.49*b*	0.54	4.04*c*	0.59	22.71	**< .001**	.54
Panning	2.88*a*	0.50	2.67*a*	0.60	3.71*b*	0.66	4.14*c*	0.63	27.24	**< .001**	.59
Space	2.64*a*	0.59	2.54*a*	0.60	3.58*b*	0.65	4.11*c*	0.65	33.40	**< .001**	.64
Dynamics	2.51*a*	0.56	2.56*a*	0.54	3.66*b*	0.62	4.14*c*	0.56	45.38	**< .001**	.70
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	**< .001**	.63
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	**< .001**	.67
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	**< .001**	.57
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	**< .001**	.51

The means that do not share the letter index (a, b, c) differ from each other at a p <.05 level—pairwise comparisons with the Šidák correction

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 using the Pearson correlation coefficient (r) was conducted (Table 5). 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.

4.3 Self-similarity matric-based analysis

After testing and analyzing the objective and subjective samples from each mix, self-similarity matrices (SSM) based on chromagrams were constructed. In the chromagram calculation process, the entire spectrum is projected onto 12 bins [91]. The method takes into account the fact that pitch consists of two components: tone height and chroma [92, 93]. The features represent the distribution of signal energy over chroma and time. The relationship between components can be defined by the following formula:

$$f = 2^{ch+h} \tag{1}$$

where ch is chroma ($ch \in [0,1]$), f is frequency, and h denotes the pitch height that indicates the octave the pitch is in.

The chroma vector sums the spectral energy into 12 bins corresponding to the 12 semitones within an octave.

The following three-step algorithm realizes the process of SSM construction:

STEP 1. The feature normalization
STEP 2. Self-similarity calculation
STEP 3. Visualization of the similarity scores

The feature normalization was performed by normalization of each column of the feature matrix. The normalized values are calculated using the following formula:

$$\hat{x}_n = \frac{x_n - \bar{x}_n}{SD} \tag{2}$$

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 nth matrix column ($n = 1, \ldots, N$). Each column of the normalized feature matrix $\hat{X}$ is compared with each other.

For the purpose of self-similarity calculation, the dot product between the feature matrix and its transpose is calculated as follows:

$$S = \hat{X}^T \hat{X} \tag{3}$$

The entries of the matrix imply the similarity scores. Each pixel in the matrix obtains a grayscale value corresponding to the given similarity score. The darkest color refers to the smallest similarity. An example of a comparison between objective and subjective analyses for "Secretariat—Over the top" is depicted in Fig. 10.

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 [94]; and visual information fidelity (VIF), treated as a full-reference image quality related to image information extracted by the human visual system [95, 96]. The results obtained are presented in Table 6.

As seen in Table 6, the "Unet" mixes are the closest to the "Pro" mixes (values highlighted in bold). Both the objective and subjective samples achieve similar results.

Table 5 Correlation between the experience in mixing and the overall ratings of mixes

		Experience in mixing
Amateur	Pearson's r	−.31
	Significance	.186
Izotope	Pearson's r	−.52
	Significance	.018
Unet	Pearson's r	.38
	Significance	.098
Pro	Pearson's r	.69
	Significance	.001

5 Summary

The main goal of this study 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 authors, was used for training purposes.

To check the validity of the hypotheses posed, multiple experiments were conducted. The first concerned the

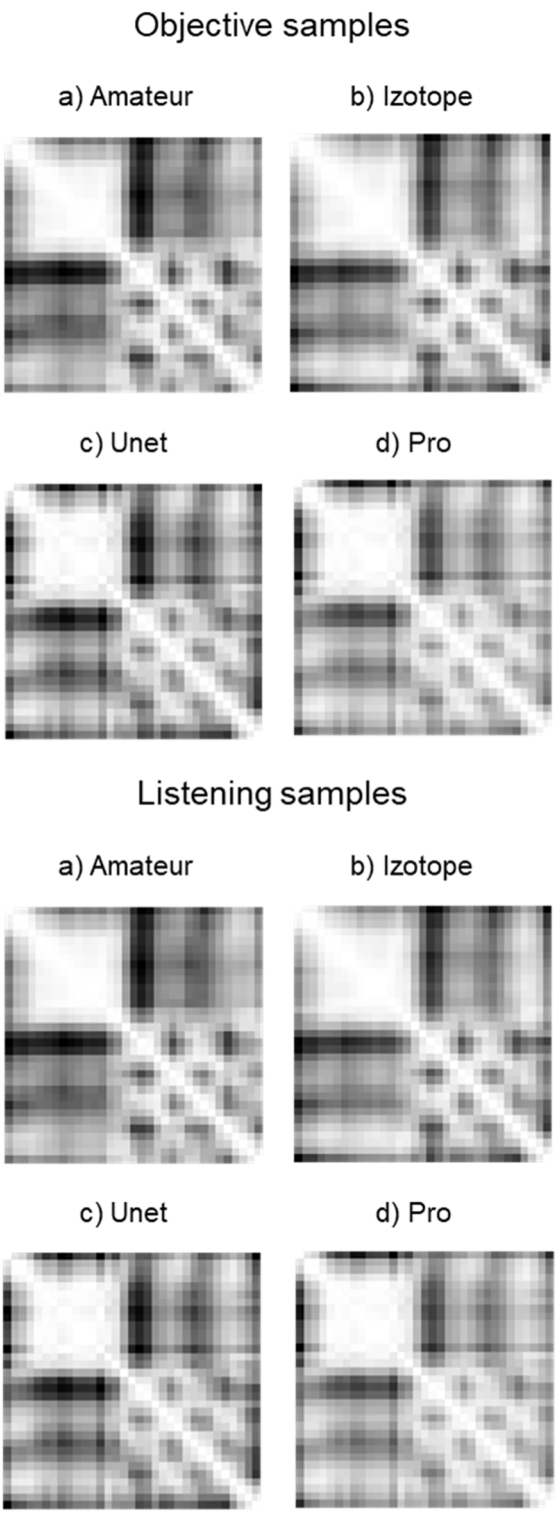

Fig. 10 Graphical representation of the SSM of "Secretariat—Over the top" objective and subjective samples

comparison of objective features of the obtained mixes. The developed system should automatically mix the input tracks so 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. It was shown that it is possible to automatically mix input tracks provided by the user, using previously trained models so that the final effect would be objectively very close to mixes prepared by a professional mixing engineer. However, this is only true for the respective audio descriptors.

All mixes created using Wave-U-Net were free of distortions or artifacts throughout the song. The overall quality can be evaluated as good or even very good (especially when compared with the Amateur mixes). The trained models behave similarly between different genres. The Authors did not find any major deviations in the final mixes when testing different songs.

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 the primary effort is on visual effects or custom music branding, where the focus is on combining songs that match the end and the beginning of tracks. These areas are open to such findings as automatizing the audio mixing process.

With regard to objective test scores, this study proposes to use a method based on self-similarity matrices, commonly used in the analysis of music signals, to assess the quality of audio mixes. The experimental results showed that the proposed method correlates closely with the subjective and objective evaluation results and can be employed as an objective measure for assessing sound quality.

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 [97–101] 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.

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

Table 6 Comparison of means for all samples

Song	Objective samples			Listening samples		
	Pro/Unet	Pro/Izotope	Pro/Amateur	Pro/Unet	Pro/Izotope	Pro/Amateur
Mean RMSE	**7.19**	18.67	16.29	**7.19**	18.68	16.29
Mean SSIM	**0.9782**	0.9213	0.9305	**0.9781**	0.9213	0.9306
Mean VIF	**0.84**	0.57	0.60	**0.84**	0.57	0.60

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.

Acknowledgements
The authors want to thank all subjective test participants.

Authors' contributions
DK conceived the methodology, took the responsibility of the experiment organization, and prepared the draft version of the paper; GK contributed to the data curation, signal analysis, and statistical validity of the results; TG contributed to the subjective test organization; BK reviewed the literature, discussed the interpretation of the results, and contributed to the final version of the manuscript. All authors reviewed the manuscript. The authors read and approved the final manuscript.

Funding
This study was partly supported by the InterPhD-2 project POWR.03.02.00-IP.08-00-DOK/16 and the Faculty of Electronics, Telecommunications and Informatics of Gdańsk University of Technology.

Availability of data and materials
There is a demo page, under this link below, where one can find all samples that were introduced to the listeners: https://drive.google.com/file/d/1dE3RDLoZar6kJD8Qej5z-JrpLCrXhpwV/view?usp=sharing.
The datasets analyzed during the current study are as follows: MUSDB18-HQ (https://doi.org/10.5281/zenodo.3338373), supplemented with individual tracks from the Cambridge database (https://www.cambridge-mt.com/ms/mtk/) and expanded by additional songs recorded by one of the authors. They are available on request as they are too large to be stored on GitHub.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors gave their consent to publish this paper.

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

Author details
[1]Telecommunications and Informatics, Multimedia Systems Department, Faculty of Electronics, Gdansk University of Technology, 80-233 Gdansk, Poland. [2]Department of Media Technology, Hamburg University of Applied Sciences, 22081 Hamburg, Germany. [3]Institute of Mathematics and Informatics, Vilnius University, LT-08663 Vilnius, Lithuania. [4]Audio Acoustics Laboratory, Telecommunications and Informatics, Faculty of Electronics, Gdansk University of Technology, 80-233 Gdansk, Poland.

Received: 25 August 2022 Accepted: 21 December 2022
Published online: 05 January 2023

References

1. S. Bennett, E. Bates, in *The Production of Music and Sound: A Multidisciplinary Critique*. Critical approaches to the production of music and sound (2018). https://doi.org/10.5040/9781501332074.0006
2. A. Case, *Mix Smart: Pro Audio Tips for your Multitrack Mix* (Focal Press, Waltham, 2011)
3. D. Chaney, The music industry in the digital age: Consumer participation in value creation. Int. J. Arts Manag. **15**(1), 42–52 (2012)
4. J. Tot, *Multitrack Mixing: An Investigation into Music Mixing Practices* (2018). https://doi.org/10.13140/RG.2.2.26537.49767
5. R. Toulson, Can we fix it? – The consequences of 'fixing it in the mix' with common equalisation techniques are scientifically evaluated. J. Art Rec. Prod. **3**, 1–14 (2008)
6. B. De Man, *Towards a Better Understanding of Mix Engineering* (PhD thesis, Queen Mary University of London, United Kingdom, 2017)
7. E. Deruty, in *2nd AES Workshop on Intelligent Music Production*. Goal-oriented mixing, vol 13 (2016)
8. H. Katayose, A. Yatsui, M. Goto, in *Int. Conf. On Automated Production of Cross Media Content for Multi-Channel Distribution*. A mix-down assistant interface with reuse of examples (2005)
9. B. De Man, J.D. Reiss, in *Innovation in Music II*, ed. by R. Hepworth-Sawyer, J. Hodgson, J. L. Paterson, R. Toulson. Crowd-sourced learning of music production practices through large-scale perceptual evaluation of mixes (Future Technology Press, United Kingdom, 2016)
10. B. De Man, R. Stables, J.D. Reiss, *Intelligent Music Production* (Focal Press, New York, 2019)
11. D. Huber, R. Runstein, *Modern Recording Techniques* (Taylor & Francis, New York, 2013)
12. V. Verfaille, M. Holters, U. Zölzer, in *DAFX—Digital Audio Effects*. Introduction (Wiley, Chichester, 2011)
13. T. Wilmering, G. Fazekas, M.B. Sandler, in *Proceedings of the AES 135th Convention, New York, NY, USA*. Audio effect classification based on auditory perceptual attributes (2013), pp. 17–20
14. T. Wilmering, D. Moffat, A. Milo, M.B. Sandler, A history of audio effects. Appl. Sci. **10**(3), 791 (2020). https://doi.org/10.3390/app10030791
15. G. Bromham, in *Mixing Music*. How can academic practice inform mixcraft? (Routledge, New York, 2017)
16. D. Reed, in *Proceedings of the 5th International Conference on Intelligent User Interfaces*. A perceptual assistant to do sound equalization (2000), pp. 212–218
17. B. De Man, J.D. Reiss, R. Stables, in *3rd AES Workshop on Intelligent Music Production, Salford, UK*. Ten years of automatic mixing (2017)
18. Audio Unity Group. https://www.audio-unity-group.com/andrew-scheps-on-mixing-100-in-the-box/. Accessed 30 June 2022
19. Pure Mix. https://www.puremix.net/video/andrew-scheps-mixing-ziggy-marley-in-the-box.html. Accessed 30 Nov 2022
20. D. Huron, Music in advertising: An analytic paradigm. Music. Q. **73**(4), 557–574 (1989). https://doi.org/10.1093/mq/73.4.557
21. D. Moffat, M.B. Sandler, Approaches in intelligent music production. Arts **8**(5), 14 (2019)

22. P. Pestana, *Automatic Mixing Systems Using Adaptive Digital Audio Effects* (Ph.D. dissertation, Universidade Católica Portuguesa, Porto, 2013)
23. P.D. Pestana, J.D. Reiss, in *53rd International Conference on Semantic Audio, London, UK*. Intelligent audio production strategies informed by best practices (Audio Engineering Society 53rd International Conference, London, 2014), pp. 1–9
24. P.E. Gonzalez, J.D. Reiss, in *10th International Conference on Digital Audio Effects (DAFx'07), Bordeaux, France*. Automatic mixing: Live downmixing stereo panner (2007)
25. P.E. Gonzalez, J.D. Reiss, in *11th International Conference on Digital Audio Effects (DAFx'08), Espoo, Finland*. Improved control for selective minimization of masking using interchannel dependancy effects (2008)
26. M. Terrell, M. Sandler, An offline, automatic mixing method for live music, incorporating multiple sources, loudspeakers, and room effects. Comput. Music. J. **36**, 37–54 (2012)
27. F. Pachet, O. Deleure, in *Audio Engineering Society Convention 109, Los Angeles*. On-the-fly multi-track mixing (2000)
28. J.D. Reiss, in *17th International Conference on Digital Signal Processing (DSP)*. Intelligent systems for mixing multichannel audio (IEEE, Corfu, 2011), pp. 1–6. https://doi.org/10.1109/ICDSP.2011.6004988
29. D. Dugan, Automatic microphone mixin. J. Audio Eng. Soc. **23**, 442–449 (1975)
30. D. Moffat, M.B. Sandler, 146th Convention, Dublin, Ireland, 2019 March 20 – 23, Automatic mixing level balancing enhanced through source interference identification (Audio Engineering Society 146th Convention, Dublin, 2019), pp 1–5
31. B. Kolasinski, in *Audio Engineering Society Convention 124, Amsterdam*. A framework for automatic mixing using timbral similarity measures and genetic optimization (2008)
32. P. Hoffmann, B. Kostek, Bass enhancement settings in portable devices based on music genre recognition. J. Audio Eng. Soc. **63**(12), 980–989 (2015). https://doi.org/10.17743/jaes.2015.0087
33. B. De Man, J.D. Reiss, in *Audio Engineering Society Convention 135*. A knowledge-engineered autonomous mixing system (Audio Engineering Society, New York, 2013), paper no. 8961
34. M.N.Y. Lefford, G. Bromham, G. Fazekas, D. Moffat, Context-aware intelligent mixing systems. J. Audio Eng. Soc. **69**(3), 128–141 (2021). https://doi.org/10.17743/jaes.2020.0043
35. M.A. Martinez-Ramírez, J.D. Reiss, in *3rd Workshop on Intelligent Music Production, Salford, UK, 15 September 2017*. Deep learning and intelligent audio mixing (Salford, 2017)
36. M.A. Martinez-Ramírez, E. Benetos, J.D. Reiss, Deep learning for black-box modeling of audio effects. Appl. Sci. **10**, 638 (2020). https://doi.org/10.3390/app10020638
37. Martinez-Ramírez M.A., Liao W.H., Fabbro G., Uhlich S., Nagashima C., Mitsufuji, Y., Automatic Music Mixing with Deep Learning and out-of-Domain Data. 2022, arXiv preprint arXiv:2208.11428
38. M.A. Martinez-Ramírez, E. Benetos, J.D. Reiss, in *23rd International Society for Music Information Retrieval Conference (ISMIR)*. Automatic music mixing with deep learning and out-of-domain data (2022). https://doi.org/10.3390/app10020638
39. C.J. Steinmetz, J. Pons, S. Pascual, J. Serrà, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Automatic multitrack mixing with a differentiable mixing console of neural audio effects (2021)
40. D. Margounakis, I. Lappa, in *Digital Tools for Computer Music Production and Distribution*. Music in video games (IGI Global, 2016), pp. 160–182. https://doi.org/10.4018/978-1-5225-0264-7.ch008
41. W. Brodsky, Developing a functional method to apply music in branding: Design language-generated music. Psychol. Music **39**(2), 261–283 (2011). https://doi.org/10.1177/0305735610387778
42. C. Hackley, in *Organising Music: Theory, Practice, Performance*. Branding and the music market (Cambridge University Press, Cambridge, 2015), pp. 127–134. https://doi.org/10.1017/CBO9781139644365.013
43. K.M. Knoferle, E.R. Spangenberg, A. Herrmann, J.R. Landwehr, It is all in the mix: The interactive effect of music tempo and mode on in-store sales. Mark. Lett. **23**(1), 325–337 (2012). https://doi.org/10.1007/s11002-011-9156-z
44. A.C. North, L.P. Sheridan, C.S. Areni, Music congruity effects on product memory, perception, and choice. J. Retail. **92**(1), 83–95 (2016)
45. E. Ovalı, in *European Proceedings of Social and Behavioural Sciences*. The effects of background music dimensions on customer attitude towards retail store. Strategic management in an international environment: The new challenges for international business and logistics in the age of industry 4.0, vol 71 (Future Academy, 2019), pp. 113–122. https://doi.org/10.15405/epsbs.2019.10.02.11
46. I. Vida, C. Obadia, M. Kunz, The effects of background music on consumer responses in a high-end supermarket. Int. Rev. Retail Distrib. Consum. Res. **17**(5), 469–482 (2007). https://doi.org/10.1080/09593960701631532
47. M.J. Terrell, A. Simpson, M. Sandler, The mathematics of mixing. J. Audio Eng. Soc. **62**(January/February), 4–13 (2014)
48. G. Wichern et al., Comparison of loudness features for automatic level adjustment in mixing (Audio Engineering Society 139th Convention, New York, 2015)
49. A. Wilson, B. Fazenda, in *2nd Workshop on Intelligent Music Production*. An evolutionary computation approach to intelligent music production, informed by experimentally gathered domain knowledge (2016)
50. S. Hafezi, J.D. Reiss, Autonomous multitrack equalization based on masking reduction. J. Audio Eng. Soc. **63**(5), 312–323 (2015). https://doi.org/10.17743/jaes.2015.0021
51. https://www.attackmagazine.com/reviews/the-best/the-best-ai-assist-plugins/. Accessed Nov 2022
52. G. Korvel, B. Kostek, in *Proceedings of Meetings on Acoustics 178ASA, San Diego, California2-6 December 2019*. Discovering rule-based learning systems for the purpose of music analysis, vol 39, No. 1 (Acoustical Society of America, San Diego, 2019), p. 035004. https://doi.org/10.1121/2.0001221
53. B. De Man, M. Mora, G. Fazekas, J.D. Reiss, in *Audio Eng. Soc. Convention e-Brief, Los Angeles, USA*. The open multitrack testbed (2014)
54. Z. Rafii, A. Liutkus, F.R. Stoter, S.I. Mimilakis, R. Bittner, *MUSDB18-HQ – An Uncompressed Version of MUSDB18* (2019). https://doi.org/10.5281/zenodo.3338373
55. A. Wilson, B.M. Fazenda, Populating the mix space: Parametric methods for generating multitrack audio mixtures. Appl. Sci. **7**, 1329 (2017). https://doi.org/10.3390/app7121329
56. F. Everardo, in *14th Sound and Music Computing Conference, July 5–8, Espoo, Finland*. Towards an automated multitrack mixing tool using answer set programming (2017)
57. D. Moffat, F. Thalmann, M. Sandler, in *4th Workshop on Intelligent Music Production, Huddersfield, UK*. Towards a semantic web representation and application of audio mixing rules (2018)
58. Ronan D., Ma Z., Mc Namara P., Gunes H., Reiss J.D., Automatic Minimisation of Masking in Multitrack Audio Using Subgroups. https://arxiv.org/abs/1803.09960. Accessed 23 Dec 2022
59. W.H. Lai, S.L. Wang, RPCA-DRNN technique for monaural singing voice separation. EURASIP J. Audio Speech Music Process. **1**, 1–21 (2022). https://doi.org/10.1186/s13636-022-00236-9
60. A.L. Benito, J.D. Reiss, Intelligent multitrack reverberation based on hinge-loss Markov random fields. Audio Eng. Soc. Int. Conf. (Semantic Audio) (AES Conference on Semantic Audio, Erlangen, 2017), pp. 1–8
61. E.T. Chourdakis, J.D. Reiss, A machine learning approach to application of intelligent artificial reverberation. J. Audio Eng. Soc. **65**(January/February) (2017). https://doi.org/10.17743/jaes.2016.0069
62. S.I. Mimilakis, E. Cano, J. Abfer, G. Schuller, in *2nd Workshop on Intelligent Music Production*. New sonorities for jazz recordings: Separation and mixing using deep neural networks (2016)
63. S.I. Mimilakis, K. Drossos, T. Virtanen, G. Schuller, in *140th Audio Eng. Soc. Conv*. Deep neural networks for dynamic range compression in mastering applications (2016)
64. M.A. Martinez-Ramirez, D. Stoller, D. Moffat, *A Deep Learning Approach to Intelligent Drum Mixing with the Wave-U-Net* (Audio Engineering Society, 2021)
65. Mixing Secrets Free. Multitrack Library, https://www.cambridge-mt.com/ms/mtk/. Accessed June 2022
66. D. Stoller, S. Ewert, S. Dixon, in *19th International Society for Music Information Retrieval Conference (ISMIR 2018), September 23–27, Paris, France*. Wave-U-net: A multi-scale neural network for end-to-end audio source separation (2018)
67. M.A. Martinez-Ramirez, J.D. Reiss, in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Modeling

nonlinear audio effects with end-to-end deep neural networks (2019), pp. 171–175. https://doi.org/10.1109/ICASSP.2019.8683529
68. Izotope software. https://www.izotope.com/en/products/. Accessed 30 Nov 2022
69. N. Zacharov, J. Huopaniemi, in *107th International Audio Eng. Soc. Convention*. Results of a round robin subjective evaluation of virtual home theatre sound systems (1998)
70. Bouraqia K., Sabir E., Sadik M., Ladid L., Quality of Experience for Streaming Services. 2019, https://arxiv.org/pdf/1912.11318.pdf
71. Brunnstrom K., Beker S.A., De Moor K., Dooms A., Egger S., Garcia M.N., Hossfeld T., Jumisko-Pyykko S., Keimel C., Larabi C., et al., Qualinet White Paper on Definitions of Quality of Experience. 2013
72. S. Kandadai, J. Hardin, C.D. Creusere, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Audio quality assessment using the mean structural similarity measure (2008), pp. 221–224. https://doi.org/10.1109/ICASSP.2008.4517586
73. K.U.R. Laghari, K. Connelly, Toward total quality of experience: A QoE model in a communication ecosystem. Commun. Mag. IEEE **50**(4), 58–65 (2012)
74. T. Zhao, Q. Liu, C.W. Chen, QoE in video transmission: A user experience-driven strategy. IEEE Commun. Surv. Tutor. **19**(1), 285–302, Firstquarter (2017). https://doi.org/10.1109/COMST.2016.2619982
75. B. De Man et al., in *15th International Society for Music Information Retrieval Conference, Taipei, Taiwan*. An analysis and evaluation of audio features for multitrack music mixtures (2014)
76. Barbedo J. G. A., Lopes A., A new cognitive model for objective assessment of audio quality. J. Audio Eng. Soc., 53, ½, 22–31, 2005
77. P. Malecki, *Evaluation of Objective and Subjective Factors of Highly Reverberant Acoustic Field* (PhD Thesis, AGH University of Science and Technology, Krakow, 2013)
78. M. Unehara, K. Yamada, T. Shimada, in *Soft Computing and Intelligent Systems (SCIS)*. Subjective evaluation of music with brain wave analysis for interactive music composition by IEC (2014), pp. 66–70
79. M. Müller, F. Kurth, in *ICASSP-88., 1988 International Conference on 5:V – V*. Enhancing similarity matrices for music audio analysis (2006). https://doi.org/10.1109/ICASSP.2006.1661199
80. D.F. Silva, C.M. Yeh, Y. Zhu, G.E.A.P.A. Batista, E. Keogh, Fast similarity matrix profile for music analysis and exploration. IEEE Trans. Multimedia **21**(1), 29–38 (2019). https://doi.org/10.1109/TMM.2018.2849563
81. Y. Shiu, H. Jeong, C.C.J. Kuo, in *AMCMM'06*. Similarity matrix processing for music structure analysis (Santa Barbara, 2006). https://doi.org/10.1145/1178723.1178734
82. F. Rumsey, The importance of loudness. J. Audio Eng. Soc. **69**(3), 211–213, Page 11 (2021)
83. R. Koenen, F. Pereira, MPEG-7: A standardized description of audiovisual content. Signal Process. Image Commun. **16**(1–2), 5–13 (2000)
84. Timbre toolbox. https://github.com/mondaugen/timbretoolbox. Accessed June 2022
85. A. Ross, V.L. Willson, *One-Way ANOVA. Basic and Advanced Statistical Tests* (SensePublishers, Rotterdam, 2017), pp. 21–24. https://doi.org/10.1007/978-94-6351-086-8_5
86. H.Y. Kim, Statistical notes for clinical researchers: Post-hoc multiple comparisons. Restor. Dent. Endod. **40**(2), 172–176 (2015)
87. What is mixing…. https://mrmixandmaster.com/what-is-music-mixing-why-it-is-important/. Accessed Nov 2022
88. Characteristics of a great mix. https://gearspace.com/board/so-much-gear-so-little-time/1251192-characteristics-great-mix.html. Accessed Nov 2022
89. Recommendations for loudness of internet audio streaming and on-demand distribution. Technical Document AESTD1008.1.21–9 (AES Technical Committee on Broadcastingand Online Delivery, 2021), https://www.aes.org/technical/documentDownloads.cfm?docID=731. Accessed 30 Nov 2022
90. J.O. Aldrich, *Using IBM SPSS Statistics: An Interactive Hands-on Approach* (Sage Publications Inc., Thousand Oaks, 2018)
91. P. Gimeno, I. Viñals, A. Ortega, A. Miguel, E. Lleida, Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. J. Audio Eng. Soc. **5** (2020). https://doi.org/10.1186/s13636-020-00172-6
92. A. Bachem, Tone height and tone chroma as two different pitch qualities. Acta Psychol. **7**, 80–88 (1950). https://doi.org/10.1016/0001-6918(50)90004-7
93. R.N. Shepard, Circularity in judgments of relative pitch. J. Acoust. Soc. Am. **36**(12), 2346–2353 (1964)
94. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
95. T.Y. Kuo, P.C. Su, C.M. Tsai, Improved visual information fidelity based on sensitivity characteristics of digital images. J. Vis. Commun. Image Represent. **40**, 76–84 (2016). https://doi.org/10.1016/j.jvcir.2016.06.010
96. H.R. Sheikh, A.C. Bovik, Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
97. M. Blaszke, D. Koszewski, in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) Proceedings*. Determination of low-level audio descriptors of a musical instrument sound using neural network (2020). https://doi.org/10.23919/SPA50552.2020.9241264
98. P. Herrera, G. Peeters, S. Dubnov, Automatic classification of musical instrument sounds. J. New Music Res. **32**(1) (2010). https://doi.org/10.1076/jnmr.32.1.3.16798
99. D. Koszewski, B. Kostek, Musical instrument tagging using data augmentation and effective noisy data processing. J. Audio Eng. Soc. **68**(1/2), 57–65 (2020). https://doi.org/10.17743/jaes.2019.0050
100. J. Liu, L. Xie, in *Intelligent Computation Technology and Automation (ICICTA)*. SVM-based automatic classification of musical instruments, vol 3 (2010)
101. A. Rosner, B. Kostek, Automatic music genre classification based on musical instrument track separation. J. Intell. Inf. Syst. **50**(2), 363–384 (2018). https://doi.org/10.1007/s10844-017-0464-5

Publisher's Note