MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

# Audio Signal Processing in the Artificial Intelligence Era: Challenges and Directions

Steinmetz, Christian; Uhle, Christian; Everardo, Flavio; Mitcheltree, Christopher; McElveen, J. Keith; Jot, Jean-Marc; Wichern, Gordon

TR2025-116 August 02, 2025

**Abstract**

Artificial intelligence ( AI ) has seen significant advancement in recent years, leading to increasing interest in integrating these techniques to solve both existing and emerging problems in audio engineering. In this paper, we investigate current trends in the application of AI for audio engineering, outlining open problems and applications in the research field. We begin by providing an overview of AI-based algorithm development in the context of audio, discussing problem selection and taxonomy. We then explore human-centric AI challenges and how they relate to audio engineering, including ethics, trustworthiness, explainability, and interaction, emphasizing the need for ethically sound and human-centered AI systems. Subsequently, we examine technical challenges that arise when applying modern AI techniques to audio, including robust generalization, audio quality, high sample rates, and real-time processing with low latency. Finally, we outline applications of AI in audio engineering, covering the development of machine learning-powered audio effects, synthesizers, and automated mixing systems, as well as spatial audio, speech enhancement, dialog separation and music generation. We emphasize the need for a balanced approach that integrates human-centric concerns with technological advancements, advocating for responsible and effective application of AI.

*Journal of the Audio Engineering Society 2025*



Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

# Audio Signal Processing in the Artificial Intelligence Era: Challenges and Directions

**Christian J. Steinmetz**[1] **Christian Uhle**[2,3] **Flavio Everardo**[4]
**Christopher Mitcheltree**[1] **J. Keith McElveen**[5] **Jean-Marc Jot**[6] **Gordon Wichern**[7]

[1]*Centre for Digital Music, Queen Mary University of London, UK*
[2]*Fraunhofer IIS Erlangen, Germany*
[3]*International Audio Laboratories Erlangen, Germany*
[4]*Tecnológico de Monterrey Puebla Campus, Mexico*
[5]*Wave Sciences LLC, USA*
[6]*Virtuel Works LLC, USA*
[7]*Mitsubishi Electric Research Laboratories (MERL), USA*

Artificial intelligence (AI) has seen significant advancement in recent years, leading to increasing interest in integrating these techniques to solve both existing and emerging problems in audio engineering. In this paper, we investigate current trends in the application of AI for audio engineering, outlining open problems and applications in the research field. We begin by providing an overview of AI-based algorithm development in the context of audio, discussing problem selection and taxonomy. We then explore human-centric AI challenges and how they relate to audio engineering, including ethics, trustworthiness, explainability, and interaction, emphasizing the need for ethically sound and human-centered AI systems. Subsequently, we examine technical challenges that arise when applying modern AI techniques to audio, including robust generalization, audio quality, high sample rates, and real-time processing with low latency. Finally, we outline applications of AI in audio engineering, covering the development of machine learning-powered audio effects, synthesizers, and automated mixing systems, as well as spatial audio, speech enhancement, dialog separation and music generation. We emphasize the need for a balanced approach that integrates human-centric concerns with technological advancements, advocating for responsible and effective application of AI.

## 0 INTRODUCTION

Technological improvements ushered in by Deep Learning (DL) algorithms, large datasets, and massive computing infrastructure have made great progress towards AI. The term AI is generally used as an umbrella term encompassing various fields beyond specific technological domains, including Machine Learning (ML), which in turn encompasses DL [1]. The AES has a long history in AI-related topics for audio, which predates the current DL boom, e.g., with topics such as semantic audio [2]. More recent overviews focus on ML and DL from both a learning algorithm [3] and human perspective [4]. In this paper, we provide a discussion on important issues that have emerged as AI-based techniques become more prominent in processing and generating audio signals.

One key concept in ML and AI is to solve problems by learning how to compute output data from input data. Desired input-output behaviour is obtained during a training phase by adjusting the parameters of the implemented functions, e.g. a Deep Neural Network (DNN) [5]. Example applications are noise reduction in speech recordings, where the signal processing algorithm is learned from pairs of noisy input signals and desired clean speech signals, or classification of musical genre of a recording that can be learned from datasets of recordings and corresponding reference for their genre.

Data-driven learning is further categorized into supervised, unsupervised and reinforcement learning [5]. The most widely used, supervised learning, is the process where the desired functionality is defined by data sets comprising input data and references (ideal output data) and loss functions for computing an evaluation criterion quantifying how well a computed output signal matches the ideal output. The learning process is numerical optimization of the trainable parameters. Starting with randomly initialized values, a loss function is computed from mismatch between reference and prediction, as shown in Figure 1.

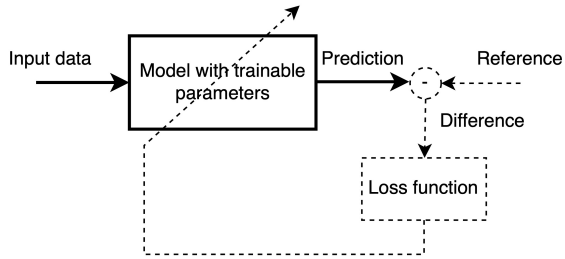


Fig. 1. Block diagram of supervised learning. Dashed lines are parts that are only used during training, all other during both, training and inference.

A gradient descent method is applied to adjust the weights such that the loss function is reduced by iteratively subtracting a fraction of the gradient of the loss function with respect to the weights. This process is also referred to as empirical risk minimization.

Data-driven concepts are important for the following discussion because they relate to challenges mentioned here: they require appropriate data. The quality of the data (how representative are the inputs and how correct are the references) determines the quality of the implementation. The noise reduction processing, for example, may not yield effective results when reference data contain artifacts or interfering sounds. The fact that similar input data can have different references causes (amongst other aspects) that the outputs of DNN are estimates of probabilities of true output. For genre classification the reference labels may stem from multiple taxonomies, musical works are often genre mixtures, and genre labeling experts have different opinions.

DNNs are trained by accumulating the evaluation metric for multiple data points in the data set, typically by averaging. This causes DNNs to perform better for input signals that are similar to training data. It is challenging to obtain good classification performance for less popular genres like Bossa Nova when DNNs have not been trained with sufficient amount of examples.

This dependence of ML on a dataset of input/output pairs for learning algorithm parameters, is in contrast to the well understood input/output mappings present in classical signal processing algorithms. In the context of audio engineering tasks, we consider three different ML problem formulations, based primarily on the types of input/output pairs used by the learning algorithm: labeling, processing, and generation, as shown in Figure 2.

Labeling is the problem formulation where a neural network is tasked with analyzing an audio recording and predicting labels. These labels can be discrete classification labels, e.g., for predicting sound events [6] or musical genre [7], or continuous regression labels, e.g., for estimating quantities such as perceptual loudness [8], musical tempo [9], or fundamental frequency [10]. Audio labeling algorithms are typically tackled using supervised learning techniques (e.g., we learn from a piece of music and its associated genre label). While labeling approaches do not

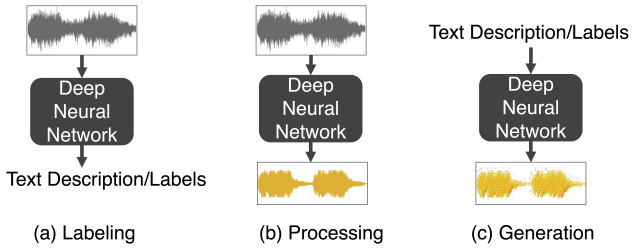


Fig. 2. Problem taxonomy categorizing ML approaches into labeling, signal processing, and generation.

produce audio, they can still be an important part of an audio engineering workflow, as certain processing decisions (e.g., a target equalization curve) can be selected based on a label estimated from an ML algorithm.

The audio processing algorithms shown in Figure 2 (b), are most-related to what we think of as classical Digital Signal Processing (DSP), where the neural network modifies certain characteristics of the input audio. An audio engineering example where DL has enabled tremendous progress is audio source separation, where a network learns to isolate certain sound signals from a mixture [11]. Combining ML methods with classical DSP components [12] has also become a workhorse technique for audio applications, especially those with processing and/or latency constraints.

The generation algorithms shown in Figure 2 (c) use generative modeling techniques to create audio signals in a manner akin to audio synthesis. This emerging class of data-driven generative audio models is beginning to allow for the creation of high-fidelity audio and music signals of remarkable complexity given only text descriptions.

This paper will present some fundamental concepts necessary for comprehending the core principles of AI, with a emphasis on ML, DL, and audio signal processing. As we move forward, our discussion will divide into two primary sections. The initial section examines the prevailing challenges encountered in this domain, analyzing both the human-related challenges in AI and the technical hurdles that require resolution. Shifting focus, the following section explores applications, paving the way for future research directions. As previously mentioned, AI is a large topic with a long history in the AES. While it is not possible to cover everything, we hope that the subset of topics covered in the following sections will provide inspiration for new audio researchers looking for interesting problems, or for experienced professionals transitioning to an AI-based audio algorithm workflow.

## 1 Human-Related Challenges

AI has emerged as a transformative technology that influences various aspects of our lives, including music, its production, and inherent facets such as DSP. As AI becomes increasingly integrated into society, it is imperative to discuss general human-AI-related concepts such as ethics, trustworthiness, explainability, and interaction, in the context of audio engineering as well. Such factors play critical

roles in shaping responsible development, deployment, and utilization of AI systems.

### 1.1 Ethics

Ethics form the foundation for ensuring that AI technologies are developed and used in a manner that aligns with societal values and norms. Ethical considerations in AI encompass a wide range of aspects, including fairness, transparency, accountability, privacy, and bias mitigation. Ethics sets the ground for what humans, for example developers, and AI should, or should not do, even starting with the question of whether a computer can create art [13]. Ethical guidelines and frameworks, along with ongoing discussions and multidisciplinary collaboration, are essential for promoting responsible AI development. Given the well-established understanding that results from a ML model are influenced by training data, our subsequent discussion focuses on ethical considerations within the audio domain, particularly regarding data-related rights.

Data-related rights address the protection of data at different levels. From the development side, privacy rights and personal data protection are essential to avoiding perpetuating discriminatory practices and biases. For training, we need to consider intellectual property (IP) rights [14]. One current challenge in dialog separation, demixing, or automated mixing, is the lack of material permitted for training. It is unlikely to obtain access to the multitrack sources behind successful commercial songs, or to an isolated dialog recording from produced content. These materials are licensed, owned, and protected by the major record labels that produce the material. Access to audio content, whether publicly available or obtained through payment, does not imply permission to train on that data.

Another popular concern is the natural uncertainty about AI replacing existing jobs [15]. Despite the same feeling that happened when drum machines emerged in the 1980s, there has not been a complete or actual replacement of human drummers. The capabilities of (ethical) AI will determine if a job remains in demand [16]. When studios firmly refused to commit to abstaining from producing AI-generated scripts, members of the Writers Guild of America recognized the imminent threat and strictly stood their ground. This labor dispute spanned more than 140 days, from May to September 2023, marking the writers' strike as the inaugural workplace confrontation between humans and AI.

### 1.2 Trustworthiness and Explainability

Trustworthiness in AI relates to reliability, robustness, and accountability of AI technology development and deployment [17, 18]. From the most general perspective, users may require assurance that AI systems will perform as intended and make decisions based on accurate, unbiased, and up-to-date information. It is understood that data protection, privacy safeguards, and secure handling of sensitive information are vital for establishing trust as discussed here.

During the training process, it is desirable to avoid discrimination or favoritism, biased training data, minimize algorithmic preferences, and promote diversity and inclusivity. In audio technology, this is a very important challenge contingent on the availability of training sets that contain representative data. In music production, avoiding biases requires data representative of multiple music genres, instrumentation choices, languages and more, across industry-standard sample rates such as 44.1, 48 or 96 kHz. Trustworthiness implies accountability of both the developers and the users of AI systems. Clear lines of responsibility, recourse mechanisms, and appropriate governance frameworks are necessary to address the potential risks.

The second way trust will be reinforced relies on the quality of results [19]. During the usage of the model, we identify two possible cases depending on whether the results are observable or verifiable immediately by the user. An example of immediate verification is the evaluation of the fidelity of an audio output signal produced by an automatic mixing solution (see Section 3.3) or the quality of dialog separation from produced content (see Section 3.6). In contrast, the effectiveness of an AI-based audio effect processor settings recommendation system is not immediately verifiable. This challenge of verifiability links closely to explainability, which seeks to address the "black-box" nature of many AI systems. Trust can be enhanced when users understand the processes behind AI outputs. For example, if a "black-box" AI-based system outputs an audio signal, the user cannot readily identify or apply a reverse-engineering process to back-trace the results. This drawback may be mitigated by implementing a white-box or explainable solution, as discussed below.

Explainable AI, or XAI [20], refers to the set of techniques and methods employed to make AI systems more transparent and interpretable. These aim to mitigate the "black box" nature of certain AI models and algorithms, which may make determinations or predictions without providing clear explanations for their derivations. In many cases, AI systems such as neural networks can achieve high accuracy and performance but lack transparency, making it challenging to understand why they arrive at specific decisions [21]. This lack of interpretability can be problematic. For instance, an AI-based audio effect processor might produce undesirable artifacts without offering means for the mixing engineer to recreate or correct the processing chain or signal flow, or even know the approach the AI took to reach the specific EQ curve or amount of compression recommended automatically by the system. Explainable AI seeks to provide insights into the inner workings of AI models, enabling users to understand how and why a particular decision was made [22]. In turn, by understanding the factors, features, or patterns that influenced the AI's output, users can gain more trust, identify potential biases, and detect errors or issues.

Trustworthy AI upholds human values by fostering a human-centric design that prioritizes enhancing human capabilities rather than replacing or undermining them. A key component of this trust is explainability, which allows users to understand and engage with AI systems effectively, bridging the gap between technological complexity and human oversight. By promoting transparency and interpretability, explainable AI empowers users to make informed deci-

sions and fosters confidence in AI systems. Frameworks such as the European Commission's Ethics Guidelines for Trustworthy AI and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems provide essential guidance for developing AI solutions that are both reliable and comprehensible, ensuring alignment with societal needs and values [23].

### 1.3 Interaction

Here, interaction refers to the collaboration and engagement between AI systems and human users. Effective interaction design goes beyond the graphical user interface, and is crucial to facilitate seamless communication and understanding between humans and AI-based tools.

The capabilities and limitations of AI are intertwined with a specific set of responsibilities. These responsibilities are delineated through levels of automation in AI, typically categorized into four tiers, which define the extent of control we are willing to entrust to AI systems [19, 24–26].

- **Automatic:** The system operates fully autonomously, without requiring human intervention. A non-expert user may, for instance, rely on the system to handle tasks entirely on its own.
- **Independent:** The AI acts as an assistant, carrying out specific delegated tasks. The user supervises the system and has the authority to override its decisions. Essential work is completed to a near-finished state.
- **Suggestive:** The system's role is limited to analyzing data and proposing recommendations. It serves to guide users by suggesting starting points or alternative options based on input analysis. The user retains complete control over the final outcome.
- **Insightive:** Offering the highest level of control, this tier empowers the user with additional insights, textual data, visualizations, and other resources to facilitate informed decision-making. It is ideal for professionals and industry experts accustomed to having comprehensive control over their work processes.

An example of automatic system is a virtual audio engineer that sets suitable recording/mixing levels for a band during a rehearsal session. A musician who is dedicated to mastering an instrument does not necessarily need to learn audio engineering techniques to meet this objective. In [27], Paul White wrote: *"There is no reason why a band recording using reasonably conventional instrumentation should not be EQ'ed and balanced automatically by advanced DAW software."* At the other extreme, an AI-based system might analyze audio signals and recommend adjustments defined by its detection that a snare drum has too much artificial reverberation effect applied, or that two audio sources cannot be heard clearly due to masking problems, leaving it to the user to execute any changes.

Figure 3 illustrates two workflows that promote AI into an assistant, according to the *suggestive* or *insightive* descriptors listed above. In Figure 3(a) the decision is performed by the user in conjunction with the system after the input analysis, whereas in Figure 3(b), the user is entirely responsible for the decision making, such as determining whether the desired output is in the solution proposals or whether the result must return as input for further analysis. Ideally, the user can either accept or decline the solution proposal by the system before an action is performed.

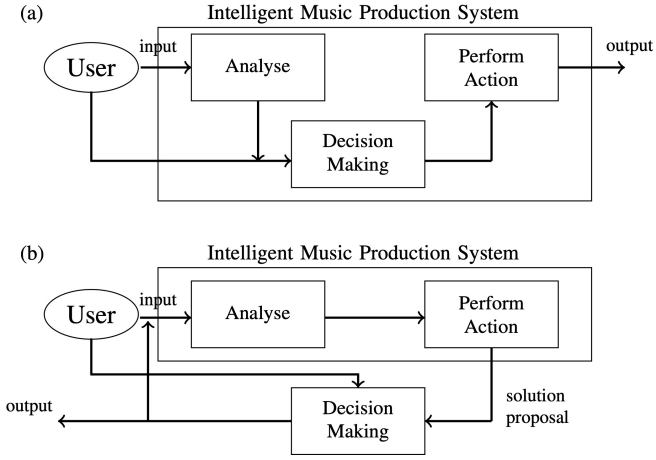


Fig. 3. Example intelligent music production system workflows (inspired by [25, 26]). (a) The decision happens within the system, in contrast to (b) where the user is responsible for deciding whether the solutions are desired or a new search with feedback is performed.

## 2 Technical Challenges

Applications of data-driven implementation and DNN are different in various aspects compared to other technical solutions that are important during development and usage, and are discussed in the following.

### 2.1 Generalization and Robustness

Neural networks typically yield larger errors when applied to data not used in training. This may cause performance drops for some inputs without apparent reason, for example when noise reduction applications achieve attenuation of different sound quality depending on spoken language. Generalization refers to the ability of a model to achieve low error for both, training data and new data [28].

A straightforward solution to obtain robust models is to train with a large variety of data to achieve domain-invariant feature representations[1]. Data augmentation refers to transforming available data to enrich the training data sets quantity and diversity, which is often used in image processing (e.g., by rotating and cropping images), but also in audio (e.g., by applying perceptual audio encoding and decoding, filtering, adding noise, etc.) [29–31]. These procedures are dataset-dependent, and thus require the use of expert domain knowledge.

A special case demonstrating the lack of robustness are adversarial examples which are created with the intent to

[1]Growing list of AI audio datasets for speech, music, and sound effects `https://github.com/Yuan-ManX/ai-audio-datasets`

obtain wrong results and uncover weaknesses, for example by applying imperceptible changes to images that result in the model outputting an incorrect result with high confidence [32]. An example from the audio field is to create adversarial inputs to fool automated speech recognition systems into outputting a malicious transcription chosen by the attacker [33]. Another type of unexpected errors are those caused by shortcut learning [34] where systems learn unintended cues to solve the task, for example by classifying image backgrounds or copyright information instead of recognizing the primary object in the picture.

### 2.2 High Sample Rates

Using very high sampling rates (e.g., 96 kHz) is a topic of debate in audio engineering, that may have some perceptual benefits [35]. While supporting variable sampling rates is a straightforward process with DSP algorithms, many DNN implementations support only a fixed sampling rate. This means that signals recorded at a sampling rate higher than the one supported by the DNN will have their high-frequency information lost or left unprocessed.

One obvious solution to this challenge is to train all models at high sampling rates, however this solution is impractical for multiple reasons. First, the availability of audio content at sampling rates greater than 44.1 kHz is extremely limited, so acquiring the data necessary to train a large audio model on 96 kHz sampling rate data can be difficult or impossible. Second, compute requirements increase dramatically at large sampling rates, as the tensor size required to contain all samples in a given time frame grows with the sampling rate, and a larger model (e.g., longer filter receptive field sizes) may be required to maintain equivalent performance. Given the large computation costs required to train state of the art models, it is quite common for these models to operate at low sampling rates, e.g., 16 kHz. While these sampling rates may be appropriate for certain speech applications where intelligibility of the speech content is the main target, they are not appropriate for most audio engineering applications.

Recently, researchers have begun exploring sampling frequency independent convolution layers for audio source separation applications [36, 37]. Here, the first convolutional layer in a network architecture is treated as an analog filter which is independent of the sampling rate of the input audio. Classical DSP techniques for digital filter design from analog prototypes are then used to convert the analog filter into a typical convolutional neural network layer during training and inference allowing the model to process data at any sampling rate (without resampling the input audio). However, none of the existing work on sampling frequency independent convolutional layers considers sampling rates greater than 48 kHz. Related sampling rate independent approach has recently been proposed for recurrent layers [38, 39].

Audio bandwidth extension based on generative modeling [40, 41] could be one potential approach for overcoming the lack of data available at high sampling rates. Further exploration of implicit neural representations [42], where an audio signal is represented as a continuous function of time rather than a discrete set of samples, could also potentially lead to compute efficient models that operate independent of sampling rate.

### 2.3 Temporal Context

Ideally, an audio ML model would not make independent decisions for each audio sample (or chunk of audio samples), but rather model dependencies across the entire signal. To better model prior knowledge of important semantic characteristics of an audio signal, typically, the signal is first transformed into a feature representation with a lower temporal frame-rate such as the short-time Fourier transform (STFT), mel frequency cepstral coefficients (MFCC), chroma, or more recently neural audio codec features [43]. Oftentimes, these feature representations have the added benefit of reducing the temporal frame-rate, and hence the amount of context that must be modeled. Additionally, many novel DL network architectures have been invented specifically to better model the temporal context of audio signals.

Stacked dilated convolutions, an approach initially popularized in WaveNet [44], effectively increase the available temporal context by increasing the receptive field in successive layers of the network where the dilated convolutions allow for compute and memory efficient operation. For recurrent neural networks (RNNs), the Dual Path RNN (DPRNN) [45] increases temporal context by maintaining two paths of recurrent layers, where the second path operates every $N$ time steps.

The transformer architecture has become the dominant type of deep learning model across application domains, and audio is no different. In addition to exhibiting state of the art performance in tasks such as audio classification [46] and music source separation [47], it also forms the backbone of many current text-to-audio generation models, which will be discussed in more detail in Section 3.7. Built on the self-attention mechanism, the transformer is unmatched in its ability to incorporate context across an entire sequence, but in a naive implementation, the complexity of the model grows quadratically with sequence length, limiting its applicability to long sequences such as audio signals. However, recent software optimizations such as key-value caching [48] and flash attention [49] are rapidly improving the efficiency of attention computation.

Given the exploding interest in large language models (LLMs) built with transformers, research on techniques to increase the context length of transformers, e.g., [50], will also likely be useful for increasing the temporal context in audio models. Additionally, a new class of architectures based on structured state space models (SSMs), which can operate in both recurrent and convolutional modes, have been shown to be incredibly powerful at modeling long sequences, including time-domain audio signals [51].

### 2.4 Real-time and Low-Latency Operation

Interacting with audio, whether playing a musical instrument, using signal processing tools for live sound and production applications, or augmenting the human auditory system with hearing-aid technology requires algorithms to

operate in real-time with low-latency. By real-time we mean that any audio input to an algorithm is processed within a given time-limit or latency requirement. For example, in the Clarity Challenge, which focused on the development of ML algorithms for hearing aids, the latency requirement was 5 ms [52]. Similarly, studies show that digital musical instruments should have a latency below 10 ms [53]. These requirements are significantly more demanding than visual systems where latency values of 30-85 ms can be tolerated [54], and present a serious challenge for audio ML systems.

We note that latency has two main components: (1) *algorithmic latency*, which is caused by constraints on the specific algorithm, e.g., overlap-add operations require all future overlapping frames containing a given audio sample to be observed before the processed version of the audio sample can be output, and (2) *hardware latency*, which is the computing time required to complete the algorithm processing. In analog hardware tools for audio processing, these two sources of latency were completely coupled, whereas digital tools allowed for buffering of audio samples, effectively decoupling algorithmic and hardware latency.

Over the past several years, there has been a growing interest in causal deep network architectures for audio processing, which aim to minimize algorithmic latency. Many architectures that operate on audio signals can be made causal in a straightforward manner such as unidirectional recurrent networks and adding the appropriate padding for dilated convolutions. The Realtime Audio Variational autoEncoder (RAVE) [55] is a good example of these techniques, although obtaining high fidelity results usually requires latency values greater than 20 ms. One interesting future direction for reducing algorithmic latency is training a neural network to predict future frames of a signal to overcome the delay from algorithms that introduce latency such as overlap-add processing [56]. However, a network that is required to perform such a difficult task as predicting the future requires significant modeling capacity and therefore computation to produce accurate results, thus increasing its hardware latency as well.

Combining classical DSP algorithms that are known to be computationally efficient with modern ML techniques, known as differentiable digital signal processing (DDSP) [12, 57], is a promising direction for overcoming hardware latency, which will be further discussed in Section 3.1. As an example, PercepNet [58], took the normally compute-heavy process of DNN-based speech enhancement, and reduced both algorithmic and hardware latency by learning perceptual band gains and comb filter taps for a DSP-inspired model.

### 2.5 Artifacts, Sound Quality, and Loss Functions

The human auditory system is a complex and non-linear function of the input stimulus it receives. Many well understood properties, such as auditory masking, have long been exploited in audio signal processing, perhaps most famously in audio coding. Additionally, the human auditory system has evolved to be particularly sensitive to certain auditory cues that when modified or missing can make processed audio sound unnatural. As a result, designers of ML systems for audio signal processing must consider potential artifacts produced by their systems.

Artifacts present in audio processed by ML models will be unique to every model, and unlike classical DSP algorithms, the cause of an artifact cannot easily be traced back to a certain property of the algorithm. Artifacts may be caused by a property of the model architecture, the data representation, or the loss function, but could also be related to lack of generalization as discussed in Section 2.1.

Artifacts are generated when loss functions do not adequately penalize them. Reduction of artifacts can be achieved by improving the design of the loss function and potentially making task specific modifications informed by human auditory perception. However, incorporating many of the best existing models of the human auditory system into loss functions used for training is non-trivial, as they may be non-differentiable, computationally costly, or lack robustness when presented with real-world stimuli.

Despite tremendous recent progress, there are still many shortcomings in both the loss functions used for network training, and the objective metrics used for network evaluation. A basic approach for constructing a loss function often involves computing the distance between the output and target signal directly in the time domain, with popular approaches such as the signal-to-distortion ratio (SDR) [59]. However, time domain metrics enforce strict adherence to the target both in magnitude and phase response, which may be overly restrictive. This motivates frequency domain metrics, which often measure distance in terms of frequency magnitude, and relax constraints on absolute phase coherence. While both time and frequency domain metrics are lacking in that they require strict coherence to their respective representations, modifications such as frequency weighting can help to further improve the perceptual relevance of these metrics [60–62].

In order to construct more perceptually relevant metrics a number of approaches have been proposed, such as deep feature losses, adversarial losses, and differentiable implementations of reference-free metrics. Deep feature losses measure the distance between representations extracted from neural networks pretrained in audio classification tasks [63–65]. Adversarial losses are similar to deep feature losses in that a neural network is used to construct the loss function, however, this network is trained in tandem with the main network, similar to generative adversarial networks (GANs) [66]. Finally, differentiable reference-free metrics, such as DNSMOS [67] and Torchaudio-Squim [68], use networks trained on perceptual evaluations, which can then be used as an objective. Future work could consider adapting techniques for enhancing language models, such as reinforcement learning through human feedback (RLHF) [69], in the context of audio quality using perceptual evaluations from human listeners.

Approaches used for evaluation are subject to fewer constraints since they need not be differentiable or as computationally efficient, and can additionally consider dataset-wide performance [70]. It is common to use traditional

signal processing-based metrics such as PESQ [71] and PEAQ [72]. However, PESQ is limited in that it is optimized for speech transmission and only considers sampling rate of 16 kHz, whereas PEAQ may only be appropriate for measuring coding artifacts, and lacks open source implementations. Improving perceptual models for other audio engineering tasks such as spatial audio [73] and overall audio quality [4] remain an important research direction. Distributional approaches such as the Fréchet Audio Distance (FAD) [74] are often employed. Unlike other approaches, FAD relies on measuring the distance between distributions of features from the system under test and the target domain. While this can capture a more holistic snapshot of the system performance, it still relies upon the representations from pretrained models, such as VGGish [75], which suffer from the same limitations as deep feature losses, although music-specific improvements were recently proposed [76]. As a result, the gold-standard in evaluation still relies on perceptual studies involving human listeners, which are often based on the ITU-R BS.1116-2 or MUSHRA test designs.

### 2.6 From Research to Practice

A challenge that remains in the field of AI and audio signal processing is bridging the gap between researchers and practitioners. This is not just about interactivity; it is a human-AI challenge that also involves technical, logistical, and design considerations to create a tool for end users from a research project. While this problem exists for all research fields, the human and technical challenges specific to audio discussed in the previous sections can make bridging this gap particularly difficult.

Making AI and audio research usable by practitioners requires more than standard software engineering best practices such as detailed installation instructions, containerization, and dependency management. For example, when these systems are made publicly available via an open source code repository, they usually require some programming knowledge to set up and are often controlled asynchronously through a command line interface. This is in stark contrast to the real-time feedback-driven workflow introduced by audio plugins which have streamlined music-making since the VST standard was introduced in 2005. An easily installable solution callable within your Digital Audio Workstation (DAW) aligns with today's production workflows and practitioner requirements of real-time processing, low-latency output, and adaptability to arbitrary input audio. Research projects that are accessible in the DAW, such as Spleeter[2] for source separation, demonstrate this and remain popular due to their ease of use, even when outperformed by newer, less accessible state of the art solutions, such as Hybrid Demucs [77].

Fortunately, in the last few years there has been a surge in general purpose deployment tools such as Google Colab and Huggingface Spaces, as well as audio specific offerings, including HARP [78], Neutone[3], RTNeural [79], and more that specialize in deploying AI audio research in the cloud or as an audio plugin in the DAW. These tools are free and provide user-friendly interfaces which further contribute to the democratization of AI audio technology.

However, making these tools available without overly restrictive technical guardrails in place such that they can also be misused creatively is a delicate balancing act. Researchers can help reduce the gap between their work and practitioners by choosing a deployment tool and considering its limitations from the onset of a project when critical decisions are typically made about latency, model sizes, sampling rates, applications, and target audience.

Another point that should be considered when bridging the gap between audio research and practice is how the capabilities of a system are communicated. While most academic works report the average performance of a system, the worst-case performance can be more important from a practitioner's point of view. Determining this can be challenging and authors may have less incentive to highlight this information when trying to publish. However, it can also be provided in the supplemental material of a paper for readers who are interested. Considering both best- and worst-case scenarios can provide a more realistic picture of an AI model's abilities, thus aiding potential users in making informed decisions about its suitability for real-world applications. Making a system easily accessible through a deployment tool like the ones discussed previously can enable others to determine the limits of the system, even if they are not explored in the original academic work. In addition to this, allowing users to provide their own inputs reduces the cherry-picking bias of pre-selected audio samples.

## 3 APPLICATIONS

In this section we provide overviews of some exciting uses of AI in audio signal processing applications. As with any widely applicable technology, the number of possible applications is vast and this list is in no way meant to be exhaustive. Some notable omissions include room acoustics, audio coding, physical modeling, and audio retrieval, among many others. It is our hope that the description of audio effects, synthesizers, automatic mixing and mastering, spatial audio, speech enhancement, dialog separation, and text-to-music generation will provide useful overviews of these areas, and also serve as potential roadmaps for using AI in new audio applications.

### 3.1 Audio Effects

Audio effects are signal processing devices used to shape the sonic characteristics of audio signals and they play a central role in audio production with applications in music, film, broadcast, and video games [80]. While there is a large body of research outlining the design and implementation of audio effects, in recent years, researchers have investigated how ML and AI may be used to address previously difficult to solve problems. These applications may include analysis tasks, such as the detection or classification of audio effects [81–83], or estimation of audio effect parameters [84, 85]. Other approaches build on this and

[2] https://github.com/deezer/spleeter

[3] https://github.com/Neutone/neutone_sdk

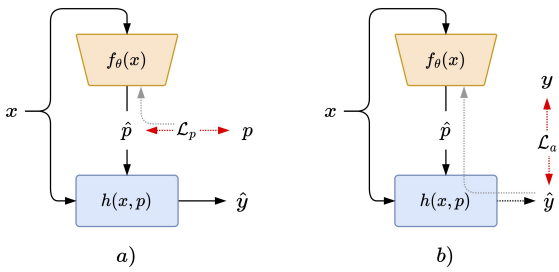


Fig. 4. Comparison of parameter and audio-based loss formulations for controlling the parameters $p$ of an audio effect $y = h(x, p)$ to process an audio signal $x$ using a neural network controller $f_\theta$. (a) Formulation using a parameter-based loss function $\mathcal{L}_p$, which requires ground truth parameters $p$. (b) Formulation using differentiable signal processing techniques to enable an audio-domain loss function $\mathcal{L}_a$ by backpropagating through $h$.

propose to automatically control audio effects by estimating optimal control parameters. This can include tasks such as automatic equalization [86, 87], noise reduction [88, 89], as well as audio production style transfer, also known as sound matching, with applications to equalization [90], dynamic range compression [91], artificial reverberation [92] and a complete signal processing chain [93].

However, applications of ML for audio are not restricted to only analysis. They also include applications where a neural network operates directly on audio signals to produce a processed version. This can involve tasks such as virtual analog modeling, where a system is designed to emulate the behavior of an analog system. Traditional approaches have leveraged measurement data for this task [94, 95]. However, studies leveraging neural networks in both black-box [96, 97] or grey-box [98, 99] models have shown superior performance and reduced need for hand-designed components.

Neural networks can also be used for the inverse problem of removing audio effects from processed recordings, both for single effect such as distortion [100], reverberation [101], and dynamic range compression [102], or entire effect chains [83, 103]. Finally, ML can also enable the construction of novel transformations through neural audio effects, with approaches including randomly weighted networks [104] and steerable networks [105].

An emerging research direction in audio effects involves leveraging techniques from differentiable signal processing [57] to construct differentiable implementations of audio effects. As shown in Fig. 4, implementing audio effects in a differentiable manner enables the use of gradient based techniques from ML. This can enable a number of applications. For example, it can facilitate reverse engineering of audio effect parameters through a gradient-based optimization process [106] and can also aid in the construction of grey-box models for virtual analog effect modeling [99]. By combining existing components such as filters and non-linearities, we can construct interpretable and efficient models in a data-driven fashion.

Differentiable audio effects also have applications in parameter estimation and control scenarios, with applications in audio production style transfer [93] and automatic mixing [107]. Directly incorporating audio effects in the computation graph while training enables the use of an audio domain loss as opposed to a parameter-based loss, which can be problematic [108]. This is critical for applications where the ground truth parameters are not known, such as in automatic mixing or style transfer applications.

Open challenges in applications of ML for audio effects include: first, the lack of neural audio representations that adequately capture information about transformations caused by audio effects [109], and second, the generalization of existing techniques to real-world use cases due to the difficulty in interfacing with commercial audio effects.

While there has been significant development in so-called general purpose audio representations [110], pretrained models have been shown to lack detailed information about audio effects [109]. This is often due to the fact that audio effects are commonly used as data augmentations during training, hence encouraging invariance to these transformations. While this may be beneficial for certain downstream tasks, such as sound event detection, it is important to capture these transformations for audio effect related tasks. As a result, more work is needed in developing audio representations that capture information about audio transformations while remaining largely invariant to the underlying content [111].

The generalization gap in existing approaches is often due to the inability to produce training data using commercial audio effects. As a result, it is common for many approaches to use simplified or basic implementations of audio effects. While this can be useful for toy problems, it prohibits the construction of systems that interface with the tools used in the audio engineering practice. This reality also impacts work in differentiable signal processing.

Due to the constraints of existing approaches, approximations must be made for explicitly differentiable approaches [93] and methods for enabling gradient based learning for black-box devices such as neural proxies [107] and gradient approximation [112] remain limited. As a result, the field could benefit from further work on generalizing and improving these methods to enable efficient and stable learning in the control of arbitrary signal processing devices.

### 3.2 Synthesizers

Synthesizers for music production are hardware devices or software instruments designed to generate sound through various signal processing techniques. There are several common types of musical instrument synthesizers, such as additive, subtractive, concatenative, wavetable, frequency modulation, granular synthesis, and more, each employing distinct methods to create and shape sound. Synthesizers offer creative freedom over every aspect of sound design and can be controlled with MIDI (Musical Instrument Digital Interface) commands for precise control and expressive performances. Since the 1990s [113, 114], and especially in the past decade [115–118], ML has been applied to synthesizers for several different tasks such as estimating their param-

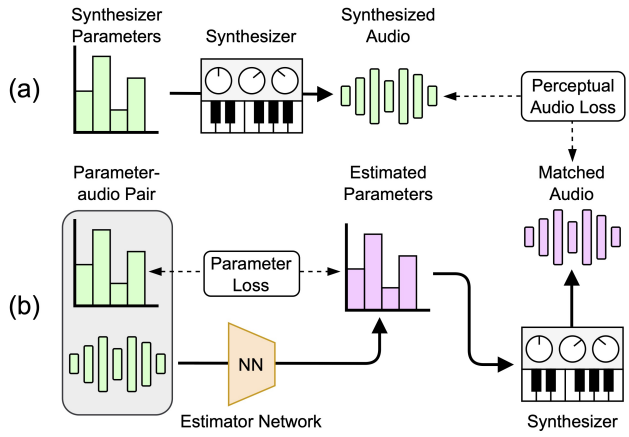


Fig. 5. Synthesizer parameter estimation for sound matching. (a) A synthesizer renders audio according to some synthesis parameters. (b) Parameter estimation using a neural network and a parameter or perceptual audio loss function.

eters, exploring new methods of synthesis, and providing novel control mechanisms. While the machine learning applied to audio synthesis research topic is vast, this section focuses on these three research directions along with their associated challenges for controllable synthesizers typically used by audio engineers and music producers.

A relatively recent research direction, illustrated in Figure 5, is sound matching using synthesizer parameter estimation. This is an analysis task involving ML systems to reverse engineer the parameter settings (also known as a patch) for creating a certain sound [119]. Learning to program synthesizers is a time-consuming process, usually obtained through inefficient trial and error and only mastered after years of experience. As a result, the parameter estimation task can serve as an educational tool for novices and as a creative tool for experts when a predicted patch is different or erroneous. Evolutionary algorithms [120] and DNNs [117] have successfully been applied to the Yamaha DX7 synth [115], Serum synth [116], modular synths [121], and more.

Several open challenges remain for the parameter estimation task. Interfacing with most existing synthesizers to automate data collection is difficult, which limits training datasets and the ability for parameter estimation systems to generalize to other synths. Different parameter settings can create the same sound, which results in contradictory gradients when training a neural network and comparing its output directly to the parameter values (also known as P-loss). Parameters also often have "dead zones" where modifying them results in no perceivable change in audio which further exacerbates this issue.

These challenges have led to recent work on learning synth parameters end-to-end [118] and directly comparing the resulting synthesized audio to the target audio using a perceptual similarity metric [122]. However, this approach comes with its own challenges, such as those related to loss functions discussed in Section 2.5 and the non-differentiability of most synths hindering gradient-based learning. This approach also expects an isolated recording of the synthesizer as target audio which may not always be obtainable. Similar to audio effects, differentiable synth implementations and neural proxies have been used to address some of these challenges, but are still at an early stage and could benefit from further work.

Another direction is using neural networks to synthesize audio directly by providing a target sound or high-level control parameters as input. This approach comes with several benefits such as fully differentiable synths enabling gradient-based learning, GPU acceleration enabling faster data generation, and novel methods of synthesis such as timbre transfer and interpolation. Recent work [12, 118] has leveraged neural networks to control the parameters of white-box differentiable digital signal processing (DDSP) synth architectures which provide better interpretability via their intermediate parameters and strong guardrails on the resulting audio. Differentiable and parameterized physical models of the vocal tract [123], resonators [124], and other instruments have also been used as the synth architecture in these systems, resulting in efficient and controllable generation of audio modalities that are beyond the capabilities of traditional DSP-based synths.

Several open challenges exist for this research direction. Neural synth architectures may be technically differentiable, but have incorrect or missing gradients [125], thus hindering gradient-based optimization. Neural synthesizers also often lack interpretable controls and are prone to producing artifacts, making them less suited for professional applications. These challenges are explained more in Sections 1.2, 1.3, and 2.5. Finally, due to their custom differentiable implementations, these novel forms of synthesis are typically less accessible and cannot be applied to existing non-differentiable synthesizers, thus preventing widespread adoption by practitioners.

Controllability of synthesizers is another recent research area. Most synthesizers have a large number of parameters that can make it difficult and tedious to discover new sounds. This has resulted in research on mapping synthesizer parameters to a latent space [126] where similar sounds are embedded close to each other. This enables a new method of generating synth patches by sampling from the latent space and "exploring" it. This approach can also be applied to novel neural synthesis methods [127], resulting in a new form of synth control called latent space exploration.

One major challenge of synth controllability is the interpretability of the latent space and reducing the dimensionality of it such that it can be visualized and explored by the user. Prior work has investigated which dimensions of the latent space are most important [128], but this area of research is still nascent. Further research needs to be done on novel control methods beyond latent space exploration that are also suitable for the workflow of audio production practitioners.

### 3.3 Automatic Mixing and Mastering

In the context of music production, both mixing and mastering without any intelligent support are very challenging tasks. After recording and editing, the mixer must perform a

set of decisions that involves technical challenges [129], such as ensuring balance between sources, minimizing masking, distributing elements in the panorama field and guaranteeing a desired stereo breadth, but also, creative or artistic decisions like the application of audio effects such as reverb or the automation of the cutoff frequency of a filter. The interaction of these mixing decisions enables a constant trade-off in trying to produce the best possible outcome [130].

The role of the mastering engineer has evolved from its early days of simply transferring music from tape to acetate disc with technical adjustments like equalization and noise reduction. Today, mastering involves not only enhancing sound quality but also ensuring consistency across different songs in an album and across various playback systems, from headphones to large event speakers. Modern mastering also focuses on meeting specific loudness standards as part of optimizing audio before distribution.

In today's digital music production, the entire process from recording to mastering can be done within a digital audio workstation (DAW). The transition from analog to digital has brought a wealth of software tools for both mixing and mastering engineers. DAWs now replicate traditional mixing consoles, in the same way audio plugins emulate physical peripherals. Modern computers, with their high computational power, handle large sessions with numerous tracks, synthesizers, and effects, all at high resolutions and sampling rates. This shift has democratized access to high-quality digital versions of analog equipment like compressors, making them more affordable than their hardware counterparts.

The integration of AI in music production has led to the rise of Intelligent Music Production (IMP; [25, 131–133]). This growing field aids mixing and mastering engineers by automating certain processes using intelligent systems. As automatic mixing enters its second decade [132, 134], the use of ML in multitrack mixing tools [135, 136] has become increasingly popular together with expert or knowledge-based systems [107][4]. Today, commercial applications in this area can produce complete mixes from individual audio sources or provide feedback on a single mix file.

AI mainly in the form of ML or DL, has demonstrated its effectiveness in research [90, 91, 107, 137] with dedicated drum mixing techniques, or even entire multitrack mixing approaches [138], as well as in various commercial products across different formats, often integrated into common daily workflows, as discussed in Section 2.6. While more companies are inclined to gather user data, it's commonplace for products to be presented within a web browser. Conversely, the most intuitive approach to introduce AI to individuals typically confined to studio environments is through audio plugins, which requires to enable further functionalities to attain visibility into activities and processes across other audio channels as shown in Figure 6, so an audio plugin for mixing purposes can *listen* to what is happening in another track at the same time. Conversely, this workflow is natural for a DAW, because of the scope and visibility of all audio information in the session.

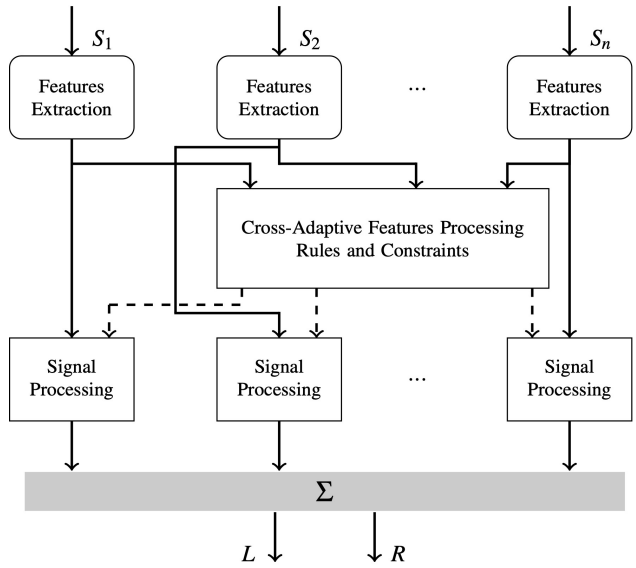


Fig. 6. Mixing processes with cross-adaptive signal processing from [134].

Exploring a potential research direction involves investigating the incorporation of explanations into ML-based solutions. Establishing trust between the user and the system remains crucial, regardless of the user's and the AI's level of participation [26]. This research aims to address this challenge by emphasizing the importance of delivering high-quality results accompanied by transparent explanations or reasons for the AI's decisions. A cross-adaptive framework can provide the desired explanations.

The evolution of AI in the realm of creative processes is evident, with a shift from emulating analog gear to exploring novel avenues. As for challenges in the automation of mixing and mastering, the application of ML and DL in specific cases, introduces both promising advancements and significant challenges in ensuring the production of reliable and valuable content. Several key considerations emerge:

**Efficiency and Speed:** DL can accelerate the mixing and mastering processes, reducing manual adjustments. However, efficiency (considering worst-case scenario as discussed in 2.6) depends on the quality and complexity of the ML model. Specific training strategies may be required for optimal performance, and the current challenge lies in accessing an ample supply of high-quality data.

**Costs:** Online or plugin-based solutions cater to hands-on producers, contrasting with professional mastering engineers who command higher fees due to established industry reputations.

**Consistency and Customization:** ML approaches ensure consistent sound quality across tracks while allowing customization. Model quality is crucial; poorly trained models may introduce inconsistencies, biases or unintended artifacts.

**Subjectivity:** Subjectivity plays a crucial role, particularly in creative decisions tied to emotional judgment. De-

[4]Automatic mixing research `https://csteinmetz1.github.io/AutomaticMixingPapers/index.html`

spite training models with industry-standard data, certain subjective decisions remain challenging to quantify, complicating the training of models for specific creative aspects. Subjectivity may enhance creative freedom but poses a tradeoff, potentially limiting the intuitive and personal touch that skilled human engineers contribute.

In leveraging ML for audio mixing and mastering, it is imperative to recognize both its exciting possibilities and inherent limitations. Striking a delicate balance between automation and human expertise in a collaborative manner stands out as a key challenge in harnessing ML for these creative processes.

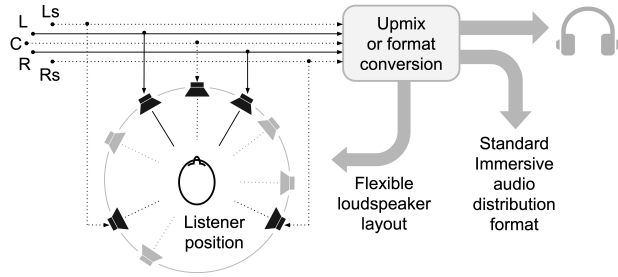


Fig. 7. Upmix or conversion of an audio signal to various immersive playback configurations or distribution formats.

### 3.4 Upmix and Format Conversion

Along with traditional two-channel stereophony, the demand for the production and distribution of audio content in surround or immersive multichannel formats has grown in the cinema, broadcast and music industries. This trend has motivated the continued development of multichannel audio signal processing methods for the conversion of recordings between different multichannel and spatial audio formats, as illustrated in Figure 7.

In traditional formats, each audio waveform channel is destined to feed discretely an individual loudspeaker or cluster within an array facing or surrounding the listener, according to a predefined geometrical layout [139]. More recent layout-agnostic formats and low-bit-rate coding specifications are deployed in professional or consumer-grade media authoring software, distribution services and playback equipment [140–142]. Layout agnosticity may be realized by representing the multichannel audio signal as an Ambisonic *spatial audio scene* or as a collection of *spatial audio objects* each assigned a fixed or time-varying perceived position on a virtual sphere centered on the listener [139].

Multichannel audio format conversion scenarios include: *Upmixing* – including the conversion of legacy mono or stereo audio content to address consumer systems, venues or distribution formats comprising a higher channel count; *Downmixing* – conversely, in some situations, it is necessary to adapt a multichannel audio signal to a reduced number of channels. An important special case is binaural audio reproduction for headphone playback, which typically involves downmixing by virtual loudspeaker array simulation, potentially following an upmixing stage aiming to enhance the listener's sense of immersion.

These operations leverage a well researched frequency-domain processing framework nowadays commonly referred to as *parametric spatial audio signal processing* [143, 144]. A common set of signal processing challenges confronts the DSP algorithm designer. *Primary-ambient decomposition* of the source signal is employed so that ambient signal portions, such as reverberation, may be processed selectively by employing a format conversion strategy that preserves the perceived diffuseness of the reproduced sound field [145–150]. Primary components, on the other hand, require distribution strategies that maximize spatial discrimination, often accomplished by estimating and re-rendering the apparent direction of arrival of each time-frequency component [145, 151–153].

Inevitably, the source audio material may present a significant degree of time-frequency overlap: at any given time and frequency, it is typically composed of ambient or reverberation components superimposed with one or more primary components emanating from different directions of arrival. Parametric methods seek to exploit the information implied by inter-channel signal differences in the time-frequency domain. They assume that spatial information is supplied for each input audio channel (such as its directional coordinates). Most critically, they implicitly assume sparsity (low overlap) in the time-frequency domain.

Accordingly, therein lies an opportunity for ML-based approaches to drastically improve input signal decomposition/reconstruction performance, in analogy with the recent progress demonstrated in the performance of musical source or stem separation methods [154, 155]. Exemplary studies explore this opportunity to develop improved solutions for ambience extraction [156–158], format conversion of primary components [159–161], or pseudo stereophony from a mono source signal [162, 163].

As observed in [164], the development of data-based signal processing algorithms for spatial audio format conversion is currently hindered by infrastructure and psychoacoustic modeling limitations. Training such algorithms may involve the construction of datasets comprising audio recordings stored in several possible multichannel formats, thus placing extra requirements on storage and computation capacity. Additionally, the many subjective attributes that underlie spatial audio perception are not readily encompassed in a differentiable loss function suitable for training supervised or non-supervised ML models, such as those discussed in Section 2.5.

In the future, bringing ML/AI techniques to bear in the development of novel spatial audio analysis/synthesis methods can unlock practical applications with transformative impact. It may enable, for instance, the realization of immersive media and experiences leveraging legacy or historic recordings (many of which accompany synchronized video media). Recent work illustrating such perspectives demonstrates the incorporation of ML techniques in wearable immersive audio [165], audio-visual processing [166, 167], spatial audio coding [43, 168, 169], or differentiable frequency-domain representation [170].

### 3.5 Speech Enhancement in Acoustic Environments

Speech Enhancement (SE), a key signal processing task in many audio systems, is utilized in numerous applications to boost the signal-to-noise ratio (SNR) and improve speech signal quality. This enhancement is typically aimed at two main objectives: serving as an immediate goal in devices like hearing aids, and as a preparatory phase in systems that depend on high SNR or speech quality, such as speech or speaker recognition systems. Common components of SE include acoustic echo suppression and cancellation, background noise reduction, bandwidth extension, suppression of competing speech, and source separation.

Recent developments in ML and neural network-based approaches have demonstrated significant progress in SE. Techniques that require pre-training, such as Generative Adversarial Networks (GANs), DNNs, and other architectures, have improved on the state-of-the-art in enhancing near-field speech in particular [171–174]. At the same time, data-driven approaches that do not need pre-training, including physics-based ML models, have advanced notably in regards to source separation for reverberant environments in the far field, addressing challenges like the Cocktail Party Problem [175].

However, applying AI and ML techniques to SE presents several considerable challenges to real world implementations, particularly those for live and battery-powered applications. For example, key requirements in many portable and wearable applications include very low power consumption and latency below 20 milliseconds or even 5 milliseconds for hearing aid applications. While hardware architectural advancements, such as the emergence of analog AI chips, have been made in response to these concerns, continued research and development efforts are still needed to achieve these goals.

Furthermore, the availability of suitable, large training datasets impacts the design of effective AI and ML solutions for some applications. Real recording datasets such as TIMIT [176], Librispeech [177], Voice Bank [178], NOISEX, and CHiME [179] are widely used but are of limited size relative to the amounts used in training models in other modalities such as text and images. Simulated datasets are widely used in research to augment the size of the training datasets, but they may not suffice for complex or dynamic applications. The effectiveness of general models derived from these simulations often hinges on their quality and realism.

In recent years, several datasets containing real, simulated, or a combination of both types of data have become publicly available, such as Amazon's MASSIVE 51-language dataset [180]. Additionally, various competitions have been held focusing on applying DL to augmented reality and hearing aids which have provided their own training datasets comprising simulated and some real recordings, such as the SPEAR Challenge [181], the Interspeech 2021 Deep Noise Suppression Challenge [182], and the Clarity Enhancement Challenge [52]. To mitigate the lack of real recording datasets suitable for training, ongoing research into learning methods that do not require large datasets is underway, with some early promise being shown by meta-learning and few-shot learning approaches [183, 184].

While all SE applications share the common goal of improving SNR and quality, the diversity and dynamic complexity of acoustic environments present both challenges and opportunities for AI and ML solutions and some research into alternative solutions is already occurring [67].

### 3.6 Dialog Separation in Produced Content

Dialog separation refers to the process of obtaining a target speech signal from produced content with interfering sounds, e.g., music, recorded ambience and sound effects during post-production. It can be applied to improve the intelligibility or reduce the listening effort of TV and movie sound when the clean speech signal is not separately available and background is mixed at too high level, or for upmixing.

An early data-driven approach for Dialog Enhancement (DE) combined conventional feature extraction with shallow Artificial Neural Networks (ANNs) [185]. Current methods use DNNs to estimate a representation of the target signal or the parameters for retrieving the target signal from the input mixture. They are mainly applied to time-domain signals [186] or Short-Time Fourier Transform (STFT) coefficients directly, except works on SE that implement strong inductive biases using signal processing methods for real-time processing at low latency [88]. Similar concepts are used for music source separation [154, 155] with applications to karaoke or music remixing and SE for communication applications.

Many methods process signals in the time-frequency domain obtained with STFT by predicting the target signal directly or by element-wise multiplication with real-valued scalars estimated by a DNN that are referred to as a mask [187]. Complex-valued parameters are used in order to restore the signal phase in addition to magnitudes of STFT coefficients [188] or phase information is estimated [189]. Current systems use architectures with jointly optimized encoder, masking and decoder. These methods are based on convolutional layers, recurrent units or attention mechanisms.

While most methods optimize a cost function without taking perceptual constraints into account, perceptually motivated cost functions have been developed based on Short-Time Objective Intelligibility (STOI) [190], Perceptual Evaluation of Speech Quality (PESQ), and Perceptual Evaluation methods for Audio Source Separation (PEASS) [191]. Other works propose to control the trade-off between sound quality and attenuation [192].

First systems are already commercialized and more work on improving separation and sound quality is continuously published. Source separation is challenging, because the variety of interfering sound is large (music, environmental noise and effect sounds), no a-priori knowledge of the microphone configuration is available, and the listeners expect high sound quality, i.e., the processing should not introduce audible artefacts. For many applications processing in real-time with low-latency is required, but better results

in terms of sound quality and separation are achieved in off-line processing.

Various data sets have been made publicly available in the past that can be used to train dialog separation [193] when combining clean speech data [194] with interfering sounds [195], but high quality data with spontaneous speech from fiction content is unfortunately not publicly available to the required extent.

Current trends are increasing depth of neural networks [189] and source separation with weakly labelled data, which is of interest to address applications with specific requirements on separated targets or signal formats for which no extensive data resources exist [196].

### 3.7 Text-to-Music Generation

Music generation is a fundamental task in music signal processing and can be traced back to early algorithm music composition that emerged in the 1950s [197]. Since then, there has been significant evolution in the application of AI for music generation with increasing sophistication [133]. In recent years, a wide range of approaches leveraging DL, and more specifically generative modeling, have been developed. The first generation of DL approaches demonstrated promise from adapting speech synthesis methods, such as WaveNet [44] and WaveRNN [198]. However, the results were notably lacking both in fidelity and controllablity, as these were unconditional models and scaling them was challenging since they operate on the waveform.

More recently, powerful large-scale generative models have demonstrated impressive performance in adjacent domains such as image generation with Stable Diffusion [199] and DALL-E [200], as well as text generation with GPT-like models [201]. These results have sparked renewed interest in audio generation using similar techniques and has led to a series of emerging generative models that adapt a similar text-prompt based generation process, commonly referred to as text-to-audio models [202].

Text-to-audio models can enable the generation of complete musical tracks based upon a user-provided text prompt, such as *"dance music with violins and cuica with odd rhythm changes and delicate dynamics."* While there has also been significant advancement in symbolic music generation systems, we will focus our discussion only on those systems that synthesize audio directly based on a user-provided text prompt. While text-to-audio has become a predominant paradigm for music generation models, there is a range of underlying training techniques and model architectures. We will further focus our discussion on two of the most common strategies, which involve leveraging diffusion or autoregressive generative models.

Autoregressive approaches, which often use transformer building blocks, treat the generation process as a sequence modeling task where the sequence is generated by predicting small subsequences in an iterative and recursive process. Examples of autoregressive transformer models include MusicGen [203] and MusicLM [204]. As addressed in Sec. 2.3, one of the major challenges of audio models is context length. This is of particular relevance for transformer based sequence models due to their quadratic scaling in relation to the sequence length, which makes training such a model at the waveform level infeasible.

To make the modeling of audio tractable, it is common to first build a temporally compressed representation on which the sequence model can be trained. In practice, this is often achieved with a neural audio codec [205, 206], a learnable model trained in a compression task to represent a waveform as a sequence of discrete codes or tokens. This transforms the task of modeling audio sequences into a form very similar to natural language enabling the application of successful techniques from this domain. Conditioning of the generation process on text prompts can be achieved either with cross-attention layers that modulate the behavior of the sequence model or by simply prepending tokens from a text encoder to the sequence of neural audio codec tokens.

On the other hand, diffusion models, which often operate either in the time domain or the time-frequency domain, model the complete sequence concurrently, but instead employ a process of iterative refinement [207]. Some diffusion models operate directly in the time domain such as Moûsai [208], however, latent-diffusion models are more common, which instead first train an autoencoder to temporally compress audio signals, and then train a diffusion model on these compressed representations. Latent diffusion models share some conceptual similarity with the use of neural audio codecs, however, they do not require discrete sequences, often easing the training process. Popular latent diffusion approaches for music generation include StableAudio [209], MusicLDM [210].

While recent models continue to demonstrate compelling generation performance, even leading to some commercially viable solutions, fundamental improvements are still needed to achieve superior fidelity of generated audio and further extend the mechanisms for control and conditioning. This may involve further advancements to the underlying architecture components, training techniques, or generative model approaches. Recent works have begun to explore more sophisticated conditioning techniques [211, 212] and improvements in neural audio codecs are rapidly developing, leading to improved audio fidelity [213].

## 4 CONCLUSION

In this paper, we have outlined the advancements in audio engineering enabled by the integration of AI. We presented an overview of current trends, challenges, and emerging applications and identified the importance of addressing both technical and human-centric challenges in the development of AI systems that are robust, ethically grounded, and user-oriented. Our discussion comprises a description of technical challenges including generalization, audio quality, and real-time processing, which require further innovation. Additionally, we identified directions for applications of AI in audio engineering. While AI has already begun to enable applications in audio engineering, we argue for a balanced approach that aligns technological progress with ethical standards and human-centric principles, fostering the responsible and effective utilization of AI in audio engineering.

## 5 ACKNOWLEDGMENTS

Christian J. Steinmetz and Christopher Mitcheltree are supported in part by EPSRC UKRI CDT in AI and Music (Grant no. EP/S022694/1).

## References

[1] S. J. Russell, P. Norvig, *Artificial intelligence: a modern approach* (Pearson) (2016).

[2] F. Rumsey, "Semantic Audio: Machines Get Clever with Music," *J. Audio Eng. Soc.*, vol. 59, pp. 882–887 (2011 Dec.).

[3] F. Rumsey, "Audio Processing—Learning from Experience," *J. Audio Eng. Soc.*, vol. 69, pp. 361–365 (2021 May).

[4] F. Rumsey, "Quality, Emotion, and Machines," *J. Audio Eng. Soc.*, vol. 69, pp. 890–894 (2021 Nov.).

[5] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press) (2016).

[6] A. Mesaros, T. Heittola, T. Virtanen, *et al.*, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83 (2021), doi.org/10.1109/MSP.2021.3090678.

[7] G. Tzanetakis, P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302 (2002), doi.org/10.1109/TSA.2002.800560.

[8] J. Schlittenlacher, R. E. Turner, B. C. Moore, "Development of a deep neural network for speeding up a model of loudness for time-varying sounds," *Trends in Hearing*, vol. 24 (2020), doi.org/10.1177/2331216520943074.

[9] H. Schreiber, J. Urbano, M. Müller, "Music tempo estimation: Are we done yet?" *Trans. Int. Soc. Music. Inf. Retr.*, vol. 3, no. 1, p. 111 (2020), doi.org/10.5334/tismir.43.

[10] J. W. Kim, J. Salamon, P. Li, *et al.*, "Crepe: A convolutional representation for pitch estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 161–165 (2018 Apr.), doi.org/10.1109/ICASSP.2018.8461329.

[11] E. Vincent, T. Virtanen, S. Gannot, *Audio source separation and speech enhancement* (John Wiley & Sons) (2018).

[12] J. Engel, L. H. Hantrakul, C. Gu, *et al.*, "DDSP: Differentiable Digital Signal Processing," in *Proc. Int. Conf. Learn. Represent.* (2020 Apr.).

[13] A. Hertzmann, "Can computers create art?" in *Arts*, vol. 7, p. 18 (2018).

[14] B. L. Sturm, M. Iglesias, O. Ben-Tal, *et al.*, "Artificial intelligence and music: open questions of copyright law and engineering praxis," in *Arts*, vol. 8, p. 115 (2019).

[15] P. K. McClure, ""You're fired," says the robot: The rise of automation in the workplace, technophobes, and fears of unemployment," *Social Science Computer Review*, vol. 36, no. 2, pp. 139–156 (2018).

[16] P. Samuelson, "Generative AI meets copyright," *Science*, vol. 381, no. 6654, pp. 158–161 (2023).

[17] L. Floridi, "Establishing the rules for building trustworthy AI," *Ethics, Governance, and Policies in Artificial Intelligence*, pp. 41–45 (2021).

[18] J. M. Wing, "Trustworthy ai," *Communications of the ACM*, vol. 64, no. 10, pp. 64–71 (2021).

[19] A. Palladini, "Intelligent audio machines," in *Keynote Talk at 4th Workshop on Intelligent Music Production (WIMP-18), Huddersfield, UK, Sep.*, vol. 14 (2018).

[20] D. Gunning, "Explainable Artificial Intelligence (XAI)," *Defense advanced research projects agency (DARPA) Project* (2017).

[21] R. Confalonieri, L. Coba, B. Wagner, *et al.*, "A historical perspective of explainable Artificial Intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391 (2021).

[22] A. Das, P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," *arXiv preprint arXiv:2006.11371* (2020).

[23] N. A. Smuha, "The EU approach to ethics guidelines for trustworthy artificial intelligence," *Computer Law Review International*, vol. 20, no. 4, pp. 97–106 (2019).

[24] B. D. Man, "Rethinking the Music Production Workflow," (2017 Nov.), URL `youtube.com/watch?v=H_OSedvWAmQ`, audio Developer Conference.

[25] D. Moffat, M. B. Sandler, "Approaches in intelligent music production," in *Arts*, vol. 8, p. 125 (2019).

[26] C. Sacristán-Ramírez, F. Everardo, Y. Burguete, *et al.*, "Al in Music: Implications and Consequences of Technology Supporting Creativity," in *What AI Can Do*, pp. 233–252 (Chapman and Hall/CRC) (2023).

[27] P. White, "Automation For The People," (2008), Sound on Sound, Vol. 23, no. 12.

[28] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer) (2007).

[29] B. McFee, E. J. Humphrey, J. P. Bello, "A software framework for musical data augmentation," in *Proc. Int. Soc. Music Inf. Retr.* (2015), doi.org/10.5281/zenodo.1418364.

[30] J. Schlüter, T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks." in *Proc. Int. Soc. Music Inf. Retr.* (2015), doi.org/10.5281/zenodo.1417744.

[31] S. Uhlich, M. Porcu, F. Giron, *et al.*, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2017), doi.org/10.1109/ICASSP.2017.7952158.

[32] I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and Harnessing Adversarial Examples," (2015), doi.org/10.48550/arXiv.1412.6572.

[33] L. Schönherr, S. Zeiler, T. Holz, *et al.*, "Robust Over-the-Air Adversarial Examples Against Automatic Speech Recognition Systems," *CoRR*, vol. abs/1908.01551 (2019), doi.org/10.48550/arXiv.1908.01551.

[34] R. Geirhos, J.-H. Jacobsen, C. Michaelis, *et al.*, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673 (2020), doi.org/10.1038/s42256-020-00257-z.

[35] J. D. Reiss, "A meta-analysis of high resolution audio perceptual evaluation," *J. Audio Eng. Soc.*, vol. 64, no. 6, pp. 364–379 (2016), doi.org/10.17743/jaes.2016.0015.

[36] K. Saito, T. Nakamura, K. Yatabe, *et al.*, "Sampling-Frequency-Independent Convolutional Layer and its Application to Audio Source Separation," *IEEE Trans. Speech Audio Process.*, vol. 30, pp. 2928–2943 (2022), doi.org/10.1109/TASLP.2022.3203907.

[37] J. Paulus, M. Torcoli, "Sampling Frequency Independent Dialogue Separation," in *Proceedings of the European Signal Processing Conference (EUSIPCO)* (2022 Sep.).

[38] J. Yu, Y. Luo, "Efficient Monaural Speech Enhancement with Universal Sample Rate Band-Split RNN," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2023 Jun.), doi.org/10.1109/ICASSP49357.2023.10096020.

[39] A. Carson, A. Wright, J. Chowdhury, *et al.*, "Sample Rate Independent Recurrent Neural Networks for Audio Effects Processing," in *Proc. Int. Conf. Digit. Audio Effects* (2024 Sep.).

[40] J. Su, Y. Wang, A. Finkelstein, *et al.*, "Bandwidth extension is all you need," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 696–700 (2021 Jun.), doi.org/10.1109/ICASSP39728.2021.9413575.

[41] Y. Li, M. Tagliasacchi, O. Rybakov, *et al.*, "Real-time speech frequency bandwidth extension," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2021 Jun.), doi.org/10.1109/ICASSP39728.2021.9413439.

[42] V. Sitzmann, J. Martel, A. Bergman, *et al.*, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473 (2020 Dec.).

[43] N. Zeghidour, A. Luebs, A. Omran, *et al.*, "Soundstream: An end-to-end neural audio codec," *IEEE Trans. Speech Audio Process.*, vol. 30, pp. 495–507 (2021), doi.org/10.1109/TASLP.2021.3129994.

[44] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499* (2016).

[45] Y. Luo, Z. Chen, T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 46–50 (2020 Apr.), doi.org/10.1109/ICASSP40776.2020.9054266.

[46] S. Chen, Y. Wu, C. Wang, *et al.*, "Beats: Audio pre-training with acoustic tokenizers," in *Proc. Int. Conf. Mach. Learn.*, pp. 5178–5193 (2023 Jul.).

[47] W.-T. Lu, J.-C. Wang, Q. Kong, *et al.*, "Music source separation with band-split rope transformer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 481–485 (2024 Apr.), doi.org/10.1109/ICASSP48485.2024.10446843.

[48] Z. Liu, J. Yuan, H. Jin, *et al.*, "KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache," in *Proc. Int. Conf. Mach. Learn.* (2024 Jul.).

[49] T. Dao, D. Fu, S. Ermon, *et al.*, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16344–16359 (2022 Dec.).

[50] H. Liu, M. Zaharia, P. Abbeel, "Ring attention with blockwise transformers for near-infinite context," in *Proc. Int. Conf. Learn. Represent.* (2024 May).

[51] K. Goel, A. Gu, C. Donahue, *et al.*, "It's Raw! Audio Generation with State-Space Models," in *Proc. Int. Conf. Mach. Learn.* (2022 Apr.).

[52] S. Graetzer, J. Barker, T. J. Cox, *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. Int. Speech Commun. Assoc.*, vol. 2, pp. 686–690 (2021 Aug.), doi.org/10.21437/Interspeech.2021-1574.

[53] A. P. McPherson, R. H. Jack, G. Moro, "Action-Sound Latency: Are Our Tools Fast Enough?" in *Proc. Int. Conf. on New Interfaces for Musical Expression* (2016).

[54] T. Kaaresoja, S. Brewster, V. Lantz, "Towards the temporally perfect virtual button: touch-feedback simultaneity and perceived quality in mobile touchscreen press interactions," *ACM Transactions on Applied Perception (TAP)*, vol. 11, no. 2, pp. 1–25 (2014), doi.org/10.1145/2611387.

[55] A. Caillon, P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011* (2021).

[56] Z.-Q. Wang, G. Wichern, S. Watanabe, *et al.*, "STFT-Domain Neural Speech Enhancement With Very Low Algorithmic Latency," *IEEE Trans. Speech Audio Process.*, vol. 31, pp. 397–410 (2023), doi.org/10.1109/TASLP.2022.3224285.

[57] B. Hayes, J. Shier, G. Fazekas, *et al.*, "A review of differentiable digital signal processing for music and speech synthesis," *Frontiers in Signal Processing*, vol. 3, p. 1284100 (2024 Jan.), doi.org/10.3389/frsip.2023.1284100.

[58] J.-M. Valin, S. Tenneti, K. Helwani, *et al.*, "Low-complexity, real-time joint neural echo control and speech enhancement based on PercepNet," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 7133–7137 (2021 Aug.), doi.org/10.1109/ICASSP39728.2021.9414140.

[59] J. Le Roux, S. Wisdom, H. Erdogan, *et al.*, "SDR–half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 626–630 (2019).

[60] K. Zhen, M. S. Lee, J. Sung, *et al.*, "Psychoacoustic calibration of loss functions for efficient end-to-end neural audio coding," *IEEE Signal Processing Letters*, vol. 27, pp. 2159–2163 (2020).

[61] A. Wright, V. Välimäki, "Perceptual loss function for neural modeling of audio systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 251–255 (2020).

[62] C. J. Steinmetz, J. D. Reiss, "auraloss: Audio focused loss functions in PyTorch," in *Digital Music Research Network One-day Workshop (DMRN+15)* (2020).

[63] F. G. Germain, Q. Chen, V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522* (2018).

[64] P. Manocha, A. Finkelstein, R. Zhang, *et al.*, "A differentiable perceptual audio metric learned from just noticeable differences," in *Proc. Int. Speech Commun. Assoc.* (2020).

[65] P. Manocha, Z. Jin, R. Zhang, *et al.*, "CDPAM: Contrastive learning for perceptual audio similarity," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 196–200 (2021).

[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27 (2014).

[67] C. K. Reddy, V. Gopal, R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 886–890 (2022 Feb.), doi.org/10.48550/arXiv.2110.01763.

[68] A. Kumar, K. Tan, Z. Ni, *et al.*, "Torchaudio-Squim: Reference-Less Speech Quality and Intelligibility Measures in Torchaudio," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–5 (2023).

[69] D. M. Ziegler, N. Stiennon, J. Wu, *et al.*, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593* (2019).

[70] M. Torcoli, T. Kastner, J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE Trans. Speech Audio Process.*, vol. 29, pp. 1530–1541 (2021).

[71] A. W. Rix, J. G. Beerends, M. P. Hollier, *et al.*, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 749–752 (2001).

[72] T. Thiede, W. C. Treurniet, R. Bitto, *et al.*, "PEAQ-The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29 (2000).

[73] P. Delgado, J. Herre, "Design Choices in a Binaural Perceptual Model for Improved Objective Spatial Audio Quality Assessment," in *155th Conv. Audio Eng. Soc.* (2023 Oct.).

[74] K. Kilgour, M. Zuluaga, D. Roblek, *et al.*, "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms." in *Proc. Int. Speech Commun. Assoc.*, pp. 2350–2354 (2019).

[75] S. Hershey, S. Chaudhuri, D. P. Ellis, *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 131–135 (2017).

[76] A. Gui, H. Gamper, S. Braun, *et al.*, "Adapting frechet audio distance for generative music evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1331–1335 (2024).

[77] A. Défossez, "Hybrid Spectrogram and Waveform Source Separation," in *Proc. ISMIR 2021 Workshop on Music Source Sep.* (2021 Nov.), doi.org/10.48550/arXiv.2111.03600.

[78] H. F. Garcia, P. O'Reilly, A. Aguilar, *et al.*, "HARP: Bringing Deep Learning to the DAW with Hosted, Asynchronous, Remote Processing," in *NeurIPS Workshop Mach. Learn. Creativity Des.* (2023 Dec.).

[79] J. Chowdhury, "RTNeural: Fast Neural Inferencing for Real-Time Systems," *arXiv preprint arXiv:2106.03037* (2021), doi.org/10.48550/arXiv.2106.03037.

[80] T. Wilmering, D. Moffat, A. Milo, *et al.*, "A history of audio effects," *Applied Sciences*, vol. 10, no. 3, p. 791 (2020), doi.org/10.3390/app10030791.

[81] M. Stein, J. Abeßer, C. Dittmar, *et al.*, "Automatic detection of audio effects in guitar and bass recordings," in *128th Conv. Audio Eng. Soc.* (2010 May).

[82] R. Hinrichs, K. Gerkens, J. Ostermann, "Classification of guitar effects and extraction of their parameter settings from instrument mixes using convolutional neural networks," in *Proc. EvoMUSART*, pp. 101–116 (2022 Apr.), doi.org/10.1007/978-3-031-03789-4_7.

[83] M. Rice, C. J. Steinmetz, G. Fazekas, *et al.*, "General Purpose Audio Effect Removal," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 1–5 (2023 Oct.), doi.org/10.1109/WASPAA58266.2023.10248157.

[84] M. Comunità, D. Stowell, J. D. Reiss, "Guitar effects recognition and parameter estimation with convolutional neural networks," *J. Audio Eng. Soc.*, vol. 69, no. 7/8, pp. 594–604 (2021 Nov.), doi.org/10.17743/jaes.2021.0019.

[85] H. Jürgens, R. Hinrichs, J. Ostermann, "Recognizing guitar effects and their parameter settings," in *Proc. Int. Conf. Digit. Audio Effects* (2020 Sep.).

[86] S. Venkatesh, D. Moffat, E. R. Miranda, "Word embeddings for automatic equalization in audio mixing," *J. Audio Eng. Soc.*, vol. 70, no. 9, pp. 753–763 (2022 Feb.), doi.org/10.17743/jaes.2022.0047.

[87] D. R. K. Balasubramaniam, J. Timoney, "Word based end-to-end real time neural audio effects for equalisation," in *155th Conv. Audio Eng. Soc.* (2023 Oct.).

[88] J.-M. Valin, "A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement," in *Int. Workshop Mult. Sig. Proc.* (2018), doi.org/10.1109/MMSP.2018.8547084.

[89] C. J. Steinmetz, T. Walther, J. D. Reiss, "High-Fidelity Noise Reduction with Differentiable Signal Processing," in *155th Conv. Audio Eng. Soc.* (2023 Oct.), doi.org/10.48550/arXiv.2310.11364.

[90] S. I. Mimilakis, N. J. Bryan, P. Smaragdis, "One-shot parametric audio production style transfer with application to frequency equalization," in *Proc. IEEE Int. Conf. Acoust. Speech*

*Signal Process.*, pp. 256–260 (2020 May), doi.org/10.1109/ICASSP40776.2020.9054108.

[91] D. Sheng, G. Fazekas, "A feature learning siamese model for intelligent control of the dynamic range compressor," in *Proc. Int. Jt. Conf. Neural Netw.*, pp. 1–8 (2019 Jul.), doi.org/10.1109/IJCNN.2019.8851950.

[92] C. J. Steinmetz, V. K. Ithapu, P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 221–225 (2021 Oct.), doi.org/10.1109/WASPAA52581.2021.9632680.

[93] C. J. Steinmetz, N. J. Bryan, J. D. Reiss, "Style transfer of audio effects with differentiable signal processing," *J. Audio Eng. Soc.*, vol. 70, no. 9, pp. 708–721 (2022 Feb.), doi.org/10.17743/jaes.2022.0025.

[94] F. Eichas, S. Möller, U. Zölzer, "Block-oriented gray box modeling of guitar amplifiers," in *Proc. Int. Conf. Digit. Audio Effects*, pp. 5–9 (2017 Sep.).

[95] T. Schmitz, *Nonlinear modeling of the guitar signal chain enabling its real-time emulation*, Ph.D. thesis, ULiège-Université de Liège, Liège, Belgium (2019).

[96] A. Wright, E.-P. Damskägg, V. Välimäki, *et al.*, "Real-time black-box modelling with recurrent neural networks," in *Proc. Int. Conf. Digit. Audio Effects*, pp. 1–8 (2019 Sep.).

[97] M. A. Martínez Ramírez, E. Benetos, J. D. Reiss, "Deep learning for black-box modeling of audio effects," *Applied Sciences*, vol. 10, no. 2, p. 638 (2020 Jan.), doi.org/10.3390/app10020638.

[98] B. Kuznetsov, J. D. Parker, F. Esqueda, "Differentiable IIR filters for machine learning applications," in *Proc. Int. Conf. Digit. Audio Effects* (2020 Sep.).

[99] A. Wright, V. Välimäki, others, "Grey-box modelling of dynamic range compression," in *Proc. Int. Conf. Digit. Audio Effects*, pp. 304–311 (2022 Sep.).

[100] J. Imort, G. Fabbro, M. A. M. Ramírez, *et al.*, "Distortion Audio Effects: Learning How to Recover the Clean Signal," in *Proc. Int. Soc. Music Inf. Retr.* (2022 Dec.).

[101] E. Moliner, J. Lehtinen, V. Välimäki, "Solving audio inverse problems with a diffusion model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–5 (2023 Jun.), doi.org/10.1109/ICASSP49357.2023.10095637.

[102] C.-B. Jeon, K. Lee, "Music De-Limiter Networks Via Sample-Wise Gain Inversion," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 1–5 (2023 Oct.), doi.org/10.1109/WASPAA58266.2023.10248055.

[103] C. Hernandez-Olivan, K. Saito, N. Murata, *et al.*, "VRDMG: Vocal restoration via diffusion posterior sampling with multiple guidance," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2024 Apr.), doi.org/10.1109/ICASSP48485.2024.10446423.

[104] C. J. Steinmetz, J. D. Reiss, "Randomized overdrive neural networks," *NeurIPS Workshop Mach. Learn. Creativity Des.* (2020 Dec.), doi.org/10.48550/arXiv.2010.04237.

[105] C. J. Steinmetz, J. D. Reiss, "Steerable discovery of neural audio effects," *NeurIPS Workshop Mach. Learn. Creativity Des.* (2021 Dec.), doi.org/10.48550/arXiv.2112.02926.

[106] J. T. Colonel, J. Reiss, "Reverse engineering of a recording mix with differentiable digital signal processing," *Journal of the Acoustical Society of America*, vol. 150, no. 1, pp. 608–619 (2021).

[107] C. J. Steinmetz, J. Pons, S. Pascual, *et al.*, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 71–75 (2021).

[108] C. Peladeau, G. Peeters, "Blind estimation of audio effects using an auto-encoder approach and differentiable signal processing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2024).

[109] S. H. Hawley, C. J. Steinmetz, "Leveraging Neural Representations for Audio Manipulation," in *154th Conv. Audio Eng. Soc.* (2023 May), doi.org/10.48550/arXiv.2304.04394.

[110] J. Turian, J. Shier, H. R. Khan, *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS Competitions and Demonstrations Track*, pp. 125–145 (2021).

[111] J. Koo, M. A. Martínez-Ramírez, W.-H. Liao, *et al.*, "Music Mixing Style Transfer: A Contrastive Learning Approach to Disentangle Audio Effects," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–5 (2023 Jun.), doi.org/10.1109/ICASSP49357.2023.10096458.

[112] M. A. M. Ramírez, O. Wang, P. Smaragdis, *et al.*, "Differentiable signal processing with black-box audio effects," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 66–70 (2021 Jun.), doi.org/10.1109/ICASSP39728.2021.9415103.

[113] A. B. Horner, *Spectral Matching of Musical Instrument Tones*, Thesis, University of Illinois at Urbana-Champaign (1993), URL `hdl.handle.net/2142/72086`.

[114] E. R. Miranda, "An Artificial Intelligence Approach to Sound Design," *Computer Music Journal*, vol. 19, no. 2, pp. 59–75 (1995), doi.org/10.2307/3680600.

[115] M. J. Yee-King, L. Fedden, M. d'Inverno, "Automatic Programming of VST Sound Synthesizers Using Deep Networks and Other Techniques," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 150–159 (2018 Apr.), doi.org/10.1109/TETCI.2017.2783885.

[116] C. Mitcheltree, H. Koike, "SerumRNN: Step by Step Audio VST Effect Programming," in *Proc. EvoMUSART*, pp. 218–234 (2021 Apr.), doi.org/10.1007/978-3-030-72914-1_15.

[117] Z. Chen, Y. Jing, S. Yuan, *et al.*, "Sound2Synth: Interpreting Sound via FM Synthesizer Parameters Estimation," in *Proc. Int. Jt. Conf. Artif. Intell.*, vol. 6, pp. 4921–4928 (2022 Jul.), doi.org/10.24963/ijcai.2022/682.

[118] N. Masuda, D. Saito, "Improving Semi-Supervised Differentiable Synthesizer Sound Matching for Practical Applications," *IEEE Trans. Speech Audio Process.*, vol. 31, pp. 863–875 (2023 Jan.), doi.org/10.1109/TASLP.2023.3237161.

[119] J. Shier, "The synthesizer programming problem: improving the usability of sound synthesizers," (2021 Dec.), available at dspace.library.uvic.ca/handle/1828/13593.

[120] S. Heise, M. Hlatky, J. Loviscach, "Automatic Cloning of Recorded Sounds by Software Synthesizers," in *127th Conv. Audio Eng. Soc.* (2009 Oct.).

[121] J. Turian, J. Shier, G. Tzanetakis, *et al.*, "One Billion Audio Sounds from GPU-Enabled Modular Synthesis," in *Proc. Int. Conf. Digit. Audio Effects*, pp. 222–229 (2021 Sep.), doi.org/10.23919/DAFx51585.2021.9768246.

[122] H. Han, V. Lostanlen, M. Lagrange, "Perceptual-Neural-Physical Sound Matching," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2023 Jun.), doi.org/10.1109/ICASSP49357.2023.10095391.

[123] C.-Y. Yu, G. Fazekas, "Singing Voice Synthesis Using Differentiable LPC and Glottal-Flow-Inspired Wavetables," in *Proc. Int. Soc. Music Inf. Retr.* (2023 Nov.), doi.org/10.5281/zenodo.10265376.

[124] R. Diaz, B. Hayes, C. Saitis, *et al.*, "Rigid-Body Sound Synthesis with Differentiable Modal Resonators," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2023 Jun.), doi.org/10.1109/ICASSP49357.2023.10095139.

[125] Y. Yang, Z. Jin, C. Barnes, *et al.*, "White Box Search over Audio Synthesizer Parameters," in *Proc. Int. Soc. Music Inf. Retr.* (2023 Nov.), doi.org/10.5281/zenodo.10364630.

[126] P. Esling, N. Masuda, A. Bardet, *et al.*, "Flow Synthesizer: Universal Audio Synthesizer Control with Normalizing Flows," *Applied Sciences*, vol. 10, no. 1 (2020 Dec.), doi.org/10.3390/app10010302.

[127] J. Engel, C. Resnick, A. Roberts, *et al.*, "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," in *Proc. Int. Conf. Learn. Represent.* (2017 Apr.).

[128] A. Caillon, *Hierarchical temporal learning for multi-instrument and orchestral audio synthesis*, Thesis, Sorbonne Université (2023), URL theses.hal.science/tel-04137258.

[129] A. Case, *Mix smart: Professional techniques for the home studio* (Taylor & Francis) (2012).

[130] M. Senior, *Mixing secrets for the small studio* (Taylor & Francis) (2011).

[131] B. De Man, R. Stables, J. D. Reiss, *Intelligent Music Production* (Routledge) (2019).

[132] B. De Man, J. Reiss, R. Stables, "Ten years of automatic mixing," in *Proceedings of the Workshop on Intelligent Music Production* (2017).

[133] E. R. Miranda, *Handbook of artificial intelligence for music* (Springer) (2021).

[134] J. D. Reiss, "Intelligent systems for mixing multichannel audio," in *Proceedings of the 17th International Conference on Digital Signal Processing (DSP)*, pp. 1–6 (2011).

[135] M. N. Lefford, G. Bromham, D. Moffat, "Mixing with intelligent mixing systems: Evolving practices and lessons from computer assisted design," in *148th Conv. Audio Eng. Soc.* (2020).

[136] D. Moffat, "AI Music Mixing Systems," *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, pp. 345–375 (2021).

[137] M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, *et al.*, "Automatic music mixing with deep learning and out-of-domain data," in *Proc. Int. Soc. Music Inf. Retr.* (2022).

[138] J. Colonel, C. Steinmetz, "Deep learning approaches to automatic mixing," (2022 Apr.), URL csteinmetz1.github.io/AutomaticMixingPapers/resources.html.

[139] A. Roginska, P. Geluso, *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio* (Focal Press) (2017 Oct.), doi.org/10.4324/9781315707525.

[140] J.-M. Jot, Z. Fejzo, "Beyond Surround Sound - Creation, Coding and Reproduction of 3-D Audio Soundtracks," in *131st Conv. Audio Eng. Soc.* (2011 Oct.).

[141] C. Q. Robinson, S. Mehta, N. Tsingos, "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," *SMPTE Motion Imaging Journal*, vol. 121, no. 8, pp. 63–69 (2012 Nov.), doi.org/10.5594/j18248XY.

[142] J. Herre, J. Hilpert, A. Kuntz, *et al.*, "MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779 (2015 Mar.), doi.org/10.1109/jstsp.2015.2411578.

[143] K. Kowalczyk, O. Thiergart, M. Taseska, *et al.*, "Parametric Spatial Sound Processing: A Flexible and Efficient Solution to Sound Scene Acquisition, Modification, and Reproduction," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 31–42 (2015 Feb.), doi.org/10.1109/msp.2014.2369531.

[144] A. Politis, S. Tervo, V. Pulkki, "Compass: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 6802–6806 (2018 Sep.), doi.org/10.1109/icassp.2018.8462608.

[145] C. Avendano, J.-M. Jot, "A Frequency-Domain Approach to Multichannel Upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740–749 (2004 Jul./Aug.).

[146] C. Faller, "Multiple-Loudspeaker Playback of Stereo Signals," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064 (2006 Nov.).

[147] M. M. Goodwin, J.-M. Jot, "Primary-Ambient Signal Decomposition and Vector-Based Localization for Spatial Audio Coding and Enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. I–9 (2007 Apr.), doi.org/10.1109/icassp.2007.366603.

[148] J. He, W.-S. Gan, E.-L. Tan, "Primary-Ambient Extraction Using Ambient Spectrum Estimation for Immersive Spatial Audio Reproduction," *IEEE Trans. Speech Audio Process.*, vol. 23, no. 9, pp. 1431–1444 (2015 May), doi.org/10.1109/TASLP.2015.2434272.

[149] C. Uhle, E. A. Habets, "Direct-Ambient Decomposition Using Parametric Wiener Filtering With Spatial Cue Control," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (2015 Aug.), doi.org/10.1109/ICASSP.2015.7177927.

[150] K. M. Ibrahim, M. Allam, "Primary-Ambient Source Separation for Upmixing to Surround Sound Systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 431–435 (2018 Apr.), doi.org/10.1109/ICASSP.2018.8461459.

[151] M. R. Bai, G.-Y. Shih, "Upmixing and Downmixing Two-channel Stereo Audio for Consumer Electronics," *IEEE Transactions on Consumer Electronics*, vol. 53 (2007 Oct.), doi.org/10.1109/TCE.2007.4341580.

[152] M. Goodwin, J.-M. Jot, "Spatial Audio Scene Coding," in *125th Conv. Audio Eng. Soc.* (2008 Oct.).

[153] E. Vickers, "Frequency-Domain Two- To Three-Channel Upmix for Center Channel Derivation and Speech Enhancement," in *127th Conv. Audio Eng. Soc.* (2009 Oct.).

[154] S. Uhlich, F. Giron, Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 2135–2139 (2015 Apr.), doi.org/10.1109/ICASSP.2015.7178348.

[155] F.-R. Stöter, A. Liutkus, N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *Latent Variable Analysis and Signal Separation*, pp. 293–305 (2018 Jun.), doi.org/10.1007/978-3-319-93764-9_28.

[156] C. Uhle, C. Paul, "A supervised learning approach to ambience extraction from mono recordings for blind upmixing," in *Proc. Int. Conf. Digit. Audio Effects* (2008 Oct.).

[157] J. Choi, J.-H. Chang, "Exploiting deep neural networks for two-to-five channel surround decoder," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 938–949 (2021 Jan.).

[158] R. T. Paez-Amaro, C. Tejeda Ocampo, E. Souza-Blanes, *et al.*, "Deep Learning Based Voice Extraction and Primary-Ambience Decomposition for Stereo to Surround Upmixing," in *154th Conv. Audio Eng. Soc.* (2023 May).

[159] S. Y. Park, C. J. Chun, H. K. Kim, "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks," in *Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 377–380 (2016 Oct.), doi.org/10.1109/ICTC.2016.7763500.

[160] H. Yang, S. Wager, S. Russell, *et al.*, "Upmixing Via Style Transfer: A Variational Autoencoder for Disentangling Spatial Images And Musical Content," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 426–430 (2022 May), doi.org/10.1109/ICASSP43922.2022.9746978.

[161] Y. Luo, "Active Barycentric Beamformed Stereo Upmixing," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 316–320 (2023 Sept.), doi.org/10.23919/EUSIPCO58844.2023.10289876.

[162] K. M. Jeon, S. Y. Park, C. J. Chun, *et al.*, "Multi-band Approach to Deep Learning-Based Artificial Stereo Extension," *ETRI Journal*, vol. 39, no. 3, pp. 398–405 (2017 Jun.), doi.org/10.4218/etrij.17.0116.0773.

[163] J. Serrà, D. Scaini, S. Pascual, *et al.*, "Mono-to-Stereo Through Parametric Stereo Generation," in *Proc. Int. Soc. Music Inf. Retr.* (2023 Aug.).

[164] M. Cobos, J. Ahrens, K. Kowalczyk, *et al.*, "An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–21 (2022 May), doi.org/10.1186/s13636-022-00242-x.

[165] B. Rafaely, V. Tourbabin, E. Habets, *et al.*, "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges," *Acta Acustica*, vol. 6 (2022 Oct.), doi.org/10.1051/aacus/2022040.

[166] S. Li, S. Liu, D. Manocha, "Binaural Audio Generation via Multi-task Learning," *ACM Transactions on Graphics*, vol. 40, no. 6, pp. 1–13 (2021 Dec.), doi.org/10.1145/3478513.3480560.

[167] R. Garg, R. Gao, K. Grauman, "Visually-Guided Audio Spatialization in Video with Geometry-Aware Multi-task Learning," *International Journal of Computer Vision*, vol. 131, pp. 2723–2737 (2023 Jun.), doi.org/10.1007/s11263-023-01816-8.

[168] A. Biswas, D. Jia, "Audio Codec Enhancement with Generative Adversarial Networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 356–360 (2020 May), doi.org/10.1109/ICASSP40776.2020.9053113.

[169] Y. Wu, R. Hu, X. Wang, *et al.*, "Distortion reduction via CAE and DenseNet mixture network for low bitrate spatial audio object coding," *IEEE MultiMedia*, vol. 29, no. 1, pp. 55–64 (2022 Jan.), doi.org/10.1109/MMUL.2022.3142752.

[170] R. Balestriero, R. Cosentino, H. Glotin, *et al.*, "Spline filters for end-to-end deep learning," in *Proceedings of the International Conference on Machine Learning*, pp. 364–373 (2018 Jul.).

[171] S. Braun, H. Gamper, C. K. Reddy, *et al.*, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 656–660 (2021 May), doi.org/10.48550/arXiv.2101.09249.

[172] F. Dang, H. Chen, P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *ICASSP*, pp. 6857–6861 (2022 Jan.), doi.org/10.48550/arXiv.2104.13002.

[173] B. J. Borgström, M. S. Brandstein, "Speech enhancement via attention masking network (SEAM-

NET): An end-to-end system for joint suppression of noise and reverberation," *IEEE Trans. Speech Audio Process.*, vol. 29, pp. 515–526 (2020 Dec.), doi.org/10.1109/TASLP.2020.3043655.

[174] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis," *Artificial Intelligence Review*, pp. 1–53 (2023 Jun.), doi.org/10.48550/arXiv.2212.00369.

[175] J. K. McElveen, N. Schiffman, S. Nordlund, *et al.*, "A Proposed Signal Processing Model Of Human Spatial Hearing Using Interaural Cross Correlation and Auditory Glimpsing To Estimate Green's Functions," in *Int. Congress on Acoustics* (2022 Oct.).

[176] J. S. Garofolo, L. F. Lamel, W. M. Fisher, *et al.*, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403 (1993 Feb.), doi.org/10.1109/LSP.2019.2953810.

[177] V. Panayotov, G. Chen, D. Povey, *et al.*, "Librispeech: an asr corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5206–5210 (2015 Apr.), doi.org/10.1109/ICASSP.2015.7178964.

[178] C. Veaux, J. Yamagishi, S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Int. Conf. Oriental COCOSDA held jointly with 2013 Conf. on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pp. 1–4 (2013 Nov.), doi.org/10.1109/ICSDA.2013.6709856.

[179] H. Christensen, J. Barker, N. Ma, *et al.*, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. of Annual Conf. of the Int. Speech Communication Assoc.* (2010 Sep.), doi.org/10.21437/Interspeech.2010-552.

[180] J. G. M. FitzGerald, C. Hench, C. Peris, *et al.*, "MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages," in *An. Meeting of the Assoc. for Computational Linguistics* (2023 Jun.), doi.org/10.48550/arXiv.2204.08582.

[181] P. Guiraud, S. Hafezi, P. A. Naylor, *et al.*, "An introduction to the speech enhancement for augmented reality (SPEAR) challenge," in *2022 Int. Workshop on Acoustic Sig. Enh. (IWAENC)*, pp. 1–5 (2022 Sep.), doi.org/10.1109/IWAENC53105.2022.9914721.

[182] C. K. Reddy, H. Dubey, K. Koishida, *et al.*, "Interspeech 2021 deep noise suppression challenge," in *Interspeech* (2021 Apr.), doi.org/10.48550/arXiv.2101.01902.

[183] S.-F. Huang, C.-J. Lin, D.-R. Liu, *et al.*, "Meta-tts: Meta-learning for few-shot speaker adaptive text-to-speech," *TSAP*, vol. 30, pp. 1558–1571 (2022 Apr.), doi.org/10.1109/TASLP.2022.3167258.

[184] J. Casebeer, N. J. Bryan, P. Smaragdis, "Meta-AF: Meta-learning for adaptive filters," *IEEE Trans. Speech Audio Process.*, vol. 31, pp. 355–370 (2022 Nov.), doi.org/10.1109/TASLP.2022.3224288.

[185] C. Uhle, O. Hellmuth, J. Weigel, "Speech Enhancement of Movie Sound," presented at *125th Conv. Audio Eng. Soc.* (2008).

[186] Y. Luo, N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE Trans. Speech Audio Process.*, vol. 27, no. 8, pp. 1256–1266 (2019), doi.org/10.1109/TASLP.2019.2915167.

[187] Y. Wang, A. Narayanan, D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE Trans. Speech Audio Process.*, vol. 22, pp. 1849–1858 (2014), doi.org/10.1109/TASLP.2014.2352935.

[188] K. Tan, D. Wang, "Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement," *IEEE Trans. Speech Audio Process.*, vol. 28, pp. 380–390 (2020), doi.org/10.1109/TASLP.2019.2955276.

[189] Q. Kong, Y. Cao, H. Liu, *et al.*, "Decoupling Magnitude and Phase Estimation with Deep ResUNet for Music Source Separation," in *Proc. Int. Soc. Music Inf. Retr.* (2021), doi.org/10.48550/arXiv.2109.05418.

[190] S.-W. Fu, T.-W. Wang, Y. Tsao, *et al.*, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE Trans. Speech Audio Process.*, vol. 26, no. 9, pp. 1570–1584 (2018), doi.org/10.1109/TASLP.2018.2821903.

[191] Y. Koizumi, K. Niwa, Y. Hioka, *et al.*, "DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 81–85 (2017), doi.org/10.1109/ICASSP.2017.7952122.

[192] C. Uhle, M. Torcoli, J. Paulus, "Controlling the Perceived Sound Quality for Dialogue Enhancement With Deep Learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 51–55 (2020), doi.org/10.1109/ICASSP40776.2020.9053789.

[193] J. Cosentino, M. Pariente, S. Cornell, *et al.*, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," in *Proc. Int. Speech Commun. Assoc.* (2020), doi.org/10.48550/arXiv.2005.11262.

[194] V. Panayotov, G. Chen, D. Povey, *et al.*, "Librispeech: An ASR corpus based on public domain audio books," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5206–5210 (2015), doi.org/10.1109/ICASSP.2015.7178964.

[195] G. Wichern, J. M. Antognini, M. Flynn, *et al.*, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Int. Speech Commun. Assoc.* (2019), 10.21437/interspeech.2019-2821.

[196] Q. Kong, Y. Wang, X. Song, *et al.*, "Source Separation with Weakly Labelled Data: an Approach to Computational Auditory Scene Analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 101–105 (2020), 10.1109/ICASSP40776.2020.9053396.

[197] J.-P. Briot, G. Hadjeres, F.-D. Pachet, "Deep learning techniques for music generation–a survey," *arXiv preprint arXiv:1709.01620* (2017).

[198] N. Kalchbrenner, E. Elsen, K. Simonyan, *et al.*, "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, pp. 2410–2419 (2018).

[199] R. Rombach, A. Blattmann, D. Lorenz, *et al.*, "High-resolution image synthesis with latent diffusion models. 2022 IEEE," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685 (2021).

[200] A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, pp. 8821–8831 (2021).

[201] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901 (2020).

[202] Y. Zhu, J. Baca, B. Rekabdar, *et al.*, "A Survey of AI Music Generation Tools and Models," *arXiv preprint arXiv:2308.12982* (2023).

[203] J. Copet, F. Kreuk, I. Gat, *et al.*, "Simple and Controllable Music Generation," in *Proceedings of the Conference on Neural Information Processing Systems* (2023).

[204] A. Agostinelli, T. I. Denk, Z. Borsos, *et al.*, "MusicLM: Generating music from text," *arXiv preprint arXiv:2301.11325* (2023).

[205] A. Défossez, J. Copet, G. Synnaeve, *et al.*, "High fidelity neural audio compression," *Transactions on Machine Learning Research* (2022).

[206] C. Wang, S. Chen, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111* (2023).

[207] J. Ho, A. Jain, P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851 (2020).

[208] F. Schneider, Z. Jin, B. Schölkopf, "Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion," *arXiv preprint arXiv:2301.11757* (2023).

[209] Z. Evans, C. Carr, J. Taylor, *et al.*, "Fast Timing-Conditioned Latent Audio Diffusion," *arXiv preprint arXiv:2402.04825* (2024).

[210] K. Chen, Y. Wu, H. Liu, *et al.*, "MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies," *CoRR*, vol. abs/2308.01546 (2023).

[211] S.-L. Wu, C. Donahue, S. Watanabe, *et al.*, "Music ControlNet: Multiple time-varying controls for music generation," *arXiv preprint arXiv:2311.07069* (2023).

[212] J. Melechovsky, Z. Guo, D. Ghosal, *et al.*, "Mustango: Toward controllable text-to-music generation," *arXiv preprint arXiv:2311.08355* (2023).

[213] H. Wu, X. Chen, Y.-C. Lin, *et al.*, "Towards audio language modeling-an overview," *arXiv preprint arXiv:2402.13236* (2024).

## THE AUTHORS


Christian J. Steinmetz


Christian Uhle


Flavio Everardo


Christopher Mitcheltree


J Keith McElveen


Jean-Marc Jot


Gordon Wichern

Christian J. Steinmetz is a PhD researcher with the Centre for Digital Music at Queen Mary University of London and his research focuses on applications of machine learning for audio signal processing with a focus on high fidelity audio and music production.

●

Christian Uhle is a chief scientist in the Audio and Media Technologies division of Fraunhofer Institute for Integrated Circuits IIS in Erlangen, Germany, and co-founder of the AES Technical Committee on Machine Learning and Artificial Intelligence (TC-MLAI).

●

Flavio Everardo is a full-time professor in Music Technology at Tecnológico de Monterrey, Puebla Campus, researcher in knowledge representation and automated reasoning within music production, AI-powered audio plugins developer, producer, mixing, and mastering engineer.

●

Christopher Mitcheltree is a PhD researcher at the Centre for Digital Music, Queen Mary University of London researching deep learning for time-varying audio systems and is also a founding member of Neutone.

●

J Keith McElveen is the founder and CEO of Wave Sciences, an audio research company and member of the AES Technical Committee on Machine Learning and Artificial Intelligence (TC-MLAI) and AES Technical and Standards Committees on Audio Forensics.

●

Jean-Marc Jot is founder and principal of Virtuel Works LLC in Aptos, California, a fellow of the AES, and co-chair of the AES Technical Committee on Spatial Audio (TC-SA).

●

Gordon Wichern is a Senior Principal Research Scientist on the Speech and Audio Group at Mitsubishi Electric Research Laboratories (MERL) and co-founder of the AES Technical Committee on Machine Learning and Artificial Intelligence (TC-MLAI).