HOCHSCHULE DER MEDIEN

Aktuelle Themen, AM3
Wintersemester 2022/2023
Dozent: Prof. Dr. Andreas Koch
Verfasser: Benjamin Pohler (45506)

„Style Transfer of Audio Effects with Differentiable Sound Processing“

Dieses Whitepaper gibt einen Einblick in die aktuelle Arbeit von Christian Steinmetz. Steinmetz beschäftigt sich mit der Thematik der „Stilübertragung von Audioeffekten“. Er entwickelt aktuell eine künstliche Intelligenz, die ein Signal (Referenzsignal) analysiert und den Produktions-Stil dieser Aufnahme auf ein anderes Eingangssignal (z.B. eine eigene Aufnahme) aufprägt. Soll also eine selbst angefertigte Aufnahme so klingen wie ein *Led Zeppelin* Song, so muss die künstliche Intelligenz beide Aufnahmen analysieren. Nach der Analyse fügt sie dem Eingangssignal so lange Audio-Effekte hinzu, bis diese stilistisch ähnlich zum Referenzsignal ist. Diese Methode, die zunächst wie Zauberei klingt, wird in dem Paper „*Style Transfer of Audio Effects with Differentiable Sound Processing*“ vorgestellt (veröffentlicht im September 2022 im Journal der Audio Engineering Society, kurz AES, u.a. mit Joshua Reiss vom *Centre for Digital Music,* Queen Mary University of London).

Die Steuerparameter, die von einem neuronalen Netzwerk vorhergesagt werden, werden anschließend von der künstlichen Intelligenz selbstständig in das zu bearbeitende Signal eingefügt. Nutzer*innen können sich darüber hinaus diese Parameter (beispielsweise in einem Programm mit grafischer Oberfläche) anzeigen lassen und diese dann selbstständig weiter bearbeiten bzw. modifizieren. Das Besondere ist also, dass die Vorhersagen nicht in mathematischen Formeln und Berechnungen verschwinden – wie es in anderen intelligenten Musikproduktionstools oftmals vorzufinden ist – sondern für Nutzer*innen interpretierbar bleiben. So wird nicht nur eine automatische Optimierung/Angleichung eines Produktionsstils von Audioaufnahmen ermöglicht, es ermöglicht auch einen Einblick in die „Black-Box“. Interpretierbarkeit fördert die Interaktion gerade für weniger erfahrene Nutzer*innen. Der von Steinmetz vorgeschlagene Ansatz wurde anhand von Sprach- und Musikaufnahmen evaluiert und hat bisher überzeugende Ergebnisse geliefert. Dieses Whitepaper soll die grundlegende Funktionsweise umreißen und kurz die Möglichkeiten und „Gefahren“ dieser Methode diskutieren.

0. Einführung

Die Hauptmotivation der Arbeit liegt laut Steinmetz in der Tatsache begründet, dass in der heutigen Zeit mehr Menschen als je zuvor (Audio-)Inhalte in den verschiedensten Bereichen produzieren. Ein Blick in aktuelle Statistiken bestätigt dies[1]. Von Inhalten für YouTube, TikTok und Instagram bis hin zu Kurzfilmen ist alles vertreten. Zusätzlich zu diesen Formaten gibt es vermehrt Menschen, die ihre eigene Musik oder Podcasts zu Hause aufnehmen und veröffentlichen. Verstärkt wurde dieser Effekt auch durch die Zeit der Corona-Pandemie[2].

Durch erschwingliche Technik stellt eine Umsetzung in hoher Qualität grundlegend keine große Hürde mehr dar, auch für Nichtprofis. Dies war zum Beispiel vor einigen Jahren noch nicht der Fall[3]. Damals waren teurere Technik und meist hochpreisige Signalverarbeitungs-Hardware sowie ausgebildete Techniker*innen notwendig, um hochqualitative Inhalte zu produzieren. Heute sind zumindest erstere Faktoren ein kleineres Problem – das Wissen um die Nachbearbeitung des Materials ist jedoch nicht trivial und nicht leicht zu ersetzen.

In damaliger Zeit wurden für das Erstellen von Aufnahmen Tonstudios mit hochpreisigem Equipment und Mitarbeitenden akquiriert, die mit dieser Technik auch umgehen konnten. Durch das digitale Zeitalter ist einiges von dem damals eingesetzten Equipment, vor allem Audioeffekte (im folgenden *Plugins* genannt), günstig als Software Lösung zu erhalten – oftmals sogar auch kostenfrei. Somit *könnte* jede*r Medienschaffende, die*der über einen halbwegs leistungsstarken Computer verfügt, hochwertige Medienprodukte erstellen. Die Eingangshürde ist also kleiner als jemals zuvor.

Das Problem, was sich nun herausstellt, ist, dass das Erstellen von Qualitätsprodukten nicht bei der Aufnahmephase endet, sondern bei der technischen Weiterverarbeitung. Wurden damals Studios gebucht, so gab es in den Studios Mitarbeitende, die wussten, wie die Technik zu bedienen ist. Die Komplexität der digitalen Signalverarbeitung und ihrer Handhabung muss mindestens beherrscht und im besten Falle durchdrungen worden sein, um diese richtig anzuwenden. Verschiedene Tools haben verschiedene Betriebs- und Arbeitsweisen, die erlernt werden müssen. Ein gewisser Grad an Fachwissen ist entsprechend Voraussetzung, selbst wenn die Tools und das Equipment an sich immer zugänglicher werden. Hier setzt Steinmetz mit seinem neuronalen Netzwerk an.

1. Hintergrund der Arbeit

Bei den hier beschriebenen „Audioeffekten/Plugins" handelt es sich um leistungs-

---

[1] Tim Fischer (2021): *8 beeindruckende Statistiken zur Creator Economy,* https://influencevision.com/blog/8-beeindruckende-statistiken-zur-creator-economy.

[2] Vgl. Sarah Oberteils (2021): *Der späte Boom der Podcasts,* https://www.faz.net/aktuell/wirtschaft/schneller-schlau/podcast-boom-die-corona-pandemie-hat-beim-erfolg-geholfen-17675766.html.

[3] Vgl. Jochen Wieloch (2016): *So richten Sie sich ein Musikstudio zu Hause ein,* https://www.welt.de/wirtschaft/webwelt/article153221951/So-richten-Sie-sich-ein-Musikstudio-zu-Hause-ein.html.

starke Werkzeuge für die Bearbeitung von Audioinhalten. Damit ist zum Beispiel die Entzerrung eines Signals mittels *Equalizer* gemeint oder das Einschränken der Dynamik mittels *Kompression*. Es gibt noch zahlreiche weitere Effekte, etwa zeitbasierte Effekte wie *Delay*, die für die folgenden Betrachtungen jedoch zunächst außer Acht gelassen werden sollen, da diese noch nicht vollständig in Steinmetz' Arbeit implementiert wurden.

Ein seit einigen Jahren weit verbreiteter Lösungsansatz die Arbeit mit Plugins zu erleichtern, ist das Erstellen von Voreinstellungen (sogenannten Presets). Diese werden oft von den Herstellerfirmen der Audioeffekte angeboten. Presets können die Verwendung solcher Audioeffekte zwar vereinfachen, aber ersetzen nicht die Fähigkeit, diese Signalverarbeitungstools auch zu bedienen. Zudem greifen einige Herstellerfirmen auf gewisse Tricks in Presets zurück, um ihre Produkte an unerfahrene Nutzer*innen zu verkaufen (beispielsweise kann das Signal mit dem Plugin-Preset lauter werden, aber nicht besser – lauter wird jedoch psychoakustisch oftmals als besser wahrgenommen[4]).

1.1 Anwendungsbeispiel und Idee

Eine Nutzerin oder ein Nutzer macht eine Aufnahme mit dem Smartphone. Diese Aufnahme soll für einen Podcast genutzt werden. Die Qualität der Smartphoneaufnahme ist für Podcastzwecke unzureichend. Um ein breiteres Publikum zu erreichen, sollte also die Audio-Qualität verbessert werden[5]. Die Nutzerin oder der Nutzer haben jedoch kein oder zu wenig Fachwissen von den Prozessen, die es braucht, die Aufnahme qualitativ so anzugleichen, dass sie mit den Podcasts, die schon auf diversen Plattformen sind, konkurrieren kann. Die Idee wäre es nun, mit einer Aufnahme des Lieblingspodcasts der Nutzerin oder des Nutzers als Referenz die eigene Aufnahme so zu bearbeiten, dass der Produktionsstil ähnlich ist. Das Eingangssignal der Nutzer*innen sollte also so klingen, als wäre es unter ähnlichen Umständen wie die des Lieblingspodcasts erstellt worden. Das meint Signalverarbeitungsprozesse wie zum Beispiel die Angleichung von Lautstärkeunterschieden (Dynamikeinschränkung, Podcasts sind oftmals sehr stark komprimiert) oder die Entzerrung von Störgeräuschen (Brummen von Klimaanlage, Entfernen von Zischlauten o.ä.).

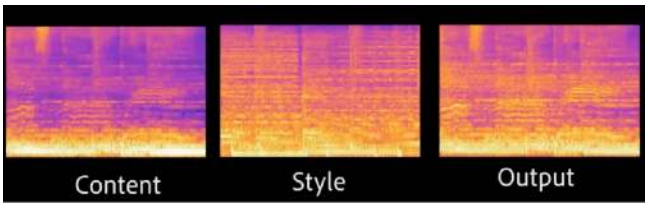

Abbildung 1: Spektral-Darstellung von Eingangssignal (Content), Referenz (Style) und resultierendes, mit Plugins bearbeitetes, Signal (Output).

Hierfür kann also ein spektrales Profil von beiden Aufnahmen (Original Aufnahme

[4] Vgl. Marko Pauli (2008): *Einfach nur laut ist nicht gleich gut,* https://www.deutschlandfunkkultur.de/einfach-nur-laut-ist-nicht-gleich-gut-102.html.
[5] Podcastle Artikel (2019): *Maximizing Audio Quality: Tips and Strategies for Podcast Creators,* https://podcastle.ai/blog/maximizing-audio-quality/.

und Referenzaufnahme) erstellt und verglichen werden. Anhand der spektralen Unterschiede müssen nun digitale Effekte so eingestellt werden, dass der Klang der Original Aufnahme in dieselbe Richtung geht wie die Referenzaufnahme (Abbildung 1 zeigt eine solche spektrale Analyse). Optimalerweise ist der generelle Klang dann so, als ob es sich um eine Aufnahme mit der selben Signalkette wie die Referenzaufnahme handeln würde.
Das System soll zudem den Nutzer*innen anzeigen, *was* eingestellt wurde. So haben sie die Möglichkeit, an den Einstellungen, die von der KI vorgenommen wurden, anhand einer Grafischen Oberfläche (GUI) noch selbstständige Veränderungen vorzunehmen (siehe Abbildung 2).

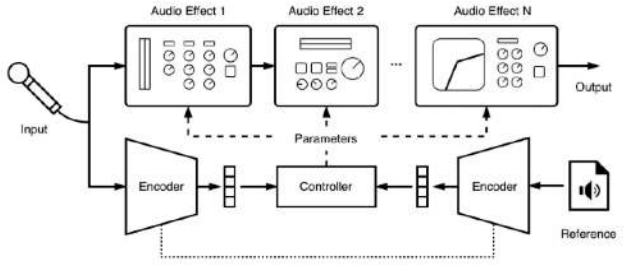

Abbildung 2: Vorhersagen aus dem Encoder füttern die Audioeffekte, die auf das Signal angewandt werden.

Es handelt sich also um eine „automatisierte Postproduktion" für Medienschaffende, welche sich dann (durch die Automatisierung) voll und ganz auf ihre Inhalte konzentrieren können.

1.2 Prozess einer Audio Produktion

Die Gestaltung des System ist, laut Autor des Papers, an den Prozess einer Audioproduktion angelehnt. Dieser sagt, dass sich dieser Audioproduktionsprozess in drei Hauptphasen aufteilt. Zunächst nennt er die Phase des *kritischen Hörens.* Es wird also eine akustische Analyse durchgeführt. In der zweiten Phase gilt es, einen *Plan* zu erstellen, wie die Aufnahme bearbeitet werden soll, um sie in einen bestimmten Stil zu bringen. Als dritte und letzte Phase ist die *Durchführung ge*nannt. Es werden die geplanten Signalverarbeitungsprozesse angewandt, um das klangliche Ziel aus Phase zwei zu erreichen. Dies ist in einer professionellen Audioproduktion bzw. Audiomischung üblicherweise so aufgeteilt[6].

Diese drei Phasen sollen nun aber von einer KI übernommen werden. Besonders bei Phase zwei haben Nutzer*innen, die wenig erfahren sind und denen es an Fachwissen mangelt, zusätzlich das Problem, dass sie oftmals auch keine Vorstellung davon haben, wie ihr Audio am Ende klingen soll. Eine solche klangliche Vorstellung ist keineswegs trivial, entwickelt sich häufig jedoch erst mit viel Erfahrung und Fachwissen in diesem Bereich.[7]

---

[6] Vgl. hierzu auch den iZotope Artikel (o. J.): *What Is Mixing in Music?,* https://www.izotope.com/en/learn/how-to-mix-music.html.
[7] Vgl. Elisabeth Kemper (2006): *Realisierbarkeit und Beurteilung ästhetischer Klangkonzepte, Hochschule für Musik Detmold,* https://www.yumpu.com/de/document/read/8248872/realisierbarkeit-und-beurteilung-asthetischer-klangkonzepte-bei-, S. 15 ff.

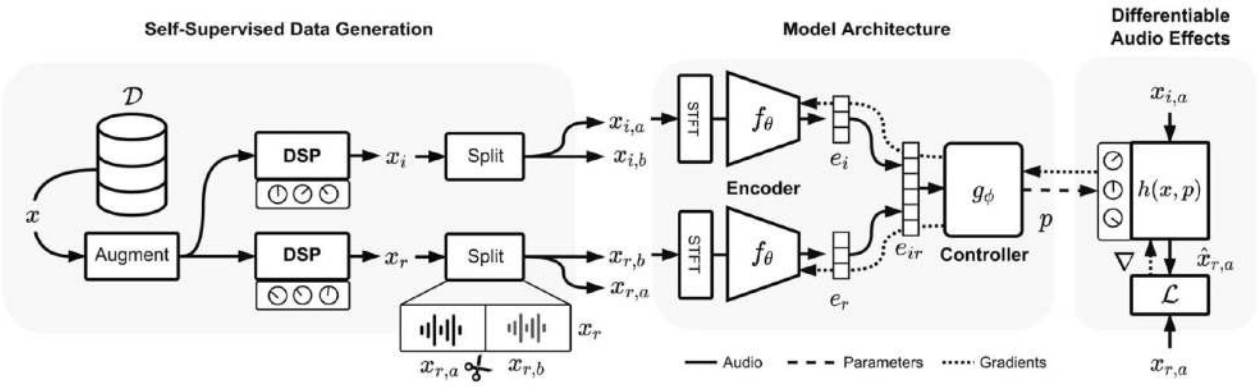

Abbildung 3: Steinmetz' Ansatz beinhaltet einen selbstüberwachten Trainingsprozess, bei dem zwei verschiedene Aufnahmestile mit zufällig konfigurierten DSP erzeugt und an gemeinsam genutzte Encoder und einen Controller weitergeleitet werden, um Parameter vorherzusagen und den Referenzstil auf Eingabesignale anzuwenden.

2. Trainingsprogramm

Eine der Besonderheiten der KI des Papers ist das vom Autor benannte „sich selbstüberwachende" Trainingsprogramm. Im folgenden wird Bezug auf Abbildung 3 genommen und die Funktionsweise des neuronalen Netzes Steinmetz' erklärt.

Zu Beginn steht ein Datensatz (D) aus verschiedensten Tonaufnahmen (x). Bei diesen Aufnahmen kann es sich um allerlei Aufnahmen handeln. Es müssen nicht zwangsläufig sauber produzierte Aufnahmen sein. Zusätzlich werden diese Aufnahmen noch weiter verfremdet (*Augment*, z.B. in Tonhöhe verändert, verlangsamt oder verschnellert, etc.) um den Datensatz anzureichern. An dieser Stelle werden die Aufnahmen dann in zwei Wege geteilt. Jeder dieser Wege wird mit randomisierten Signalverarbeitungsprozessen versehen (*DSP*, z. B. Equalizer), wodurch sie weiter diversifiziert werden. Somit werden klanglich zwei verschiedene Aufnahmen (x_i als Input Signal und x_r als Referenzsignal) generiert. Bei x_i und x_r handelt es sich um das selbe Signal mit verschiedenen Effekten.

Beide Signale werden nun in der Hälfte der Zeit geteilt (*Split*), womit vier Aufnahmen, also $x_{i,a}$ sowie $x_{i,b}$ und $x_{r,a}$ sowie $x_{r,b}$ generiert werden. Somit entstehen jeweils zwei verschiedene Zeitabschnitte. Dies ist wichtig, damit der Lernprozess der KI verbessert wird und das Netzwerk sich selbstständig kontrollieren kann. $x_{i,a}$ und $x_{r,b}$ werden in den Encoder überführt (wo der Vorhersageprozess für die Effektparamter p gesteuert wird), während $x_{i,b}$ und $x_{r,a}$ den Encoder umgehen.

Im Encoder werden für $x_{i,a}$ Effektvorhersagen getroffen anhand des Stils des Zeitabschnittes von $x_{r,b}$. Die zuvor am Encoder vorbeigeführten Aufnahmen $x_{i,b}$ und $x_{r,a}$ werden anschließend für einen Abgleich herangezogen. Hierbei kontrolliert die KI selbstständig, ob die getroffene Vorhersage für das Signal $x_{i,a}$ mit dem Signal von $x_{r,a}$ vergleichbar ist oder ob der Stil noch zu ähnlich zu $x_{i,b}$ ist.

Der Encoder operiert also nicht mit nahezu identischen Signalen, die nur verschiedene Effekte haben, sondern mit einem ähnlichen Signal (da gesplittet) mit verschiedenen Effekten. Das fördert den Lernprozess der künstlichen Intelligenz, da sich das System immer auf eine neue Aufnahme einstellen muss und nicht zwei vom Material komplett unterschiedliche Signale erhält, womit der Lernerfolg aufgrund fehlender Ankerpunkte gemindert werden könnte. Wenn nach dem Training komplett unterschiedliche Signale in das System geführt werden und mit unterschiedlichen Signalen eine Stilangleichung stattfinden soll, ist dieser Prozess für die „eingelernte KI" förderlich. Die KI wird also mit einfacheren Aufgaben angelernt, um dann mit komplexeren Signalen arbeiten zu können.

2.1 Beispiel des Lernprozesses

Es soll das Podcastbeispiel, das zu Beginn dieses Whitepapers angeführt wurde, herangezogen werden: Es handelt sich um gesprochenes Wort. Zuerst wird die Aufnahme dupliziert und mit verschiedenen Effekten versehen (zum Beispiel Tonhöhen-Veränderung für das Inputsignal und gefiltert für das Referenzsignal). Nun wird diese Aufnahme genau in der Mitte geteilt, um sowohl vom Input als auch vom Referenzsignal je einen a und einen b Abschnitt zu erhalten. Bei der Parametervorhersage im Encoder wird nun der a-Teil des Inputsignals und der b-Teil des Referenzsignals überführt. Die KI muss nun anhand vom b-Teil (Referenz) Effektparamter für die verschiedensten Audioeffekte/Plugins für den a-Teil so vorhersagen, dass sie klanglich sehr ähnlich zum b-Teil (Referenz) angepasst werden kann. Da es sich um zwei unterschiedliche Zeitabschnitte handelt, die sich bestenfalls inhaltlich nicht wiederholen, muss die KI sich also immer neu auf jeglichen Signalinput einstellen. Da es sich aber nicht um allzu unterschiedliche Signale handelt, ist die KI zu Beginn des Lernens nicht „orientierungslos", was die Lernrate gerade am Anfang steigert. So wird das System dazu ermutigt, nicht einfach anhand eines Vergleiches des selben Quellsignals die Parameter vorherzusagen, sondern eine Analyse von Grund auf anzufertigen. Das Ziel des Trainings ist es, dass das neuronale Netz mit völlig unterschiedlichen Signalen arbeiten kann. Je besser es gelernt hat, desto besser wird dies der Fall sein. Zudem handelt es sich um eine Art „Trick", sodass das System nie von der Annahme ausgehen kann, gleiche Inhalte zu erhalten.

Nach diesem Prozess werden die vorhergesagten Parameter in den *Controller* überführt und im Bereich der *Differentiable Audio Effects* (Plugins) auf das Signal $x_{i,a}$ angewandt. Im nächsten Schritt kann der Controller mittels Analyse von $x_{i,a}$ und $x_{r,b}$ vergleichen, inwiefern die Vorhersagen zutreffen und mittels Fehlerfunktion zum Encoder zurückgeben. Mittels des Gradientabstiegsverfahrens wird sich dann an das Referenzsignal iterativ angenähert, bis der gewünschte Stil erreicht wurde.

Als Vorteil des Verfahrens stellt sich heraus, dass bereits existierende digitale Audioeffekte bzw. Plugins, die auf dem Markt sind, genutzt werden können. Denn bei den Parameter aus den Vorhersagen handelt es sich um gestalterische Prinzipen um ein klangliches Ziel zu erreichen. Diese können auf jedem Plugin gleicher Machart erreicht werden, auch wenn diese von anderen Firmen hergestellt wurden. Ein Equalizer (egal von welcher Firma) ist dafür da, den Frequenzverlauf zu entzerren. Es handelt sich also um ein Funktionsprinzip, das im Grunde bei jedem Plugin ähnlich funktioniert.

Audio Effekte von renommierten Firmen weisen eine hohe Qualität mit wenig Artefakten auf, da sich die Hersteller*innen seit einiger Zeit mit der Entwicklung von digitalen Signalverarbeitungsprogrammen beschäftigen. Durch ein gutes Graphical User Interface (GUI) sind auch für wenig erfahrene Nutzer*innen nachvollziehbare Werte gegeben, die interpretiert werden können.

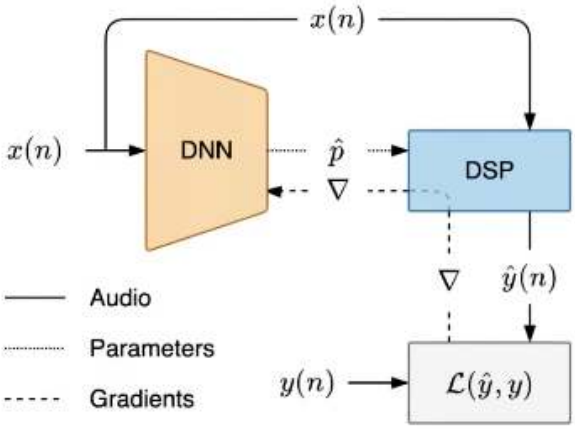

Abbildung 4: Verallgemeinerte Darstellung.

2.2 Verallgemeinerte Darstellung der Methode

Das Konzept aus Abbildung 3 ist in Abbildung 4 zum Verständnis vereinfacht dargestellt und soll nun folgend kurz erläutert werden. Das Inputsignal $x(n)$ wird einmal durch das Deep Neural Network geführt (Abbildung 4, *DNN*) und einmal zu Vergleichszwecken direkt in die Digital Signalprocessing Domäne (*DSP*). Bei dieser DSP Sektion kann es sich um eine Verkettung von verschiedenen Audioeffekten handeln. Das Modell hat nun die Aufgabe, Parameter für das Inputsignal hervor zu sagen ($\hat{p}$), die die DSP so konfiguriert, dass ein neues Signal $\hat{y}(n)$ erzeugt wird, das so nah wie möglich an dem Zielsignal bzw. Referenzsignal $y(n)$ liegt. $L(\hat{y}, y)$ vergleicht nun die Signale und gibt über das Gradientenabstiegsverfahren die Verlustfunktion bzw. die Fehlerfunktion durch das DSP zurück an das DNN. Durch den iterativen Prozess wird der Prozess solange wiederholt, bis das DSP so eingestellt ist, dass $\hat{y}(n)$ und $y(n)$ quasi identisch sind. So ist der Stiltransfer abgeschlossen und das Input Signal hat den Stil des Referenzsignals aufgeprägt bekommen.

Der DSP Block (blau) könnte natürlich auch mit irgendeiner „Black Box" mittels komplexer mathematischer Formeln realisiert werden. Da dies aber für die Nutzer*innen nicht nachvollziehbar ist, hat sich Steinmetz zur Aufgabe gemacht, dass an dieser Stelle jegliche DSP eingesetzt werden kann (Plugins). Diese können gelesen und verstanden werden (je nach Komplexität

des DSP-Effektes) und vor allem modifiziert werden. Das hebt dieses System besonders hervor.

3. Diskussion und Schlussfolgerungen

Die von Steinmetz vorgeschlagene Methode ermöglicht es nicht nur unerfahrenen Nutzer*innen einen schnellen Einstieg in die Audiowelt zu ermöglichen, es steckt auch weiteres Potential dahinter. So könnte die Methode auch von professionellen Anwender*innen genutzt werden, um schneller produzieren zu können und somit mehr Zeit für die kreative Arbeit zu haben.

Für viele in der Audiobranche könnte dies jedoch auch eine Gefahr darstellen, da eventuell ein Kundenstamm wegfällt, der bis dato professionelle Audiotechniker*innen aufgesucht hätte.

Als weiterer positiver Effekt für unerfahrene Nutzer*innen ist der einfach zu erreichende Qualitätssprung zu nennen, sodass sich am Markt von der Konkurrenz abgehoben werden kann. So würde der Markt zunächst nicht (wie es momentan der Fall ist) mit qualitativ minderwertigem Material geflutet werden. Das Potential hinter den Aufnahmen könnte so von Rezipient*innen besser bewertet werden.

Die Methode von Steinmetz ermöglicht eine automatische, adaptive und intelligente Audioproduktion in der der Produktionsstil einer Aufnahme auf eine andere übertragen wird. Die Methode wurde sowohl bei Sprach- als auch bei Musikaufnahmen getestet und mit anderen Signalverarbeitungsmethoden verglichen. Es können die von Steinmetz und seinem Forschungsteam publizierten Ergebnisse unter https://csteinmetz1.github.io/Deep-AFx-ST/ angehört und selbst bewertet werden.

4. Quellen

4.1 Primärliteratur

Christian Steinmetz, Nicholas J. Bryan und Joshua D. Reiss (2022): *Style Transfer of Audio Effects with Differentiable Signal Processing* (Paper).

Tim Fischer (2021): *8 beeindruckende Statistiken zur Creator Economy,* https://influence-vision.com/blog/8-beeindruckende-statistiken-zur-creator-economy,zuletzt am 15.03.2023 abgerufen.

Elisabeth Kemper (2006): *Realisierbarkeit und Beurteilung ästhetischer Klangkonzepte, Hochschule für Musik Detmold,* https://www.yumpu.com/de/document/read/8248872/realisierbarkeit-und-beurteilung-asthetischer-klangkonzepte-bei-, zuletzt am 15.03.2023 abgerufen.

Sarah Oberteils (2021): *Der späte Boom der Podcasts,* https://www.faz.net/aktuell/wirtschaft/schneller-schlau/podcast-boom-die-corona-pandemie-hat-beim-erfolg-geholfen-17675766.html, zuletzt am 15.03.2023 abgerufen.

Marko Pauli (2008): *Einfach nur laut ist nicht gleich gut,* https://www.deutschlandfunkkultur.de/einfach-nur-laut-ist-nicht-gleich-gut-102.html, zuletzt am 15.03.2023 abgerufen.

Podcastle Artikel (2019): *Maximizing Audio Quality: Tips and Strategies for Podcast Creators,* https://podcastle.ai/blog/maximizing-audio-quality/, zuletzt am 15.03.2023 abgerufen.

Jochen Wieloch (2016): *So richten Sie sich ein Musikstudio zu Hause ein,* https://www.welt.de/wirtschaft/webwelt/article153221951/So-richten-Sie-sich-ein-Musikstudio-zu-Hause-ein.html, zuletzt am 15.03.2023 abgerufen.

iZotope Artikel (o. J.): *What Is Mixing in Music?,* https://www.izotope.com/en/learn/how-to-mix-music.html, zuletzt am 15.03.2023 abgerufen.

Alle ***Abbildungen*** sind dem Paper von Steinmetz entnommen.

4.2 Sekundärliteratur

Jesse Engel et. al. (2020): *DDSP, Differentiable Digital Signal Processing* (Paper).

David Moffat, Mark B. Sandler (2019): *Approaches in Intelligent Music Production* (Paper).

Wang et. al. (2021): Differentiable Signal Processing with Black Box Audio Effects (Paper).

5. Anhang

Hörbeispiele zum Paper *Style Transfer of Audio Effects with Differentiable Signal Processing:* https://csteinmetz1.github.io/DeepAFx-ST/, zuletzt am 15.03.2023 abgerufen.

Video einer Vorstellung von *Style Transfer of Audio Effects with Differentiable Signal Processing* des Autors Steinmetz: https://www.youtube.com/watch?v=Kcm1Xj-pyG8, zuletzt am 15.03.2023 abgerufen.