Acoustic Intelligence

acoustic studies düsseldorf

Herausgegeben von
Dirk Matejovski und Kathrin Dreckmann

Band 5

Acoustic Intelligence

Hören und Gehorchen

Herausgegeben von
Anna Schürmer, Maximilian Haberer
und Tomy Brautschek

d|u|p
düsseldorf university press

Dieser Sammelband wurde mit freundlicher Unterstützung durch die Anton-Betz-Stiftung der Rheinischen Post e. V. und der Philosophischen Fakultät der Heinrich-Heine-Universität Düsseldorf verwirklicht.

ISBN 978-3-11-072720-3
e-ISBN (PDF) 978-3-11-073079-1
e-ISBN (EPUB) 978-3-11-073086-9
ISSN 2702-8658
e-ISSN 2702-8666

Library of Congress Control Number: 2022936369

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

d|u|p düsseldorf university press ist ein Imprint der Walter de Gruyter GmbH

Druck und Bindung: CPI books GmbH, Leck
Covergestaltung: Silvia Sunderer, Kommunikation & Design, Berlin
Titelbild: Martin Wietell
Lektorat: Anna Schürmer, Maximilian Haberer, Tomy Brautschek, Markus Radermacher
Satz: Elisabeth Stanciu, Markus Radermacher
Redaktionelle Mitarbeit: Emily Glavan

dup.degruyter.com

Inhalt

Dirk Matejovski und Kathrin Dreckmann
Vorwort —— VII

Anna Schürmer, Maximilian Haberer, Tomy Brautschek
***Acoustic Intelligence*. Hören und Gehorchen —— 1**

Teil 1: Panakustik

Intermezzo
Sean Dockray, James Parker, Joel Stern
(Gegen) die kommende Welt hörender Maschinen —— 41

Elena Ungeheuer
Hören macht Macht. *Acoustic Intelligence* und ihre Potentiale —— 47

Intermezzo
Artemi-Maria Gioti
***Machine Listening* als aktiver Prozess in interaktiven Kompositionen —— 71**

Rolf Großmann
Hören, was die Maschine hört —— 83

Intermezzo
Jonathan Sterne
Was bedeutet es, *Machine Listening* als eine Form des Hörens zu begreifen? —— 95

Teil 2: Monitor

Pedro J S Vieira de Oliveira und Shintaro Miyazaki
Maschinelle Intelligenz?! Stimmbiometrie als hörend-bestimmende Medien-Techno-Logie —— 101

Intermezzo
Alexander Schubert
***AV3RY* & *CRAWLERS* oder: „Unheimliche" Postdigitalität —— 125**

Christoph Borbach
„A Wall of Sound“. Das Unterwasserschallsignalwesen als Hörregime und techno-akustische Einkerbung des Ozeans —— 133

Intermezzo
Nik Nowak
***Schizo Sonics*. Installation und Interview —— 155**

David Waldecker und Axel Volmar
Die zweifache akustische Intelligenz virtueller Sprachassistenten zwischen verteilter Kooperation und Datafizierung —— 161

Intermezzo
Sean Dockray, James Parker, Joel Stern
Lektionen, wie man (nicht) gehört wird —— 183

Teil 3: Agent

Alan Fabian
Musik(tabellen)formulare: Musikmach(t)dinge von Musikklangverwaltungen —— 191

Intermezzo
Sean Dockray, James Parker, Joel Stern
Improvisation und Kontrolle. Interaktive (Musik-)Systeme —— 215

Max Alt und Jens Gerrit Papenburg
***Streamability*. Überlegungen zu einer Ästhetik des Musik-Streaming —— 227**

Intermezzo
Johannes Kreidler
#dirigate. Über akustische Regimes in meiner Musik —— 247

Lukas Iden, Sophia Tobis, Malte Pelleter
„Recognized by sound“ oder: *Acoustic Intelligence* im Trackmodus —— 257

Intermezzo
Transhuman Art Critics
***Mycelium Melody* —— 281**

Beiträger*innen —— 283
Register —— 289

Dirk Matejovski und Kathrin Dreckmann

Vorwort

Der vorliegende Sammelband, der bereits der fünfte innerhalb der Reihe *acoustic studies düsseldorf* ist, beschäftigt sich wieder einmal mit einem über die Fachgrenzen hinaus relevanten Thema. Denn ohne Zweifel bildet die akustische Dimension von Überwachung und sozialen Disziplinarordnungen ein zentrales Thema innerhalb der globalen Medienkultur.

Sprachassistent*innen wie *Siri* und *Alexa* oder klangoperierende Spracherkennungs- und Diktiersoftware, algorithmengesteuerte Streamingplattformen wie *Spotify* und *Deezer* oder auch durch Sprachaufnahmen gesteuerte personalisierte Werbeanzeigen in Social Media sind nur einige Beispiel für die Verbindung zwischen KI und Strategien medialer Kontrolle.

Die auch für den vorliegenden Band charakteristische Perspektive, soziokulturelle und ästhetische Prozesse von auditiven Medientechnologien her zu verstehen, bildet eine spezifische Forschungsperspektive ab, die seit langem durch eine Gruppe von Wissenschaftler*innen am Institut für Medien- und Kulturwissenschaft der Heinrich-Heine-Universität Düsseldorf in Forschung und Lehre präsent ist.

Unter dem Titel *Acoustic Intelligence – Hören und Gehorchen* versammelt dieser Band 5 einerseits Beiträge einer von den Herausgeber*innen veranstalteten Tagung, andererseits legt die multiperspektivische Anlage der verschiedenen Texte die Ausgangsbasis für weitere Forschungen in diesem Feld vor.

Und so wird das auch in diesem Band thematisierte Wechselverhältnis zwischen akustischen, sozialen und ästhetischen Dispositiven sowohl in der weiteren Arbeit der Düsseldorfer Forschungsgruppe als auch in zukünftigen Publikationen thematisiert werden.

https://doi.org/10.1515/9783110730791-203

Rolf Großmann

Hören, was die Maschine hört

Abstract: Technische Medien wie Musikautomaten, Grammophone oder Musikcomputer haben einen bedeutenden Anteil an der Konstruktion unserer auditiven Wirklichkeit. Die mit ihnen erfahrbaren Wahrnehmungswelten entstehen in einer Gemeinschaftsleistung von Mensch und Maschine. In dieser komplexen Situation kann es hilfreich sein, den Medieninput genauer zu betrachten: Wenn es etwas außerhalb der technisch medialen Konfiguration gibt, das als ihr Input die Voraussetzung für jeden Output bildet, lässt sich seine technosensorische Aneignung als ‚Hören' der Maschine beschreiben. Der Beitrag stellt zunächst grundsätzliche Überlegungen zum Verhältnis von Reproduktionsmaschinen und ‚Realität' an, um schließlich einige künstliche Intelligenzen genauer zu betrachten.

Schlüsselwörter: Technikkultur der Musik, Digitale Musik, Algorithmische Komposition, AI und Musik, Medienästhetik

Der Mensch befiehlt, die AI gehorcht. Siri oder Alexa schalten das Licht ein und säuseln untertänige Antworten, *AIVA* (s. u.) komponiert im gewünschten Genre den Backgroundsound für ein beliebiges Video. Doch mit ein wenig Distanz betrachtet ist in unserer alltäglichen Soundkultur erst einmal das Gegenteil der Fall: Wir hören auf die Maschine. Immer, wenn wir unsere *Earbuds* einsetzen und unsere Streaming Plattform aktivieren, immer, wenn wir auf dem Sofa unserer Stereoanlage lauschen oder im Club dem Flow des DJs folgen, hören wir, physikalisch betrachtet, einer Maschine beim Erzeugen von Schallwellen zu. Elektrische Spannungsschwankungen werden übertragen, prozessiert, verstärkt und bewegen eine Membran, die schließlich hörbare Schallwellen erzeugt. Wir hören also – im Falle eines handelsüblichen Lautsprechers – ein elektromagnetisch bewegtes Stück Pappe. Eine triviale Tatsache, die jedoch gerade wegen ihrer Selbstverständlichkeit jenseits der Bewusstseinsschwelle angesiedelt ist. Das technische Medium, Marshall McLuhan hat es bereits 1964 beschrieben, verschwindet mit zunehmender Vertrautheit. Es sei dahingestellt, ob das Medium die „Message" ist, in jedem Fall erzeugt es physikalisch gesehen das Wahrnehmungsangebot für die Rezeption.

Damit eine solche Verdrängung und ein Verschwinden des Mediums möglich ist, bedarf es der Annahme, dass auf der entgegengesetzten Außenseite des Mediums, der Input-Seite, eine medienunabhängige Realität oder – im Falle einer

https://doi.org/10.1515/9783110730791-005

Klangsynthese – eine generative Aktivität existiert, auf die sich unsere Aufmerksamkeit unter Ausblendung des medialen Dazwischen richtet. Wir gehen also davon aus, dass es eine direkte Beziehung zwischen einem akustischen Ereignis oder einer gestaltenden Vorgabe des Inputs und dem Medien-Erzeugnis auf der Output-Seite gibt. Erst diese Erwartung in die Vermittlung einer vermeintlich außermedialen Wirklichkeit lässt uns das Medium vergessen. Friedrich Kittler macht auf diese Bezüge aufmerksam, wenn er mit Rekurs auf Lacan das Grammophon das „Reale" und dem Film das „Imaginäre" zuordnet. (Kittler 1986) Diese Zuordnung bleibt jedoch dem Lacan-Universum verhaftet und verkennt in der unterschiedlichen Etikettierung die grundlegenden Abbildfunktionen der beiden Medientypen: Dass im Film durch Selektion und Inszenierung des Abzubildenden und durch die folgende Montage Raum und Zeit nicht bruchlos abgebildet werden, lenkt zwar den Blick auf die visuelle Wirklichkeit einer vorgestaltenden Imagination, dies gilt jedoch für die Phonographie gleichermaßen. Bereits das Grammophon nimmt seit seinen Anfangstagen ausschließlich an das Medium angepasste Klanginszenierungen nach Maßgabe einer imaginierte Hörwirklichkeit auf. Die damals verwendeten Stroh-Geigen sind etwa für ein direktes Hören höchst unangenehm, während sie – transformiert durch Membran, Nadel und Materialität des Tonträgers – eine musikalische Balance der Instrumente erst ermöglichen. Dabei ist diese Relation von Realität und Medienkonstrukt, die technische Transformationsprozesse beinhaltet und die jeweils in komplexe kulturelle Praxen medientechnischer Produktion und Rezeption eingebettet ist, entscheidend für die kognitive Formung der Rezeptionsergebnisse als individuelle Kommunikate. Sie bildet das Dispositiv des Mediums.

Die dabei wirksamen Voraussetzungen und Transformationsprozesse sind schwer zu beschreiben und analysieren. Bevor wir also einige künstliche Intelligenzen genauer betrachten, gilt es, grundsätzliche Überlegungen zum Verhältnis von Reproduktionsmaschinen und Realität vorauszuschicken. Schon die Annahme einer medienunabhängigen Realität bringt Probleme. Spätestens seit Kant wissen wir, dass ein reales Apriori außerhalb der Wahrnehmung zwar eventuell existiert, jedoch – wenn überhaupt – nur begrenzt durch die Voraussetzungen der Wahrnehmung, einschließlich ihrer Verarbeitung erfahrbar ist. Auch auf technischer Ebene ist dies evident: So erfasst bei der digitalen Aufnahme ein Mikrofon auch akustische Ereignisse, die über das hörbare Frequenzspektrum hinausgehen und bei der digitalen Rasterung erhebliche Schwierigkeiten bereiten. Wenn diese Frequenzen nicht vorher nach Maßgabe des menschlichen Wahrnehmungsvermögens und der angewendeten Abtastrate herausgefiltert werden, bilden sich Alias-Frequenzen, welche die unhörbaren

Spektren des Ultraschalls in den hörbaren Bereich spiegeln.[1] Technische Medien haben also wie die menschliche kognitive Wahrnehmung ihren Anteil an der Konstruktion von Realität, die durch sie erzeugten Wahrnehmungsangebote sind bereits vorgeformt. Dieser ohnehin komplexe Sachverhalt wird noch komplexer durch die grundsätzlich technokulturelle Genese und Situiertheit technischer Apparate, die zugleich Ergebnis menschlicher Kulturen sind.

Vor diesem Hintergrund könnte es hilfreich sein, und das ist die Grundidee dieses Beitrags, den Transformationsprozess nicht aus der Sicht der Rezeption, sondern mit Blick auf den Medieninput zu betrachten. Drehen wir also das Modell um: Wenn es etwas außerhalb der technisch medialen Konfiguration gibt, das als ihr Input die primäre Voraussetzung für jeden Output bildet, lässt sich seine technosensorische Aneignung als ‚Hören' der Maschine beschreiben. Natürlich ist dies zunächst nicht mehr als eine einfache Metapher. Inwieweit Grammophone, Tape Recorder, Digital Audio Workstations, Streamingportale oder künstliche Intelligenzen (AIs) im menschlichen Sinne hören, lässt sich für jede neue Technologie neu verhandeln. Eine solche Zuschreibung setzt immer auch ein Bild eines Menschen voraus, dessen Sinnesorgane und Kognition maschinenähnliche Prozesse ausführen. Eine Annahme, der man folgen kann, aber nicht muss – für unseren Kontext ist dieser Aspekt nebensächlich: Hier geht es nicht primär um eine Diskussion der Vermenschlichung technischer Prozesse, sondern in einem ersten Schritt um das Bewusstsein einer Teilnahme technischer Apparate an der Konstruktion unserer auditiven Wirklichkeit sowie in einem zweiten Schritt um die Beschreibung von Differenzen, Relationen und Konfigurationen maschinellen ‚Hörens'. Dass einem fortschreitenden Agens der Maschine in einer ‚gemeinsam' gebildeten Hörwirklichkeit meist auch menschliche Attribute zugeschrieben werden, liegt jedoch nahe.

1 Analoge und digitale Reproduktionsmaschinen

> Historically, machines have listened to human musicians via two methods: symbolic data and signal data.
>
> (Magnusson 2019, S. 159)

Die sich um die Jahrhundertwende zum 20. Jh. etablierenden Reproduktionsmaschinen erhalten ihren Input zumeist durch Lochbänder bzw. -platten oder durch Membranen. Die gebräuchlichen Eingabevorrichtungen von Musikauto-

1 Aliasing gemäß des Nyquist-Shannon Theorems.

maten und phonographischen Apparaten werden zwar gleichermaßen mit Begriffen wie (*auf-*)*zeichnen*, *lesen* oder *schreiben* belegt, unterscheiden sich jedoch in ihrer ‚Wahrnehmung' der Außenwelt. Bei der Melographie wird etwa ein Papierband zur pneumatischen Steuerung von Musikautomaten perforiert, während bei phonographischen Apparaten aller Art – entweder direkt beim mechanischen Nadelton oder indirekt über elektrische Spannungsschwankungen beim Mikrofon – eine Membran zur Klangaufzeichnung genutzt wird.

Im Falle der Musikautomaten passt die Metapher des Hörens weniger, es handelt sich dabei meist um eine maschinenlesbare Transkription symbolischer Tonhöhen-Notation bzw. bei der Melographie von Künstlerrollen um die Aufzeichnung von Bewegungsabläufen des Spiels mechanisierter Instrumente (wie etwa der Klaviermechanik). Die Membran dagegen weist tatsächlich eine Parallele zum hörenden Ohr auf, sie setzt per Resonanz die Mechanik für eine weitere Verarbeitung der Schwingungen in Bewegung. Beide Verfahren sind keineswegs neutrale Abbildungen einer äußeren Realität, sondern formatieren den Input im Sinne ihrer technischen und kulturellen Konfiguration. Neben der rein technischen Formatierung erfolgten immer auch eine Inszenierung und Manipulation dieses Inputs (s. o.). Die mit dem Medium abzubildenden Wahrnehmungsangebote sind bereits im Wissen um ihre Formatierung gestaltet. Im Falle eines Steuerungscodes wie bei den Künstlerrollen für Musikautomaten wird die Vorstellung einer angemessenen Wiedergabe von Anordnungen von Tönen (einem interpretierten ‚Werk' im Sinne traditioneller Notation) zur primären Gestaltungsvorgabe, für die Phonographie geht dem technischen Hören ein imaginiertes, ideales Klangbild voraus. Zusätzlich wird in diesem Sinne der Input zumeist während den weiteren Schritten der medientechnischen Übertragung und Speicherung in diesem Sinne manipuliert. Wenn wir den Reproduktionsapparaten zuhören, hören wir also eine spezifische Aufführung dessen, was die Maschine nach Maßgabe einer gestalterischen kulturellen Praxis liest oder hört: Die Aufführung einer vorformatierten und vorinszenierten musikalischen Wirklichkeit.

Dies gilt auch für die digitalen Medienmaschinen, die unser heutiges musikalisches Leben dominieren. Die beiden genannten Inputmodi, Steuerungscode einerseits und Abbildung akustischer Schwingungen andererseits, sind auch hier wiederzufinden. Diese Eingaben werden nun in zählbare diskrete Einheiten gewandelt, damit sie durch eine Zahl repräsentiert und digital verarbeitet werden können. Während die Steuerungscodes ohnehin als diskrete Notenevents vorliegen, werden die kontinuierlichen phonographischen Schwingungen gerastert und vermessen, so dass auch Schwingungsformen in einem arbiträren

symbolischen Kode abgebildet werden können. Aus den Lochstreifen werden MIDI-Daten[2], aus der analogen wird digitale Phonographie.

Digitale Maschinen definieren sich durch Raster und Regel. Die Rasterung bestimmt ihr Verhältnis zur Außenwelt des Systems, ihre internen Programmstrukturen ermöglichen ihre spezifische Funktionalität. Durch das Zusammenspiel beider Elemente werden sie zur Büromaschine, zum Wettersimulator, zum Smartphone, zur Drummachine oder zur Digital Audio Workstation. Im Bereich des Auditiven entsteht durch die digitale Phonographie eine neue Situation: Steuerungsdaten und Audiodaten können in gemeinsamen digitalen Umgebungen transformiert werden. Darüber hinaus können aufgrund ihres arbiträren symbolischen Codes auch andere Eingabedaten aus völlig anderen Umgebungen, etwa Börsendaten oder Textnachrichten, in Audiomaschinen nach „Audioregeln" verarbeitet werden. Solche Sonifikationen sind zwar bereits durch Neuverschaltung der elektronischen Signale in der analogen Phonographie möglich, werden jedoch in digitalen Umgebungen hochgradig vereinfacht, sie sind per Datenmapping konfigurierbar und durch Parameter steuerbar.

Auch die kulturelle Vorformatierung und Inszenierung ändert sich durch Begleiterscheinungen digitaler Technologie wie Miniaturisierung, Mobilisierung und Vernetzung. Als Input werden neue auditive Bereiche erschlossen oder weiterentwickelt. *ASMR* etwa setzt das bereits mit dem Aufkommen des Mikrofons seit den 1920er Jahren bekannte Phänomen medialer Intimität (*Crooning*) fort, indem die größtmögliche sensorische Annäherung zum Prinzip der auditiven Inszenierung wird.[3]

2 Intelligente Maschinen

Maschinen scheinen dem menschlichen Hören verwandt zu sein, wenn sie ihre internen Zustände durch ihre ‚Erfahrungen' (aufgrund des jeweiligen Inputs und Datenbanken) verändern, also ‚lernen' und über so etwas wie ‚Intelligenz' verfügen. Sie transformieren aktiv ihren Input, ihnen werden kognitive Eigenschaften zugeschrieben. Tatsächlich sind Begriffe wie ‚Erfahrung' auch in gebräuchlichen Definitionen des *Deep Learning* zu finden:

2 Im Musical Instrument Digital Interface (MIDI) Protokoll wurden 1983 Steuerungsdaten für digitale Instrumente standardisiert.

3 Siehe dazu Großmann 2020, S. 436 f.

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
>
> (Tom M. Mitchell 1997 in *Machine Learning*, zit. nach Briot 2020, S. 64.)

Betrachten wir also die historisch gesehen zunächst letzte Station des maschinellen Hörens, in der lernende Algorithmen mit musikalischen Phänomenen verbundene Daten verarbeiten. Auf ihrer Inputseite steht in einer funktionsentscheidenden Anfangsphase auch eine prozessbegleitende Reihe menschlicher Eingaben: Anders als regelbasierte Systeme, deren innere Struktur ein anwendungsspezifisches Set von Regeln zur Verarbeitung enthält, sind lernende Systeme auf ein ‚Training' angewiesen, in der die Bewertung der Performance der Maschine – also die Passung ihrer Mustererkennung vor dem Hintergrund der Zielvorgabe – in den Prozess rückgekoppelt wird.[4] Mit dem Ende der Trainingsphase entfällt dieser Input, die Bewertung ist Teil der Algorithmik des neuronalen Netzwerks geworden. Die Implementation von Regeln erfolgt entsprechend nicht direkt als Programmstruktur, sondern durch Selbstorganisation des Systems. Die entstehende interne Struktur bildet so die Trainingseingaben in einer von außen nur begrenzt einsehbaren funktionalen Repräsentation ab.

Während sich dieser Teil des Inputs bei den im Folgenden betrachteten Beispielen für lernende Systeme kaum unterscheidet, bestehen auf der Inputseite der zu verarbeitenden Daten große Differenzen. Daraus resultieren zwei grundlegende Fragen: Was gelangt durch die Sensorik aus der äußeren Welt ins Innere der Maschine? Und: Wie wird dieser Input systemintern repräsentiert?

AIVA ist ein „Artificial Intelligence Virtual Artist" sowie ein kommerzielles Projekt, das per AI „emotional soundtrack music" komponiert.[5] Anfänglich auf klassische Musik fokussiert, generiert es inzwischen ebenso personalisierte *soundalikes* anderer Genres wie Jazz, Rock, Pop oder Electronic. *AIVA* hört keine Klänge, sondern liest Noten bzw. Dateien im MIDI-Format oder in ähnlichen Formaten, in denen Töne als Events in Notenparametern beschrieben sind. Wie CEO Pierre Barreau in einem TED-Talk erklärt, extrahiert das neuronale Netzwerk aus „30,000 scores, written by the likes of Mozart and Beethoven" (Barreau 2018) Muster, die mithilfe zufälliger oder gesteuerter Parameteränderungen

4 Die Unterteilung zwischen regelbasierten und lernenden Systemen ist in unserem Kontext hilfreich, vereinfacht jedoch die komplexe Situation sich gegenseitig durchdringender Verfahren. Einen tieferen Einblick geben hier Fernández und Vito 2013, S. 518 ff.

5 *AIVA*. The Artificial Intelligence composing emotional soundtrack music. https://www.aiva.ai/. Zugegriffen am 05. Oktober 2021. Das Projekt ist u. a. EU gefördert (Horizon 2020) und bei der Verwertungsgesellschaft SAGEM als Urheber (Komponist) eingetragen.

zur Erzeugung neuer MIDI-Daten genutzt werden können. Damit steht *AIVA* in der Tradition der Verarbeitung von *symbolic representations*, wie sie von vielen Pionierprojekten in der AI-Music Szene genutzt wird.[6] Für den Input solcher Projekte ist es i. d. R. nicht erforderlich, dass die Dateien speziell für das Projekt eingespielt oder aus Audiomaterial transkribiert werden,[7] da im Internet unzählige MIDI-Dateien zur Verfügung stehen. Daneben können auch komplette *Open Source*-Datenbanken wie etwa *Mutopia* oder die *Weimar Jazz Database* genutzt werden.[8]

Die Maschine hört bzw. liest damit standardisierte Steuerungsdaten, deren konkreter Klang in den Daten nicht repräsentiert ist. So enthält eine MIDI-Datei die üblicherweise in einer Partitur gegebenen Informationen in maschinenlesbarer Form (relative Tonhöhe, Anschlagstärke, Zuordnung der jeweiligen Stimme zu einem genannten Instrumentencharakter, optionale Modulationsparameter). Da MIDI-Dateien auch die zeitliche Anordnung der einzelnen Daten-‚events' enthalten, können sie – anders als die Notenschrift – eingespielte Tonfolgen zeitlich genau abbilden. Je nach Relevanz dieser Daten für das Projekt kann eine Reduktion durch Quantifizierung dieses Parameters beim Einlesen oder bei der Speicherung der MIDI Dateien in einem gerasterten Zeitschema wie Metrum und Takt vorgenommen werden.

Mit der Entscheidung, symbolische Daten in Form parametrisierter Notenevents zu nutzen, ist auch eine weit reichende kompositorische Vorentscheidung verbunden, die bis in das Musikverständnis des jeweiligen Projekts reicht.

> We believe that the essence of music (as opposed to sound) is in the compositional process, which is exposed via symbolic representations (like musical scores or lead sheets).
>
> (Briot 2020, S. 21)

Wenn also auf der Inputseite genrespezifische symbolische Notenwerte gelesen werden, kann sich auch das in vielen Trainingsvorgängen entwickelte Modell (das operative ‚Wissen' der Maschine) ausschließlich auf dieses Symbolsystem beziehen. Damit fällt die Maschine aufgrund der fehlenden Daten über die klin-

6 Andere ähnlich konfigurierte kommerzielle Projekte sind etwa *amper* oder *Jukedeck*, das 2019 von Bytedance, der Muttergesellschaft von Tiktok übernommen wurde.

7 Für das Telekom Projekt *Beethoven X – The AI-Projekt* (2021) zur Konstruktion einer Sinfonie aus den Handschriftfragmenten Ludwig van Beethovens mussten allerdings neben vorhandenen genrespezifischen Daten die Handschriftenfragmente Beethovens zunächst in maschinenlesbare Notenevents übertragen werden.

8 *Mutopia* (o. J.) ist ein Open Source Projekt mit dezentralen Beiträgen; die *Weimar Jazz Database* (2013–2020) wurde im Rahmen des *Jazzomat Research Project* an der Hochschule für Musik Franz Liszt Weimar erstellt.

gende Aufführungspraxis auf einen Teil des Wissensstands des Genres zurück. Ein ähnlich reduziertes Wissen entsteht bei der Lektüre von Partituren ohne Berücksichtigung ihres performativen und kulturellen Kontexts. Dies ist gleichermaßen problematisch für „alte Musik", deren Notationssysteme in weitreichende kulturelle Praxen eingebunden sind, wie für aktuelle Genres, die in phonographischen Medien gestaltet werden und für die eine Abbildung auf Notenevents eine erhebliche Verkürzung bedeutet. Es entsteht ein von seiner historischen und kulturellen Einbettung entkoppelter Mustergenerator, der Symbolstrukturen für eine maschinelle oder menschliche Aufführung generiert.

Ein künstlerischer Einspruch der ebenfalls mit AI-Prozessen arbeitenden Künstlerin Holly Herndon auf solche Verfahren lautet folgerichtig:

> Wenn du Kunst aus ihrem Kontext nimmst, verliert sie ihre Substanz. [...] Das Ergebnis [im Hinblick auf Popmusik, RG] ist Mood-Music, die dich letztlich an gar nichts mehr denken lässt. So etwas bringt das Medium Musik kein Stück weiter. Kunst sollte die Umwelt und Zeit reflektieren, aus der sie stammt.
>
> (Herndon 2020)

Ihr eigenes „AI-baby" *SPAWN* hört für das mit einer Mensch-Maschine Konfiguration verwirklichte Album *Proto*

> [h]auptsächlich Stimmen. Die eines kleinen Gesangsensembles, aber auch unsere eigenen und die von Freunden. [...] Man kann die Arbeit des neuronalen Netzwerks gut in dem Album-Opener *Birth* nachvollziehen, der auf meiner Stimme basiert.[9]
>
> (Ebd.)

Die Baby-Metapher wurde bewusst gewählt, um sowohl den Entwicklungsstand der Technologie als auch den Stand des Wahrnehmen und Lernens der AI zu kennzeichnen. (Herndon 2019b, TC 00:01:26)

SPAWN hört anders und Anderes als *AIVA* & Co. Es hört phonographisches Material, Audiosamples, deren Obertöne analysiert wurden und als Spektrogramm vorliegen. In der Regel werden diese Spektrogramme, die Audiodaten als hochauflösende Pixelgrafiken repräsentieren, im *Deep Learning*-Prozess wie Bilder verarbeitet (Briot 2020, S. 23). Um in der Logik der Wahrnehmungsmetaphorik zu bleiben, werden demnach von der Maschine Spektralbilder analysierter Audio Signale angeschaut, die AI hört, indem sie ‚sieht'. An dieser Stelle wird klar, dass bereits vor dem eigentlichen *Machine Learning* in der Repräsentation des Inputs

9 Eine ausführliche Kritik zur Nachkonstruktion alter Musik findet sich in Herndon 2019a, TC 00:01:09.

verschiedene Transformationen stattfinden, die mit den üblichen Bezeichnungen menschlichen Wahrnehmens nur schwer vergleichbar sind.

Die Beschränkung auf stark individualisierte Vokalsamples eröffnet hier einen völlig anderen Erfahrungsraum der Maschine als bei dem vorausgehenden Beispiel. Sie generiert ein spezialisiertes Klangwissen, das etwa in Call und Response oder im Layering als Maschinenwissen mit Originalstimmen korrespondieren kann. In diesem Zusammenspiel können Differenzen und Gemeinsamkeiten des menschlichen und maschinellen Erfahrungsraums unmittelbar ästhetisch reflektiert werden. Ihre kompositorische Arbeit, so Herndon, liegt dabei in beträchtlichem Maße in der Produktion des Datasets für das Voicemodel, also im gesungenen und gesprochenen Input sowie im Training des Modells. Anhand von Proben der Resynthese wurde die Ausführung der Inputs noch während des Eingabeprozesses angepasst, so dass eine weitere Rückkopplung entstand. Im Ergebnis ist *SPAWN* innerhalb gewisser Grenzen in der Lage, die eigene Stimme Herndons oder die Stimmen eines Gesangsensembles parametergesteuert zu resynthetisieren. Die AI bleibt hier Teil eines individuellen Künstlerprojekts, aus dem ihr kompletter Input stammt, der schließlich in transformierter Form wieder in das Projekt einfließt.

Eine ähnliche, jedoch stark erweiterte Konfiguration verwendet Alexander Schubert in seiner multimedialen Arbeit *Convergence* (2020)[10]. Die Arbeit basiert auf aufgenommenem Audio- und Videomaterial des Hamburger Ensemble Resonanz, das als Input eines *Machine Learning*-Prozesses dient. Es geht dabei wie bei Herndon nicht um ein Lernen und Anwenden vertrauter Kompositionsregeln mit dem Ziel der Erzeugung künstlicher Kunstwerke, die wie menschliche Schöpfungen klingen, sondern um die Erfahrbarkeit der Mensch-Maschine-Relation in einer gemeinsamen künstlerischen Performance. Wie sieht und hört die Maschine menschliche Aktionen? Ein zentraler Ausgangspunkt ist – in Schuberts Worten – „to kind of take a look at how a machine classifies and categorizes an input" (Schubert 2021b: TC 00:03:58). Die Arbeit reflektiert die Differenz maschineller „Wahrnehmung" zu unserer eigenen Wahrnehmung und Kognition in der ästhetischen Gegenüberstellung ihrer unterschiedlichen Resultate. Sie zielt also nicht auf die möglichst überzeugende Fortsetzung einer genrespezifischen Komposition, sondern auf die Differenzerfahrung zwischen Mensch und Maschine in einer kooperativen Praxis.

Aufgrund der Komplexität der Arbeit soll hier nur der auditive Teil beschrieben werden. Auch hier wird eine selbst konstruierte Programmumgebung genutzt, die zum Teil aus Open Source Parts zusammengesetzt ist. Es geht nun in

10 Siehe Schubert 2021a.

Convergence nicht wie bei *Proto* um die eigene musikalische Performance, sondern um die Erarbeitung und Präsentation einer technoästhetischen Umgebung, die von Akteuren (hier vom Hamburger Ensemble Resonanz) ‚bespielt' werden kann. Bereits die vorausgehende Kategorisierung des Audioinputs nimmt eine Abstraktion vor, die zwischen möglichen Aktionen und Klängen unterscheidet. Zwei Spielweisen von Streichinstrumenten stehen drei klanglichen Bereichen der Stimmperformance gegenüber (Schubert 2021b, TC 00:21:23):

1. konventionelle Streichtechniken
2. erweiterte Streichtechniken
3. Singen
4. Sprechen
5. Schreien

Die dadurch entstehenden Datenmodelle repräsentieren als übergeordnete Kategorien Spiel und Gesangstechniken und nicht primär den individuellen Klang einer Person oder eines Ensembles, der allerdings ebenfalls in den phonographischen Proben erhalten bleibt und für die Performance genutzt werden kann. Auf einer weiteren Ebene spielt der Input bereits mit Kategorien wie Expressivität und Emotionalität, die insb. in den Datasets der Modelle zwei und fünf in die Mustererkennung eingehen. Dieser auditive Baukasten kann schließlich im Zusammenspiel mit den zugehörigen Videos und der Live Performance der Akteure eingesetzt werden. Zuhörer*innen sehen und hören eine Assemblage aus Medienphänomenen der Abbildung und Transformation.

Es ist also bei den beiden letztgenannten Beispielen nicht der Geist Ludwig van Beethovens oder Franz Schuberts (Huawei 2019), der per Maschinenintelligenz wieder zum Leben erweckt wird, im Mittelpunkt eines musikalischen *Uncanny Valley* sein Unwesen treibt und den Blick auf Mensch und Maschine verschleiert. Im Gegenteil: Auch hier hören wir – wie in allen vorausgehenden Beispielen – dem Hören einer Maschine zu, die nun jedoch nicht als Werkzeug oder Medium verschwindet, sondern als Teil künstlerischer Arbeit ästhetisch erfahren werden soll. Transformationen auditiver Phänomene bis in unsere Wahrnehmung, wie sie bei allen Medien geschehen, werden dabei ganz im Sinne McLuhans sichtbar und hörbar:

> The effects of technology do not occur at the level of opinions or concepts, but alter sense ratios or patterns of perception steadily and without any resistance. The serious artist is the only person able to encounter technology with impunity, just because he is an expert aware of the changes in sense perception.
>
> (McLuhan 1964, S. 18)

Medienverzeichnis

Literatur

Briot, Jean-Pierre, Gaëtan Hadjeres und François-David Pachet. 2020. *Deep Learning Techniques for Music Generation. Computational Synthesis and Creative Systems*. Cham: Springer Nature Switzerland AG.

Deutsche Telekom. 2021. *Beethoven X – The AI-Projekt*. Telekom. https://www.telekom.com/de/konzern/themenspecials/special-250-jahre-beethoven/beethovens-unvollendete. Zugegriffen am 05. Oktober 2021.

Fernández, Jose David und Francisco Vico. 2013. AI Methods in Algorithmic Composition. A Comprehensive Survey. *Journal of Artificial Intelligence Research* 48: 513–582.

Großmann, Rolf. 2020. The Instrument as Medium. Phonographic Work. In *The Bloomsbury Handbook of Sound Art*, hrsg. Sanne Krogh Groth und Hoger Schulze, 436–445. New York u. a.: Bloomsbury Academia.

Herndon, Holly. 2020. Interview: *Künstliche Intelligenz kann etwas so zutiefst Menschliches wie Ekstase auslösen* von Fabian Peltsch. *musikexpress*. https://www.musikexpress.de/holly-herndon-im-interview-kuenstliche-intelligenz-kann-etwas-so-zutiefst-menschliches-wie-ekstase-ausloesen-1290787/. Zugegriffen am 05. Oktober 2021.

Huawei. 2019. Uraufführung in London: HUAWEI komplettiert Schuberts *Unvollendete*. Huawei. https://consumer.huawei.com/de/press/news/2019/urauffuehrung-in-london-huawei-komplettiert-schuberts-unvollendete/. Zugegriffen am 05. Oktober 2021.

Kittler, Friedrich. 1986. *Grammophon Film Typewriter*. Berlin: Brinkmann & Bose.

Magnusson, Thor. 2019. *Sonic Writing. Technologies of Material, Symbolic, and Signal Inscriptions*. New York, NY: Bloomsbury Academic.

McLuhan, Marshall. 1964. *Understanding Media. The Extensions of Man*. New York: McGraw-Hill.

Mutopia. o. J. *Mutopiaprojekt*. https://www.mutopiaproject.org/index.html. Zugegriffen am 05. Oktober 2021.

Romero, Juan, Anikó Ekárt, Tiago Martins und João Correia, Hrsg. *2020. Artificial Intelligence in Music, Sound, Art and Design*. 9th International Conference, EvoMUSART 2020 Seville, Spain. Proceedings. Cham: Springer Nature Switzerland AG.

Weimar Jazz Database. 2013–2020. *Jazzomat Research Project*. Hochschule für Musik Franz Liszt Weimar. https://jazzomat.hfm-weimar.de/dbformat/dboverview.html. Zugegriffen am 05. Oktober 2021.

Audio und Video

Alexander Schubert. 2021a. *Convergence*. Ensemble Resonanz @Kampnagel/Eclat, 00:34:25. *YouTube* Videostream. https://www.youtube.com/watch?v=o5UXkJWJciQ&t=1194s. Zugegriffen am 05. Oktober 2021.

Alexander Schubert. 2021b. *Alexander Schubert – Presentation „Convergence“*. Eclat Festival Presentation Series, 01:25:49. *YouTube*. Videostream. https://www.youtube.com/watch?v=laoV7cGXUNo. Zugegriffen am 05. Oktober 2021.

Holly Herndon. 2019a. *The One Song Holly Herndon Wishes She Wrote*, 00:03:29. *YouTube*. Videostream. https://www.youtube.com/watch?v=wT9ycCUCaV4. Zugegriffen am 05. Oktober 2021.

Holly Herndon. 2019b. *Holly Herndon – Birthing PROTO*, 00:06:04. *YouTube*. Videostream. https://www.youtube.com/watch?v=v_4UqpUmMkg. Zugegriffen am 05. Oktober 2021.

Holly Herndon. 2019c. *Birth*, 00:01:15. *YouTube*. Videostream. https://youtu.be/ZFe0Bcngp64. Zugegriffen am 05. Oktober 2021.

Pierre Barreau. 2018. *How AI could compose a personalized soundtrack to your life*, 00:08:20. *TED*. Videostream. https://www.ted.com/talks/pierre_barreau_how_ai_could_compose_a_personalized_soundtrack_to_your_life#t-116274. Zugegriffen am 05. Oktober 2021.