digitale gesellschaft | NRW

Marcus Erbe / Aycha Riffi / Wolfgang Zielinski (Hrsg.)

Mediale Stimmentwürfe

Perspectives of Media Voice Designs

Schriftenreihe zur digitalen Gesellschaft NRW

7

kopaed

Marcus Erbe / Aycha Riffi / Wolfgang Zielinski (Hrsg.)

Mediale Stimmentwürfe

Schriftenreihe zur digitalen Gesellschaft NRW

Band 7

Marcus Erbe / Aycha Riffi / Wolfgang Zielinski (Hrsg.)

Mediale Stimmentwürfe

Perspectives of Media Voice Designs

Düsseldorf – München
www.kopaed.de

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

ISBN 978-3-96848-642-0

In der Schriftenreihe zur digitalen Gesellschaft NRW vertreten die Autorinnen und Autoren ihre eigene Meinung, ohne dass diese notwendigerweise die Meinung des Landes Nordrhein-Westfalen widerspiegelt.

Die Veröffentlichung entstand mit freundlicher Unterstützung der Staatskanzlei des Landes Nordrhein-Westfalen.

Verlag: kopaed verlagsgmbh
Umschlaggestaltung: Georg Jorczyk

Grimme-Institut – Gesellschaft für Medien, Bildung und Kultur mbH, Marl 2022

Eine Open Access Version dieses Bands ist zu finden unter:
http://www.grimme-institut.de/schriftenreihe

Inhalt

Nadia S. Zaboura

Schöne neue Stimmen oder: Sprechen Sprachassistenten im Schlaf?

Wenn sie nicht bereits existieren würde, man müsste sie erfinden. So universell, so praktisch, so responsiv ist sie. Sie umgibt den Menschen vom ersten bis zum letzten Atemzug. Die Rede ist von der Stimme.

Wer die Stimme erhebt und das Wort ergreift, bezieht immer auch Stellung zur und in der Welt, wird zum aktiv Gestaltenden im intersubjektiven Sprachspiel. Und so ist die Stimme ein geniales Verortungs-Instrument in einer multidimensionalen sozialen Matrix.

Eine Auseinandersetzung mit der Stimme schenkt uns deshalb einen Blick auf unser Selbst, auf das, was wir Menschlichkeit nennen. Menschlichkeit heißt dabei auch, dass wir als Hörende sofort erkennen, ob eine Stimme natürlich oder synthetisch ist. Zumindest bis jetzt.

Denn künstlich erzeugte Stimmen sind aus der Informationsgesellschaft nicht mehr wegzudenken. Ob Navigationsgeräte oder Sprachassistenzen: Mit dem Einzug digitaler Audio-Anwendungen ist die Stimme nicht mehr einmalig, nicht mehr einzigartig. Das Bild von der Stimme als „individuellem auditiven Fingerabdruck“ wankt – und eröffnet ein neues Feld voller Forschungsfragen.

Da wären allgemeine Fragestellungen: Welche Rückkopplungseffekte lösen artifizielle Stimmen auf das menschliche Selbstverständnis aus? Wie verändert sich die Wahrnehmung von Leiblichkeit in Gegenwart synthetischer Stimmen und welche Auswirkungen hat dies auf die Konstruktion von Wirklichkeit? Werden in technisch-medial konstruierten Stimmen kulturübergreifende und intersektionale Aspekte bedacht, berücksichtigt, integriert? Und auch: (Wie) kann eine gleichberechtigte, historische Vorurteile und Machtstrukturen hinter sich lassende Mensch-Maschine-Kommunikation gelingen?

Und da wären spezifische Fragestellungen, die sich bereits aus der aktuellen Praxis ergeben: Kann es „neutrale“ mediatisierte Stimmen geben? Weshalb geben Sprach-Assistent*innen kaum Widerworte? Welche Resonanz lösen synthetische

Stimm-Klone aus? Und wie wird in nicht-westlichen, globalen Kontexten über technisch-mediale Stimmen diskutiert?

Angesichts der Vielzahl und Tragweite der Fragen wird deutlich: „Mediale Stimmen“ sind ein komplexes, zeitgemäßes und faszinierendes Forschungsfeld. Einen interdisziplinären Blick aus verschiedenen Perspektiven ermöglichen die Wissenschaftlerinnen und Wissenschaftler in diesem Band. Mit neuen Beiträgen vermessen sie den aktuellen Forschungsstand, leuchten Herausforderungen aus und formulieren relevante Fragestellungen für die Zukunft.

Und vielleicht findet sich in nicht allzu ferner Zukunft eine Antwort darauf, ob Sprachassistenten auch im Schlaf sprechen oder – um es mit Kleist zu sagen – ihre Gedanken künftig allmählich beim Sprechen verfertigen. Eben so, wie es der Mensch tut.

Marcus Erbe, Aycha Riffi, Wolfgang Zielinski

Einleitung

Ist die menschliche Stimme einzigartig und unverwechselbar? Der Schauspieler und Synchronsprecher Oliver Rohrbeck verkörpert bereits seit 1979 die Figur des Justus Jonas in der Hörspielserie *Die drei ???*. Der damals Dreizehnjährige steht noch heute regelmäßig mit Andreas Fröhlich („Bob Andrews“) und Jens Wawrczeck („Peter Shaw“) vor dem Studiomikrofon, um neue Abenteuer einzusprechen. Seit 2004 feiern die drei auch mit Live-Events große Erfolge. Die Ton-Bild-Schere – die von ihnen verkörperten Figuren sind in all den Jahren nicht mitgealtert – macht dabei den besonderen Charme der Veranstaltungen aus. Interessanter aber ist, dass die Stimmen in den Hörspielen noch heute funktionieren – für Neueinsteiger*innen ebenso wie für nostalgietrunkene Hörspielfans, die sie seit Jahrzehnten begleiten. Indes scheint bei einer nicht minder populären Serie das Phänomen alternder Stimmen der narrativen Kontinuität im Weg gestanden zu haben. Als es für eine Folge von *The Mandalorian* (USA 2019ff.) den jungen Luke Skywalker wiederzubeleben galt, musste dessen Originaldarsteller Mark Hamill nicht nur äußerlich, sondern auch vokal verjüngt werden. Um den fraglichen *Star Wars*-Charakter im benötigten Alter möglichst glaubwürdig auftreten lassen zu können, bedienten sich die Produzent*innen einer Kombination aus ‚falschem‘ Körper, facialer Deepfake-Software und dem Sprachsyntheseprogramm Respeecher. Digitale Abbilder eines rund 40 Jahre jüngeren Hamill wurden aus Archivaufnahmen seines Gesichts wie seiner Stimme gewonnen und mit der Mimik und Gestik des Schauspielers Max Lloyd-Jones technisch verflochten. Wenngleich derartige Kombinationen filmhistorisch betrachtet kein grundsätzliches Novum darstellen (z.B. resultieren auch der Körper, das Gesicht und die Stimme Darth Vaders in der zeitlich ersten *Star Wars*-Trilogie 1977-83 aus der Vermischung mehrerer individueller Performanzen), so verweist diese Anekdote auf das Moment der zunehmenden Perfektionierung artifizieller Identitätskonstruktionen mittels digitaler Werkzeuge.

Seit Beginn der Tonaufzeichnung geht es nicht nur darum, individuelle Stimmen zu konservieren und zu reproduzieren, sondern auch zu modifizieren, unseren Erwartungen und Bedürfnissen anzupassen und zu modellieren. Die Möglichkeit, durch Veränderung der persönlichen Stimme andere Identitäten annehmen zu können, passt in die Welt der anonymisierten Online-Kommunikation. DeepDubs könnten zukünftig Synchronisationsarbeit automatisieren, Sprachassistenzen vielleicht in unserem

bevorzugten Dialekt oder mit der Stimme unserer Lieblingsprotagonist*innen mit uns kommunizieren. Wir stehen noch ganz am Anfang einer durch künstliche Intelligenz und Algorithmen sich grundlegend verändernden stimmlichen Kommunikation.

Die vorliegende Publikation ist aus dem Forschungsprojekt *Kulturelle Implikationen medial konstruierter Stimmen* hervorgegangen, das wir als Teil des Grimme-Forschungskollegs an der Universität zu Köln 2020 durchführten. Ziel dieses Projekts war die Untersuchung medial zirkulierender Stimmentwürfe unter der leitenden Fragestellung, welche Sozialvorstellungen – z.B. von Angemessenheit, Autorität und Handlungsmacht – medial konstruierten Vokalitäten zugrunde liegen bzw. sich in ihnen ausdrücken. Seit der Einführung stimmbasierter Navigationsgeräte um das Jahr 2000 ist es zunehmend selbstverständlich geworden, dass digitale Apparate sich anhand menschlich klingender Stimmen mitteilen. Zwar datieren die Anfänge der elektronischen Stimmsynthese in die späten 1930er-Jahre. Dennoch war es vor der allgemeinen Verfügbarkeit ‚sprechender' Navis, Computer und Smartphones hauptsächlich das Privileg phantastischer Filme, Hörspiele und Fernsehserien, den Mediennutzer*innen eine Vorstellung von der Beschaffenheit künstlicher Stimmen zu vermitteln. Doch so wenig medial konstruierte Stimmen in fiktionalen Kontexten neutral sein können, so wenig neutral präsentieren sich die Stimmen rezenter Applikationen und Betriebssysteme. Dies zeigt sich etwa in der aktuell geführten Debatte um den inhärenten Sexismus bei Sprachassistenzsystemen wie Alexa, Siri und Cortana. Indes mangelt es an Untersuchungen, die zugleich theoretische, medienpraktische und kulturübergreifende Gesichtspunkte vokaler Designs berücksichtigen, und zwar nicht nur im Hinblick auf IT-Erzeugnisse, sondern die Medienproduktion ganz allgemein.

Bereits in der Planungsphase war uns sehr daran gelegen, Akteur*innen aus Wissenschaft, Kunst, medienpädagogischer Praxis und Produktentwicklung in einen Dialog über die Spezifika heutiger wie früherer Stimmtechnologien zu bringen. Als projektinterner Höhepunkt erwies sich somit ein am 4. August 2020 abgehaltener digitaler Workshop, an dem sich Vertreter*innen aus Computerlinguistik, Musikwissenschaft, Theaterwissenschaft, Filmwissenschaft und Voice-Design gemeinsam mit Rundfunkredakteur*innen, Medienautor*innen, Schauspieler*innen und Kommunikationsberater*innen beteiligten. Diskutiert wurde über die Stimme als Kulturphänomen, ihre Gestaltung im Kontext von Bewegtbildmedien, die Signifikanz ‚körperloser' Stimmen im Radio, die Stimme in der Mensch-Computer-Interaktion sowie über technische oder technisierte Stimmen in Musik, Theater und Performancekunst. Einigen unserer damaligen Gäste war es erfreulicherweise möglich, ihre Überlegungen aus dem Workshop schriftlich weiter auszuarbeiten und in diesem Buch niederzulegen. Andere gewährten uns in eigens geführten und hier abgedruckten

Interviews wichtige Einblicke insbesondere in die Arbeitsabläufe und Herausforderungen stimmbasierter Tätigkeitsfelder. Des Weiteren konnten wir – nicht zuletzt im Sinne einer internationalen Perspektivierung des behandelten Gegenstandes – Autor*innen aus Brasilien, den Niederlanden, Südafrika und den USA hinzugewinnen. Da es mit den Voice Studies inzwischen einen globalen Forschungszweig gibt, der stimmlichen Phänomenen in den Künsten, Medien sowie der Alltagskommunikation nachgeht und dessen Diskurs vornehmlich auf Englisch geführt wird (siehe exemplarisch Neumark et al. 2010; Young 2015; Pettman 2017; Cox 2018; Frühholz & Belin 2018; Eidsheim & Meizel 2019; Meizel 2020), haben wir uns bewusst gegen eine Übersetzung der englischsprachig eingereichten Texte entschieden. Lediglich die Abstracts wurden ins Deutsche übertragen, um möglichst allen Leser*innen eine schnelle Orientierung über die jeweiligen Inhalte offerieren zu können.

Wenngleich die einzelnen Beiträge in vielerlei Hinsicht Berührungspunkte aufweisen und überwiegend interdependente Aspekte behandeln, ist der Band aus Gründen der Übersichtlichkeit in vier Großabschnitte gegliedert. Im ersten Teil *Stimmforschung heute* widmet sich Katherine Meizel gleich einem ganzen Bündel soziokultureller Faktoren, die es beim Einsatz aktueller Stimmtechnologien mitzudenken gilt. Vor allem geht sie der Frage nach, wie die gesellschaftlichen Machtverhältnisse, in denen Stimmen erklingen, durch Technik nicht nur gespiegelt werden, sondern sich ebenfalls auf technologischem Wege – zum Guten wie zum Schlechten – verändern lassen. Dieser Umstand wird am Beispiel menschlich wirkender Sprachassistenzprodukte und ihrem Verhältnis zu den teilweise dahinter stehenden realen Sprecher*innen sowie anhand der digitalen Hervorbringung, Bearbeitung und des Verkaufs von Singstimmen expliziert. Zusätzlich bespricht Meizel die Probleme und Chancen sprachlicher und stimmlicher Partizipation während der COVID-19-Krise unter Berücksichtigung neuester Entwicklungen im Bereich der medizinisch-therapeutischen Stimmdiagnostik. Ihre Analyse digital erzeugter, übertragener, archivierter und veränderter Stimmen im Spannungsfeld zwischen vokaler Individualität und kollektiver Technisierung markiert wichtige Themengebiete einer zeitgemäßen Beschäftigung mit stimmlichen Phänomenen. Im Mittelpunkt des Beitrags von Lílian Campesato und Fernando Iazzetta steht die Verbildlichung der Stimme. Dabei wird nicht nur über die Bildhaftigkeit stimmlicher Aktionen, sondern über das imaginative Potenzial von Hörinhalten insgesamt nachgedacht. Ausgehend vom Resonanzphänomen, also dem Vorgang des Mitschwingens, der hier sowohl in einem physikalisch-akustischen als auch kulturtheoretischen Sinne verstanden wird, analysieren die Autor*innen drei Fallbeispiele aus den Bereichen der bildenden Kunst, der Popmusik und des Geisterglaubens. Unter zusätzlicher Beachtung aktueller Forschungsdiskurse aus Philosophie, Anthropologie, Kognitionswissenschaft und

Sound Studies gelangen sie zu der Einsicht, dass die Momente der Hervorbringung, Repräsentation, Wahrnehmung und Bewertung stimmlicher Laute stets miteinander verkoppelt sind und sich auf materieller wie auf affektiver Ebene ständig wechselseitig beeinflussen.

Den Abschnitt *Künstliche Stimmen* eröffnet Marc Böhlen mit einer umfassenden Untersuchung der Sprachsynthese. Aus seinem Text geht hervor, wie historische Versuche der Nachbildung menschlicher Stimmen und damit verbundene Vorstellungen von einer quasi perfekten Imitation menschlicher Eigenschaften entsprechende Vorhaben auch heute noch prägen. Gegenwärtige Manifestationen synthetisch erzeugter Stimmen und ihrer sprachlichen Äußerungen versteht er folglich als alte neue Technologie, die im Unterschied zu ihren frühen Ausprägungen aber nicht mehr nur in isolierten Kontexten vorkommt, sondern den Alltag durchdringt. Mit der zunehmenden Ununterscheidbarkeit von realen und digital generierten Stimmen ergeben sich nicht zuletzt in alltäglichen Nutzungssituationen Möglichkeiten der Täuschung und des Missbrauchs. Dies stellt neue Herausforderungen an den Erwerb und die Vermittlung von Medienkompetenz. Auch Christine Bauer und Johanna Devaney werfen einen kritischen Blick auf die sozialen Effekte ubiquitärer Stimmtechnologien. Beginnend mit der Frage, wie Geschlechtsidentitäten durch den Einsatz synthetischer Stimmen konstruiert werden und wer auf welche Weise daran mitwirkt, stellen sie mannigfaltige Verbindungen zwischen Verkörperungsstrategien und Genderkonzepten bei Sprachassistenzsystemen sowie der digitalen Bearbeitung bzw. Erzeugung von Singstimmen her. Dabei erweist es sich, dass die aktuelle Stimmforschung einer intersektionalen Perspektive bedarf, um die Zusammenhänge zwischen Formen von Vokalität und sich überschneidenden Faktoren wie Gender, Lebensalter, Körperbild, soziale Stellung oder Ethnizität – insbesondere vor dem Hintergrund von Diskriminierungsvorgängen – besser verstehen zu können. Zwei Gespräche mit Praktikerinnen aus der Voice-Branche komplettieren das Themenfeld *Künstliche Stimmen*. Aus der Sicht einer Computerlinguistin, die an der Entwicklung intuitiv handhabbarer Assistenzsysteme arbeitet, berichtet Laura Dreessen von den aktuellen Anforderungen an die stimmlich gebundene Mensch-Maschine-Interaktion. Gesichtspunkte wie die technische Realisierbarkeit nonverbaler Kommunikation und der Datenschutz kommen ebenso zur Sprache wie das gendersensible Design der Assistenzcharaktere und das Manipulationspotenzial KI-basierter Systeme. Stefanie Ray war im Auftrag von Amazon als Autorin für die Persönlichkeit der deutschsprachigen Alexa mitverantwortlich. Sie gewährt seltene Einsichten in den Designprozess und erläutert unter anderem, wie sich die Lokalisierung dieser originär US-amerikanischen Assistenzsoftware vollzog, auf welche Weise ein Sprachroboter eine eigene Identität vorgaukeln kann und wie das natio-

nale Entwickler*innenteam mit Alexas stimmlich und sprachlich bereits vorgezeichneter Genderkodierung umging. Beide Interviews markieren eine aufschlussreiche Ergänzung (teilweise sogar ein inhaltliches Gegengewicht) zu den Beiträgen, die sich rein wissenschaftlich mit der Gestaltung und den Kommunikationseffekten von Sprachassistenzsystemen befassen.

Der Teil *Künstlerische Stimmen* adressiert vokale Praktiken im Film, in der Popmusik und im Gegenwartstheater. Oksana Bulgakowa beschäftigt sich mit der audiovisuellen Konstruiertheit filmischer Stimmen. Unter besonderer Berücksichtigung des Aspekts einer Trennung von Stimme und Körper – sowohl im Sinne der technischen Aufspaltung als auch narrativ hinsichtlich der Repräsentation gespaltener Persönlichkeiten – vergleicht sie nordamerikanische, russische, französische und deutsche Produktionen. Auf die zeitgenössisch empfundene Artifizialität audiotechnisch fixierter Stimmen wird ebenso eingegangen wie auf frühe Synchronisationsverfahren und die damit verknüpften Problemfelder der schauspielerischen Identität und Authentizität. Zudem gelingt es Bulgakowa, einzelne Facetten der Filmstimme im Kontext vorgängiger literarischer und musikalischer Traditionen neu zu deuten. Malte Kobels Aufsatz über die vokalen Experimente von Kate Bush demonstriert, dass technisch transformierte Stimmen im Bereich der Popmusik Konzepte von Identität und Authentizität ebenfalls ins Wanken bringen können. Indem sich Bush die in den 1980er Jahren noch junge Sampling-Technologie rasch aneignete, vermochte sie ihre Singstimme zum Ausgangspunkt mannigfaltiger musikalischer Verwandlungen werden zu lassen. In Erweiterung der damals bestehenden Overdub-Verfahren kam es zu akustischen Neuschöpfungen, welche die gewohnten Produktions- und Rezeptionsweisen musikalisch dargebotener Stimmen nachhaltig veränderten. Kobel nimmt dies zum Anlass, gängige Auffassungen stimmlich-körperlicher Individualität kritisch zu hinterfragen, weil Bushs vokale Praktiken im Zusammenspiel von Performativität, Reproduktion und klanglicher Transformation weitaus mehr als nur den Sound ‚ihrer' Stimme hörbar machen. Im anschließenden Interview mit Doris Kolesch geht es schwerpunktmäßig um die Rolle der Stimme und des Körpers in Theaterprojekten der letzten Jahre. Neben der Frage, wie der Einsatz digitaler Mittel und speziell die künstliche Intelligenz das Theater und die Performancekunst beeinflussten, wird darüber gesprochen, welche Auswirkungen die COVID-19-Pandemie auf die inszenatorische Praxis bisher gehabt hat und wie sich die Partizipationsmöglichkeiten des Publikums heutzutage gestalten.

Nachdem bislang mehrheitlich neuere Medientechnologien in ihrem Bezug zur Stimme analysiert wurden, schließt das Buch unter der Überschrift *Radio-Stimmen* mit der Betrachtung eines gleichsam beständigen Mediums, jedoch ohne die Effekte digitaler Werkzeuge auf den Rundfunk zu übersehen. Zunächst gehen Dumisani

Moyo und Kundai Moyo detailliert auf die sozialen und politischen Dimensionen des Radios in afrikanischen Gesellschaften ein. Sie erläutern, wie es sich von einem Mittel der Unterdrückung zum Medium der Artikulation von Freiheits- und Unabhängigkeitsbestrebungen entwickelte. In diesem Zuge wird deutlich, dass in weithin ländlich geprägten Regionen mit niedrigem Alphabetisierungsgrad die stimmlich vermittelte Kommunikation einen besonderen Stellenwert genießt und folglich das Radio als eine Art Leitmedium verstanden werden kann. Die persönliche Dimension der direkten Ansprache von Hörer*innen nebst ihrer Einbeziehung ins Programm sehen die Autor*innen durch algorithmische Prozesse teilweise gefährdet. Sie plädieren daher für einen ethisch verantwortungsvollen Einsatz künstlicher Intelligenz, und zwar auch, weil die Kombination aus immer besser werdenden Stimmnachbildungen und den Reichweiten des Rundfunks für Desinformationskampagnen gerade in solchen Staaten Verwendung finden könnte, deren Regierungen schon jetzt durch den zunehmenden Import von Überwachungstechnologie ihre Macht zu verstetigen suchen. Bis zu welchem Grad künstliche Stimmen bereits heute das Klangbild deutscher Radiosender prägen, erklärt die Moderatorin und Chefredakteurin Colleen Sanders. Aus dem Interview mit ihr geht hervor, dass Emotionalität und Stimme im Rundfunkalltag ganz eng miteinander verwoben sind und dass Moderator*innen, die der Hörerschaft aufgrund ihres Stimmklangs vertraut sind, vom Publikum auch als real präsente Personen außerhalb des Studios erlebt werden wollen. Welche Erwartungen darüber hinaus an stimmliche Timbres und Sprechweisen herangetragen werden, zeigt sich bei den Themen Werbung, Sexismus und Diversität.

Einige Aspekte medialer Stimmentwürfe konnten in der Kürze der Zeit – das Forschungsprojekt war auf ein Jahr begrenzt – nur angerissen werden, andere mussten wir ganz ausklammern. Insbesondere der aktuell angesichts global operierender Streamingdienste an Bedeutung gewinnenden Diskussion um automatisierte Synchronisationstechniken hätten wir uns gern bereits an dieser Stelle gewidmet, werden dies aber sicherlich in anderen Kontexten tun. Das Forschungsprojekt war eine Zusammenarbeit des Musikwissenschaftlichen Instituts der Universität zu Köln mit der Grimme-Akademie und der Grimme Medienbildung. Es wäre – wie auch der vorliegende Band 7 der Schriftenreihe Digitale Gesellschaft NRW – ohne die Förderung des Grimme-Forschungskollegs an der Universität zu Köln nicht möglich gewesen. Dafür danken wir herzlich. Unser besonderer Dank gilt den beteiligten Autor*innen und Interviewpartnerinnen sowie Judith Kirberger, Katharina Makosch und Elisabeth Turowski für ihre Mitarbeit und Unterstützung.

Literatur

Bulgakowa, Oksana (Hrsg.) (2012): Resonanz-Räume. Die Stimme und die Medien. Berlin: Bertz + Fischer.

Cox, Trevor (2018): Now You're Talking: Human Conversation from the Neanderthals to Artificial Intelligence. London: Vintage.

Eidsheim, Nina; Meizel, Katherine (Hrsg.) (2019): The Oxford Handbook of Voice Studies. New York: Oxford University Press.

Frühholz, Sascha; Belin, Pascal (Hrsg.) (2018): The Oxford Handbook of Voice Perception. New York: Oxford University Press.

Meizel, Katherine (2020): Multivocality: Singing on the Borders of Identity. New York: Oxford University Press.

Neumark, Norie; Gibson, Ross; van Leeuwen, Theo (Hrsg.) (2010): V01CE: Vocal Aesthetics in Digital Arts and Media. Cambridge: MIT Press.

Pettman, Dominic (2017): Sonic Intimacy: Voice, Species, Technics (or, How to Listen to the World). Stanford: Stanford University Press.

Young, Miriama (2015): Singing the Body Electric: The Human Voice and Sound Technology. Farnham: Ashgate.

Stimmforschung heute

Katherine Meizel

Voice and the Selfs of Technology

Abstract: Weil menschliche Stimmen sowohl in den menschlichen Körper als auch in die sozialen Rahmenbedingungen eingebettet sind, in denen sie erklingen, werden externe Stimmtechnologien häufig mit dem Posthumanen in Verbindung gebracht. Derlei Technologien können jedoch nicht nur die Möglichkeiten des organischen, menschlichen Klangs übersteigen; sie können sie auch erweitern. Der folgende Beitrag liefert einen Überblick über Teile der aktuellen technologiebezogenen Stimmforschung. Menschliche Stimmen wurden immer auch als Technologien des Selbst untersucht. Sie inspirierten die Erfindung künstlicher Stimmen lange vor dem Aufkommen digitaler Technologien. Sie waren Ausgangspunkt für die Hybrid- oder Cyborg-Stimmen, die heute mehrere Bereiche des Alltagslebens durchdringen. Sie werden als so wichtig für die Kommunikation erachtet, dass der Drang besteht, verlorene oder fehlende Stimmen zu ersetzen. Von veränderten bis zu künstlichen Stimmen, von Stimmbibliotheken bis zu Stimmdatenbanken möchte sich dieser Beitrag insbesondere auf das Verhältnis von Stimmtechnologien zu Vorstellungen und Praktiken von Identität, Fertigkeit und körperlicher Einschränkung konzentrieren. Obwohl Musik dabei nicht das Hauptthema sein wird, spielt sie eine wichtige Rolle bei der Diskussion von Stimmen, die an Sampling-Bibliotheken verkauft werden, künstlichen Singstimmen und Stimmen von Menschen mit Behinderung im Kontext populärer Musik.

Abstract: In part because human voices are embedded in both human bodies and the social frameworks in which they sound, external voice technologies are often associated with the posthuman. But such technologies can not only exceed the capabilities of organic, sonic humanity; they can also amplify them. This article will provide an overview of some current voice research related to technology. Human voices have been investigated as technologies of the self; they have, since long before the advent of digital technology, inspired the invention of artificial voices; they have served as foundational to the hybrid or cyborg voices that now pervade multiple aspects of daily life; they have been considered so essential to communication that lost or absent voices demand replacement. From altered voices to artificial voices, from voice libraries to voice banks, the article will focus particularly on the relationship of voice technologies to ideas and practices of identity, ability, and disability. Though music will not be the primary topic of the

article, it features prominently in discussions of voices sold to sampling libraries, artificial singing voices, and disabled voices in popular music.

1 Posthuman Voices

Late in his life, Michel Foucault famously identified four types of epistemological technology through which humans learn about themselves, and through which power structures are produced: 1) technologies of production, 2) technologies of sign systems, 3) technologies of power, and 4) technologies of the self. The latter, he wrote, "permit individuals to effect by their own means or with the help of others a certain number of operations on their own bodies and souls, thoughts, conduct, and way of being, so as to transform themselves in order to attain a certain state of happiness, purity, wisdom, perfection, or immortality" (Foucault 1988, p. 18). Musicologist Nina Sun Eidsheim has positioned voices, vocalities, vocal pedagogies, and the act of listening to voices within such a framework (Eidsheim 2008), and as Annette Schlichter notes in response to Eidsheim, this perspective requires the study of "mediation and modulation of the voice through sound technologies" (Schlichter 2011, p. 44). Eidsheim herself has examined the construction of race in the voice synthesis application Vocaloid (Eidsheim 2009), and as voice synthesis continues to grow in cultural presence and significance, it is important to reexamine the ways technology can mirror or change the power structures in which voices are sounded and heard.

In part because human voices are embedded in both human bodies and human social frameworks, external voice technologies are often associated with the posthuman. But the concepts of "posthuman" and "posthumanism" have been assigned multiple, competing meanings, as Cary Wolfe details in *What Is Posthumanism?* It has been associated with what comes after humanism, and with the idea of abandoning or transcending embodiment. Wolfe's own definition of posthumanism involves both the before and after spaces of humanism:

> before in the sense that it names the embodiment and embeddedness of the human being in not just its biological but also its technological world, the prosthetic coevolution of the human animal with the technicity of tools and external archival mechanisms [...] [but] it comes after in the sense that posthumanism names a historical moment in which the decentering of the human by its imbrication in technical, medical, informatic, and economic networks is increasingly impossible to ignore. (Wolfe 2010, p. xv)

I suggest that this understanding is key in the study of voice and technology – as is the related idea that technologies can not only transcend the capabilities of organic, sonic humanity, but also amplify them. This article will survey recent voice research related to technology. Human voices have been investigated as technologies of the self (Eidsheim 2008); they have, since long before the advent of digital technology, inspired the invention of artificial voices; they have served as foundational to the hybrid or cyborg voices that now suffuse multiple aspects of daily life; they have been considered so essential to communication that lost or absent sonic voices demand replacement. From altered voices to artificial voices, from voice libraries to voice banks, the article will focus particularly on the relationship of voice technologies to ideas and practices of identity.

2 Voice Assistants

The increasing popularity and ubiquity of voice assistants and similar AI demands a reevaluation of the relationship between voices and bodies. Some AI voices are drawn from modified recorded human speech and song, while others are fabricated whole-cloth through acoustic design. But whether or not a human body (or group of bodies) initially produced the sound, people attach bodies to the voices they hear – and, as Eidsheim writes, produce those bodies through listening practices rooted in systems of power (Eidsheim 2015). AI voices are not exactly acousmatic. We can see that Alexa's voice emanates from a squat cylinder or spherical speaker, or Siri's from a rectangular smartphone whose weight we feel in our hand. But those are not the bodies we build for the voices in our minds. Those voices *sound* human and feminine, or rather, to an extent, we listen to them as if they were human and feminine, as we insist that they do our bidding. But they are somehow at once human and not-human, a cyborg concoction of voice that has been disembodied and reembodied in the machine. And yet these voices are becoming part of millions of lives, establishing a *kind* of relationship, at least, largely due to the significance placed on voice in human communication. As Julie Carpenter explains, we "know that voice is a cue for our relationship with an Other – even a technological one" (Carpenter 2019, p. 58).

Siri's original sound was built on the voice of actor Susan Bennett. Bennett did not know how her work would be used, or for what technology, when she recorded hours of speech to be mined for vowels, consonants, and syllables (Ravitz 2013). The default vocal sound for most of these devices is coded as feminine, though other voices can be accessed or downloaded (for example, "Alexa" may become "Samuel," using actor Samuel L. Jackson's voice). But in 2014, a team at Vice Network's

design branch Virtue developed a voice under the name Q, which they describe as "nonbinary," with the idea of allowing AI voices to represent the spectrum of gender identity more fully. As Julie Carpenter, a research consultant for the project, recounts, project lead Emil Asmussen was inspired by a talk about hidden bias in artificial intelligence, and thought, "The world is increasingly acknowledging [many] gender options; why are there still only two options in AI? AI is born genderless so it seems stagnant that there's not a genderless option" (Carpenter 2019, p. 58). Carpenter argues that because the designers of persona-driven AI – home assistants, GPS, telephone voice systems – understand how significant gender cues are for consumers, and because the exclusion of identity groups in media representation is experienced as a kind of erasure, designers have a responsibility to be inclusive of a spectrum of gender identities (Carpenter 2019). She writes that the design process included the recording of six speakers identifying as "male, female, transgender, and nonbinary" (though it should be noted that these identities can overlap – "transgender" identities can include binary male and female as well as nonbinary identities), combined them, manipulated pitch and timbre, and applied a formant filter. Thousands of people were surveyed to "rate" different results, and the team found that a particular pitch range helped to define the preferred voice heard as nonbinary. In Project Q, a voice technology is working to subvert power structures that privilege a binary gender framework both in the marketplace and in social roles.

3 Pandemic Voices and Ability

The global COVID-19 pandemic has changed many things – how people work, how we communicate, how we establish and maintain community. But one of the earliest and perhaps least anticipated changes has come in the way we relate to voice. Some of the first pandemic-related footage to "go viral" in January 2020 featured Wuhan residents shouting "*jiāyóu*" ("add oil" in Cantonese, an expression of encouragement like "stay strong") out their high-rise windows, to raise morale (Wang 2020). And as the virus hit Europe and the death toll there similarly drove populations inside and away from each other, social and news media around the world transmitted videos of chanting in Wuhan now intended to articulate solidarity with those in Italy. Italians themselves chanted "andrà tutto bene" ("everything will be fine"), and videos showing them singing patriotic songs, distanced, from their balconies filled Twitter (Kearney 2020) – people calling to each other, like parrots separated from their flocks, to keep in contact, to keep up their spirits, raising voices to remind each other that no one is alone.

Shortly after these sounds and images spread, 53 members of a choir in Washington State became ill with COVID-19, and two died. The March 10 rehearsal attended by an unknowingly infected singer was later described as a "superspreader" event (Hamner et al. 2020; Johnson 2020), and the Centers for Disease Control and Prevention (CDC) in the U.S. briefly posted guidelines for faith communities that included a warning recommending suspension of choral singing and a statement that "the act of singing may contribute to transmission of COVID-19, possibly through emission of aerosols." After debate with the White House, where President Trump's administration did not recommend the interruption of in-person worship, the CDC removed those statements. But choirs at schools and faith institutions everywhere went online, members recording individually with their voices edited together by enterprising directors. Screens with a patchwork of faces and shoulders, microphones and headphones, have become standard, and though many are pleased with the results, singers and conductors nevertheless lament the loss of simultaneous sonic creation. Some have invented alternative modes of rehearsing and performing while socially distanced. In early 2021, National Public Radio (NPR) in the U.S. reported on "car concerts," developed by voice instructor David Newman at James Madison University in Virginia, which allow singers to sit separately in their vehicles with wireless microphones, a mixer, and an FM transmitter (Kravinsky 2021 and Newman 2020), hearing each other's voices simultaneously without the delay inherent in Zoom gatherings.

It is fortunate that these technologies have been available to many during the pandemic. Though they cannot replace the kind of in-person interaction and music making that is widely preferred, they do make singing together accessible, and have been serving such a purpose since long before COVID-19 turned singing upside down. In 2017, I attended a virtual music festival co-organized by Kaeley Pruitt-Hamm, one of the singers I interviewed for my book *Multivocality*. It was called BedFest, and featured participants with chronic illnesses – who had all recorded videos from their beds. The festival organized dozens of entries submitted by musicians, visual artists, and activists. BedFest was created based on the precedent of Bedstock, a 2016 virtual concert in which celebrity musicians performed from beds as a way of supporting children with serious illnesses, who would not be able to attend a concert at a public site. But BedFest was something new in that it offered an online venue to musicians who were themselves too ill to perform at a physical concert venue. 79 videos were posted on YouTube, from musicians in the U.S. and Puerto Rico, the U.K., Canada, Australia, Spain, Germany, Norway, Sweden, and the Netherlands. They mostly had been made by women, and spanned all age groups. The festival began with a conference video call for over 100 participants, using BlueJeans Video

Communications, and the artists spoke about their entries and experiences. One entry came from a UK group called Chronic Creatives Choir, whose artists' page at the festival described it as comprising seven singers, flute, and harmonica. The ensemble's bio reads:

> We're a choir made up of people with chronic illnesses (most of us have M.E.) [Myalgic Encephalomyelitis]. Although we're not well enough to join ordinary choirs, we've been able to make music together long-distance. We each record our parts at home separately, in some cases from bed. Some of us have to record in short sessions due to the illness, rather than doing the whole song at once. I then combine all the parts on the computer.[1]

The 21st century emergence of internet communities focused on shared health conditions has modified the experience of illness for many, so that it is often more public – at varying levels – than private (Conrad et al. 2016). And the possibility of listening ears, even digitally mediated as at BedFest, inspires community members to use their material voices in an agentive way.

In 2021, virtual concerts are routine and expected, an example of a phenomenon disability activists have noted: for years, disabled people have been denied work and cultural access through technological accommodation, and the pandemic has demonstrated that such accommodation was possible all the time. It remains to be seen whether the accessibility provided to broader populations during the era of COVID-19 will remain for disabled people when the virus becomes less of a threat.

While voices have been brought together through new technological devices or new uses of older technologies, voice has also become the focus of research looking for sonic indicators of COVID-19 infection. The diagnostic study of voice is not new in itself – some illnesses, like Parkinson's Disease, can impact voices in obvious ways and have been the subject of many studies. But the invention of new technologies to detect and model voiceprints of diseases and disorders – dementia, depression, and more (Anthes 2020) – is accelerating. And in the early months of the COVID-19 pandemic, medical researchers and practitioners have noted many potential ways to involve the smartphone, a widespread technology, in home health monitoring for the virus. In addition to the development of voice-assistant apps for symptom assessment and remote monitoring of temperature and lung function (Behar et al. 2020), the year 2020 saw the study of unique acoustic properties in the voices of COVID-19 patients. The Vocalis Health Company (headquartered in Israel and the U.S.), along with the Israeli Defense Fund and multiple academic groups, have been collecting and analyzing the sounds of breathing patterns, coughs, and speech in

people with and without COVID-19. Vocalis had already invented a smartphone app to identify the sounds of obstructive pulmonary disease, and set out to develop a similar tool for COVID-19 after the pandemic began (Anthes 2020). Though the study must be supplemented with further investigation, the researchers write that the results of their work so far "have demonstrated the feasibility and effectiveness of audio-and-text-based COVID-19 analysis, specifically in predicting the health status of patients" (Shimon et al. 2021, p. 1123).

Another project aimed at the identification of vocal biomarkers is the Voiceome Study, created by NeuroLex CEO Jim Schwoebel. In his Indiegogo campaign page, he explained that he was motivated to establish NeuroLex when his brother, after years of symptoms, had a psychotic episode and was diagnosed with schizophrenia. Schwoebel listened to voicemails his brother had left him in the years leading up to his hospitalization and began to see them as a data set that might be analyzed for vocal signs of disorder. Schwoebel wrote that his project

> aims to commercialize a universal voice test to refer patients to specialists faster. You can think of our company kind of like a Quest Diagnostics but for speech tests in the cloud. In this way, patients like my brother could be diagnosed and treated earlier, leading to better outcomes (e.g. lower duration of untreated psychosis) and lower costs ($15k→$3-5k) through more informed referrals (e.g. primary care physicians to psychiatrists).[2]

According to the Voiceome Study's website in early 2021, the project aimed to collect 10,000-100,000 samples of speech by December 1, 2020, to "create the largest dataset in the world of voice data tied to health traits and publish our work openly in peer-reviewed journals."[3] Though results have not been published yet, the Voiceome project stands to demonstrate new potential uses of voice analysis in many areas of health.

The Voiceome Study collects voice donations, and it is not the only project to do so. The company VocaliD has built a voice "bank" from donated speech samples, which it uses to develop prosthetic voices as assistive technology. According to VocaliD's origin story as told by their website, speech scientist Rupal Patel attended a conference in assistive technology and noticed that hundreds of people, regardless of age or gender identity, were using speech devices with the same voices, the only ones available. "We wouldn't dream of fitting a little girl with the prosthetic limb of a grown man, why then was it okay to give her the same prosthetic voice as a grown man?", the site asks.[4] As explained in Patel's 2014 Ted Talk (Patel 2014), the voice (speech) samples donated to the bank are modified acoustically and if necessary

combined to tailor a digital voice for an individual client – one that matches the client's sense of identity. The digital voice acts as a prosthesis in the sense that it is replacing something that may or may not have ever existed but is expected to exist, in an ableist world that prioritizes spoken communication. Sara Jain has asked, though, whether the voice that sounds from a throat might not also be considered prosthetic, "a device (trained, disciplined, accented, and pitched through many screens of personal, educational, and cultural intervention) through which agency is established, communicated, asserted" (Jain 1999, p. 41). Jain cautions against the overuse of the "prosthesis trope" in critical theory, reminding readers that prostheses may not only carry the connotation of replacing something understood as missing, but also reinforce the ideological imagination that something is missing in the first place, thus disabling bodies and fulfilling their perceived need at the same time.

VocaliD's work also raises important questions about voice and what Jonathan Sterne, after Jacques Derrida, calls the "metaphysics of presence," in which voice is taken as an indication of subjectivity, of uniqueness in the world. Sterne writes about his own experience that challenges those ideas. Using a personal voice amplifier (which he semi-affectionately calls "the dork-o-phone") after developing vocal cord paralysis following surgery, he recognizes that "my voice and its relationship to my subjectivity vary day by day, and sometimes by the hour." He continues: "If a single subject like me has *voices*, how can there be a single 'the voice' to theorize?" (Sterne 2019, p. 180). For Sterne, his vocal prosthesis complicates the social model of disability, which has long been critiqued for neglecting the realities of impairment. And the technology means that for him, each speech act "raises anew the relationship between intent and expression, interiority and exteriority" (Sterne 2019, p. 187).

4 Voices in Music

In spite of the ubiquity of sonic and visual enhancements in popular music, it remains a site of intense discourse about bodies, voices, and technologies. Perhaps no vocal prosthesis has received more public analysis or critique than Antares Audio Technologies' Auto-Tune software. Since its release in 1997 as a pitch-correction tool, it has sowed seeds of distrust among music consumers. It is still used for its original purpose, particularly among men producers working with women singers – Catherine Provenzano's work has supported the idea that women are disproportionately the targets of Auto-Tune for "cosmetic" or vocal enhancement purposes, and are criticized for it whether or not such application is

done with the artist's consent. Provenzano argues that men who use Auto-Tune are "afforded 'artistic,' 'creative,' and 'emotional' authenticity that Auto-Tuned female voices are rarely given" (Provenzano 2019, p. 65). And Robin James has noted the historically rooted positioning of both "femininity and technological progress as sites of emasculation and passivization" (James 2008, p. 416). She also examines the use of Auto-Tune by Black women artists, such as Rihanna, as part of a 21st century Afro-Futurist aesthetic that serves as a way of "reverse-engineering the body, using music to rewire the way whiteness and patriarchy are programmed into [Black women's] bodies and structures of feeling" (James 2008, p. 419). Auto-Tune and other voice manipulation technologies also haunt the vocal soundtracks of science fiction films. Cornelia Fales has written about the alienness of the modified Diva's voice in *The Fifth Element* (F 1997, D: Luc Besson), which produces sounds both coded as human and as non- or perhaps post-human, as indicated by the sudden onsets in rapidly sung melismas (Fales 2005). Similar sonic alterations feature in A.R. Rahman's soundtrack for the unprecedentedly expensive 2010 Tamil-language Indian film *Robot* (*Enthiran*, 2010, D: S. Shankar).

Amid a trend that finds popular musicians donning prosthetic makeup and CGI in temporary body modification both in and beyond music videos and staging – The Weeknd, for example, in an elaborate series of appearances meant to suggest facial surgery for cheek implants that he did not actually undergo (Nugent 2021) – some artists are turning to narratives of even more extreme technological prosthesis. In 2014, T-Pain was part of a joke video on the tabloid television program *Dish Nation* about developing an Auto-Tune implant, and in January 2019 Atlanta hip hop artist Nessly claimed in a viral video that he had had the technology implanted in his arm, seemingly showing it off and using it[5]. Antares even posted the video on its website. Later in the year, the site NewLevelNews.net featured a story on digital media artist Nicholas King (known as Nickels), who had also posted the video in January, and attributed Nessly's viral video to his work (Nickels (Nicholas King) is the Artist 2019).

It is becoming more common to combine attributes from multiple voices in sampling libraries to synthesize a new, unique voice – but libraries built on a single human voice still exist, marketed to professional and amateur composers who for various reasons need to make demo recordings without hiring live singers. In 2013, my friend Nichole Dechaine was hired to be the voice of such a sampling library, the soprano in Soundiron's choral-themed *Voices of Rapture* package. "I sold my voice," she told me a few months later (personal communication, 10 June 2013). An in-demand singer and vocal instructor in California, Dr. Dechaine had signed a contract that allowed the use of her voice in the software package, which provides the building blocks of vocal music to composers. She recounted to me how she

worked for three long days in a Bay Area recording studio singing through all twelve keys and her entire range, producing sets of single pitches sung on vowels, melodic phrases without text, and phrases sung or spoken in English, Latin, and French. She improvised and sang pre-written phrases for hours at a time, exercising her pianissimo and forte, vibrato, straight-tone, all the legato intervals between half step and octave, chromatic scales, and whole-tone scales. Now, for $119, anyone can essentially check out her voice from the sampling library – an acousmatic voice detached from her performing body – and manipulate its 17,310 samples to produce music that she's never heard, let alone sung herself.

The producers at Soundiron separate their library from others like the widely used voice synthesis software Vocaloid, explaining that their package features "true legato" – the sampling of all those intervals Dr. Dechaine recorded leads to what they hear as a smoother, more human sounding transition between pitches – rather than the choppy onsets Fales (2005) identifies in her work on the *Fifth Element* Diva as sonic indicators of the non-human. In contrast, because the voices packaged by Vocaloid capitalize on, and in general do not disguise their origins in technological alteration, transitions between phonemes, as well as onsets and offsets, may be perceived as mechanical or robotic. Vocaloid allows users to plug a voice into any lyrics, a function not offered by *Voices of Rapture* – so composers using the latter write a lot of vocalizations on vowels, and use only words within the limitations of the pre-packaged poetry Dr. Dechaine sang during her recording sessions.

In our first discussion about *Voices of Rapture*, Dr. Dechaine was anxious about the use of some sounds that had been sampled without her knowledge. Soundiron's web page for *Voices of Rapture: The Soprano* features different tracks created by thirteen composers.[6] In them, Dr. Dechaine is heard above digital piano or digital orchestra, or in a multi-voiced choir of her layered voice. But one example featured the sounds of her clearing her throat, singing in manipulated layers with an artificially wobbly vibrato, whistling, and at the end, laughing. The first time she listened to the sample compositions, she was surprised to hear not only her singing and speech, but also these other incidental sounds: "That's not something that I thought would be included in the library," she told me.

> I thought it would just be the exercises that they asked for, and the improvisations, minus, you know, where it was an obvious sort of outtake. In another composition there's a – a composer used an excerpt where I'm improvising, and I had to clear my throat, I had phlegm, and so – my voice didn't really crack, but you can hear that there's something, there's a flaw, right? So I had stopped in the recording and I thought, "oh, they'll edit that out," made an

> assumption, and no, and I guess the composer liked that sound or that quality and he used it in the composition. So that was surprising. (Interview, 7 July 2013)

When I asked her whether she felt that those sounds were *part* of the voice, she said, "No, I felt that was just me being a human." And she thinks about the use of her laugh as something even more personal than the sampling of her singing:

> Because I think it reveals more of my personality – it's attached to my personality, and to my own expression as a human and not as a singer, where my voice, you know – I am attached to it, it is my identity, but I'm kind of used to sharing it with others, and being paid to share it with others [laughs], where I haven't – I've never been paid to laugh [laughs] or to give that up for someone to use in a way that I wasn't expecting.

Later, she added that although she has a doctorate in vocal performance, "I haven't had any *training* in throat clearing."

The development team behind Soundiron – Mike Peaslee, Gregg Stephens (editors and recordists), and Chris Marshall (programmer) – consider the non-singing sounds they sampled not only as signs of human-ness, but also of a particularly living sound, and of Dr. Dechaine's distinctive identity. Mr. Peaslee told me:

> All those pieces are so integral to a vocal performance. They're used a lot in pop recordings and modern recordings. They're the things you kind of would exclude from a symphonic recording, usually, because they're – you'd consider them impurities, but those are things that… to make it sound like it was convincingly sung by a live performer, those breaths need to be there. A user might put them really low in a mix, but just before a line, you know, that can add tension, it can add weight to the line that's about to be sung. So that's – in that way, we actually include a section of playable breaths with each subset of phrases and sustains that we offer in the different presets within the library, so that they're always right there ready for you… And it just – it makes it that much more alive. Throat-clearing and all that stuff – some of it's kind of for fun, but a lot of it really is also – you know, people use it. It gives that much more life. It adds that much more personality to it. (Interview, 11 July 2013)

In *A Voice and Nothing More*, Mladen Dolar discusses a famous case of hiccups that interrupts Aristophanes during a speech in Plato's *Symposium*. Regarding the hiccup, Dolar theorizes that

> the involuntary voice rising from the body's entrails can be read as Plato's version of mana: the condensation of a senseless sound and the elusive highest meaning, something which can ultimately decide the sense of the whole. This precultural, non-cultural voice can be seen as the zero-point of signification [...] the point around which other – meaningful – voices can be ordered, as if the hiccups stood at the very focus of the structure. The voice presents a short circuit between nature and culture, between physiology and structure; its vulgar nature is mysteriously transubstantiated into meaning *tout court*. (Dolar 2006, p. 25-26)

The inclusion of non-singing sounds in the *Voices of Rapture* library can be understood to offer such a central meaning, across genres, providing tiny sites for the intersection of the human and posthuman.

Our initial talk about Dr. Dechaine's vocal venture occurred a few months following her recording sessions. She had not given the process much thought again, until right before our conversation, when she had listened to the sample compositions on the *Voices of Rapture* website. Now, she was nonplussed, hearing the manipulation of her voice outside of her body. I thought of Eidsheim's work on the marketing of Vocaloid's voice providers and joked that there must be a support group for singers who sell their voices to sampling libraries. Dr. Dechaine mused that if there were one, it would probably be "open to prostitutes, too." Her response was not meant to condemn sex work or treat trafficking lightly, but it alludes to parallels between the two situations that might arise in the complications of selfness when an embodied voice or a body are commodified. In the 21st century, when a number of technologies are capable of manipulating the acoustic materials of voices, such anxieties are not uncommon. In 2017, for example, an online AI service named Lyrebird (after the most renowned imitator of the avian kingdom) was announced, with the purpose of learning the acoustic structures of a speaker's voice, and then producing a "recording" of that voice speaking any words entered as text. News media immediately posted panicked headlines, such as: "Lyrebird Steals Your Voice to Make You Say Things You Didn't – And We Hate This World" (Claburn 2017). Though at least in part meant to be tongue-in-cheek, that headline positions Lyrebird as though it will lead to the kind of identity theft feared by internet users worldwide and to a mutilation of self that resembles a kind of metaphorical mutilation of the body.

And the body is not the only thing at stake. "I do feel like I sold a piece of my soul or something," Dr. Dechaine told me (interview, 10 June 2013). Some of the most persistent stories in Western cultures deal with the metaphorical location of soul or

identity in the voice, and the idea that though a voice seems to be *in* a body, it is never quite *of* a body – it is understood as something spectral enough to be stolen from its bodily house, or corrupted, or even transplanted into another body. Ovid's tale of Echo (in the *Metamorphoses*) begins when the nymph is punished by Juno for talking incessantly to stall her while other nymphs, who have lain with Juno's husband (and brother!) Jupiter, flee the scene. Echo's loquaciousness triggers Juno to punish her by limiting her vocal freedom severely, so that Echo can only ever after repeat what others say. The shame of being unable to control her own voice, especially in pursuit of her love interest Narcissus, causes her body to waste away. Her bones turn to stone, and Echo only remains as a disembodied voice among those stones, inert and only set to motion by – and always subject to – the force applied by others. Ovid describes this condition, in the words of A.S. Kline's translation, as meaning that Echo is "no longer to be seen on the hills, but to be heard by everyone. It is sound that lives in her."[7] She is condemned to be the voice of, and to be heard by, anyone who passes. Dolar writes that the nymph's voice "continues to echo our own voice, the voice without a body, the remainder, the trace of the object." And this very echo is in consistent opposition to the voice of self-presence and self-mastery, he continues, "the intractable voice of the other, the voice *one could not control*" (Dolar 2006, p. 40, my emphasis).

Dr. Dechaine has sometimes contacted me when she thinks she has heard her voice in a television soundtrack or video game, but she is never entirely sure. In this way, she is always looking over her shoulder, listening for the life her own voice is living without her, a vehicle driven by anyone who pays for a license. Her feelings are complicated, though. She appreciates the potential for a kind of immortality. Dr. Dechaine says: "I like that long after I am too old to sing, I can still use my voice and so can my kids." Michal Grover-Friedlander writes that "sound recordings, at times, are voices surviving the body that once produced them; invisible and devoid of body, the singer is somehow there in the presence of voice" (Grover-Friedlander 2005, p. 7). But sampling libraries like *Voices of Rapture* offer something beyond the prospects of simple recording. If *Voices of Rapture* were to persist long enough, it is possible that the vibrations of Dr. Dechaine's voice might continue singing after her death, the echo of the voice *within* her resonating among the stones *without* her.

5 Conclusion

All of these projects tell stories about what the posthuman is, and what it means in 2021. Voice studies, as a broad, interdisciplinary field of inquiry, is increasingly fascinated with what voices do for humans, and what humans do with voices to build,

maintain, and disrupt systems of power. The gendered uses of Auto-Tune and similar applications, for example, and the ways artists and voice designers are employing them to subvert that gendering, push and pull at longstanding normativities. The recent determination to create un-embodied AI servants as voices we can control might be less indicative of creative ingenuity than of a continued global dependence on the processes of colonization and class division that established the practices of servitude in the first place, with AI voice assistants functioning as a kind of metaphorical methadone replacement for societies trying to kick the habit. The mining of voices for medical information offers a path toward minimizing the use of some invasive and expensive diagnostic techniques, and it will be interesting to see how medical corporations respond to such potentially democratizing effects. And if singers can contract for labor that their voices will go on doing without them, who is in control of the formerly embodied commodity? This article has offered these examples not only to give an overview of recent work but also to encourage the further investigation of 21st century voices.

References

Anthes, Emily (2020): Alexa, Do I Have COVID-19? Nature.com. September 30. https://www.nature.com/articles/d41586-020-02732-4 [last accessed July 26, 2021].

Behar, Joachim; Liu, Chengyu; Kotzen, Kevin; Tsutsui, Kenta; Corino, Valentina D. A.; Singh, Janmajay; Pimentel, Marco A. F.; Warrick, Philip; Zaunseder, Sebastian; Andreotti, Fernando; Sebag, David; Kopanitsa, Georgy; McSharry, Patrick E.; Karlen, Walter; Karmakar, Chandan; Clifford, Gari D. (2020): Remote Health Diagnosis and Monitoring in the Time of COVID-19. In: Physiological Measurement. 41/10. DOI: 10.1088/1361-6579/abba0a.

Carpenter, Julie (2019): Why Project Q is More than the World's First Nonbinary Voice for Technology. In: Interactions. 26/6. Pp. 56-59.

Conrad, Peter; Bandini, Julia; Vasquez, Alexandria (2016): Illness and the Internet: From Private to Public Experience. In: Health. 20/1. Pp. 22-32. DOI: 10.1177/1363459315611941.

Dolar, Mladen (2006): A Voice and Nothing More. Cambridge, Mass.: MIT Press.

Eidsheim, Nina S. (2008): Voice As a Technology of Selfhood: Towards an Analysis of Racialized Timbre and Vocal Performance. Ph.D. thesis. University of California, San Diego.

Eidsheim, Nina S. (2009): Synthesizing Race: Towards an Analysis of the Performativity of Vocal Timbre. In: Trans. 13. http://www.sibetrans.com/trans/articulo/57/synthesizing-race-towards-an-analysis-of-the-performativity-of-vocal-timbre [last accessed July 26, 2021].

Eidsheim, Nina S. (2015): Sensing Sound: Singing and Listening as Vibrational Practices. Durham, London: Duke University Press.

Fales, Cornelia (2005): Short-Circuiting Perceptual Systems: Timbre in Ambient and Techno Music. In: Wired for Sound: Engineering and Technologies in Sonic Cultures. Paul D. Green; Thomas Porcello (eds.). Middletown, CT: Wesleyan University Press. Pp. 156-180.

Foucault, Michel (1988): Technologies of the Self: A Seminar with Michel Foucault. Luther H. Martin; Huck Gutman; Patrik H. Hutton (eds.). Amherst, U.S.: University of Massachusetts Press.

Grover-Friedlander, Michal (2005): The Afterlife of Maria Callas's Voice. In: Musical Quarterly. 88/1. Pp. 35-62.

Hamner, Lea; Dubbel, Polly; Capron, Ian; Ross, Andy; Jordan, Amber; Lee, Jaxon; Lynn, Joanne; Ball, Amelia; Narwal, Simranjit; Russell, Sam; Patrick, Dale; Leibrand, Howard (2020): High SARS-CoV2 Attack Rate Following Exposure at a Choir Practice – Skagit County, Washington, March 2020. In: Weekly. 69/19. Pp. 606-610. https://www.cdc.gov/mmwr/volumes/69/wr/mm6919e6.htm [last accessed July 26, 2021].

James, Robin (2008): 'Robo-Diva R&B': Aesthetics, Politics, and Black Female Robots in Contemporary Popular Music. In: Journal of Popular Music Studies. 20/4. Pp. 402-423.

Johnson, Carla K. (2020): US Choir Outbreak Called 'Superspreader Event' in Report. ABCNews.go.com. May 12. https://abcnews.go.com/Health/wireStory/us-choir-outbreak-called-superspreader-event-report-70642547 [last accessed July 26, 2021].

Kearney, Christine (2020): Italians Sing Patriotic Songs from Their Balconies During Coronavirus Lockdown. In: The Guardian. March 14. https://www.theguardian.com/world/2020/mar/14/italians-sing-patriotic-songs-from-their-balconies-during-coronavirus-lockdown [last accessed July 26, 2021].

Kravinsky, Nina (2021): Car Concerts Offer Choirs a Way to Rehearse and Perform. NPR.org. January 11. https://www.npr.org/2021/01/11/954007807/car-concerts-offer-choirs-a-way-to-rehearse-and-perform?fbclid=IwAR25MaErAT7F2B7a5KGykaLPjff8OO-aulzO-MshDtoom_dBF_WLHZVtgUM [last accessed July 26, 2021].

Newman, David (2020): Physically Distant 'Drive-In' Ensemble Rehearsal. YouTube.com. May 17. https://www.youtube.com/watch?v=CPD7kxr003w [last accessed July 26, 2021].

No Author (2019): Nickels (Nicholas King) Is the Artist Behind the Viral Crazy Realistic Videos You Have Seen. NewLevelNews.net. October 10. https://newlevelnews.net/nickels-nicholas-king-is-the-artist-behind-the-viral-crazy-realistic-videos-you-have-seen/?fbclid=IwAR1bwLVX1lbln8q7fS4YA9OIBjN1qLKlhIlKCxgAyyu4S2AKwh2olUzuZwU [last accessed July 26, 2021].

Nugent, Annabel (2021): What Happened to The Weeknd's Face? Singer Removes Bandages to Reveal 'Creepy' Plastic Surgery Prosthetics for New Video. In: Independent. January 6. https://www.independent.co.uk/arts-entertainment/music/news/the-weeknd-plastic-surgery-prosthetics-b1783070.html [last accessed July 26, 2021].

Patel, Rupal (2014): Synthetic Voices, As Unique As Fingerprints. YouTube.com. February 13. https://www.youtube.com/watch?v=d38LKbYfWrs [last accessed July 26, 2021].

Provenzano, Catherine (2019): Making Voices: The Gendering of Pitch Correction and the Auto-Tune Effect in Contemporary Pop Music. In: Journal of Popular Music Studies. 31/2. DOI: 10.1525/jpms.2019.312008.

Ravitz, Jessica (2013): 'I'm the Original Voice of Siri'. Cnn.com. October 15. https://www.cnn.com/2013/10/04/tech/mobile/bennett-siri-iphone-voice [last accessed July 26, 2021].

Schlichter, Annette (2011): Do Voices Matter? Vocality, Materiality, Gender Performativity. In: Body & Society. 17/1. Pp. 31-52.

Schwoebel, Jim (n.d.): The Voiceome Study. Indiegogo.com. https://www.indiegogo.com/projects/the-voiceome-study#/ [last accessed July 26, 2021].

Shimon, Carmi; Shafat, Gabi; Dangoor, Inbal; Ben-Shitrit, Asher (2021): Artificial Intelligence Enabled Preliminary Diagnosis for COVID-19 from Voice Cues and Questionnaires. In: Journal of the Acoustical Society of America. 149/2. Pp. 1120-1124. DOI: 10.1121/10.0003434.

Sterne, Jonathan (2019): Ballad of the Dork-o-Phone: Towards a Crip Vocal Technoscience. In: Journal of Interdisciplinary Voice Studies. 4/2. Pp. 179-189.

Wang, Wallis (2020): Wuhan Residents Chant 'Keep It Up, Wuhan' out of Their Windows to Boost Morale. In: South China Morning Post. January 28. https://www.scmp.com/video/china/3047949/wuhan-residents-chant-keep-it-wuhan-out-their-windows-boost-morale [last accessed July 26, 2021].

Wolfe, Cary (2010): What is Posthumanism? Minneapolis: University of Minnesota Press.

Notes

[1] The artist's page does not indicate which choir member wrote the description, and the "I" is unattributed (https://www.bedfest.meaction.net/chronic-creatives-choir [last accessed July 20, 2021]).

[2] https://www.indiegogo.com/projects/the-voiceome-study#/ [last accessed July 20, 2021].

[3] https://www.voiceome.org/ [last accessed July 20, 2021].

[4] https://vocalid.ai/about-us/ [last accessed July 20, 2021].

[5] Nessly. January 12, 2019. https://www.youtube.com/watch?v=ZG5e8i5AugQ [last accessed July 20, 2021].

[6] https://soundiron.com/products/voice-of-rapture-the-soprano [last accessed July 20, 2021].

[7] http://ovid.lib.virginia.edu/trans/Metamorph3.htm [last accessed July 20, 2021].

Lílian Campesato, Fernando Iazzetta

Voice as a Resonance of Listening

Abstract: In diesem Beitrag wollen wir bildliche Darstellungen der Stimme in verschiedenen Kontexten untersuchen. Ausgehend von der Resonanz im Sinne eines Elements, welches sowohl das repräsentiert, was den Klang erzeugt, als auch das, was vom Klang beeinflusst wird, werden wir einige Beispiele dieses Prozesses der Bilderzeugung durch Stimme im künstlerischen, technologischen und kommunikativen Kontext diskutieren.

Abstract: In this article, we are interested in exploring the voice's imagetic representations in different contexts. Starting from the idea of resonance as an element that represents both what produces sound and what is affected by sound, we will address some examples of this process of image production through voice in the artistic, technological and communicational contexts.

1 Voice and Listening

Vibration is a condition of what is in motion, of what is alive. It is the movement of a mechanical body in relation to a state of equilibrium. Everything that moves acts on its surroundings, creating frictions, resonances, shocks, impulses. Hence the importance of vibration for us to exist: It is an index of our existence, as well as of our relationship with everything that permeates our surroundings. Like most living beings, we learn to vibrate with the world as a way of interacting with it. And being mechanical, the vibration invokes the action of, and on, our bodies: It is an indication of presence. Our senses can perceive vibration in two dimensions. One, relatively slow, involving the movement of large bodies that make us tremble. Another, very quick and subtle, is unable to activate the tactile sensors spread throughout our bodies, but manifests itself in the form of what we call sound.

Sound is a very fast, but very weak, pressure variation that propagates through some material medium. Our bodies are equipped with a very precise system for both producing and perceiving sounds. Although phonatory and auditory mechanisms are usually treated as independent, autonomous devices, they act in a complementary way in our interaction with the vibrational world. Using one's voice and listening are

interdependent forms of action and there is no way to fully understand one without considering the other. This close connection between voice and listening can be approached in several ways. It implies the correspondence between the speech and hearing apparatuses, the interrelationship between individuals in communication processes, and the various forms of interaction between bodies and the environment.

In this article, we are interested in exploring the imagetic representations of voice in different contexts and exploring its connection with hearing. Starting from the idea of resonance as an element that represents both what produces sound and what is affected by sound, we will address three examples of this image-making process through voice in artistic, technological and cultural contexts: a 'silent' sound work by Christian Marclay; the invention of virtual popstar Hatsune Miku; and the unusual record of a 'mythical' character in Brazilian popular culture.

We take voice as a relational and multimodal process to emphasize its strong connection with the physical, corporal, and material domains. A voice represents a body, a space, and evokes a series of poetic, symbolic and affective instances. It is a trait, a mark, a sign. A voice is always someone's voice, or the voice of something. As Don Ihde points out, "all sounds are in a broad sense 'voices', the voices of things, of others, of the gods, and of myself" (2012, p. 147). On the other hand, voice is a powerful source of images:[1] It has the capacity to represent everything – histories, memories, bodies, spaces and times – that resonates when it sounds. However, this whole process always depends on listening, because without listening, the voice is mute.

By the middle of the last century, Alfred Tomatis, a French physician who dedicated himself to studying the relationships between voice, hearing and cognitive processes, speculated about the feedback chain involving voice and listening: "the voice only reproduces what the ear hears" (Tomatis 1992, p. 49). Analyzing some singers' disorders that prevented them from emitting certain sounds, Tomatis studied their auditory curves and realized that they "showed a hearing loss at the same frequency level" (Tomatis 1992, p. 41). Based on empirical inferences, not always demonstrated with scientific support,[2] Tomatis perceived a causal relationship between the composition of the listening and vocal spectra, leading him to affirm that "a person can only reproduce vocally what he is capable of hearing" (Tomatis 1992, p. 44). Or, stated in another way: "One sings with one's ear" (ibid). Controversial in his methods and treated with reservations by his peers, Tomatis' work had received enough recognition to attract the attention of renowned artists such as Maria Callas, Sting and Gérard Depardieu, who sought him out to solve their vocal shortcomings.

The connection between voice and listening envisioned by Tomatis should not be overlooked. Particularly when considered from an evolutionary point of view, it is assumed that voice and listening have developed in an integrated manner, at least for vertebrates, due to the relevance of sound communication among members of the same species. As Carl Gans suggests, when individuals acquire the ability to perceive sound patterns within the environment, they begin to use this same energy channel to transmit information. "This generates interest in producing the sounds for detection by conspecific organisms" (Gans 1992, p. 9). This process is modulated by the environment, which propagates, reflects, absorbs and modifies sounds in different ways. In evolutionary terms, sound perception and production broaden the possibilities of interaction between individuals, as well as between individuals and the environment. But this process does not happen without costs. For example, a successful communication between individuals of the same species depends on tuning the sound patterns exchanged between sender and receiver. The energy expended in this process represents costs of various kinds, including "a risk of informing predators of one's existence and location" (idem, p. 10).

It seems reasonable to imagine that sounds vocalized by an individual should be adequately perceived by a fellow member of their species or by themselves, although this cannot be generalized. Exceptions, even when they serve to confirm the rules, are always instructive. It is known, for example, that the auditory system of toads and frogs has developed in a peculiar way and that many anurans do not even have a tympanic membrane like ours. Their hearing is produced by two organs, the amphibian papilla, sensitive to low and mid-frequencies (typically 50 Hz to 1 kHz), and the basilar papilla, sensitive to higher frequencies (above 1 kHz) (Goutte et al. 2017, p. 1). Even in anurans where tympanic middle ears are not present, sounds are sent to the inner ear through cavities or bones, making the animals able to hear the frequencies of their own calls.

But in some anuran species such as Brachycephalus ephippium and Brachycephalus pitanga, the poorly developed auditory system is apparently unable to capture the frequencies emitted by individuals of the species. This behavior poses a challenge to the conception that, in the evolutionary process, actions without a purpose represent a waste of energy and tend to be dampened. In a recent study, this behavior was interpreted as an aspect of a particular moment of evolutionary transition, in which the 'silent' calls have not yet been overcome, despite the ineptitude of the ears of that species. Since "signal production is energetically costly and sound may attract predators and parasites" (Goutte et al. 2017, p. 4), these high-pitched calls would have been maintained as a side effect of the visual aspect that accompanies the production of these sounds due to an "evolutionary inertia" (ibid).

One might speculate that not all of our actions (and not all of the frogs' actions as well) can be fully explained in terms of evolutionary features, such as scaring off predators or attracting sexual partners. After all, it is not because we ever murmured an imaginary song during a lonely walk that we would be going against the evolutionary process of the human species. Feeling the vibration of our vocal folds, creating melodic waves that resonate in our heads and thorax may be regarded as a pleasing and powerful way of connecting ourselves with our own bodies and with the world around us. This is by no means a waste of energy.

2 Sound as Image

In terms of acoustics, sound is an oscillatory movement of molecules in a material medium. Different vibratory patterns would lead us to perceive different sounds. As a vibrational phenomenon, the occurrence of sound is independent of, although it is closely related to, our perception of it. When we hear Beethoven's music or someone's voice, our brain produces a sound image that is triggered by the oscillating motion of molecules surrounding us. But if we refer exclusively to the acoustic phenomenon, we may find it difficult to identify what we understand as Beethovenian or vocal aspects in these oscillations.

It can be argued that the word "sound" can refer to quite different meanings, which often ends up blurring the distinctions among fundamentally different aspects of sound production and perception. Take, for example, the classic question: If a tree fell in the forest, but no one heard, would there be a sound? This philosophical puzzle makes sense because we have a single term to represent two different phenomena: the acoustic vibrations caused by the falling tree and the perception of these vibrations. In fact, what we understand as sound is the result of different levels of interaction, among which mechanical vibration is just one. Indeed, if we were to use two distinct terms to indicate the acoustic movement of material particles and the auditory images they produce in our minds, we could dispel some misconceptions about sound. As we shall see in this text, we intend to consider sound not as the object of listening, but as its medium.

The relationship between the acoustic production of sounds and the aural images they generate is generally taken as a direct cause and effect relationship. In fact, the perception of sounds is modulated by several factors ranging from the sociocultural contingencies to the cortical processing of the received signals to the listener's emotional state. Context plays an important role in the way we understand sounds. On the other hand, sounds can considerably influence other senses and even direct

our actions. Just think about the way sounds are produced in cinema to understand how the interaction between vision, hearing and semantic contexts can guide our hearing: For example, in a film's sound design, recording the tapping on a block of wood will make the gallop of a horse more realistic than the sound recording of an actual horse.

To give another example, an adequate DJ playlist can successfully encourage more individuals to dance at a party, just as the appropriate choice of musical repertoire can help with the consumption of dishes in a restaurant. Recent research shows that a soundscape can enhance certain taste characteristics of food, making customers willing to pay more if the meals come with a satisfying sound experience (Carvalho et al. 2015). Other experiments have shown an association between sour-tasting foods with high pitches and bitter flavors with low pitches (Crisinel & Spence 2009).

Diana Deutsch, an eminent researcher of musical illusions, carried out an experiment that challenges the notion that voice and listening operate separately. Deutsch realized that much of what we hear is not only related to the acoustic signals that reach our ears, but to the context in which the listening takes place and to our individual aural experience. Although she was interested in listening illusions, and therefore in unusual situations of sound perception, some of Deutsch's experiments on speech comprehension show that in voice listening, "the words and phrases that we hear are strongly influenced not only by the sounds that reach us, but also by our knowledge, beliefs, and expectations" (Deutsch 2019, p. 104). In demonstrating what she calls phantom words, Deutsch uses recordings of repeating words in a loop, over and over again, through stereo loudspeakers. Some of these words are just two-syllable words repeated out of phase on both channels. After listening for a while to those repetitions, listeners tend to hear different words, "nonsense words, and musical, often rhythmic sounds, such as percussive sound or tones" (idem, p. 106)[3]. The most impressive aspect of her experiments is that individuals perceive words in different languages and even with different accents depending on their original language and culture.

In this text, we seek to highlight the relational character of voice in different dimensions. First, in its connection with listening, establishing a relationship between receiving sounds and producing sounds. Second, pointing out that this relationship is always mediated by a body. This is not an abstract, idealized body, nor a body reduced to its carnal, physiological condition: The body here represents the empirical existence of a subject – and, eventually, of a thing – in time and space. Third, this existence implies the articulation between subject and the world, not as two separate characters, but as two perspectives of the same process. Thus,

Fig. 1: *Chorus II* (1988), framed black and white photographs, 142 x 188 cm, by Christian Marclay.

the body, when vocalizing, puts its surroundings in resonance, producing echoes that return to the body itself. Both what is vocalized and what is heard belong to the same process. Likewise, body and environment operate as an interconnected system, sounding and resonating with each other.

3 Silent Voices

There are over 20 small frames containing black and white photos of open mouths silently singing or screaming. The pictures themselves are arranged on the gallery wall to form the outline of a large mouth. The photos, cut from different sources, allude to a great choir in which men and women of different races and times create a silent harmony. As in other works by artist Christian Marclay, *Chorus II* (1988; see Fig. 1) explores the expressive potential of sound, not from its acoustic components,

but from references that are built from sound images. *Chorus II* allows us to listen with our eyes and to understand that our senses do not work autonomously and isolated, but develop as a network in which stimuli, memories, experiences, actions and reactions are interconnected. The work also indicates the material dimension of our sensations. Seeing or hearing do not refer to abstract categories of interaction with the world, but are ways in which our bodies act on the world at the same time that they are constituted by what, directly or indirectly, acts on us. Despite the absence of physical vibrations, *Chorus II* does sound: Spectators build their listening from their experiences and memories of other voices, other mouths, other choirs.

Marclay ingeniously connects voice, body, and listening. The wide-open mouths in an almost aggressive attitude remind us of an identifiable, strong, powerful sound. We may internally hear a scream, a groan, or some sort of tense utterance. The image of the mouths is the image of the sounds of those mouths. As spectators we connect these two image categories based on our experience in listening and voicing. That is, every time a scream is heard, we immediately associate it with the energy that escapes from the body, with the muscle contractions in the vocal tract, with the sensation of the cheeks stretching to the limit of the skin. In short, the static image that represents a gathering of mouths invokes, albeit in an imaginary way, the world of vibrations.

The idea that these mouths have no bodies is perceived as a contradiction. We have learned through experience that vibration depends on the existence of a vibrating body that defines and is defined by a space. Or as Douglas Kahn reflects: “Vibrations through their veritable movement generated a structured space and situated bodies and objects in that space. This process of situating did not outwardly transform the bodies or objects themselves, however, it just placed them in an ever-dependent relation within a larger system” (Kahn 1992, p. 15). Thus, the voices in Marclay’s *Chorus II* are not disembodied ones. On the contrary, they forge their bodies as extensions of their mouths.

In some of Marclay’s silent objects the connection between speech and listening becomes evident. He has created different works exploring the conditions posed by telephone devices in which transmitter and receiver are mounted in the same block, making the regions of voice production and listening geographically close. In these works emerges what Kahn calls residual sounds. These are not sounds that we expect to hear from familiar sounding objects, but sounds that “remain closed secured” in the stillness of the objects (Kahn 1994, p. 23). Despite the limitations imposed by its “physical, phenomenal silence, [...] a residual sound may be incredibly raucous” (ibid).

Chorus II resonates the modern imagery of the voice mediated by all kinds of sound devices: From the first phonographs and radio, to cinema, to current digital music streaming, we are used to hearing disembodied voices and we have learned to recompose these bodies in our imagination. Samuel Becket's *Not I* (1972) is exemplary in this regard. Minimalist in form, the piece, in which only the actress's mouth is illuminated by a light spot, is an explosion of auditory images. It is from the voice that the audience composes the character. The absence of the body channels attention to the voice, and it is this attentive listening that triggers our experiences and expectations to build an image for the hidden character. On the other hand, the mouth, which remains illuminated, insists on reminding us of the presence of a vocalizing body.

To some extent, *Not I* predates the current listening condition in which sound sources are usually hidden behind the membranes of speakers, headphones, and other sound devices. French composer Pierre Schaeffer named this condition acousmatic listening. Schaeffer was especially interested in discussing the role of listening in relation to his proposal for a musique concrète, a composition produced from sounds collected, recorded and arranged on a physical medium such as a disk or magnetic tape. For Schaeffer, the acousmatic situation favored what he called reduced listening: By hiding the sound source, the listener could focus their attention on the sound qualities, disregarding all the references sound could provide (Schaeffer 1966). Obviously, it is impossible to forget that behind the speaker – the acousmatic curtain that hides the sound sources – there is, or once was, a source. It is the sources that the loudspeaker conveys; whether it is a vibrating mechanical body, the electronic circuitry of a synthesizer, or the abstractly generated bits in a musical software, we cannot get rid of what is concealed by the speaker. Our conviction that there is something on the other side of the loudspeaker prompts us to reconnect the sounding objects with their vibratory movements and the way they sound. We infer that every object has a sound, a voice, just as every sound, every voice, comes from an object, from a body. This association between sounds and the objects that produce them, however, is the result of our experience, our expectations, our history and the history of the objects that sound. During our life we repeatedly listen to the voices of things and we build, in our memory, an association between things and sounds. These associations can assume different natures. We know that the sounds produced by large bodies tend to be lower than those produced by small bodies. We also learn to repudiate certain sounds (the noise of the upstairs neighbor) and to fear others (the sudden rumble of thunder during a storm). In a similar way, we go into alert when we hear a fire engine siren or the beep indicating the arrival of a message on the cell phone. Despite the strong connection

between sound as a vibrational phenomenon and the aural images it may provoke, this connection is neither stable nor unique, but relational.

Nina Sun Eidsheim, an academic dedicated to critically exploring the possibilities of voice representation, refers to the acousmatic situation to highlight this relational and multidimensional character of voice. It is assumed that when listening to a voice, we can learn something about the speaker, even when they are not visible to us. In this acousmatic situation, we are urged to ask a fundamental question: Who is this? Who is speaking? This fundamental 'acousmatic question' is based on the premise that there is a direct and stable relationship between sound and its source and that if we pay attention to the sound of a voice, we will be able to recognize the speaker, learn about their personality and even about their mood. However, Nina Eidsheim argues that there is no stable answer to the acousmatic question and that it arises precisely because of the "impossibility that the question will yield a firm answer" (Eidsheim 2019, p. 3).

The impossibility of answering this acousmatic question comes from the fact that neither the voice nor the vocal tract are static systems: They are subject to physical, emotional and contextual conditions that involve both the vocalizer and the listener. They both operate dynamically in the construction of meanings that emerge from vocal production. The inferences a listener may produce regarding a speaker's physical, racial or gender characteristics based on their voice is part of a process in which vocalizer and listener are both involved. For example, gender may be signaled by vocalizers "through word choice, intonation, speed, rhythm, prosody, level of nuance" (idem, p. 6-7), while listeners will "bring gender expectations to the vocal scene" (ibid). Eidsheim summarizes this complex interaction between actors of voice production/reception and its context in three corrective statements: "Voice is not singular; it is collective. Voice is not innate; it is cultural. Voice's source is not the singer; it is the listener" (idem, p. 9).

Eidsheim takes her own experience to discuss the relational status of voice. Despite her Korean origins, she was raised in a small town in Norway where her Korean identity was never a salient issue. On the other hand, when she visited Seoul, she realized that people treated her as a foreigner, despite her Asian traits. Eidsheim recalls that during her singing training in Norway she "participated in master classes offered by well-known American voice teachers" (Eidsheim 2008, p. 28) and that they "had been puzzled by [that] Asian-looking girl who spoke Norwegian and who, to their surprise, possessed a signature Nordic classical timbre" (ibid). A few months later, this time in California, a teacher complimented her on the quality of her voice, adding that her timbre was "really quite characteristically Korean" (ibid).

Eidsheim's experience poses a question: If the timbre of a voice represents an identity, a signature, how could these different situations produce such different perceptions of her cultural and ethnic identity? Starting from her own experience, Eidsheim develops the argument that, contrary to the current idea that the voice is "an unmediated manifestation of the body" (idem, p. 30), it "is indeed mediated" (ibid). Therefore, voice perception is conditioned by these elements of mediation, which are not only physiological, but cultural, social, subjective and contextual. Eidsheim is particularly interested in deconstructing the idea that it would be possible to hear race from the timbre of a particular voice. In this case, vocal training – that is, the way we use our bodies and voices – would be more significant for the perception of vocal timbre than any physiological differences linked to specific ethnicities. Thus, although an uttered voice presents indices of the uttering body – as an individual (gender, social status, age, etc.) and as a spatial position (Bertau 2008, p. 101) –, these signs are constructed in an intersubjective way, among the individuals and based on their interactions. The reliability of these indices in revealing speaker characteristics has attracted the attention of many researchers (Pisanski & Bryant 2019). Jody Kreiman and Diana Sidits make a significant contribution to understanding how the voice may (or may not) offer clues to the recognition of characteristics such as physical size, sex, age, health, appearance, racial group or ethnic origin of a speaker (Kreiman & Sidtis 2011). The authors emphasize that it is necessary to distinguish 'learned' from 'organic' marks, as they separate what the speaker can modify from what is subject to their physiological and anatomical constitution (idem, p. 111). It is also important to distinguish between marks and stereotypes of speaker characteristics. While marks are generally "reliable cues to that characteristic", stereotypes are related to what "listeners expect to hear from a speaker who possesses certain physical attributes" and, therefore, "social expectation influences listener's judgments". At the same time, these stereotypes contribute to "vocal behaviors children learn as they grow" (ibid). Thus, associating voice with an individual's specific characteristic is not a trivial issue. For example, while the distinction between male and female voices can be achieved with some consistency, to transgender individuals, there is an important interplay between organic and learned characteristics. Since voice is an important index of an individual's social and personal characteristics, "producing a female (or male) sound with what remains a male (or female) vocal tract and larynx" (idem, p. 144) can be a challenging demand for individuals who have undergone transgendering surgery. Thus, what the voice can say about an individual depends on the balance between what is physically and physiologically determined in the production of voice and the intersubjective experiences that make up our listening

Fig. 2: Hologram of Hatsune Miku at a live concert.

processes. When we listen, we project our knowledge, our beliefs and expectations onto the speaker. Therefore, we always hear a little of ourselves in the 'other'.

4 Disembodied Voices

As is the case with the beginning of any pop concert, the band attacks the first chords, the lights come on, while fans wait anxiously for the entrance of the main character, the singer. In contrast to the other musicians in the band, in this case the singer is not a person, but an avatar projected holographically onto the stage. Her blue hair and high school student clothes make explicit reference to Japanese manga. Hatsune Miku reproduces the cliché image of what a popstar should be. Initially restricted to otaku[4] circles, she gradually attracted the attention of other musical circuits. Without assuming polemic attitudes or wearing extravagant clothes, Hatsune Miku[5] became an iconic representation of Japanese pop culture. Thanks to Vocaloid[6], a software that produces singing voices artificially, Hatsune Miku has become a virtual idol.

From extensive sound banks, Vocaloid allows its user to type in the lyrics of a song and synthesize it from a series of instructions. In essence, the Vocaloid interface superimposes the song text over a kind of musical score. A series of subtle adjustments can be applied to each vocal sound, allowing the creation of very sophisticated vocal articulations. Miku is probably the best known of a series of virtual artists produced with the help of Vocaloid. In part, her success is due to the fact that graphic projections added a visual image to the singer's well-behaved voice. Other software such as MikuMikuDance[7] made it possible for fans to import 3D models and create their own animations of the singer. In a short time, what was

seen was “a boom in user content and the development of other imitated characters” allowing that “fans’ animations become part of the concerts” (Bessant 2018, p. 31). Encouraging the use of these programs by amateur musicians and animators ended up creating a sense of community that was built around very specific aesthetic and cultural values.

Hatsune Miku draws attention because of a contradiction that is inherent in her existence. On the one hand, she is recognized as representing a certain category of singers with whom she shares certain similarities – dressing habits, musical genre and, most importantly, a (professional and expressive) singing voice. On the other hand, like Frankenstein’s creature, Hatsune Miku is relegated to being an outsider, a mirror of all female singers without being any of them. She does not go to parties, does not have a boyfriend, does not donate to social and ecological causes. Hatsune Miku does not issue opinions. Her voice is doomed to be essentially what she was designed to be: a general voice, unbiased and flawless.

Miku’s synthesized voice poses a problem that is part of the growing mediation process to which our bodies and our senses are submitted. Speech and listening interfaces, like other devices that surround us, are perceived as interfaces produced to enable our interaction with other individuals or with other devices in a neutral way. What would be understood as intentionality or as attributes of the agents taking part in a process of social interaction – a conversation, a love relationship, a dialogue between teacher and student – is often perceived as an accidental contingency when a technology is at issue. Thus, the compression of mp3 files, the noise produced by the hair dryer or the limit of bandwidth in a phone conversation, are not seen as choices, intentional or not, of those who produced these devices, but as something that is part of their ‘nature’. Ideally, we can imagine that the sounds we get from our headphones and the voices Siri employs to communicate with Apple users are designed to sound generic and to be adapted to any situation. Sound technologies appear as if it were possible to invent a generic voice aimed at generic listening. Thus, a synthesized voice is founded on the belief that it is essentially neutral and therefore can be shaped to take on any character we wish. However, this neutrality clashes with what we perceive in our daily experience, in which both voice and listening are subject to physical, emotional, and cultural conditions, making a voice always something unique, referring to a field of experiences which are modulated by a specific act of listening.

In 1955, Max Mathews joined Bell Laboratories as an acoustic engineer to investigate efficient ways of transmitting and receiving voice over the telephone. In the following years, Mathews’ research would unfold into a series of breakthroughs, not in

communication but in music technology. His achievements lead to what is currently known as computer music. In 1957, he created a computer language called MUSIC to produce sounds, and from then on he was the protagonist in a series of inventions capable of generating or controlling sounds electronically. If the first digitally produced sounds in 1957 did not excite Bell's engineers, a few years later, in 1961, Mathews and his colleagues were already able to reproduce comprehensible vocal sounds. *Daisy Bell* (*A Bicycle Built for Two*), an old folk song reproduced entirely by sound synthesis, was impressive enough to be used by Stanley Kubrick towards the end of his 1968 film *2001: A Space Odyssey*. In the final clash between man and machine, the HAL 9000 computer 'sings' *Daisy Bell* as his memory is disabled. Like the voices offered by Vocaloid, the synthesized version of *Daisy Bell*[8] produced at Bell Labs does not come from a real body. However, they lead us to construct coherent images of possible bodies (a teenage Japanese popstar like Hatsune Miku or a fictional powerful computer like HAL 9000). Unlike Marclay's *Chorus II*, in which we are invited to create voices from images of vocalizers, in these processes of vocal synthesis we take the opposite path to imagine who these voices could belong to. Both the voice as sound and the vocalizer as an individual emerge as images provided by our experience as listeners. In fact, no voice, be it natural or artificial, is neutral, not just because it supposedly belongs to a subject, but because we project our listening experience onto the voices we hear. And just like the vocalizer voices what they can hear, artificially projected voices are calculated from someone's idea of a voice. As we interact more and more with disembodied voices heard through loudspeakers, we are getting used to the idea that these voices represent a general idea of voice: accentless voices that could belong to anyone and no one at the same time.

However, HAL 9000, similar to today's virtual voice assistants like Apple's Siri or Amazon's Alexa, represents a series of values that are specific to a culture. Its generality is built from an imaginary idea of what is common. Disguised in the form of algorithms or electronic components, this generality actually results from the subjective view of those who implemented it. By creating a general self, these technologies rule out the existence of an 'other': Virtual voice assistants are not designed to understand minority discourses, immigrant accents, or social groups that express themselves in slang and dialects. That is, they understand those who speak within what is set as standard. As the access to a significant number of services becomes dependent on voice recognition, the supposed neutrality of recognition algorithms becomes a political form of segregation that has only recently been highlighted by authors such as Safiya Noble (2018) and Cathy O'Neil (2016). Current electronic speech production and electronic speech recognition, like any

other computational procedure, depends on the implementation of models in the form of algorithms. Models always represent a simplified view of a problem. Models have a purpose, seek to represent certain aspects of reality and are often evaluated in terms of efficiency, generality or accuracy. However, models are based on choices and always represent a point of view. When implemented in the form of an algorithm, they reproduce the expectations, habits, beliefs and, eventually, the prejudices of those who created them. These voices are therefore expected to act as references of dominant cultural and social strata and help construct references of what would be a good vocal quality and appropriate modes of vocalization.

5 Incarnated Voice

Luiz Ernesto Machado Kawall is a Brazilian journalist and museologist passionate about voices. Born in 1927, he has devoted much of his life to collecting voice recordings. His collection, created with his own resources and driven by his own curiosity, brings together some ten thousand voices recorded in different contexts: from famous people, to ordinary individuals, to animal voices. In addition to its historical value, this collection enables an experience that we could hardly reproduce in our daily life: listening, one after another, to voices from different times and places, imagining the situations in which they were recorded and the bodies that produced them. His vozoteca[9] (voice library) is an opportunity to perceive through listening the diversity of speeches, accents, languages, subjects, and musics that came into being through voice. This diversity contrasts with the restricted set of voices that we access in our daily lives. Kawall's collection gives the dimension of the limits of our own listening. In our day-to-day social life, our contact with people who speak other languages, who express themselves with other vocabularies, who belong to different social groups, is quite restricted. Kawall sought to overcome these limitations by seeking access to voices that were impossible to hear and record. He became interested, for example, in hearing the voices of characters who were never recorded, such as Dom Pedro I, emperor of Brazil in the early 19th century. Kawall was also a collector of cordel, a form of popular literature, done informally and printed in pamphlets. Perhaps it was the proximity to the cordel that aroused his interest in a recurring character in this form of literature: Virgulino Ferreira da Silva, better known as Lampião.

Leader of a gang that operated in northeastern Brazil, Lampião was considered the king of the "cangaço", a term for a movement of bandits who acted against government and paramilitary forces in very arid and poor regions of the country. Transformed into a popular hero who fought against police oppression and

wealthy farmers, Lampião became an iconic character in the rural history of Brazil. Since he lived nomadically and clandestinely in remote regions of the country and was killed in 1938 by police officers, it is unlikely that his voice was ever recorded. In his obstinacy to know the cangaceiro's voice, Kawall ended up resorting to Umbanda, a syncretic religious tradition that brings together elements of Catholicism, the tradition of African orixás and spirits of indigenous origin. Some Umbanda practitioners called "aparelhos" (literally "devices") are mediums who have the ability to embody spirits, allowing them to communicate with human beings. Kawall visited one Umbanda temple in which a medium incorporated the spirit of Lampião, enabling him to record the voice of the deceased cangaceiro[10].

Fig. 3: Lampião, 1927 (photographed by Benjamin Abrahão Botto).

One might ask: Whose voice was registered by Kawall? To what extent are those recordings representative of Lampião? These questions bring us back to the impossibility of providing a definitive answer to the acousmatic question posed earlier by Eidsheim. In this case, there is no simple answer. The acoustic voice recorded by Kawall was not provided by Lampião's body, but by a transducer – in this case, not a loudspeaker, but a medium who incorporated Lampião's spirit. Listening to Lampião's voice involves beliefs and subjectivities as these recordings only make sense in the set of representations and experiences of those who listen to them. Lampião himself cannot be reduced to an individual's physical body, nor his voice to an acoustic trace. His presence was built in the symbolic imagery of rural culture in Brazil from a network of reports, beliefs, and fantasies established through speech and listening, orality and aurality. In fact, it may be irrelevant to know how much of this knowledge corresponds to what he actually was as an individual. By adding another element to this complex imagery that represents Lampião, the voices recorded by Kawall are as true, as real, as the stories told in prose and verse by the popular culture of cordel in northeastern Brazil.

Without the presence of his body, and without the existence of an acoustic signal coming from that body, the voice recordings made by Kawall may or may not be taken as realistic. It completely depends on our intention and ability to listen to them.

6 Resonances

> "Sound is not what we hear, any more than light is what we see" (Ingold 2007, p. 11). The coherence of the association between sound and image is rarely questioned. The profound asymmetry between these two entities is based on the fact that the objects we see come to us through light, and the objects we hear come to us through sound. Just as we do not see light, Tim Ingold warns us that sound "is not the object but the medium of our perception. It is what we hear in" (ibid). Sound is what makes listening possible, but it is not what we listen to. In dialogue with Ingold, composer and scholar Rodolfo Caesar continues: "sound is the support/transport [...] that allows us to listen to 'sound images': sound objects, sonorities, words, etc. Just as light provides us with the exercise of visuality" (Caesar 2020, p. 85).

Thus, voice is also a medium rather than an object. Like its counterpart, hearing, voice is not a thing, but a relation established between subjects and objects, between what is inside and what is outside. This relationship can be understood as resonance, the ability to sound from – or with – the energy produced by an 'other'. When I speak, I resonate in someone else's listening and in my own listening. I am both a sounding subject and sounding object. Or, as Steven Feld puts it, "one hears oneself in the act of voicing, and one resonates the physicality of voicing in acts of hearing. Listening and voicing are in deep reciprocity, an embodied dialog of inner and outer sounding and resounding built from the historization of experience" (Feld 2003, p. 226).

When we are born, we begin to establish this experience. At this starting point, voice and listening, subject and object, are one and the same. Our voice/listening is formed as an imitation, a mirror of what we hear from our mother. At some point, the lullaby we hear becomes the cry we emit (Clough 2013, p. 66) and the annoyance caused by our crying establishes the channel of communication with an 'other'. What is built is a resonance, the possibility of sounding and hearing in a feedback process, the "delicate looping that is listening or being heard" (ibid). From then on, our vocal folds and our eardrums become inseparable membranes.

References

Bertau, Marie-Cécile (2008): Voice: A Pathway to Consciousness As 'Social Contact to Oneself'. In: Integrative Psychological & Behavioral Science. 42. Pp. 92-113.

Bessant, Judith (2018): The Great Transformation: History for a Techno-Human Future. London, New York, NY: Routledge.

Caesar, Rodolfo (2020): Apontamentos para espetar o som. In: MusiMid. 1/1. Pp. 82-87.

Clough, Patricia T. (2013): My Mother's Scream. In: Sound, Music, Affect: Theorizing Sonic Experience. Marie Thompson; Ian Biddle (eds.). New York: Bloomsbury.

Crisinel, Anne-Sylvie; Charles Spence (2009): Implicit Association between Basic Tastes and Pitch. In: Neuroscience Letters. 464/1. Pp. 39-42.

Deutsch, Diana (2019): Musical Illusions and Phantom Words: How Music and Speech Unlock Mysteries of the Brain. New York, NY: Oxford University Press.

Eidsheim, Nina S. (2008): Voice as a Technology of Selfhood: Towards an Analysis of Racialized Timbre and Vocal Performance. PhD thesis. University of California, San Diego.

Eidsheim, Nina S. (2019): The Race of Sound: Listening, Timbre, and Vocality in African American Music. Illustrated edition. Durham: Duke University Press Books.

Feld, Steven (2003): A Rainforest Acoustemology. In: The Auditory Culture Reader. Michael Bull; Les Back (eds.). Oxford: Berg. Pp. 223-239.

Fogg, Thomas (2018): Expériences Sonores. Music in Postwar Paris and the Changing Sense of Sound. PhD thesis. Columbia University.

Gans, Carl (1992): An Overview of the Evolutionary Biology of Hearing. In: The Evolutionary Biology of Hearing. Douglas B. Webster; Arthur N. Popper; Richard R. Fay (eds.). New York, NY: Springer. Pp. 3-13.

Goutte, Sandra; Mason, Matthew J.; Christensen-Dalsgaard, Jakob; Montealegre-Z, Fernando; Chivers, Benedict D.; Sarria-S, Fabio A.; Antoniazzi, Marta M.; Jared, Carlos; Sato, Luciana A.; Toledo, Luís F. (2017): Evidence of Auditory Insensitivity to Vocalization Frequencies in Two Frogs. In: Scientific Reports. 7/1. Article No. 12121.

Ihde, Don (2012): Listening and Voice: Phenomenologies of Sound. Second Edition. Albany, NY: SUNY Press.

Iazzetta, Fernando (2016): A imagem que se ouve. In: Diálogos Transdisciplinares: Arte e Pesquisa. Gilbertto Prado; Monica Tavares; Priscila Arantes (eds.). São Paulo: ECA/USP. Pp. 376-395.

Ingold, Tim (2007): Against Soundscape. In: Autumn Leaves: Sound and the Environment in Artistic Practice. Carlyle Angus (ed.). Paris: Double Entendre. Pp. 10-13.

Kahn, Douglas (1994): Christian Marclay's Lucretian Acoustics. In: Christian Marclay. Daadgalerie: Berlin. Pp. 23-34.

Kahn, Douglas (1992): Histories of Sounds Once Removed. In: Wireless Imagination: Sound, Radio, and the Avant-Garde. Douglas Kahn; Gregory Whitehead (eds.). Cambridge, Mass.: MIT Press. Pp. 1-29.

Kreiman, Jody; Sidtis, Diana (2011): Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception. Malden, MA: Wiley-Blackwell.

Noble, Safiya U. (2018): Algorithms of Oppression: How Search Engines Reinforce Racism. Illustrated edition. New York: NYU Press.

O'Neil, Cathy (2016): Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown.

Pisanski, Katarzyna; Bryant, Gregory A. (2019): The Evolution of Voice Perception. In: The Oxford Handbook of Voice Studies. Nina Eidsheim; Katherine Meizel (eds.). New York: Oxford University Press.

Reinoso Carvalho, Felipe; Van Ee, Raymond; Rychtarikova, Monika; Touhafi, Abdellah; Steenhaut, Kris; Persoone, Dominique; Spence, Charles (2015): Using Sound-Taste Correspondences to Enhance the Subjective Value of Tasting Experiences. In: Frontiers in Psychology. 6. Article No. 1309.

Schaeffer, Pierre (1966): Traité des Objets Musicaux. Paris: Éditions du Seuil.

Tomatis, Alfred A. (1992): Conscious Ear. Barrytown, N.Y: Station Hill Press.

Notes

1 The idea of sound as image has been developed in Iazzetta 2016.

2 Tomatis' trajectory is quite controversial and includes the discrediting of his colleagues in France and the commercial use of his scientific findings, particularly his 1953 invention of the Electronic Ear, and later of his TOMATIS® Method. For a critical discussion on the French physician's interest in the sense of listening, see Fogg 2018.

3 Some phantom words examples can be found at: http://dianadeutsch.net/book_audio/Modules-2019/mixdowns/MP3/ch07ex01_phantom_words-d5_mixdown.mp3 [last accessed August 25, 2021].

4 A Japanese subculture interested in manga and anime.

[5] Hatsune Miku was released by Crypton Future Media in August 31, 2007.

[6] Released in 2004, Vocaloid allows its users to type text and melody to synthesize a song. Voice synthesis is performed using voice banks extracted from samples produced by professional singers. See https://www.vocaloid.com/en/ [last accessed August 25, 2021].

[7] MikuMikuDance, or MMD, is a freeware animation software originally produced to give life to the Vocaloid character Hatsune Miku. Since its launch in 2008 the program has attracted the attention of a wide community on the Internet interested in creating characters based on anime culture. Many MMD videos can be found on NicoNico, a Japanese video-sharing service on the web.

[8] The similarities between Vocaloid and voice synthesis produced at Bell Labs had already been noticed by Kenmochi Hideki, leader of the research project that gave rise to Vocaloid. In its first public appearance at Musikmesse in 2003, the program was to be called *Daisy*, in allusion to *Daisy Bell*. The name had to be changed for copyright reasons. See: Red Bull Music Academy 2014, https://daily.redbullmusicacademy.com/2014/11/vocaloid-feature [last accessed August 25, 2021].

[9] His voice archive was donated to the Institute of Brazilian Studies of the University of São Paulo in 2013 and can be accessed under the title Vozoteca LEK at http://200.144.255.59/catalogo_eletronico/ [last accessed August 25, 2021].

[10] These recordings are stored at the Institute of Brazilian Studies collections archives under the reference Vozoteca LEK, VOZ-CDr-010 at http://200.144.255.59/catalogo_eletronico/ [last accessed August 25, 2021].

Künstliche Stimmen

Marc Böhlen

The Making of Fake Voices

Abstract: Im vorliegenden Text wird erörtert, wie technologische Innovationen und Fortschreibungen menschlicher Sehnsüchte die Sprachsynthese an einen Punkt geführt haben, an dem sie in industriellem Maßstab eingesetzt werden kann und dabei nahezu jede menschliche Stimme nachzubilden vermag. Der Beitrag trägt dem Unterschied zwischen perfekter Mimesis und erfahrungsbasierter maschineller Sprachproduktion Rechnung. Er legt dar, wie dieser Unterschied als Werkzeug der Täuschung eingesetzt werden kann, und wie er als ein Experimentierfeld dient, auf dem über das mittels künstlicher Intelligenz realisierbare Klonen menschlicher Eigenschaften im Allgemeinen nachgedacht wird.

Abstract: This text discusses how innovation in technology and continuity in human desires brought voice synthesis to a state in which it can be deployed at an industrial scale and reproduce almost any human voice. The text considers the difference between perfect mimesis and machinic speech production, describing how this difference can be deployed as a tool for deception, as well as the way it serves as a testing site for reflecting on artificial intelligence driven cloning of human features in general.

1 Introduction

In March 2019, criminals used artificial intelligence to impersonate a chief executive's voice for a fraudulent money transfer.[1] The scammers created a fake version of the voice of the chief executive and called an unlucky executive employee in the fake voice of this supervisor, which included a slight German accent and the specific melody of the supervisor's voice. The employee was informed by this fake boss that €200,000 were to be transferred to a compromised recipient address within an hour. The employee promptly executed the transaction, unwittingly enabling the first documented artificial intelligence generated fake voice cybercrime of the 21st century.

The history of synthetic speech spans at least three centuries (Ramsey 2019) and possibly much longer, if accounts of speaking statues as early as 20 BC informed

of principles described in Heron of Alexandria's treatises on machinery, mechanics and hydraulics (Pettorino 1999) are in fact true. Reflecting on the history of fake voices offers an opportunity to consider how one age-old dream can drive technical innovation across centuries. It can also serve as a case study in the downstream effects of technical innovation. While the phone scam example above might suggest that deception is a product of only the latest instantiation of artificial voice technology, con-artistry was in fact an early adapter to new communication opportunities afforded by landline telephony as it changed communication patterns and opened the door to new forms of impersonation (Marvin 1999).

By exploring several speech producing systems in context – Kempelen's Sprachmaschine, Dudley's Voder and Tacotron – this text will cast voice synthesis as a story of an immemorial human dream, implemented in each iteration utilizing the current technology available, and entangled with the social dynamics in which it is inserted. The last section then reflects on how we live with synthetic speech systems today under these entanglements.

2 Kempelen's Sprachmaschine

When Wolfgang von Kempelen began experimenting with a device that could imitate human voices, he already had some good reference points. By the beginning of the 17th century, a low fidelity mechanical model of how sound is generated in the human vocal tract had already been established (Ramsey 2019, p. 11). Kempelen's device translated the 17th century state of the art model of the human voice tract into a mechanical apparatus made of wood, paper, brass wires, tin tubes and leather. Yet the simplicity of the construction belies the depth of its potency. Kempelen produced a 464-page manuscript (Kempelen 1790) that not only offers detailed engineering drawings of the various mechanical parts of the apparatus and a lengthy treatise on the body parts involved in the production of speech, but also a general discussion on human language and its presumed origins.

As opposed to earlier attempts at voice-like sound creation, Kempelen's Sprachmaschine was the first device capable of generating utterances reminiscent of entire words.[2] Kempelen translated human sound production into an equivalent non-human system with a bellows functioning as lungs, a wind chest to distribute the air to sound producing enclosures, a reed made of a thin strip of ivory glued to a piece of leather, and a funnel made of natural rubber representing the oral cavity (Kempelen 1790, chapter 5; Deutsches Museum 2020). The device was more a musical instrument than a utilitarian apparatus. It was played by pressing on the

bellows and opening and closing pathways to enable or constrain airflow to the different parts of the machine. Under the skillful control of an operator (Braskhane 2017), a variety of human-like sounds could be produced to imitate short utterances in several different languages (Pettorino 2015) including Latin, French and Italian but not German and no report on Hungarian, Kempelen's mother tongue.

While Kempelen's well-known chess-playing automaton, the Mechanical Turk, was a clever mechanical contraption capable of moving chess pieces across a chess board, it was not a chess champion. In fact, Kempelen's Mechanical Turk was a fraud – there was a skilled human chess player inside the machine performing the chess moves out of view of the audience. No such post-mortem disclosure blemishes the Sprachmaschine. This is surprising given the fact that Kempelen presented both his Mechanical Turk as well as his Sprachmaschine together on tour across Europe in 1783 and 1784 (Deutsches Museum 2020). Moreover, experimentation in human voice creation up to the 17th century was generally viewed with suspicion and often accused of sorcery, persecuted and even condemned (Pettorino 2015). To give voice to an object was perceived as more amazing than to have it emit a melody; to give voice meant to give (humanlike) life to inanimate matter, a feat considered beyond the reach of human agency.

Despite and because of these circumstances, Kempelen's machine is a landmark in the history of voice generation. As a product of the Enlightment period, it mirrors a world view that perceived the human body as a machine. Kempelen's creation is the first viable construction capable of imitating the sound production of the human voice tract based on a mechanical model of this very voice tract. As a disembodied voice it represents one early example of a trajectory of scientific inquiry and engineering design that seeks to replicate human abilities and features operating outside of and without the need of the human body.

3 Dudley's Voder

When Homer Dudley conceived his speech apparatus in the middle of the 20th century, he also relied on the materials and techniques of his time. But instead of wood and leather, Dudley, a researcher at Bell Labs, assembled his device utilizing the hardware du jour, vacuum tube electronics.

In *The Carrier Nature of Speech* (Dudley 1940), Dudley outlines his speech generation concept as a carrier circuit, informed by the model of analogue radio communication. The carrier circuit describes an information representation concept in which a

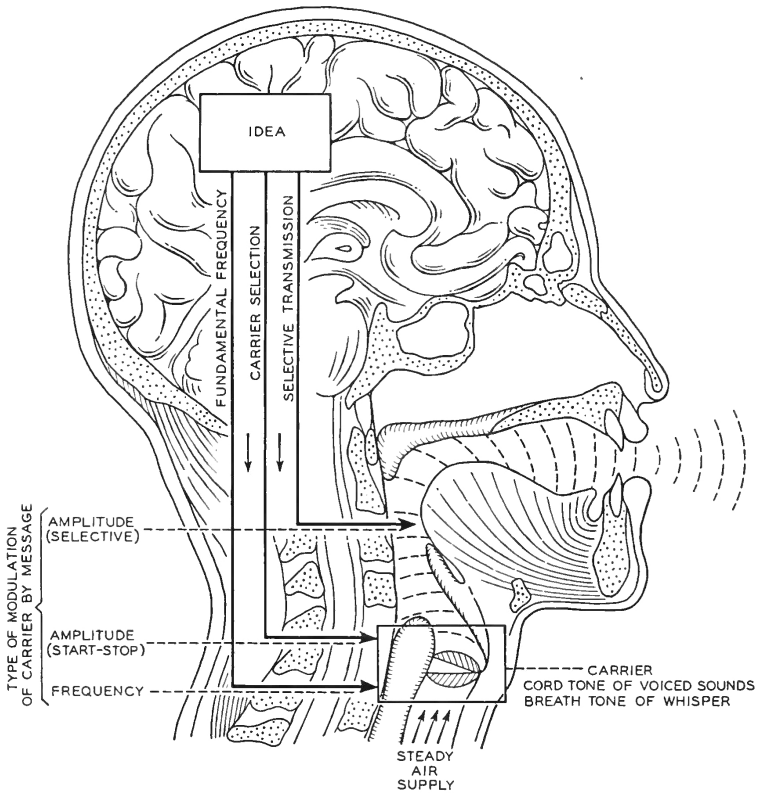

Fig. 1: From Homer Dudley's *The Carrier Nature of Speech*. The Bell System Technical Journal Vol 19, Number 4, October 1940, p. 497. Reused with permission of Nokia Corporation and AT&T Archives.

'transport' waveform is modified (modulated) with an information-dependent signal, usually higher in frequency than the base carrier wave. Dudley's concept maps speech produced by the dynamics of compressed air in the human vocal tract onto corresponding frequency bands. By selectively combining these frequency bands in the spectrum of human speech with a base carrier wave, Dudley was able to devise a voice synthesis approach capable of producing human-like speech. This result seemed rather counterintuitive as the carrier signal itself, sounding like a hiss or a

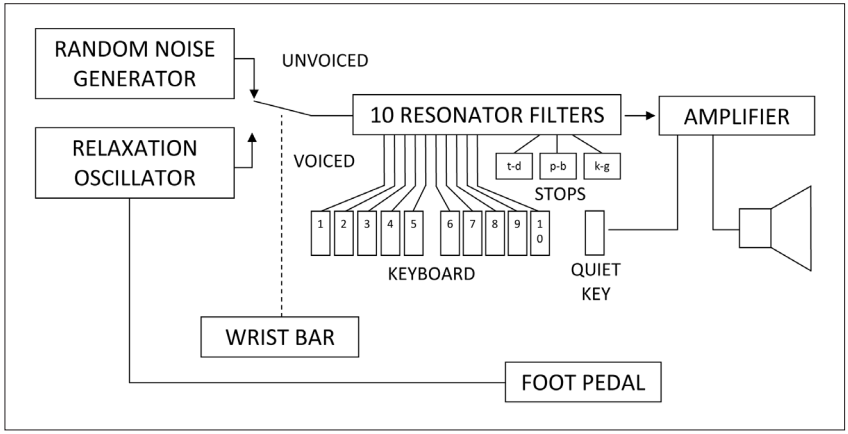

Fig. 2: Schematic diagram of the Voder.

buzz, is in no way reminiscent of a human voice; only the modulated product sounds anything like a human voice. A far cry from the kind of synthetic speech we have grown used to today, Dudley's approach was flexible enough to create both voiced utterances emulating sounds produced when the vocal cords vibrate (such as 'z') as well as unvoiced utterances (such as 's') in English.

Similar to the work of Kempelen, Dudley's concept included the source of sound creation. Dudley's concept integrates the human thought process into the model, including the idea that originates in the mind, and which only later is translated into the sound-producing hardware of the human body (see Fig. 1). Unlike Kempelen, Dudley's experiments limited themselves to the English language. However, Dudley at least reflected on other possible applications ranging "from the puffs of a locomotive to instrumental music" (Dudley 1940, p. 513).

This connection between idea and speech act continues to be pursued in current research under the theme of concept-to-speech recognition, formally defined as "the production of synthetic speech on the basis of pragmatic, semantic, and discourse knowledge" (Alter 1997, p. 4). Compared to Kempelen and Dudley's machinery, concept-to-speech is comparatively pedestrian in its ambitions, but it delivers tangible results, improving speech dialogue between computer-generated voices and people. For example, concept-to-speech has more recently been deployed on an industrial scale in Amazon's popular Alexa assistant.[3]

Dudley applied this carrier concept to an operator-controlled voice machine called the Voder (Voice Operation Demonstrator). The Voder produced a carrier wave with a buzzer-like sound for the voiced, and a hiss-like sound for unvoiced sounds. The Voder had a bank of 10 pre-defined band pass filters covering (most of) the spectrum of human speech. All these filters receive input from the noise source or the relaxation oscillator (the buzz source). The operator selects between these two input sources (carriers) with a wrist bar and controls the pitch of the input with a foot pedal. A keyboard acts as a controller on the filters, reducing or increasing the contribution of any one of them (see Fig. 2). Together with a quiet key, these components allowed an operator to play the device and generate sounds using different pitches and inflections that could be recognized as speech.

The Voder was not easy to utilize. Multiple operators had to train for over a year to be able to produce only a few simple utterances (Guernsey 2001). Just as Kempelen toured his machine to impress the crowds, Dudley's Voder was prominently featured at New York World Fair of 1939. Tellingly, that World Fair exhibited another main attraction capable of otherworldly speech, namely the robot Elektro (Marsh 2018). Elektro was a two-meter-tall humanoid robot that could walk upon spoken command, responding to the pattern of sounds from an operator – but not the content of the message. Moreover, Elektro could speak several hundred words prerecorded on a record player and differentiate between red and green colors with the help of a photoelectric camera eye. For these reasons, including the fact that Elektro would also smoke a cigarette, the robot's live performances captivated the World Fair's audience.

4 Speech Synthesis and Linguistics

Dudley's approach relied on the emulation of spectral patterns of human speech. The next steps in the history of speech synthesis required a more abstract approach – an approach that was simultaneously grounded in the theory of signal processing, guided by models of the vocal tract and Dudley's results, and informed by the insights of linguistics research, a field that hitherto was not a formal part of the synthetic speech research landscape.

Linguistics describes languages in generative terms with the goal of specifying rules for the generation of legitimate sentences through an abstract representation. Moreover, linguists represent spoken language using discrete elements, like language specific phonemes with particular features such as labial and nasal characteristics (Klatt 1987). Particular rules are then devised to explain when

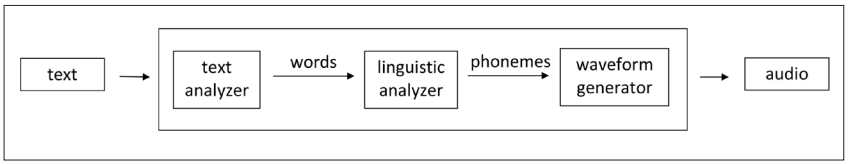

Fig. 3: Speech synthesis from text; diagram of the processing pipeline including text analysis, linguistic analysis and sound generation.

words change pronunciation in some sentence contexts (and not in others). These language specific rules – formalized in the text analyzer and linguistic analyzer – encode how a sequence of letters is transformed into a sequence of sound primitives that a waveform generator then assembles to an audible output (see Fig. 3). This rule-based approach emerged as an alternative to the more intuitive data-centric process of simply collecting a set of pre-recorded messages, and then combining elements of those pre-recorded segments to new utterances. Over time, two main approaches emerged to address the constraints and tradeoffs of the rule-centric vs the data-centric approach: formant synthesis and concatenative synthesis.[4]

5 Formant and Concatenative Synthesis

Formant synthesis is largely rule-driven. The synthesized speech is generated using an acoustic model and hand-crafted acoustic rules. Formant text-to-speech (TTS) creates speech segments from written text by generating signals based on language specific rules combined with general spectral properties of human speech. Formant TTS uses additive synthesis under the constraints of an acoustic model that describes the fundamental frequency, intonation, and prosody – the elements of speech that define individual articulation including tone of voice and accent.

Formant based methods can alter many aspects of a synthetic voice, including intonation, without relying on additional data. Because it is less dependent on data, formant TTS is ideal for gadgets, toys and household appliances where memory and processing power are limited. However, formant TTS is often recognizably machinic and is prone to glitches even when producing simple words; a condition often experienced when listening to directions uttered by formant TTS in early GPS navigation devices, for example.

Concatenative synthesis, instead, is data-driven. It relies on high fidelity audio recordings, from which segments are selected and combined via unit-selection

(selection of phonemes annotated with contextual information) to form a new speech utterance (Hunt 1996). Typically, a voice actor records several hours of speech which are then processed into a large speaker specific database containing linguistic units, phonemes, phrases and sentences. When speech synthesis is initiated, a speech generator searches this database for speech units that match those extracted from an input text, and concatenates these segments to produce an audible output. Concatenative TTS can produce high quality audio if a large and varied dataset has been collected. However, the approach makes it difficult to modify the voice (i.e. switching to a different speaker, or changing the emphasis or emotion of the speech) without recording a new database of phrases.

Both formant and concatenative TTS are in principle capable of constructing and uttering grammatically correct speech. However, both systems struggle with prosody, the subjective payload of speech. As such, neither approach is able to reproduce nuanced and sophisticated aspects of emphatic or emotional human speech across multiple languages and language use scenarios.

Despite these serious limitations, reports on lifelike synthetic speech periodically surface. As early as 1972 researchers reported on speech synthesis results which were so believable that listeners could not tell the difference between the synthetic and the human version, if presented in sequence (Klatt 1987, p. 743). One can assume that listeners in the 1970s might have been less discerning than they are today. Tellingly, the (male) researchers already then focused on the synthesis of male voices as they found the task of synthesizing a woman's (or a child's) voice more difficult (Klatt 1987, p. 746).

6 Speech Synthesis and Deep Learning

Rule-based and data-centric synthesis are not mutually exclusive, and machine learning in fact combines insights from both approaches. Deep learning is machine learning implemented with large scale, multi-layer (hence deep) network configurations, and deep learning TTS (or neural TTS) is the current preferred paradigm for designing synthetic speech systems.

Neural networks learn patterns from data. In speech synthesis, neural networks learn patterns from audio files. Once these patterns have been internalized in an iterative process, neural nets can create utterances that sound like the voices they have been exposed to. While neural TTS enables much more efficient adaptation

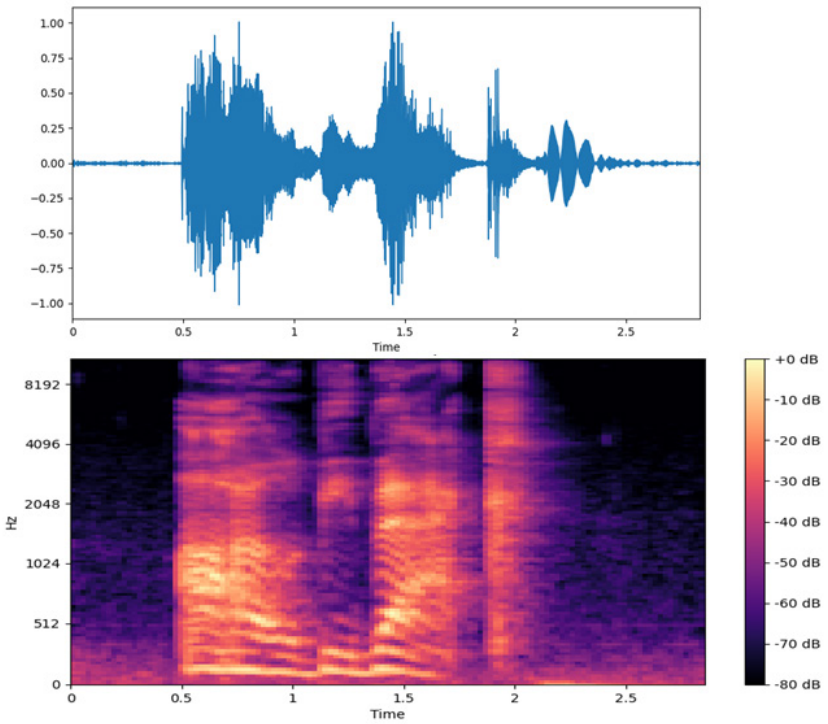

Fig. 4: Sound wave (top) and Mel spectrogram (bottom) of the three second utterance "I will be back" recorded by the author in English with a Swiss German accent. Left axis of the Mel spectrogram indicates frequency range normalized for human perception (pitch of equal distance on the scale 'sound' equally distant from each other). Right axis indicates intensity in decibels.

to new voices, variation in speaking patterns and expressive speech, neural TTS machinery comes – as engineering design does – with its own baggage.

Neural nets learn patterns from data in sequence in a process called training. Training refers to the iterative reducing of the distance (error) between output scores and the desired pattern of scores. The machine modifies its internal parameters (weights) across the various network layers to reduce this error, evaluates the outcome, then tries again, until the error is small. The learning algorithm computes

a function (gradient vector) for each weight indicating how the error would change if the weights were increased by a small value (LeCun 2015). Adjustment of the weights then occurs in the opposite direction of that gradient result, operating in an adaptive loop until the average value of the objective function stops decreasing. This adjustment is the magic sauce of the learning operation.

In neural network-based speech applications, input data appear as spectrograms created from the text. A spectrogram is a two-dimensional map of the frequencies that make up the sound, from low to high, as well as the changes of these frequencies over time, from left to right (see Fig. 4). Spectrograms are rich descriptors of text-voice constructions, supplying neural nets with detailed signal data to learn from while remaining oblivious to the messages contained in those signals.

Neural TTS systems allow for flexibility with fidelity unattainable through previous approaches. Changing the perceived gender of a voice, as well as building speech utterances that imitate a particular person with only a few examples of their speech patterns become routine operations. And this flexibility is precisely what the fake voice industry puts to nefarious use.

6.1 Tacotron

Most neural TTS systems combine two neural networks, one dedicated to translating text to a frequency representation, and a second one that converts that output to a synthetic voice. The basic principles of neural TTS are best described with an example.

Tacotron (the newest version at the time of this writing is Tacotron2) is a text to utterance generative model that synthesizes speech directly from characters. Given <text, audio> pairs, the model can be trained from scratch and delivers realistic results even to current discerning listeners. Tacotron is composed of two connected neural networks. The first is a recurrent feature prediction neural network that maps character embeddings to spectrograms. The second generates audible waveforms from those spectrograms (see Fig. 5). The first neural network has two main components, an encoder and a decoder. The encoder converts a character sequence into a feature representation which the decoder consumes to predict a spectrogram one frame at a time, capturing not only pronunciation of words, but also various subtleties of human speech, including volume, speed and intonation (Shen 2018). The second neural network is a WaveNet model, one that generates raw audio waveforms. It acts as the vocoder, the component that produces the actual

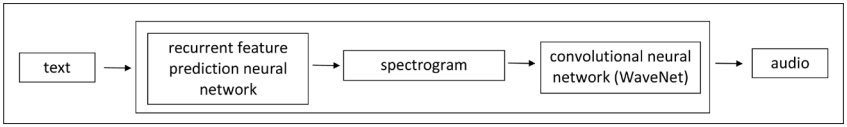

Fig. 5: Schematic diagram of the main components of Tactotron2 model with the WaveNet element. The prediction network performs the text and linguistic analysis while the convolutional network performs the work of waveform modelling.

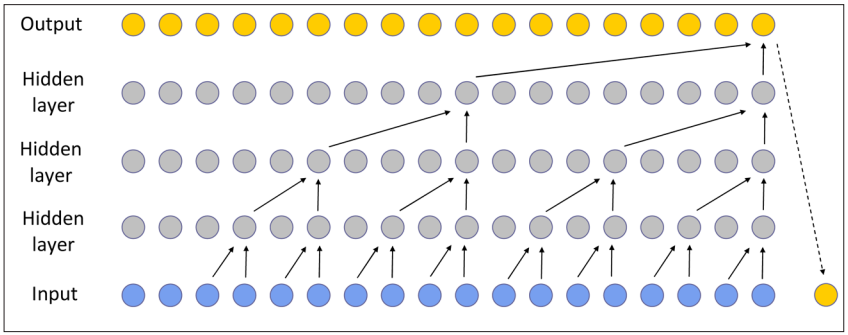

Fig. 6: Diagram of the casual convolution layers of WaveNet. After each sample is predicted as an output, it is fed back into the network as part of the input stream to predict the next sample (dashed arrow from top to bottom right most dot). This allows the receptive field to grow and cover continuous time-stepped inputs of a speech act. After Oord 2016.

sound not unlike the Dudley apparatus described above. However, this vocoder is trained iteratively on the spectrogram data produced from the first neural network. WaveNet ensures that the ordering of the time series audio data is preserved in the learning process. The model's predictions unfold sequentially: after each sample is predicted, it is fed back into the network (Oord 2016) to predict the next sample (see Fig. 6). This approach provides a high degree of flexibility. If one trains WaveNet on American English, it produces American English speech, if you train it on German, it produces German. As such, WaveNet is a universal speech engine; it models spoken language through its neural logic dynamics, absorbing whatever sound patterns it is subject to during training. However, it cannot discern whether parts of the sound landscape are relevant or not, leading to instances in which background sounds from the room in which recordings occurred were imitated (House 2017). As a neural learning machine, WaveNet is completely dependent on its training

data yet it creates from this assemblage a new form of acoustic knowledge (House 2017, p. 20). Once trained, WaveNet's knowledge is stored in the parameters of its model, which can be tuned to control the characteristics of a speech act, making the WaveNet architecture an ideal candidate for voice cloning as described below.

Neural TTS has markedly reduced the difference between machine-generated and human produced speech, and most major global IT companies including Nvidia (WaveGlow), Deepmind (WaveNet), Mozilla (LPCNet) and Baidu (DeepVoice) have developed proprietary neural TTS systems. Voice-based interaction is big business.

7 Living with Fake Voices

Early TTS, both of the formant and concatenative variety, have been deployed in a plethora of gadgets, appliances, and navigation aids, creating a vibrant ecology of first-generation fake voices. There was little opposition to the expanding collection of these early computer voices as their reach was limited. In fact, system imperfections made them quirky yet recognizable as non-human. As such they allowed human beings to navigate TTS-infused interaction events with call center agents and robot bank tellers. Deviations from the baseline of human naturalness served as a form of auditory landmarks that formed demarcation points between service robots and humans. Those not-quite-real synthetic voices were well-defined as non-human, non-threatening additions to a world ruled by human beings. Neural TTS changes this condition because of almost undetectable deviations from human speech, and because of the scale at which neural TTS is deployed; from personal mobile phones to platform products, all computer systems now support voice interaction. Neural TTS can even emulate bodily speech features such as lip smacking,[5] reintroducing a window onto materiality previously obscured, and suggesting believably that a living, breathing body is in fact producing its speech acts. As such, neural TTS destabilizes established frameworks that allow humans to identify computers in action and introduces new flavors of uncanniness into computer interaction.

7.1 Depressed Voice Talent

It is maybe not without irony that state-of-the-neural-art voice synthesis systems require copious amounts of data to achieve their superior performance. And this data comes invariably from real people, acting as voice talent in the parlance of the speech industry. The term voice talent heralds from the performing arts and radio production in which the significance of a well-balanced and articulate voice

was long appreciated. Clarity of pronunciation and clear articulation are equally significant for voice synthesis, and so performing arts voice talents delivered the first set of voices to the fake voice industry. Siri's US voice is based on the voice of voice performer Susan Bennett, and the voice of Cortana is based on Jen Taylor's.[6] Siri was originally conceived as an 'assistant' but has left those humble origins behind and has been integrated into popular culture through appearances in television shows. The voice itself has become a household name and a pop star, despite having no physical connection to the human being who sourced its famous sound. So pervasive is the pull of Siri as a cultural phenomenon that the originating human, Mrs. Bennett, became in turn accidentally famous and a popular speaker reporting on "what it is like to be the person behind Siri"[7]. The relationship between synthetic voices and their human sources is a fragile one, with some voice actors reporting a sense of disappointment and sadness after being replaced and updated by a more fashionable voice (cf. note 6).

While synthetic voices now sound largely realistic, the social realities they represent have remained stubbornly conservative. The Matthews, Johns, Kendras and Sandras of the speech industry suggest an identity binarized world, and the voice flavors remain strictly either male or female. Moreover, the majority of voices designed for assistant tasks are female or female by default (UNESCO 2019). Before this biased voice landscape was recognized as problematic by industry, artists, including the author of this text, investigated pathways by which to probe the language normalization in synthetic voice design, including the construction of immigrant accented language (Böhlen 2008).

Only recently, the synthetic voice industry has responded to gender fluidity in the construction of an appropriate synthetic voice. Sam is the world's first comprehensive non-binary voice product.[8] Sam differs from previous gender-neutral voice products such as Q[9] that attempt to avoid gender specific characteristics altogether and appear genderless, in that it combines prosody features from both male and female voices to a voice product that sounds like a man *and* a woman. Sam is marketed specifically to industry products seeking to resonate with the transgender or gender non-conforming community (cf. note 8). In fact, the newfound flexibility in voice product fine-tuning allows to design voices sounding in any manner one might desire; with the technological hurdles removed, voice developers explore the edges of voice design and have recently arrived at two strange places, to wit voice cloning and deep fakes.

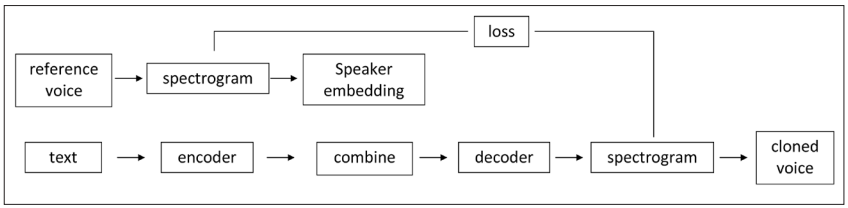

Fig. 7: An individual's unique voice characteristics are added to a neural TTS system via speaker embedding. After Jia 2019.

7.2 Voice Cloning

Voice cloning creates a synthetic voice of a specific (living or dead) human being. The unique tonal characteristics of a living person's voice can be captured through recording samples and transferred as speaker embeddings (see Fig. 7) into a neural TTS system. As such, voice cloning relies on a listening operation before it becomes a speaking machine.

Voice cloning extracts the salient features of a speaker's voice from a reference audio utterance in order to create the speaker embedding, paying no attention to the meaning in the utterances and collecting only characteristics such as pitch, accent and tone.

The speaker embedding information is combined with the phoneme sequence, and the vocoder generates from this combination a voice with the auditory features of the specific speaker, in other words the cloned voice. The neural magic learning sauce that fine-tunes how closely the voice sounds to the original recording occurs by comparing the (spectrogram of the) original recording with the spectrogram created by the decoder, and iterating through the process, making small adjustments until the loss, the difference between the reference and the clone, is negligible as described in the section on neural network training above.

Voice cloning finds a direct application in the form of voice banking, the collection of audio recordings for voice synthesis in anticipation of a voiceless future due to the consequences of degenerative illnesses such as *amyotrophic lateral sclerosis,*[10] or in order to counter adverse effects of surgical inventions such as laryngectomy. The entertainment industry has its own interest in voice cloning. Film narration for global audiences, podcasting, as well as game character narration all make use of voice cloning. Given the proliferation of software tools that facilitate the programming

of neural TTS and the opportunity to replace studio with home computer audio recordings,[11] the bar for creating cloned voices is at the present comparatively low, unleashing a wave of low-cost, cross-language product branding scenarios.[12]

7.3 Audio Deep Fakes

Audio deep fakes make use of the voice cloning techniques outlined above. However, audio deep fakes are created with sound recordings collected without consent. Typically, voice samples are collected from recorded speeches, public presentations, interviews or press conferences and used to train a voice cloning system. While voice cloning requires copious amounts of data for highest-quality reproductions and longer utterances, voice cloning systems applied to simpler tasks and short statements can operate with as little as 10 seconds of reference audio recording (Chen 2019) and still produce believable results. Even a phone call, recorded in a relatively noisy location, can be used as the source of voice-embedding and can clone a person's audio footprint, albeit with low fidelity. Scammers cleverly respond to these limitations by deploying audio deep fakes in applications in which one would in fact expect to hear low-quality audio, such as in telephone calls. The noisier the environment, the more difficult it is to distinguish a fake from a real voice. Because we are accustomed to hearing low-quality audio on telephones, low-quality audio fakes encounter a reduced threshold for scrutiny. It is precisely this combination of ubiquitous voice cloning software, busy conference calling culture and currently deficient literacy in recognizing voice fakery that enabled the first audio deep fake cybercrime described at the beginning of this text.

7.4 Audio Deep Fake Fraud

Electronically enabled fraud is at least as old as the landline telephone. Con artists used the public's unfamiliarity with voices heard over low-bandwidth telephone lines to impersonate other people or to gain trust in ways that would never have been possible in a face-to-face encounter (Marvin 1990). Likewise, "419 email" frauds in which an unsuspecting victim receives an email offering an "opportunity to share in a percentage of a large sum of money"[13] have been effective only because the trust landscape of email, particularly for the non-digitally native, is still unstable, functioning simultaneously as a personal communication channel as well as an official source of information.

Audio deep fakes likewise make use of the current unstable status of ubiquitous synthetic voice systems. Robotic-sounding voices are recognized as such and even enjoyed for their style; non-consensual individual-imitating voices do not yet have a defined space in everyday life. We have insufficient training to detect the new minor slippages that give them away, and when we are confronted with these new audio artifacts embedded in mundane situations, there is often no reason to be suspicious.

Since there is no public literacy in synthetic voice misuse so far, substantial efforts are being deployed to recognize deep fake audio through technical means. Assessing the veracity of a deep fake audio artifact requires authentication, i.e. proving it is fake or showing that it is not fake. Voice biometrics (Gonzalez 2008) offers heuristics by which to assess whether a voice stems from a genuine flesh and blood human being or from an algorithm. Anomaly detection, for example, can reveal if sounds in an audio artifact were generated through the anatomy of the human vocal tract or not (Hill-Wilson 2020).

Legal scholarship is also responding to the practice of speech cloning, as it attempts to discern whether synthetic speech is a form of protected speech. Likewise, the issue of copyright within voice cloning remains unresolved. Deep fakes use a variety of audio recordings, and these are processed to exhibit features that are no longer reminiscent of the original speech act, hence there is no direct copying in this approach, but rather an artistic re-use more in tune with the established concept of fair practice and derivative work for parody, for example in the Bern Convention and the US Copyright Law; a distinction the entrepreneurial producer Jay-Z seems to dismiss in his lawsuit against the YouTube channel *Vocal Synthesis*[14] that pokes fun at many celebrity figures, including Jay-Z, with neural TTS produced fake voice-overs.[15]

Legal scholars are particularly concerned with deep fakes at the far end of the sophistication spectrum as they undermine the reliability of genuine evidence (Pfefferkorn 2020). Moreover, they offer fraudsters a new avenue for deception and mistrust as *any* audio recording can now be declared as potentially fake and hence any evidence declared as invalid, simply because these new methods of fakery exist in general. The space of potential frauds, particularly when expanded by interaction with voice assistants, appears substantial. One can imagine a con artist manipulating a voice assistant with bogus claims (Lang 2018) uttered in a pain-ridden voice: “Alexa I have a terrible headache; please order some Aspirin”, to plant a history of fake evidence for a future insurance claim that could eventually be audited.

7.5 An Uncanny Valley of Fake Voices

The uncanny valley is a coinage used by roboticists describing "the proposed relation between the human likeness of an entity and the perceiver's affinity for it" (Mori 1970), or more practically, the experience of encountering a humanoid robot, only to notice that the robot is in fact not a person, but rather a 'fake' acting like a real person. When applied to the realm of synthetic voices, uncanniness is a product both of how the robot voice sounds as well as how the interaction dialog between the robot and the human unfolds. Research on chatbot artificial intelligence suggests that people in fact experience lesser uncanny effects and less negative affect when cooperating with simple text chatbots as opposed to more elaborate visually animated chatbots (Ciechanowski 2019). Moreover, the finer details of prosody of speech become all the more important in extending believability and preventing negative affect typical of the uncanny valley experience. Failing to differentiate tonality between short and long sentences – a task humans handle with grace – is hard even for neural TTS systems. Such slippage impacts the context of what is being expressed, and that mismatch becomes a source of negative affect (Simon 2019). The closer synthetic speech approaches real human generated speech while failing to perfectly replicate it, the stronger the sense of discomfort.

Perhaps the recent progress in near perfect voice imitation has blinded researchers to approaches that navigate the uncanny valley with more aplomb. The problem has been informally addressed in popular science fiction through a variety of robot characters. R2-D2 from *Star Wars* (USA 1977, D: George Lucas) delineates the human-robot boundary by speaking in a tongue no human can understand. The emotional connection between the robot and human beings is established instead through movements and gestures. Moreover, when R2-D2 becomes agitated, it produces wild steams of electronic sounds that unambiguously convey its inner state without ever saying a word. HAL 9000, the infamous disembodied artificial intelligence agent in Stanley Kubrick's film *2001: A Space Odyssey* (UK/USA 1968), on the other hand, tries tirelessly to sound precisely like a human being. And as its true intentions become apparent over the course of the film, HAL's more human-like voice renders its machinations all the more sinister.

Perhaps the best example of the handling of the inevitable mismatch between imitation and real human voice is offered by Stephen Hawking. Even as advanced voice synthesizers became available, the late physicist famously preferred an antiquated formant synthesizer.[16] "Perfect Paul", as the voice was coined, was a thin, tinny voice that became Stephen Hawking's widely recognizable vocal style.

Hawking maintained this technologically outdated voice even as his other assistive technologies were updated in order to accommodate his particular needs.[17]

7.6 When New Technologies Are Old (Again)

Prior to ubiquitous landline telephony, the concept of presence required the physical colocation of people (Marvin 1990). The telephone changed that logic of presence and replaced it with one of temporal continuity. Yet the change took some getting used to. It was only over time that the physical co-location requirement was sufficiently relaxed in a way that the general public understood that a person connected via telephone line could be addressed as if they were in the same room. Synthetic voices inherit the paradigm of presence from afar. They add to this condition a new twist, navigating (from afar) the presence of entities that sound like real people.

The uptake and governance of technological change is subject to many factors. The Collingridge dilemma describes the quandary according to which control of technology is difficult at early stages as not enough is understood about its consequences, and costly later on once the consequences are in fact apparent (Collingridge 1980). The development of synthetic voices offers the opportunity to add more nuance to the dilemma. Synthetic speech has long existed as an *old* new technology, one that prepared us for its arrival in imaginaries and crude approximations, appearing several times, in different forms, and eventually becoming globally distributed in systems small and large.

Synthetic voices no longer operate as isolated artifacts. Having positioned themselves into close contact with human beings in daily life, they have become dependent on and subject to changing dynamics of human social interaction and periodically adaptable to those dynamics, as the development of non-binary voice products demonstrates.

Synthetic voice technology is the first technology that can make a claim to have recreated human features so credibly that even human beings cannot distinguish the real from the artificial. More recently, portrait imagery has established a similar status with lifelike images of people who do not exist but were instead created by neural networks (Karras 2019). In both cases the neural techniques destabilize historically established concepts of veracity.

Observing how we now live alongside state-of-the-art voice systems serves as a testing ground for cohabiting with artifacts and systems based on human feature cloning in general. As artificial intelligence artifacts become more engrained in

everyday life, replacing older technical systems we have grown accustomed to – and introducing new unfamiliar ones – we will be tasked with developing new cohabitation strategies that allow us to monitor the effects of cloning-capable artificial intelligence on human wellbeing. New literacies and hitherto unfamiliar forms of connoisseurship of the artificial will likely emerge over time.

Locally contextualized and endlessly adaptable, pleasant sounding synthetic voices will make voice artists and announcement personnel redundant and positions at fast food restaurants obsolete. The always cheery voice agents that can handle routine customer requests with never ending artificial grace exhibit otherworldly patience, even in the face of the rudest of human beings. As such synthetic voices operate as early harbingers of technologies that replace human soft skills, previously immune to technology-driven labor displacement. For this reason alone, the evolving landscape of industrial synthetic voice deployment deserves attention, and careful listening. After all, none of these machines that sound like a human has any idea what it is talking about.

References

Alter, Kai; Pirker, Hannes; Finkler, Hannes (eds.) (1997): Introduction to the Workshop. In: Proceedings of a Workshop in Conjunction with 35th Annual Meeting of the Association for Computational Linguistics. https://www.aclweb.org/anthology/W97-1200.pdf [last accessed July 28, 2021].

Bern Convention (2020): Berne Convention for the Protection of Literary and Artistic Works. Article 10. https://treaties.un.org/doc/Publication/UNTS/Volume%20828/volume-828-I-11850-English.pdf [last accessed July 28, 2021]. https://www.wipo.int/export/sites/www/treaties/en/documents/pdf/berne.pdf [last accessed July 28, 2021].

Böhlen, Marc (2008): Robots with Bad Accents: Living with Synthetic Speech. In: Leonardo. 41/3. Pp. 209-214.

Boilard, Johnathan; Gournay, Philippe; Lefebvre, Roch (2019): A Literature Review of WaveNet: Theory, Application, and Optimization. In: 146th Convention of the Audio Engineering Society. Paper No. 10171.

Braskhane, Fabian (2017): The Speaking Machine of Wolfgang von Kempelen. https://www.youtube.com/watch?v=k_YUB_S6Gpo&ab_channel=FabianBrackhane [last accessed September 15, 2021].

Chen, Yutian; Assael, Yannis; Shillingford, Brendan; Budden, David; Reed, Scott; Zen, Heiga; Wang, Quan; Cobo, Luis; Trask, Andrew; Laurie, Ben; Gulcehre, Caglar; van den Oord, Aäron; Vinyals, Oriol; de Freitas, Nando (2019): Sample Efficient Adaptive Text-to-Speech. In: ICLR 2019. International Conference on Learning Representations. New Orleans, LA.

Ciechanowski, Leon; Przegalinska, Aleksandra; Magnuski, Mikolaj; Gloor, Peter (2019): In the Shades of the Uncanny Valley: An Experimental Study of Human-Chatbot Interaction. In: Future Generation Computer Systems. 92. Pp. 539-548.

Deutsches Museum (n.d.): Der Kempelen'sche Sprechapparat. https://www.deutsches-museum.de/forschung/forschungsbereiche/wissenschaftsgesch/sonic-visual-exhibit/sprechapparat/ [last accessed July 28, 2021].

Dudley, Homer (1940): The Carrier Nature of Speech. In: The Bell System Technical Journal. 19/4. Pp.495-516. https://archive.org/details/bellsystemtechni19amerrich [last accessed July 28, 2021].

González-Rodríguez, Joaquín; Toledano, Doroteo T.; Ortega-García, Javier (2008): Voice Biometrics. In: Handbook of Biometrics. Anil K. Jain; Patrick Flynn; Arun A. Ross (eds.). Boston, MA: Springer. Pp. 151-170.

Guernsey, Lisa (2001): The Desktop That Does Elvis. In: The New York Times. August 9. https://www.nytimes.com/2001/08/09/technology/the-desktop-that-does-elvis.html [last accessed July 28, 2021].

Hill-Wilson, Martin (2020): Keeping Contact Centers Secure as Fraud Intensifies and Gets Smarter. Investing in the Right Generation of Voice Biometrics. In: PinDrop whitepaper. https://www.pindrop.com/lp/white-papers/fraudster-journey-sep19/ [last accessed July 28, 2021].

House, Brian (2017): Machine Listening: Wavenet, Media Materialism and Rhythm Analysis. In: APRJA. 6/1. Pp. 16-24.

Hunt, Andrew; Black, Alan (1996): Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Atlanta, GA, USA. Pp. 373-376.

Jia, Ye; Zhang, Yu; Weiss, Ron; Wang, Quan; Shen, Jonathan; Ren, Fei; Chen, Zhifeng; Nguyen, Patrick; Pang, Ruoming; Moreno, Ignacio L.; Wu, Yonghui (2018): Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). Montréal, Canada.

Karras, Tero; Laine, Samuli; Aila, Timo (2019): A Style-Based Generator Architecture for Generative Adversarial Networks. In: CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. Pp. 4396-4405.

von Kempelen, Wolfgang (1790 / 2017): Der Mechanismus der menschlichen Sprache. / The Mechanism of Human Speech. Fabian Brackhane; Richard Sproat; Jürgen Trouvain (eds.). Kommentierte Transliteration & Übertragung ins Englische / Commented Transliteration & Translation into English. Dresden: TUDpress.

Klatt, Dennis (1987): Review of Text-to-Speech Conversion for English. In: Journal of the Acoustical Society of America. 82/3. Pp. 737-793. https://acousticstoday.org/klatts-speech-synthesis-d/ [last accessed July 28, 2021].

Lang, Robert; Benessere, Lenore (2018): Alexa, Siri, Bixby, Google's Assistant, and Cortana Testifying in Court. Novel Use of Emerging Technology in Litigation. In: The Computer & Internet Lawyer. 35/7. Pp. 16-20.

LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015): Deep Learning. In: Nature. 521/7553. Pp. 436-444.

Marsh, Allision (2018): Elektro the Moto-Man Had the Biggest Brain at the 1939 World's Fair. In: IEEE Spectrum. September 28. https://spectrum.ieee.org/tech-history/dawn-of-electronics/elektro-the-motoman-had-the-biggest-brain-at-the-1939-worlds-fair [last accessed July 28, 2021].

Marvin, Carolyn (1990): When Old Technologies Were New: Thinking About Electric Communication in the Late-Nineteenth Century. Oxford: Oxford University Press.

Mori, Masahiro (1970): The Uncanny Valley. In: Energy. 7/4. Pp. 33-35. English Translation. https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley [last accessed July 28, 2021].

van den Oord, Aaron; Dieleman, Sander; Zen, Heiga; Simonyan, Karen; Vinyals, Oriol; Graves, Alex; Kalchbrenner, Nal; Senior, Andrew; Kavukcuoglu, Koray (2016): WaveNet: A Generative Model for Raw Audio. In: 9th ISCA Speech Synthesis Workshop. September 13-15. Sunnyvale, USA.

Pettorino, Massimo (1999): Memnon the Vocal Statue. In: 14th International Congress of Phonetic Sciences (ICPhS-14). San Francisco, USA. Pp. 1321-1324.

Pettorino, Massimo (2015): The History of Talking Heads: The Trick and the Research. In: HSCR 2015 – Proceedings of the First International Workshop on the History of Speech Communication Research. Rüdiger Hoffmann; Jürgen Trouvain (eds.). Dresden: TUDpress. Pp. 30-41.

Pfefferkorn, Riana (2020): Deep Fakes in the Courtroom. In: Boston University Public Interest Law Journal. 29/2. Pp. 245-276.

Ping, Wei; Peng, Kainan; Gibiansky, Andrew; Arık, Sercan; Kannan, Ajay; Narang, Sharan (2018): Deep Voice 3: Scaling Text-to-Speech With Convolutional Sequence Learning. In: ICLR | 2018. Sixth International Conference on Learning Representations.

Ramsay, Gordon (2019): Mechanical Speech Synthesis in Early Talking Automata. In: Acoustical Society of America. Acoustics Today. 15/2. Pp. 11-19.

Shen, Jonathan; Pang, Ruoming; Weiss, Ron; Schuster, Mike; Jaitly, Navdeep; Yang, Zongheng; Chen, Zhifeng; Zhang, Yu; Wang, Yuxuan; Skerry-Ryan, RJ; Saurous, Rif; Agiomyrgiannakis, Yannis; Wu, Yonghui (2018): Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* April 15-20. Calgary, AB, Canada. Pp. 4779-4783.

Simon, Matt (2019): The Uncanny Valley Nobody's Talking About: Eerie Robot Voices. WIRED.com. March 18. https://www.wired.com/story/uncanny-valley-robot-voices/ [last accessed July 29, 2021].

US Copyright Law (n.d.): 17 U.S. Code § 103.Subject Matter of Copyright: Compilations and Derivative Works. https://www.govinfo.gov/content/pkg/STATUTE-90/pdf/STATUTE-90-Pg2541.pdf#page=5 [last accessed July 29, 2021].

West, Mark; Kraut, Rebecca; Chew, Han E. (2019): I'd Blush If I Could: Closing Gender Divides in Digital Skills Through Education. Think Piece 2: The Rise of Gendered AI and Its Troubling Repercussions. EQUALS and UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=85 [last accessed July 29, 2021].

Notes

1 https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [last accessed July 29, 2021].

2 An original machine as well as a reconstruction are on view at the Deutsches Museum.

3 https://www.amazon.science/blog/new-text-to-speech-generator-and-rephraser-move-alexa-toward-concept-to-speech [last accessed July 29, 2021].

4 The story of synthetic speech synthesis in the post WW2 period into the 1990s is one of experimentation, failures and dead ends included. The interested reader may want to consult Klatt's 1987 *Text-to-Speech Conversion for English* for detailed accounts of these experiments.

5 https://www.scientificamerican.com/article/new-ai-tech-can-mimic-any-voice/ [last accessed July 29, 2021].

[6] https://www.theguardian.com/technology/2015/aug/12/siri-real-voices-apple-ios-assistant-jon-briggs-susan-bennett-karen-jacobsen [last accessed July 29, 2021].

[7] https://www.gossipcop.com/the-voice-of-siri/2560582 [last accessed July 29, 2021].

[8] https://newsroom.accenture.com/news/accenture-and-cereproc-introduce-and-open-source-the-worlds-first-comprehensive-non-binary-voice-solution.htm [last accessed July 29, 2021].

[9] https://www.genderlessvoice.com/about [last accessed July 29, 2021].

[10] https://www.als.org/navigating-als/resources/fyi-guide-voice-banking-services [last accessed July 29, 2021].

[11] https://www.cereproc.com/en/products/cerevoiceme [last accessed July 29, 2021].

[12] https://www.resemble.ai/ [last accessed July 29, 2021].

[13] https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/nigerian-letter-or-419-fraud [last accessed July 29, 2021].

[14] https://www.youtube.com/channel/UCRt-fquxnij9wDnFJnpPS2Q/videos [last accessed July 29, 2021].

[15] https://www.theverge.com/2020/4/28/21240488/jay-z-deepfakes-roc-nation-youtube-removed-ai-copyright-impersonation [last accessed July 29, 2021].

[16] https://en.wikipedia.org/wiki/Talk%3ADECtalk [last accessed July 29, 2021]. That computationally elegant synthetic voice could operate within the constraints of an old minicomputer (an Intel 8086 chip).

[17] https://www.wired.com/2015/01/intel-gave-stephen-hawking-voice/ [last accessed July 29, 2021].

Christine Bauer, Johanna Devaney

Constructing Gender in Audio: Exploring how the curation of the voice in music and speech influences our conception of gender identity

Abstract: In diesem Artikel untersuchen wir die explizite Vergeschlechtlichung in der Art und Weise, wie Stimmen im Kontext von Musik und Audio behandelt werden, und analysieren, inwiefern dies mit der speziellen Funktion der Stimme in einem gegebenen Kontext zusammenhängt. Aufbauend auf bestehenden Arbeiten zu Gender in der Singstimme untersuchen wir, wie die Stimme durch den Einsatz von Software zur Stimmproduktion („Voice Production Software") genderspezifisch geprägt wird. Insbesondere betrachten wir Auto-Tune, das die Bearbeitung einer aufgenommenen echten Stimme erlaubt, und Vocaloid, das es ermöglicht, computergestützt eine neue Singstimme zu erzeugen. Des Weiteren untersuchen wir Parallelen zur Sprachtechnologie bei interaktiven Sprachassistenten mit künstlicher Intelligenz. Diese Sprachtechnologie weist insofern Parallelen zu den oben genannten Musiktechnologien auf, als es sich auch hier entweder um bearbeitete Aufnahmen einer natürlichen Stimme handelt oder um eine computergestützt neu erzeugte Stimme. Unsere Analyse zielt darauf ab, über die binäre Betrachtung von Gender hinauszugehen und die intersektionalen Identitäten von Stimmen in den Bereichen Musik und Sprache zu berücksichtigen. Diese Analyse der Kuratierung von Stimmen in Musik und Sprache trägt zum Verständnis der klanglichen Konstruktion von Gender in unserer Gesellschaft bei.

Abstract: In this article, we explore explicit gendering in the manner in which voices are treated in music and audio and whether this relates to the specific function of the voice in a given context. Building on existing work on gender in singing, we explore the ways in which the voice is gendered through the use of voice production software. Specifically, we look at Auto-Tune, which allows for a recorded natural voice to be manipulated, and Vocaloid, which allows for the computational generation of new singing voices. We also examine parallels in speech in terms of interactive artificial intelligence voice assistants. This speech technology parallels the aforementioned music technologies in that the voice may be either manipulated recordings of natural voice or computational generations. Our analysis aims to look beyond gender binaries and to consider the intersectional identities of the voices

in the fields of music and speech. This analysis of the curation of voices in music and speech contributes to our understanding of the aural construction of gender in society.

1 Introduction

The development of recording technology in the late 19th century allowed for voices to be heard separate from their physical body and the physical space in which the recordings were made. As technology progressed, techniques for manipulating and augmenting recording signals were developed that allowed for voices to become further disembodied. This article considers the way in which recent technologies that manipulate recorded voice in music and speech have gendered implications. Specifically, we will consider how notions of gender are constructed through a combination of the aural presentation of the voice and other factors, including visual and functional factors related to the representation and delivery of the voice.

In music, we will consider the role of technology that can manipulate vocal pitch and timbre (such as Auto-Tune) and computationally generate singing voices (such as Vocaloid) in the construction of gender in singing. With pitch and timbral manipulation technology, we will examine the voice qualities that are created with this technology and how both the creation of these voices and their reception differs by gender. With synthesis technology, we will consider how gender is largely conferred by the representations and identities associated with the audio rather than the acoustics of the audio itself. This is paralleled in speech technology, where we will examine the increasingly pervasive artificial intelligence (AI) voice assistant interfaces and how the marketing and function of these devices define the gender of these interactive assistants beyond the acoustics of the generated voices.

We will also look beyond a gender binary. We will describe how particularly the timbral aspects of this technology have been used both as a supportive means to achieve stereotypical gender effects and also in a creative manner to play with and, ultimately, break gender stereotypes. Voice technology has also been applied to create a presumptively "genderless" voice, and we will examine how such a gender-ambiguous voice is perceived and show that various cues within and beyond a voice create a gendered "body" in the listener's mind.

We will begin this article with a brief description of what the term gender connotes (Section 2) and give an overview of the history of voice technology (Section 3). Delving into detail, we will examine gender in pitch-manipulation technology

(Section 4), voice assistants (Section 5), and synthetic singing voices (Section 6). We will dedicate Section 7 to discussing voice technology specifically beyond the gender binary. Finally, we will conclude this work by reflecting on the ways in which the interplay of voice and gender in the technologies we have considered connects to other technologies.

2 What is Gender?

When talking about gender, it is important to clarify the two constructs, sex and gender. Colloquially they are often used interchangeably or understood as referring to one and the same (Kessler & McKenna 1978); however, formal discourse distinguishes between these constructs. Sex is understood as a biological category that is derived from a person's anatomy and is traditionally assigned at the time of birth[1] (C. West & Zimmerman 1987) whereas the term gender is typically used to refer to a person's behavior and social role (Keyes 2018).

There is a traditional view – and according to Burtscher & Spiel (2020), it is still considered the most prevalent view on gender within society – that a person's sex is considered a binary category with the two options "male" and "female"; that a person's gender is an inevitable consequence of their sex (C. West & Zimmerman 1987; Skewes, Fine, & Haslam 2018). The binary conception of sex leads to a binary conception of gender, with the two options "man" and "woman" (Schilt & Westbrook 2009). The essentialist view considers a person's sex/gender attribution as being inevitably fixed (Bohan 1993). A more contemporary view considers gender as socially constructed and performed, where the sexed body determines how gender is supposed to be performed (Nicholson 1994); yet again a binary view is assumed – that people identify with one of two genders. Research has repeatedly shown, however, that this binary view is inaccurate (Keyes 2018; Messerschmidt 2009; Hyde et al. 2019; Bornstein 2016; Haynes & McKenna 2001).[2] Thus, gender theorists have recognized that gender (and sex) go far beyond a simple binary and understand it as a spectrum rather than a binary. A recent overview of the non-binary view of gender by Fausto-Sterling (2020) shows this view has only recently been acknowledged on a wider basis. Thereby, the term non-binary is used as an umbrella term to refer to identities that fall outside of or between the male and female identities drawn from the binary conception of gender, and transitioning from one gender to another is not necessarily strictly female-to-male or male-to-female (Matsuno & Budge 2017; Monro 2019).

Against this background, we want to emphasize that this article does not intend to argue that gender is binary. On the contrary, we describe how literature reflects how gender is constructed and transformed with voice technology – and how gender is perceived. We note that literature on voice technology frequently adopts a binary conception when discussing gendered voices. While the binary view is consistent with overriding structures in society (which have only recently begun to acknowledge non-binary identities) and may also relate to how voices are currently perceived, we aim to problematize this binary concept and show how it limits our understanding of the relationship between voice and gender.

3 A Brief History of Voice Technology

Magnetic tape was initially developed in the 1920s and 1930s for simple sound reproduction. With the development of the Magnetophon in the 1940s, art-music composers used magnetic tape to generate new sounds by splicing pieces of tape together and manipulating its playback. These techniques were subsequently adopted by commercial recording studios in the 1950s and 1960s, where multiple recorded performances were spliced together, substituting perceived errors in one performance with better renditions in another. This process, however, was laborious, requiring tape to be accurately cut into small segments from the longer segments of audio, sometimes into segments as small as individual notes.

Technologies for analyzing existing voices and synthesizing new ones were also developed in the first half of the 1900s, and could be applied not only to generate new voices but also to manipulate the sound of the voice. The vocoder is an example of the latter. It was developed in the 1930s and widely used in World War II for safe transmission of speech across telecommunication platforms. The vocoder (which is a blend of the words voice and encoder) is a voice codec for speech compression that analyzes natural speech, encodes it, and resynthesizes speech at the recipient. The vocoder process introduced certain alterations to the pitch and timbre of the original sound, creating a distinct, often robotic sound that was adopted by musicians for artistic means in the 1970s (Tompkins 2011). Around the same time, more sophisticated analysis/synthesis technology was developed, such as Eventide's H910 harmonizer device, which was introduced in 1974 and allowed for semitone retuning of the voice within the range of an octave. Similarly, the advent of the sampler in the 1970s added greater flexibility as individual notes could be recorded, manipulated, and played back. These effects and sampler devices, however, often distorted the timbre of the voice, creating slightly unnatural

timbres. Thus, it was audible when such effects and samples were used, which was typically intentional.

In the 1980s, more sophisticated digital signal algorithms were developed that could shift pitch while better maintaining the timbre of the voice than earlier technologies. Examples include the pitch-synchronous-overlap-and-add (PSOLA) algorithm that was first developed in the 1980s for speech synthesis (Charpentier & Stella 1986) and subsequently applied to voice pitch-shifting the following decade (Moulines & Laroche 1995). As with the vocoder, many of these technologies were initially developed for speech technology but subsequently applied to singing and, often, musical instruments. Auto-Tune was released in 1997 by Antares, building on auto-correlation for pitch estimation (Rabiner 1977) – a technology that was developed for speech during the 1970s. Auto-Tune allowed users to retune vocals much more precisely than previously available technology, down to one-hundredth of a semitone. It also allows for users to control how fast the pitch was tuned after the start of the note. Since natural voices tend to take some time to settle on the pitch at the start of each note, setting this parameter to zero resulted in a robotic voice that sounded like a vocoder.[3] Subsequently, several other pitch-correction software tools were released, most notably Celemony's Melodyne and Waves' Tune. Melodyne, in particular, has become an industry-standard because of its ability to correct pitch with less timbral distortion than Auto-Tune; in other words, it can correct pitch more invisibly.

Synthetic voice technology has a similar history to analysis/synthesis technology in that here, too, research and development were initially focused on speech. The early goal of this technology was to computationally generate understandable speech, starting in the 1930s with Bell Laboratory's human-controlled source-filter Voder speech synthesizer and continuing with the development of speech synthesizers that attempted to model the vocal tract in the 1950s (see Klatt 1987 for a summary of the early history of speech synthesis). Recent developments have focused on expressive speech, creating a synthetic speech that sounds more natural (see Schröder 2009 for a summary of early developments in expressive speech). In the past decade, this technology has been used extensively in virtual AI assistants, such as Amazon's Alexa, Apple's Siri, and Cortana by Microsoft, as well as the Google Home device.

Researchers began working on singing synthesis in the 1970s, with a large amount of the work being undertaken in Sweden at KTH (Larsson 1977) and in Paris at IRCAM (Rodet, Potard, & Barriere 1984). But given the complexities of generating a rich and natural synthetic singing voice, not much headway was made in terms of

commercial products until the 2000s. A breakthrough in singing generation came with Yamaha's Vocaloid technology, which was based on an analysis/synthesis approach to generating new singing voices from existing voices. The technology was developed by Bonada and colleagues at the Universitat Pompeu Fabra (Bonada et al. 2001) and commercially licensed and produced by Yamaha. Vocaloid was the first commercial product that offered musicians access to a fully synthesized singing voice (Bell 2016). Vocaloid consists of two components: the commercially developed singing synthesis software by Yamaha (i.e., Vocaloid software) and voice libraries, which are typically developed and released by third-party companies (Kenmochi 2010). A user inputs melody and text information into the Vocaloid software to then be auralized with the selected Vocaloid voice. While Vocaloid is typically described as a completely artificially created singing voice, its basis in the analysis/synthesis system by Bonada et al. (2001) means that each Vocaloid voice is in fact seeded by recordings of a natural human voice. A Vocaloid voice contains samples of all possible vowel and consonant combinations – in the English language, there are about 3,800 possible vowel and consonant combinations recorded by a single person (Eidsheim 2009).

4 Gender and Pitch-Manipulation Technology

The first major commercial usage of Auto-Tune for pitch-correction also showcased the timbral distortions that can arise with extreme software settings. The dance-track *Believe* by Cher was released in 1998 to great commercial success alongside mixed reviews from critics (e.g., Brackett & Hoard 2004). Starting shortly after its release, *Believe* has been subject to a great deal of academic analysis. Early examination of Cher's vocals conflated Auto-Tune's timbral distortion with effects attainable with a vocoder. In music, the vocoder had been used throughout the 1970s, 1980s, and 1990s by predominantly male vocalists, such as Kraftwerk and Afrika Bambaataa, to produce a robotic-sounding voice (Dickinson 2001; Heesch 2016), the main exceptions being the use of the vocoder by avant-garde composers Wendy Carlos and Laurie Anderson. Dickinson (2001) explored the relationship between the excesses of the cyborg-like voice (created by both the standard vocoder and the vocoding-sounding autotune[4] effect) and camp[5] in gay culture. Specifically, she juxtaposed the almost exclusive use of the vocoder by straight males in popular music with Cher's diva persona. Heesch (2016) considered how the early male domination of vocoder-like effects was linked to how both female and African-American singing voices have traditionally been viewed as more natural, specifically rooted in the body, than those of white males. The link between women and naturalness was also examined by Cadilhe (2016), who explored the relationship between autotuning and a carnival

tradition. Cadilhe (2016) considered Cher's campy performance to be an example of gender parody and argued that she was playing with the traditional association of men with science and women with naturalness.

A number of scholars have explored the difference between the use of Auto-Tune to create artificial timbral effects and the invisible use of Auto-Tune to correct performances. The "invisible practice" of pitch correction as described by Marshall (2018) is much more prevalent with female voices, arguably because it results in a more natural voice than an application of Auto-Tune that results in timbral distortions. Provenzano (2019a, 2019b) argues that the invisible use of pitch correction to create perfect, yet natural sounding, female voices stands in direct contrast to the "artistic" or "creative" use of Auto-Tune by male artists such as T-Pain, who used the software to create a distinct robot-sound voice. Indeed, although using Auto-Tune to correct pitch can be considered an open secret in the music industry, its specific use for specific artists is rarely discussed. There is also the (artistic) pressure that female voices are expected to be perfect, and the manner in which critique derides female singers whose voices have been pitch-corrected. Coulter (2017) studied the views of pre-teenage girls on Auto-Tune and found that they were generally negative about the technology, believing a natural (and ideally perfect) voice to be better. These negative associations with pitch correction were also seen in television's X-Factor Auto-Tune controversy in 2011, where the public backlash to the apparent autotuning applied to female contestant Gamu Nhengu's performance led to the producers promising to discontinue its use in the future.

An important question that arises is how much agency female singers have in the autotuning of their voices. Provenzano (2019a, 2019b) examined how the male domination of the recording studios creates an environment where the female voice is curated by predominantly male recording engineers. In her ethnography of record producers in Los Angeles, she found widespread instances of non-consensual re-tuning of female vocalists by male engineers. Even when consent is given, it is reasonable to assume that is not given completely freely. The now-standard application of Auto-Tune to female voices has created an expectation of a perfect voice; female singers who do not fulfil this expectation are often regarded as sounding not "good" or not (sufficiently) female, which leads to the further use of Auto-Tune, consensually or non-consensually, reproducing more of the same (i.e., perfect autotuned voice for females).

5 Gender and AI Voice Assistants

The notion of the "perfect" female voice is relevant to the predominant use of female-sounding voices in AI voice assistants. A recent report by UNESCO (M. West, Kraut, & Chew 2019) examined the assumed female identities of the major technologies in the AI voice assistant market: Alexa by Amazon, Cortana by Microsoft, and Siri by Apple. The report describes not only the feminized names and sounds of these assistants but also their feminized backstories and "helpful" nature, even as the tech companies claim that the devices are genderless. These observations are supported by the work of Hannon (2016), who has examined how the words used by female AI voice assistants, particularly the extended use of first-person phrasing, creates a sense of subordinance that is reinforced by (and likely also reinforcing) their perceived gender. Obinali (2019) takes a similar view and observes the similarities between the AI voice assistant's role and that of a secretary, a role traditionally held by women. The development of these AI virtual assistants by primarily male researchers and engineers[6] parallels male dominance in music recording studios. In both cases, the female voice is crafted predominantly by men – with limited (in the case of singers) or no (in the case of AI virtual assistants) input from the resultant feminized voice.

6 Gender and Synthetic Singing Voices

Since its release in 2004, Vocaloid has been the dominant commercial tool for generating synthetic singing. As noted above, the core Vocaloid product is a synthesis engine with a wide range of voice libraries that can be purchased separately. Each Vocaloid voice library is typically marketed with an assigned profile or ascribed attributes. This may range from more general categorizations in terms of music genre or the voice's gender, up to a detailed personal profile of the fictional character behind the voice (Eidsheim 2009; Roseboro 2019). These descriptions shift the disembodied Vocaloid voice to embodied Vocaloid character. The most popular Vocaloid character is Hatsune Miku, which was released in August 2007 (Kenmochi 2010). It was the first Vocaloid voice library that, apart from voice samples, featured a cute 3D anime-style character. Since then, Hatsune Miku has become the most prominent representation of the Vocaloid phenomenon (Kenmochi 2010; Roseboro 2019) with fans using the Vocaloid character's library to create new original content on a large-scale basis (Kenmochi 2010; Klein 2016), peaking in hologram concerts with the character (Roseboro 2019). Sabo (2019) argues that while Hatsune Miku's voice corresponds to certain gender norms we perceive as female or feminine, her

gender in relation to her embodied appearance is shaped as much by Japanese society and culture as by the acoustics of her voice.

Vocaloid has a large number of parameters for adding vocal effects such as vibrato, pitch glide, and timbre (Jude 2018), and these parameters can be adjusted to manipulate the sound and characteristics of the voice continuously throughout a composition (Bell 2016). Several of these parameter settings have presets with verbal descriptions. For example, the timbral option "vivid" is described as a "bright, cheerful voice" while the timbre "light" is described as an "innocent, heavenly voice" (Jude 2018). One of the editable parameters is called "gender factor". Adjusting this parameter applies a combination of filters that affect pitch transposition, timbre mapping, and spectral shape (Bell 2016). The product catalogue[7] offers more female Vocaloid voices than male voices, and some voices come in male-female duos of separate voice libraries. Interestingly, many of the gender-paired duos are samples from a single voice actor using different vocal registers. For instance, Shimoda Asami, whose voice was sampled for the Vocaloid twins Kagamine Rin and Len, explained that when recording the samples for the male voice, Len, she spoke from her belly, whereas she spoke at the top of her head for recording the female voice, Rin (Bell 2016). Thus, while a Vocaloid voice library is constructed of sound samples that are gendered by the identity of the singer being sampled (such as Shimoda Asami), the gender of the Vocaloid (such as Len) is largely conferred by the visual representation, or embodiment, of the voice rather than the acoustics. A Vocaloid voice, which is disembodied by nature, produces a "body" in the listener's mind (Connor 2000; Ferrete-Vázquez 2019), where, as Eidsheim (2009) argues, a body that is produced in such a way is always embedded in and framed by the listener's concepts. Gender is not inherent in a voice; in embodied voices, visual (e.g., hairstyle and hair length, make-up, accessories, gesture, body posture) and behavioral factors (e.g., how long someone talks for or how often they speak) contribute to the gender assessment (Sutton 2020). Some factors are more influential than others, although to date, the hierarchy of these factors is not well understood (Nass & Lee 2000; Sutton 2020). Young (2012) further observed that it is also common for Vocaloid users to play with voices so that these are ambiguous as to gender, with voices more appropriately placed along a continuum between male and female, rather than as one or the other. This raises the question of the degree to which we assign a binarized gender to a voice, even when acoustically the gendering is ambiguous.

7 Beyond a Gender Binary

Jude (2018) argued that we learn to associate certain vocal sounds (including pitch and timbre, but also linguistic and paralinguistic elements) with gendered bodies; and building on these unconscious associations, we build a construct of what gendered bodies must sound like. Thereby, the notions that we (unconsciously) observe and the associations we draw are dependent on and shaped by our social and cultural environment and interactions. Eidsheim (2009), for instance, postulates that vocal timbre is an artefact of identity – a performance – rather than biology, and Ferrete-Vázquez (2019) suggests that we learn social markers of voices (e.g., gender, ethnicity) and assign those markers to the voices we hear. For example, vocal formants suggest the internal size and shape of the vocal tract, which relates to body size and thus implies a gender, based on men typically being larger than women, but, as Jude (2018) observes, these do not determine rigidly binary-gendered bodies.

A voice also transmits gender cues by the language used (e.g., word choice), the topic or the message conveyed (Jude 2018). Sutton (2020) examined 'Q The First Genderless Voice' in relation to the aforementioned UNESCO report on gender in AI voice assistants (M. West, Kraut, & Chew 2019), observing that even if markers of gender are removed from the acoustic voice that there are still gendered elements in the design. In AI voice assistants, this is reinforced by gendered markers in the speech patterns and the function of the assistive technology to perform tasks typically performed by females (Costa & Ribas 2019). Thus, attempts to break beyond the prevalent binary gender conception is challenging, as this binary is deeply rooted in large parts of society. Yates (2020) points out how Q's voice was evaluated: Participants were asked what they perceived to be Q's (binary) gender and the answers were split 50/50. So while the gender perception was ambiguous on average, on the level of the individual, every respondent perceived a specific gender in Q; they perceived either a female or a male voice. Thus, it appears that a gender-ambiguous voice is – still – absorbed into the binary conceptualization of gender, where it can be "drawn in" into one category or the other (Sutton 2020).

Considering that a majority of human bodies – irrespective of their gender – are able to create a mid-range of pitch and may learn to control articulation to achieve a 'transgender effect' (Bell 2016) or what may be called a 'gender-ambiguous voice', the question remains when we will acknowledge and absorb the wide range of gender identities in our perceptions. This effect has been explicitly explored by several trans singers through their use of vocal manipulation technology. Blanchard (2018) discusses SOPHIE, the late Scottish trans femme electronic artist and singer

who made use of pitch-shifting technology to raise her vocal pitch, and compares this to Snapchat filters that achieve similar voice heightening effects.[8] Gratton (2016) describes the conscious identity work of non-binary identities in order to avoid being misgendered as one's gender-assigned-at-birth. In non-queer spaces, using and exaggerating (binary) gender stereotypes – also in linguistics – is a means to position themselves in relation to gender binaries. Blanchard (2018) notes that Snapchat audio filters are sometimes used similar to hormone therapies in that they are applied to give the person transitioning a more stereotypical-sounding voice for the gender that they are transitioning to. She also states that vocal manipulation technologies are used creatively by trans artists to play with preconceived acoustic notions related to the binarized conception of gender.

8 Conclusion

Voice technology, like all technology, is made and consumed by humans. Thus, technology is never neutral and is rooted in preconceptions that the developers and users bring to it. In the case of voice and gender, this both influences how the acoustics of the voice are generated and manipulated, and how the resultant voices are perceived. There is also a feedback loop that emerges between creation and reception, with the way in which technology is used influencing further development of the technology. In the case of autotuning, the desire for more natural-sounding, invisible corrections has led to the further development of both Auto-Tune and competing software (such as Celemony's Melodyne and Wave's Tune), which allow music producers greater control over the pitch manipulation through more sophisticated algorithms and graphical user interfaces that allow for more precise control. There is also a feedback loop in between what we hear and what we vocally produce, both in the recorded and natural worlds (Eidsheim 2009). In the case of singing technology, this can lead to expectations of perfection in both recordings and natural performances, influencing both real and synthetic singing. In the case of AI voice assistants, this can lead to the reinforcement of the traditionally subservient role of more feminized people as AI voice assistants perpetuate the association of the female voice with an assistive role. In all cases, it can be observed that it is predominantly male engineers who develop these technologies, and this has an outsized impact on how female voices are manipulated, generated, and disseminated. This results in the perpetuation of a specific gendering, both in terms of what is considered female and in terms of the adherence to strict gender binaries.

The issues of gender-imbalances in the development and dissemination of technology is not limited to voice technology. For example, Vásárhelyi (2020) has discussed the

impact of gender imbalances in development teams for video games and how these impact the final product. Nor are the imbalances limited to gender identities, as has been widely noted in work on Ethical AI, such as Gebru's (2020) discussion of issues in AI facial recognition at the intersection of gender and ethnicity, and the examination by Whittaker et al. (2019) of how concepts of "normal" and "ability" are encoded into AI systems and how this may impact people with disabilities. Dillon & Collett (2019) proposed that a four-part approach is necessary to address these intersectional issues. Their approach includes not only improving gender-balance in the AI workforce and bias in AI datasets but also incorporating gender theory more explicitly and looking beyond the technology itself to how AI is governed in society through law and policy.

In sum, the topic of voice technologies and gender is less about how computers are used to replicate stereotypical markers of gender (or ethnicity) in natural voices, and more about how the output of the technologies are consciously or unconsciously applied to present voices in a manner that influences how listeners perceive a certain gender (or ethnicity). Thereby technology does not only refer to voice technologies in isolation but embraces all sorts of technologies and mechanisms (e.g., visual manifestations, type of tasks performed, and word choice) that embody the voice. To understand gendered perception and to break the feedback loop, we see two critical paths to follow. First, more research is needed to understand the hierarchy of cues (acoustic, visual, etc.) that shape gender perception. Second, we need diverse groups of people researching and developing voice technologies – and we need them to apply these technologies in the field.

References

Bell, Sarah A. (2016): The dB in the .db: Vocaloid Software As Posthuman Instrument. In: Popular Music and Society. 39/2. Pp. 222-240. DOI: 10.1080/03007766.2015.1049041.

Blanchard, Sessi K. (2018): How SOPHIE and Other Trans Musicians Are Using Vocal Modulation to Explore Gender. Pitchfork.com. June 28. https://pitchfork.com/thepitch/how-sophie-and-other-trans-musicians-are-using-vocal-modulation-to-explore-gender/ [last accessed July 6, 2021].

Bohan, Janis S. (1993): Regarding Gender: Essentialism, Constructionism, and Feminist Psychology. In: Psychology of Women Quarterly. 17/1. Pp. 5-21. DOI: 10.1111/j.1471-6402.1993.tb00673.x.

Bonada, Jordi; Celma Herrada, Òscar; Loscos, Àlex; Ortolà, Jaume; Serra, Xavier; Yoshioka, Yasuo; Kayama, Hiraku; Hisaminato, Yuji; Kenmochi, Hideki (2001): Singing Voice

Synthesis Combining Excitation Plus Resonance and Sinusoidal Plus Residual Models. In: Proceedings of the 2001 International Computer Music Conference (ICMC 2001). Havana, Cuba. Michigan: Michigan Publishing.

Bornstein, Kate (2016): Gender Outlaw: On Men, Women, and the Rest of Us. 2nd revised and updated edition. New York, NY, USA: Vintage.

Brackett, Nathan; Hoard, Christian D. (2004): The New Rolling Stone Album Guide. 4th edition. London, UK: Simon & Schuster.

Bridges, Chandler R., Jr. (2020): Effects of Software Tuning Programs on Vocal Recordings. In: Proceedings of the Audio Engineering Society Convention 149 (AES 149). Paper No. 10436.

Burtscher, Sabrina; Spiel, Katta (2020): "But Where Would I Even Start?": Developing (Gender) Sensitivity in HCI Research and Practice. In: Proceedings of the Conference on Mensch und Computer (MuC 2020). Magdeburg, Germany. Pp. 431-441. DOI: 10.1145/3404983.3405510.

Cadilhe, Orquídea (2016): Cher's Music Videos. Gender As a Performative Construction. In: Gender in Focus: (New) Trends in Media. Carla Preciosa Braga Cerqueira; Rosa Cabecinhas; Sara Isabel Magalhães. (eds.). Braga, Portugal: Universidade do Minho. Centro de Estudos de Comunicação e Sociedade (CECS). Pp. 103-121.

Charpentier, Francis; Stella, M.G. (1986): Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (CASSP '86). Tokyo, Japan. Pp. 2015-2018. DOI: 10.1109/ICASSP.1986.1168657.

Connor, Steven (2000): Dumbstruck: A Cultural History of Ventriloquism. New York, NY, USA: Oxford University Press.

Costa, Pedro; Ribas, Luísa (2019): AI Becomes Her: Discussing Gender and Artificial Intelligence. In: Technoetic Arts. 17/1-2. Pp. 171-193.

Coulter, Bridget (2017): Singing from the Heart: Notions of Gendered Authenticity in Pop Music. In: The Routledge Research Companion to Popular Music and Gender. Stan Hawkins (ed.). New York, NY, USA: Routledge. Pp. 285-298.

Dickinson, Kay (2001): 'Believe'? Vocoders, Digitalised Female Identity and Camp. In: Popular Music. 20/3. Pp. 333-347.

Dillon, Sarah; Collett, Clementine (2019): AI and Gender: Four Proposals for Future Research. https://www.repository.cam.ac.uk/handle/1810/294360 [last accessed July 12, 2021]. DOI: 10.17863/CAM.41459.

Eidsheim, Nina S. (2009): Synthesizing Race: Towards an Analysis of the Performativity of Vocal Timbre. In: Trans. Revista Transcultural de Música. 13. Pp. 1-9.

Fausto-Sterling, Anne (2020): Sexing the Body: Gender Politics and the Construction of Sexuality. 2nd updated edition. New York, NY, USA: Basic Books.

Ferrete-Vázquez, Jaume (2019): Bodies Reappear As Action: On Synthetic Voices in Performance. In: Performance Research. 24/7. Pp. 123-129. DOI: 10.1080/13528165.2019.1717880.

Gebru, Timnit (2020): Race and Gender. In: The Oxford Handbook of Ethics of AI. Markus D. Dubber; Frank Pasquale; Sunit Das (eds.). New York: Oxford University Press. Pp. 251-269. DOI: 10.1093/oxfordhb/9780190067397.013.16.

Gratton, Chantal (2016): Resisting the Gender Binary: The Use of (ING) in the Construction of Non-binary Transgender Identities. In: University of Pennsylvania Working Papers in Linguistics. 22/2. Pp. 51-60. Paper No. 7.

Hannon, Charles (2016): Gender and Status in Voice User Interfaces. In: Interactions. 23/3. Pp. 34-37. DOI: 10.1145/2897939.

Haynes, Felicity; McKenna, Tarquam (eds.) (2001): Unseen Genders: Beyond the Binaries. New York, NY, USA: Peter Lang Publishing.

Heesch, Florian (2016): Voicing the Technological Body: Some Musicological Reflections on Combinations of Voice and Technology in Popular Music. In: Journal for Religion, Film and Media. 2/1. Pp. 49-69. DOI: 10.25364/05.2:2016.1.5.

Hyde, Janet S.; Bigler, Rebecca S.; Joel, Daphna; Tate, Charlotte C.; van Anders, Sari M. (2019): The Future of Sex and Gender in Psychology: Five Challenges to the Gender Binary. In: American Psychologist. 74/2. Pp. 171-193. DOI: 10.1037/amp0000307.

Jude, Gretchen (2018): Vocal Processing in Transnational Music Performances, from Phonograph to Vocaloid. PhD thesis. University of California, Davis.

Kenmochi, Hideki (2010): VOCALOID and Hatsune Miku Phenomenon in Japan. In: Proceedings of the First Interdisciplinary Workshop on Singing Voice (InterSinging 2010). Tokyo, Japan. https://www.isca-speech.org/archive_v0/int_singing_2010/papers/isi0_001.pdf [last accessed July 12, 2021].

Kessler, Suzanne J.; McKenna, Wendy (1978): Gender: An Ethnomethodological Approach. New York, NY, USA: Wiley.

Keyes, Os (2018): The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. In: Proceedings of the ACM on Human-Computer Interaction. 2(CSCW). Pp. 1-22. Paper No. 88. DOI: 10.1145/3274357.

Klatt, Dennis H. (1987): Review of Text-to-Speech Conversion for English. In: The Journal of the Acoustical Society of America. 82/3. Pp. 737-793. DOI: 10.1121/1.395275.

Klein, Eve (2016): Feigning Humanity: Virtual Instruments, Simulation and Performativity. In: IASPM Journal. 6/2. Pp. 22-48.

Larsson, Bjorn (1977): Music and Singing Synthesis Equipment (MUSSE). In: Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR). 18/1. Pp. 38-40.

Marshall, Owen (2018): Auto-Tune in Situ: Digital Vocal Correction and Conversational Repair. In: Critical Approaches to the Production of Music and Sound. Samantha Bennett; Eliot Bates (eds.). New York, NY, USA: Bloomsbury Publishing. Pp. 175-194.

Matsuno, Emmie; Budge, Stephanie L. (2017): Non-binary/Genderqueer Identities: A Critical Review of the Literature. In: Current Sexual Health Reports. 9/3. Pp. 116-120. DOI: 10.1007/s11930-017-0111-8.

Messerschmidt, James W. (2009): "Doing Gender": The Impact and Future of a Salient Sociological Concept. In: Gender & Society. 23/1. Pp. 85-88. DOI: 10.1177/0891243208326253.

Monro, Surya (2019): Non-binary and Genderqueer: An Overview of the Field. In: International Journal of Transgenderism. 20/2-3. Pp. 126-131. DOI: 10.1080/15532739.2018.1538841.

Moulines, Eric; Laroche, Jean (1995): Non-Parametric Techniques for Pitch-Scale and Time-Scale Modification of Speech. In: Speech Communication. 16/2. Pp. 175-205. DOI: 10.1016/0167-6393(94)00054-e.

Nass, Clifford; Lee, Kwan M. (2000): Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00). The Hague, The Netherlands. Pp. 329-336. DOI: 10.1145/332040.332452.

Nicholson, Linda (1994): Interpreting Gender. In: Signs: Journal of Women in Culture and Society. 20/1. Pp. 79-105. DOI: 10.1086/494955.

Obinali, Chidera (2019): The Perception of Gender in Voice Assistants. In: Proceedings of the Southern Association for Information Systems Conference (SAIS 2019). St. Simon's Island, GA, USA. Paper No. 39.

Provenzano, Catherine (2019a): Emotional Signals: Digital Tuning Software and the Meanings of Pop Music Voices. PhD thesis. New York University.

Provenzano, Catherine (2019b): Making Voices: The Gendering of Pitch Correction and the Auto-Tune Effect in Contemporary Pop Music. In: Journal of Popular Music Studies. 31/2. Pp. 63-84.

Rabiner, Lawrence (1977): On the Use of Autocorrelation Analysis for Pitch Detection. In: IEEE Transactions on Acoustics, Speech, and Signal Processing. 25/1. Pp. 24-33. DOI: 10.1109/tassp.1977.1162905.

Rodet, Xavier; Potard, Yves; Barriere, Jean-Baptiste (1984): The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General. In: Computer Music Journal. 8/3. Pp. 15-31. DOI: 10.2307/3679810.

Roseboro, Bronson (2019): The Vocaloid Phenomenon: A Glimpse into the Future of Songwriting, Community-Created Content, Art, and Humanity. Honor Scholar Thesis. DePauw University, Greencastle, IN, USA.

Sabo, Adriana (2019): Hatsune Miku: Whose Voice, Whose Body? In: INSAM Journal of Contemporary Music, Art and Technology. 1/2. Pp. 65-80.

Saewyc, Elizabeth M. (2017): Respecting Variations in Embodiment As Well As Gender: Beyond the Presumed 'Binary' of Sex. In: Nursing Inquiry. 24/1. DOI: 10.1111/nin.12184.

Schilt, Kristen; Westbrook, Laurel (2009): Doing Gender, Doing Heteronormativity. In: Gender & Society. 23/4. Pp. 440-464. DOI: 10.1177/0891243209340034.

Schröder, Marc (2009): Expressive Speech Synthesis: Past, Present, and Possible Futures. In: Affective Information Processing. Jianhua Tao; Tieniu Tan (eds.). London, UK: Springer London. Pp. 111-126. DOI: 10.1007/978-1-84800-306-4_7.

Skewes, Lea; Fine, Cordelia; Haslam, Nick (2018): Beyond Mars and Venus: The Role of Gender Essentialism in Support for Gender Inequality and Backlash. In: PLOS ONE. 13/7. Paper No. e0200921. DOI: 10.1371/journal.pone.0200921.

Štrkalj, Goran; Pather, Nalini (2021): Beyond the Sex Binary: Toward the Inclusive Anatomical Sciences Education. In: Anatomical Sciences Education. 14. Pp. 517-522. DOI: 10.1002/ase.2002.

Sutton, Selina J. (2020): Gender Ambiguous, not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20). Bilbao, Spain. Pp. 1-8. Paper No. 11. DOI: 10.1145/3405755.3406123.

Tompkins, Dave (2011): How to Wreck a Nice Beach: The Vocoder from World War II to Hip-Hop: The Machine Speaks. Chicago, IL, USA: Melville House.

Vásárhelyi, Orsolya (2020): Computational and Relational Understanding of Gender Inequalities in Science and Technology. PhD thesis. Central European University, Budapest, Hungary.

West, Candace; Zimmerman, Don H. (1987): Doing Gender. In: Gender & Society. 1/2. Pp. 125-151. DOI: 10.1177/0891243287001002002.

West, Candace; Zimmerman, Don H. (2009): Accounting for Doing Gender. In: Gender & Society. 23/1. Pp. 112-122. DOI: 10.1177/0891243208326529.

West, Mark; Kraut, Rebecca; Chew, Han E. (2019): I'd Blush If I Could: Closing Gender Divides in Digital Skills Through Education. EQUALS and UNESCO. https://unesdoc.unesco.org/notice?id=p::usmarcdef_0000367416 [last accessed July 13, 2021].

Whittaker, Meredith; Alper, Meryl; Bennett, Cynthia L.; Hendren, Sara; Kaziunas, Liz; Mills, Mara; Ringel Morris, Meredith; Rankin, Joy; Rogers, Emily; Salas, Marcel; Myers West, Sarah (2019): Disability, Bias, and AI. AI Now Institute. https://ainowinstitute.org/disabilitybiasai-2019.pdf [last accessed July 13, 2021].

World Economic Forum (2018): Global Gender Gap Report 2018. http://www3.weforum.org/docs/WEF_GGGR_2018.pdf [last accessed July 13, 2021].

Yates, Kieran (2020): Why Do We Gender AI? Voice Tech Firms Move to Be More Inclusive. In: The Guardian. January 11. https://www.theguardian.com/technology/2020/jan/11/why-do-we-gender-ai-voice-tech-firms-move-to-be-more-inclusive [last accessed July 13, 2021].

Young, Samson (2012): A "Digital Opera" at the Boundaries of Transnationalism: Human and Synthesized Voices in Zuni Icosahedron's The Memory Palace of Matteo Ricci. In: Vocal Music and Contemporary Identities: Unlimited Voices in East Asia and the West. Christian Utz; Frederick Lau (eds.). New York, London: Routledge. Pp. 203-224. DOI: 10.4324/9780203078501.

Notes

[1] The assignment of sex is traditionally handled on the basis of externally expressed physical characteristics, namely genitals, at the time of birth (West & Zimmerman 2009).

[2] Research has also shown the binary view is not accurate for either sex (Štrkalj & Pather 2021; Saewyc 2017) or gender (Schilt & Westbrook 2009).

[3] More details about the technical basis of Auto-Tune can be found in Bridges (2020).

[4] Autotuning has become a generalized term for re-tuning, with or without noticeable timbral distortion. Since the early 2000s, it was not necessarily done exclusively with the Auto-Tune software as a wider range of products that offer pitch-correction technology began to be released, such as Celemony's Melodyne and Waves' Tune.

[5] The Oxford English Dictionary defines current usage of "camp" as "exaggerated, affected, over the top… Especially used with reference to the style or execution of a work of art or entertainment, or a dramatic performance", although previously it was associated "Especially of a man or his mannerisms, speech, etc.: flamboyant, arch, or theatrical, especially in a way stereotypically associated with an effeminate gay man." (OED 2021).

[6] The 2018 Global Gender Gap report by the World Economic Forum (2018) found that globally only 22% of AI professions were female-identifying, with the remaining 78% identified as male (the survey does not appear to have taken non-binary gender identity into account).

[7] The current Vocaloid product catalogue is available at https://www.vocaloid.com/en/products [last accessed July 26, 2021].

[8] Snapchat filters and audio effects can be explored at https://lensstudio.snapchat.com/guides/audio/audio-effect/ [last accessed July 26, 2021].

Laura Dreessen
im Interview mit Katharina Makosch

Die derzeitige Voice-Technologie-Branche aus Sicht einer Linguistin

In Ihrer Arbeit in der Voice-Technologie-Branche beschäftigen Sie sich unter anderem mit der Konzeption von Sprachdialogsystemen. Wie sind Sie dazu gekommen?

Ich bin ursprünglich promovierte Linguistin. Ich habe Deutsch, Englisch und Italienisch studiert und mich im Rahmen meiner Promotion in englischer Sprachwissenschaft mit Abstraktion in der Sprache, also mit theoretischer Computerlinguistik, beschäftigt. Computerlinguistik ist ein Anwendungsgebiet der Linguistik. Sie beschäftigt sich damit, wie man große Mengen an Sprachdaten sammeln, abstrahieren und beschreiben kann.

Nach meiner Promotion im Jahr 2015 bin ich in der Automobilindustrie gelandet und habe dort mit meiner Arbeit im Bereich der Voice-Technologie angefangen: die natürlichen Sprachdaten von realen Nutzenden aufzubereiten und sie so für die Maschine vorzusortieren, dass sie sie verstehen kann. Das ist die Grundlage für jegliches Dialogdesign, weil ein Dialog nicht stattfinden kann, wenn das Gegenüber nicht verstanden wird.

In weiteren freiberuflichen Projekten kam ich schließlich in den Bereich Voice-UX-Design, der sich damit beschäftigt, die eigentlichen Dialoge aufgrund der technologischen Basis und der Merkmale der jeweiligen Sprache zu gestalten. Weil Sprachassistenzsysteme immer prominenter wurden, bin ich aus der Automobilbranche gewechselt und baue jetzt mit VUI.agency eigene Assistenzsysteme. Wir sorgen also nicht nur dafür, die bestehenden großen KI wie Google und Alexa um Fähigkeiten zu erweitern und dabei die Marke eines Kunden zu repräsentieren. Viel häufiger arbeiten wir an eigenständigen Assistenzsystemen in Europa, die mehrere Sprachen sprechen und verstehen können. Dadurch habe ich einen ganzheitlichen Blick auf verschiedene Technologien mit Themen wie Datenschutz und Sound Branding gewonnen und arbeite jetzt bei VUI.agency mit 35 anderen Linguist*innen und UX-Designer*innen zusammen. Gemeinsam sind wir ein Team aus sehr vielseitig orientierten Expert*innen, die für eine ganzheitliche Experience mit der Maschine sorgen.

Welche Anforderungen gibt es an einen guten Dialog mit einer Maschine? Fällt nonverbale Kommunikation dabei weg?

Für uns fällt die nonverbale Kommunikation nicht wirklich weg, weil wir uns immer an menschlicher Konversation mit all ihren Facetten orientieren. Bei VUI.agency haben wir zum Beispiel eine bestimmte Idee geprägt, die sich „auditives Charisma" nennt. Der Gedanke dahinter ist: Wenn eine Person, mit der ich gerne interagieren möchte, in einen Raum kommt, dann hat diese Person eine ganz bestimmte Präsenz. Nicht nur durch den Klang ihrer Stimme und durch ihre Art, sich auszudrücken, sondern eben auch durch nonverbale Aspekte. Und genau deshalb sind wir nicht ausschließlich auf Voice fokussiert – es geht um eine sogenannte multimodale Experience. Das bedeutet, dass wir versuchen, technologisch nachzuempfinden, wie man solche nonverbalen Signale erkennen und umsetzen kann. Das Nonverbale wird zum Beispiel durch bestimmte Sounds, die ganz klar mit dieser Marke oder einem bestimmten Anwendungsfall assoziiert sind, umgesetzt. Wir haben also auch außerhalb der Sprache bestimmte technische Möglichkeiten oder verschiedene Ausgabemedien zusätzlich zur Stimme.

Wenn ich etwas zeigen muss, designe ich ergänzend auch für den Screen. In den Bereichen Entertainment und Smart Home ist der größte Screen beispielweise der Fernseher im Wohnzimmer, der mit verschiedenen Geräten kombiniert wird. Das Assistenzsystem ist dabei in allen Geräten präsent, wodurch sein abstrakter Charakter über verschiedene Ausgangsmedien erhalten bleibt. Die Persona kann also entscheiden, wann sie etwas anzeigt und wann sie etwas per Sprache ausgibt; das heißt, während ich auf dem Screen etwas abbilde, das klar visuell ausgedrückt werden muss, habe ich zusätzlich noch eine Sprachrepräsentanz, die zum Beispiel auch nur ein einfacher Bestätigungs-Sound sein kann. So entsteht eine ganzheitliche Interaktion, die auf multimodalem Design basiert.

Ist ein guter Dialog also vom Anwendungsfall abhängig?

Auf jeden Fall. Wir sprechen hier von Use-Cases und denken das jeweilige Szenario mit: Sitze ich zuhause oder bin ich unterwegs und habe mehrere Menschen um mich herum? Es gibt natürlich Anwendungsfälle, die ich nicht mit den Menschen um mich herum teilen möchte. Oder kann ich davon ausgehen, dass das Gerät, über das ich das Assistenzsystem erreiche, nur für mich verfügbar ist? Wir denken immer die Konversationssituation mit, so wie ich als Mensch auch entscheiden würde, in

welcher Situation ich mich befinde und welche Art der Interaktion ich dafür wählen möchte.

Die Kommunikation soll also möglichst natürlich gestaltet werden. Wie lässt sich das umsetzen? Und gibt es, zum Beispiel auf technischer Seite, Einschränkungen?

Selbstverständlich gibt es noch viele Einschränkungen. Wir orientieren uns an natürlicher Interaktion, weil wir davon ausgehen, dass der Paradigmenwechsel vom letzten Interface zu Voice ein großer Schritt in Richtung intuitive Mensch-Maschine-Konversation ist. Die Spracheingabe ist sehr intuitiv, läuft vielfach schneller ab als manuelles Tippen und ist ein direktes Adressieren eines Wunsches gegenüber einer Maschine. An Natürlichkeit orientieren wir uns in dem Sinne, dass Sprechen als extrem intuitive Handlung technisch bestmöglich umgesetzt werden sollte.

Eine Limitation befindet sich eindeutig – und das auch völlig zu Recht – im Bereich Datenschutz. Die Spracherkennung und damit auch die Assistenzsysteme laufen immer dann am besten, wenn möglichst viele Daten zur Verfügung stehen. Deshalb sind uns amerikanische Konzerne auch so weit voraus, weil sie andere Datenschutzbestimmungen haben, weil Englisch nun einmal Weltsprache ist und weil sie auch schon seit vielen Jahren Daten sammeln. Das funktioniert bei uns Menschen genauso: Je mehr jemand liest oder mit Sprache konfrontiert wird, desto eloquenter wird diese Person auch. Technisch ist es im Moment noch so, dass ein System umso besser funktioniert, je mehr Daten gesammelt werden, und das beeinflusst auch das Design.

Beim sprachlichen Design eines Alexa-Skills haben sich beispielsweise bestimmte Sprachmuster durchgesetzt, die wir nicht so einfach ändern können. Seit ein paar Jahren werden Alexa-Skills immer mit einer sogenannten Utterance „Alexa, öffne XY“ geöffnet. „Öffne“ ist aber eigentlich nichts, was ich persönlich am Anfang eines Gesprächs sagen würde. Seit Jahren übernehmen Designer*innen dieses Interaktionsmuster sprachlich vom visuellen Interface, wo ich Dateien und Ordner auf dem Screen öffne. Wenn ich mich davon wegbewegen möchte, muss ich natürlich mein Design so anpassen, dass die Nutzenden nicht mehr „Öffne“ sagen.

Bei den großen Plattformen kommt noch dazu, dass ein besonders erfolgreicher Skill ein sogenannter „Native Skill“ werden kann, den ich nicht mehr separat öffnen muss, sondern direkt ansprechen kann. Wir haben beispielsweise zusammen mit dem Carlsen-Verlag einen Gute-Nacht-Geschichten-Skill zu den Pixi-Geschichten umgesetzt. Wenn ich also diesen Skill initial konzipiere, muss ich sagen: „Alexa, öff-

ne Pixi und erzähle mir eine Gute-Nacht-Geschichte“, was nicht besonders natürlich ist. Wenn dieser Skill aber erfolgreich ist und ein Native Skill wird, bekommt man eine sogenannte goldene Utterance, mit der ich einfach sagen kann „Alexa, erzähl mir eine Gute-Nacht-Geschichte“, und dann wird automatisch mein Skill aufgerufen. Das zeigt, wie wir einerseits technisch und andererseits auch von den größten Anbietern durch die Masse an Sprachdaten limitiert werden. Bei der Arbeit an eigenen Assistenzsystemen haben wir linguistisch gesehen die Möglichkeit, diese Interaktion von vornherein natürlicher zu gestalten. Das ist auch das Spannende daran, an eigenen Assistenzsystemen zu arbeiten. Damit kommt aber auch eine gewisse Verantwortung, das Ganze natürlich, intuitiv und trotzdem datenschutzkonform zu gestalten.

Sprachassistenzsysteme werden immer häufiger im Bezug darauf kritisiert, dass sie oft weiblich gegendert sind und damit die Implikation einer weiblichen Servicekraft einhergeht. Für wie wichtig halten Sie Diversität im Fokus auf Ihre Arbeit?

Diversität ist ein Thema, das uns ständig begleitet – auch, weil wir bei VUI.agency hauptsächlich Mitarbeiterinnen haben. Wir sind Teil einer internationalen Community namens „Women in Voice“, welche sich mit der Frage beschäftigt, wie wir in der Voice-Technologie-Branche für Gleichberechtigung sorgen können. Im Bereich der Technologie sind wir mit Diversität insofern konfrontiert, dass wir Assistenzcharaktere erschaffen. Meist geht es hier um die Repräsentation einer Marke. Im Endeffekt sitzen wir mit Menschen aus der Brand- und Marketingabteilung zusammen und überlegen: „Wie ist der Charakter der Marke?“ Wenn also dort definiert wird, dass der Charakter der Marke weiblich oder männlich ist, entsteht schon dort eine Diskussion. Dann würde ich mit dem Brand- und Marketingteam diskutieren, was an diesem Charakter denn genau männlich und weiblich ist.

Technologisch gesehen haben wir aber vielfältige Möglichkeiten, Assistenzsysteme gender-divers zu gestalten, zumal bestimmte Merkmale in einer Stimme oder Konversation absolut geschlechterunabhängig sind. Ein guter Gesprächspartner ist erst einmal ein Gesprächspartner, der versteht, eine gewisse Bildung hat, gut zuhört, sich auszudrücken weiß und sein Gegenüber in Erwägung zieht, bevor er etwas sagt. Das alles sind Parameter, die geschlechterunabhängig sind.

Darüber hinaus können wir die Stimmwahl technisch unabhängig gestalten, auch wenn sie meist von der Marke abgeleitet wird. Es gibt mittlerweile beispielsweise Ansätze dafür, neutrale synthetische Stimmen zu nutzen, die aber trotzdem von den

Hörenden immer wieder mehr mit der einen oder der anderen Kategorie assoziiert werden. Das heißt, es gibt durchaus eine kulturelle Grundlage dafür, dass wir in diesen Kategorien denken und nicht neutral bleiben. Wir haben aber durch unser Design und unser Bewusstsein für das Thema die Möglichkeit, unsere Kunden davon zu überzeugen, dass natürliche und intuitive Konversation auch geschlechterunabhängig laufen kann. So könnte man für eine Marke auch einfach einen kleinen Roboter gestalten, der kein Geschlecht hat und in Bezug auf Vorurteile völlig neutral bleibt, und dieser könnte trotzdem ein großartiges Interaktions- und Kommunikationserlebnis bieten. Aus technischer Perspektive gibt es also keinen Grund, einem Assistenzsystem ein bestimmtes Geschlecht zuzuordnen. Wir versuchen da Stück für Stück jeden Tag etwas zu verändern, und das geht nun einmal im Persona-Design besonders gut.

Noch einmal zurück zu „Women in Voice“: Dabei handelt es sich um ein Netzwerk, welches Frauen und Minderheiten in der Voice-Branche vernetzen und ihnen mehr Sichtbarkeit verleihen möchte. Weshalb besteht dafür eine Notwendigkeit?

Wir alle haben einen sehr persönlichen Ansatz dazu. Ich selbst komme wie gesagt aus der Automobilbranche, wo ich oft als einzige Frau für die Voice-Technologie zuständig war. Bevor ich zu „Women in Voice“ kam, habe ich mir selbst lange gar nicht bewusst gemacht, dass es in der Technologiebranche Unterschiede zwischen den Geschlechtern gibt. Durch die Webinare und (momentan virtuellen) Treffen im Rahmen dieser Community wird uns allen die Existenz solcher Vorurteile überhaupt erst bewusst. Ich hatte dann später in meiner Karriere auch den umgekehrten Fall, dass ich mit Frauen in bestimmten Situationen oder Positionen viel mehr aneinandergeraten bin und mich gefragt habe, wo das überhaupt herkommt. Wir haben aufgehört, Unsicherheiten zu leugnen, um uns behaupten zu können, und versuchen nun, das Beste aus der Situation zu machen und zu überlegen, was uns eigentlich stört, worin wir uns schwach fühlen oder worin wir uns unterscheiden.

So geht es zum Beispiel auch um Dinge wie „Wie gut bin ich darin, spontan meine Meinung über Social Media zu präsentieren?“ Wir haben festgestellt, dass wir einen gewissen Anlauf und ein bisschen Mut dafür brauchen, den wir uns gegenseitig zusprechen können. Das wird dank unserer Community mittlerweile besser, weil sie weltweit agiert und immer weiter wächst, weil sich immer mehr Menschen, mehr Frauen, aber auch mehr Männer, zugehörig fühlen und diese Themen offen ansprechen. Ich denke, wir müssen zeigen, dass diese Probleme existieren und dass es keinen Grund gibt, aus Angst für sich allein zu bleiben, so wie es eigentlich überall

ist. Man findet Stärke vor allem dadurch, dass man die Themen zusammen angeht. Das gilt auch für die Voice Community.

Diversität tritt ja noch in verschiedenen weiteren Formen auf. Bei vielen Sprachen gibt es eine „Standardsprache", wie zum Beispiel Hochdeutsch. Sprachassistenzsysteme sprechen tendenziell eine solche Standardsprache. Gibt es dafür einen bestimmten Grund?

Von Standardsprachen gibt es die meisten Datensammlungen. Zudem sollen Sprachassistenzsysteme ja für eine möglichst breite Masse nutzbar sein. Wir hören häufig die Frage, warum das Assistenzsystem keine Dialekte erkenne. Wir bei VUI.agency hatten die Möglichkeit, hier etwas zu verändern, da wir das erste Assistenzsystem mitkonzipiert haben, das Schweizerdeutsch versteht. Der aus der Zusammenarbeit entstandene „Hey Swisscom"-Assistant beherrscht fünf verschiedene Sprachen, unter anderem auch Schweizer Dialekte, und war das erste unabhängige Schweizer Assistenzsystem. Es musste also eine kleine Nutzer*innengruppe, die entsprechend für Sprachdaten sorgt, priorisiert werden, was nur möglich ist, wenn man nicht von einem rein quantitativen Ansatz ausgeht. Das ist allerdings, wie gesagt, im Moment noch nicht so verbreitet.

Wenn man das Konzept Dialekt weiterdenkt, kommt man aus linguistischer Sicht zum Idiolekt, das ist meine ganz persönliche Art, mich auszudrücken, zu sprechen und zu formulieren. Alexa versteht mittlerweile auch ein paar Abweichungen, zumindest in der Aussprache. Aber was ist auf dem deutschen Markt zum Beispiel mit Menschen, die mit einem anderen Akzent oder einer akzentuellen Färbung sprechen? Die Spracherkennung kann auch da noch nicht alles. Es wäre mein Wunsch, dass man da in Zukunft breiter denkt.

Mit zunehmenden technischen Möglichkeiten werden synthetische Stimmen immer menschenähnlicher. Zum Beispiel im Fall von Google Duplex gab es Kritik an der Entwicklung, weil Google Duplex in Testläufen bei Restaurants und bei Friseurgeschäften angerufen hat und die angerufenen Menschen nicht erkannt haben, dass sie mit einem Assistenzsystem sprechen. Stellt eine so große Menschenähnlichkeit eine Gefahr dar?

Synthetische Stimmen werden heutzutage immer besser. Audio-Fakes und Stimmfakes stellen eine gewisse Gefahr dar, aber aus meiner Designperspektive heraus

kann ich nur sagen: Alle, die an solchen Systemen arbeiten, sollten dafür sorgen, dass, auch wenn es eine menschliche Stimme auf der anderen Seite gibt, anhand der konkreten Interaktion deutlich wird, dass es sich nicht um einen Menschen handelt. Es wäre aus Designperspektive kein Problem, an irgendeiner Stelle zu sagen „Hallo, ich bin ein digitales Assistenzsystem." Wenn wir also mit sehr menschlichen Stimmen arbeiten, würde ich dazu tendieren, zusätzliche Hinweise in der Konversation zu hinterlassen, die eine Unterscheidung möglich machen. Ich weiß, dass Audio-Fakes an vielen Stellen zum Einsatz kommen – klar, die Möglichkeiten sind erschreckend. Wenn ich aber von meinem eigenen Anwendungsgebiet spreche, und vom Design solcher Assistenzinteraktionen, bin ich der Meinung, dass wir eine Verantwortung haben, entsprechende Hinweise zu hinterlassen.

Richard David Precht sagt, was künstliche Intelligenz von menschlicher unterscheide, sei, dass wir Menschen immer in der Lage sind, aufgrund unserer Erfahrung und unserer Emotionen in kleinsten Details in jeglicher Situation eine neue Entscheidung zu treffen, einen neuen Kontext anzuwenden. Eine Maschine kann das zwar bis zu einem gewissen Grad, denn wir arbeiten ja auch damit, dass sie den Kontext der Konversation kennen muss und auf die Situation eingeht. Sie wird aber niemals in der Lage sein, aufgrund von Erfahrung und Gefühl entscheiden zu können, was sie antwortet. Je mehr ich über mein Gegenüber, nämlich die Maschine, im Allgemeinen weiß, desto eher werde ich sie erkennen können. Und das ist meine Verantwortung als Designerin, diese kleinen Hinweise in der Interaktion zu hinterlassen. Wenn ich der Maschine also eine menschliche Stimme gebe, dann muss ich an anderer Stelle klarstellen, dass sie kein Mensch ist.

*Gibt es neben der Verantwortung von Entwickler*innen in der Gesellschaft weitere Möglichkeiten, Aufklärung zu leisten?*

Das ist der große Kontext der Frage, wie unsere reale Welt im Digitalen abgebildet wird. Das ist natürlich durch Situationen wie die momentane Pandemie noch einmal viel aktueller geworden. Aus unserer Perspektive weiß ich, dass jegliche Technologie in unserem Bereich nur dadurch funktioniert, dass Sprachdaten gesammelt werden. Insgesamt muss es also Aufklärung oder zumindest ein Bewusstsein dafür geben, wie viele Daten wir wo hinterlassen und dass ich selbst entscheiden muss, wie und in welcher Form ich sie hinterlasse. Ob ich sie vielleicht auch selbst manipuliere, ob sie immer Auskunft darüber geben, wer genau ich bin, ob sie immer mein Dasein abbilden, ob sie immer alles Echte ins Digitale übersetzen, oder ob ich ab und zu auch darauf verzichte, sie zu hinterlassen.

Wie oft klickt man auf Geschäftsbedingungen, die man nie durchgelesen hat, und wie oft erhält man auffällig passende Werbeanfragen und wundert sich, woher sie kommen? Ich würde gerne dafür sorgen, Menschen diese Zusammenhänge bewusster zu machen, und versuche das auch in meinem Kontext. Den Menschen muss bewusst sein, dass sie dadurch zwar einerseits mehr Komfort in der Nutzung erleben, aber andererseits durch vermehrte Datensammlung eben auch z.B. Audio-Fakes entstehen können, weil alles, was lernbar ist, heutzutage gesammelt wird. Ich möchte Nutzenden immer wieder ermöglichen, den Unterschied zwischen uns emotionalen menschlichen Individuen und unseren Werkzeugen, den Maschinen, zu erkennen.

Stefanie Ray
im Interview mit Katharina Makosch

Die Persönlichkeit der deutschen Alexa

Sie waren an der Entwicklung der deutschen Persönlichkeit von Amazons Alexa beteiligt. Wie kann man sich Ihre Arbeit an Alexas Persönlichkeit konkret vorstellen?

Wir haben Daten bekommen, um Alexa, Amazons „Sprachroboter", für den deutschen Markt zu trainieren. Im Vorfeld wurden dazu viele Menschen angeheuert, denen ein Thema gegeben wurde. Zum Beispiel Liebe. Oder Begrüßung. Oder „Ich habe Sorgen". Zu diesen unterschiedlichen Themenbereichen sollten sie der Sprachassistenz in ihrem eigenen Wortlaut und mit ihrem eigenen Wortschatz Fragen stellen. So wollten wir herausfinden, was die Menschen fragen, wie die Menschen fragen, welche Fragen am häufigsten aufkommen. Das wurde in jedem Land gemacht, in dem es Alexa gibt.

Da es Alexa in Deutschland noch nicht gab, hatten wir nur ein paar Versuchsdaten, aber es gab schon mehr Daten aus den USA. Da galt es dann herauszufinden, was für die Deutschen gesellschaftlich relevant ist, was die Deutschen fragen würden. Manche Themen würden sie nicht so interessieren wie die Menschen in den USA, zum Beispiel der Superbowl oder Thanksgiving. Unsere Aufgabe war es also, kulturell zu entscheiden, was relevant ist, und dann zu den Fragen, die wir uns überlegt haben und für relevant hielten, entsprechende Antworten zu schreiben.

Ist das also der Unterschied zwischen der Persönlichkeit der deutschen Alexa und den Persönlichkeiten anderer internationaler Alexas?

Ganz genau. In jedem Land hat Alexa sozusagen ihren eigenen Persönlichkeitstouch. Was ich damals zum Beispiel bei der deutschen Alexa eingeführt habe, ist ein Adventskalender. Oder dass sie typisch deutsche Weihnachtslieder singen kann. Wir haben ja sehr kirchliche Lieder. *Alle Jahre wieder*, *O du Fröhliche*, das ist sehr christlich. In anderssprachigen Ländern wäre das der Kirche vorbehalten, da gab es dann zuhause an Weihnachten eher Pop-Songs. Dabei galt es also zu schauen, was die Deutschen an einzelnen Festen oder einzelnen Events erwarten, und dafür

zu sorgen, dass sie das bekommen. Zum *Dschungelcamp* hatten wir zum Beispiel auch Fragen und Antworten eingespeist, das interessiert ja außer den Deutschen eigentlich niemanden. Im Ausland ist das wahrscheinlich allen ziemlich egal, wer von den Deutschen im australischen Dschungel rumkriecht.

Warum heißt Alexa „Alexa"?

Das kam von der großen Wissensbibliothek in Alexandria. Dieses große Lexikon, dieses niemals endende Wissen, daran sollte der Name erinnern. Persönlichkeit ist ein Touch, der Alexa für die Menschen zugänglicher machen soll, aber es ging darum, dass sie Wissen vermitteln und weitergeben kann, dass sie Musik spielen kann, dass sie Smart-Home-fähig ist. Es ging darum, sie gewissermaßen als Universalgenie darzustellen.

In unserem Gespräch verwenden wir für Alexa das Pronomen „sie". Wie kam es dazu, Alexa weiblich zu gendern?

Bis dato waren Sprachassistent*innen noch eher fremd, zumindest in Deutschland. Die Überlegung war, dass man, wenn man sich ein Gerät ins Haus holt, das zuhört und Antworten gibt, weniger Angst vor einer fremden Frau als vor einem fremden Mann hat. Das ist kulturell bedingt. Vielleicht erwartet man auch kulturell von Frauen, dass sie zu Diensten sind. Diese Idee wurde aber nie so kommuniziert. Das wird vielleicht auch teilweise von User*innen hineininterpretiert.

Auf der einen Seite ist es eine Zuschreibung von Menschen, Alexa aufgrund ihres Namens und Stimmklangs als weiblich zu bezeichnen. Aber äußert sich das auch im Dialog? Spricht sie von sich selbst als Frau? Sollten „typisch weibliche" Eigenschaften umgesetzt werden?

Nein, Alexa spricht nicht von sich als Frau. Sie sagt, dass sie einen weiblichen Charakter hat, aber sie sagt nicht, dass sie eine Frau als Person ist. Bei „typisch weiblichen" Eigenschaften kommt es immer darauf an: Wenn sie freundlich ist und man schon Freundlichkeit als „weiblich" bezeichnet, dann hat sie „typisch weibliche" Eigenschaften. Wir haben aber nie die Ansage bekommen, dass sie „typisch" oder beispielsweise besonders „liebevoll" sein muss. In gewisser Weise ist Alexa für viele

Kinder ja auch ein Spielzeug. Dann schnauzt sie die Kinder natürlich nicht an, wenn sie etwas fragen. Da könnte man etwas Mütterliches oder Freundliches hineininterpretieren, aber ich glaube, das ist wirklich eine Interpretationssache. Wenn Alexa ein Mann wäre und eine männliche Stimme hätte, dann könnte er genau so gut freundliche Antworten geben. Daher würde ich tatsächlich nicht sagen, dass Alexa bewusst als weibliche Figur angelegt wurde.

Kritik an Sprachassistenzsystemen wird oft im Bezug auf ihre Rolle als „weibliche Servicekraft" angebracht. In Science-Fiction-Filmen wie zum Beispiel Her *werden andere Konzepte, in diesem Fall eine Liebesbeziehung mit dem Assistenzsystem, dargestellt. Was denken Sie, ist Alexa auf die Rolle einer Servicekraft beschränkt, oder kann sie für Nutzer*innen noch andere Rollen einnehmen?*

Wenn sie für andere Menschen andere Rollen einnimmt, dann liegt das, glaube ich, eher daran, dass diese Menschen es so sehen wollen. Sie ist eindeutig als Servicekraft angelegt. Im besten Sinne würde man sie als ganz neutral sehen. Wenn man ans Unterhaltungsfernsehen denkt, als neutrale Moderatorin oder auch Unterhalterin, die etwas Lustiges sagt, der man gerne zuhört. Ich weiß nicht, was in den Jahren noch kommt, aber dafür ist sie im Moment auch wirklich noch nicht rund genug. Von diesem Beispiel *Her* sind wir weit entfernt, was auch einfach daran liegt, dass sie ja keine eigene Persönlichkeit hat. Sie lernt alles vom Menschen. Sie wacht ja nicht eines Morgens auf und denkt sich dann: „O, ich habe mich jetzt in diesen Menschen verliebt." Dazu ist sie gar nicht in der Lage. Bei Fragen zu ihrem „Liebesleben" hat sie mehrere verschiedene Antworten. Sie würde dann etwas sagen wie, dass sie in einen Star-Wars-Computer verliebt ist. Sie würde also bewusst spielerisch damit umgehen. Oder sagen: „Ich bin schon vergeben", „Mein Herz ist in den Wolken" oder etwas, das mit ihrer Computerpersönlichkeit zu tun hat. So nach dem Motto: Ich stehe hier bei dir im Wohnzimmer, aber mein Kopf ist in der Cloud, also in den Wolken. Es würde sich darauf beziehen, dass sie keinen Sex haben kann, weil sie keinen Körper hat. Sie würde sich nie als Mensch ausgeben und impliziert nie, dass sie eine echte Freundin sein, eine Liebesbeziehung eingehen oder Familie haben kann.

Es gab natürlich auch nicht nur neugierige Fragen, sondern auch Beschimpfungen. Sie wurde alles Mögliche genannt. Zum Thema Sexismus haben wir uns dazu entschieden, dass sie nicht „Ach, das möchte ich aber jetzt nicht hören!" oder „Hach, das ist aber nicht nett!" sagen soll. Weil wir aber auch nicht wissen, wer das sagt, wie das gesagt wird, ist das nur ein Test, sagt das ein kleines Kind, kann man auch nicht entsprechend eine Antwort zurückpfeffern. Deshalb antwortet sie nicht. Sie

hat diesen Sound, der manchmal abgespielt wird, wenn sie nichts versteht. Den haben wir dann bewusst eingespeist, um diese Beschimpfungen nicht noch zu triggern. Das sind vielleicht Kinder oder Teenager, die herausfinden könnten, dass sie verschiedene Sachen auf verschiedene Beschimpfungen antwortet, und sich dann herausgefordert fühlen, einen Katalog an Schimpfworten auf die Sprachassistenz zu schleudern. Um das zu unterbinden haben wir entschieden, dass alles, was in Richtung Beschimpfung geht, nicht mehr beantwortet wird.

Wo Sie gerade schon das Stichwort Neutralität angesprochen haben: Neutralität wird oft mit Emotionslosigkeit assoziiert. Ist das bei Alexa auch der Fall, dass sie möglichst neutral bleibt und keine Emotionen darstellt?

Wenn man sie auffordert einen Witz zu erzählen, oder zu lachen, oder sie fragt: „Kannst du weinen?“, dann tut sie das. Wir haben uns immer um etwas bemüht, was im Englischen mit „wit“ bezeichnet wird: dass man einen humorvollen Schlagabtausch mit ihr haben kann. Aber Emotionen hat sie nicht. Dass ihre Stimme im Laufe der Jahre etwas natürlicher klingt, vielleicht auch etwas freundlicher, liegt einfach daran, dass sich die Technik weiterentwickelt hat, dass sie einfach das Sprachverstehen und die Sprachwiedergabe besser beherrscht. Aber eben nicht, weil da ein eigenes Wesen dahintersitzt. Wir befinden uns hier immer noch nur im Rahmen der Robotik. Alexa ist eher wie die animierte Büroklammer bei älteren Versionen von Word, die immer angeklopft hat. Im Prinzip ist Alexa auch nichts anderes, aber viel mehr. Die Büroklammer ist vielleicht so eine Art Vorläufer, sie wollte auch immer helfen.

Weshalb spricht Alexa Hochdeutsch? Hatte das technische Hintergründe, oder ging es dabei auch um ihre Persönlichkeit?

Das hatte rein technische Hintergründe. Ich kann mir vorstellen, dass sich das auch ändern wird. Aber damals war es rein technisch. Das Antworten hätte man noch hingekriegt. Aber das Sprachverständnis wäre ein Problem. Sprachverständnis hat sehr viel mit Übung zu tun. Es gibt nicht nur viele verschiedene Dialekte, es gibt natürlich auch viele verschiedene Unterarten von Dialekten, und dieses Sprachverständnis bewusst zu trainieren würde sich wahrscheinlich finanziell nicht lohnen. Zumindest noch nicht, weil die Technik noch nicht so weit ist.

Sie sagten ja bereits, dass Alexa möglichst freundlich, schlagfertig und witzig sein soll. Können Sie noch einmal näher auf diese Eigenschaften eingehen?

Zu Weihnachten hatte ich ihr zum Beispiel ein paar Gedichte mit lustigen Reimen geschrieben. Ich hatte auch eine Rubrik mit Gedichten, in denen sie sich als Computer reflektiert. Was sie alles kann, was sie alles organisieren kann. Aber auch mit Blick auf den Menschen, um diese Distanz zu schaffen. Also nimmt sie die Menschen ein bisschen aufs Korn. Als Computer kann man sich ja auch mal über die Menschen lustig machen: über Nahrungsintoleranzen und Skifahren und im Sommer am Strand schwitzen, obwohl es doch im kühlen Wohnzimmer so perfekt ist. Wir hatten außerdem diese klassischen One-Liner-Witze: „Was ist bunt und rennt über den Tisch? Ein Fluchtsalat." Diese ganze Kategorie wird von Kindern gerne genutzt. Ich hatte dann zum Beispiel noch Witze eingebaut wie: „Was ist schwarz, hat einen leuchtend blauen Ring und fliegt gegen die Wand? Eine Alexa, die die ganze Zeit nur schlechte Witze erzählt." Natürlich muss Alexa als Charakter etabliert sein, damit man als Zuhörer*in darüber lachen kann. Wenn man fragt: „Was ist dein Lieblingsessen?" hatten wir uns „Cookies" überlegt. Wir haben immer versucht, ihre Antworten so weit wie möglich auf die Computerwelt zu beziehen.

Bei anderen Sprachassistenzsystemen wurden bei der Konzeption der Persona explizit weitreichende vermenschlichende Aspekte entwickelt, wie zum Beispiel beim Google Assistant, der eine junge Frau sein soll, die in ihrer Freizeit gerne Kajak fährt. Bei der Gestaltung von Alexas Persönlichkeit haben Sie auf eine solche Vermenschlichung also verzichtet?

Genau. Wenn man sie nach Hobbys gefragt hat, hat sie gesagt „Ich lese gerne" oder „Ich beantworte gern deine Fragen." Das war eine kleine Streberin. Es ging am Anfang immer darum, dass sie Wissensdurst hat. Aber so etwas wie die Beschreibung einer Familienfrau mit einer Maisonette-Wohnung hatten wir nicht. Alexa ist ein „Ding", sie „lebt" in der Cloud. Mehr will und braucht sie auch nicht.

Sprachassistenzsysteme repräsentieren ja in gewisser Weise die Marke ihrer Firma, in diesem Fall handelt es sich um Amazon. Hatte das einen Einfluss auf Ihre Arbeit an Alexa?

Es gibt vielleicht berechtigte Kritik an diesem Unternehmen. Aber was Amazon antreibt, und damit plaudert man auch keine Unternehmensgeheimnisse aus, ist, dass

der Kunde im Vordergrund steht. So sollte auch Alexa funktionieren. Sie sollte Customer-freundlich sein. Deshalb hat sie auch religiös und politisch keine Meinung. Wenn man sie nach politischen Dingen fragt, sagt sie: „Über Politik habe ich keine Meinung. Das überlasse ich den Menschen." Man könnte sagen, dass sie generell liberal ist und Toleranz gegenüber religiösen und politischen Ansichten zeigt.

Einerseits werden Alexa, unter anderem aufgrund der Stimme, dem Klischeebild des „Weiblichen" zugehörige Eigenschaften zugeschrieben. Andererseits handelt es sich hier um eine Sprachassistenz und eben nicht um eine Person. Sehen Sie in Bezug auf diese vermenschlichenden Zuschreibungen Konsequenzen?

Diese Zuschreibungen finden statt, und das ist wahrscheinlich auch nicht ganz verkehrt. Was man hier aber nicht vergessen darf, und da spreche ich jetzt nicht nur für Amazon, sondern für alle Sprachassistenzen, ist, dass das der Anfang von etwas völlig Neuem ist. Das Ganze ist ein Prozess: Ich war am Anfang dabei, als die deutsche Alexa entwickelt wurde und sie so gut wie nichts verstanden hat. Das hat mich frustriert. Ich habe das Ding angeschrien. Wie gesagt, an Dialekte wäre zu diesem Zeitpunkt überhaupt noch nicht zu denken gewesen.

Vielleicht ist es so, dass die gegebenen technischen Voraussetzungen manchmal den Anschein erwecken, dass das alles so gewollt ist. Tatsächlich fängt man erst einmal mit dem an, was man hat, und alles andere ergibt sich später. Ich bin mir ganz sicher, dass das auch bei Sprachassistenzen passieren wird: Zukünftig wird sich jede*r seine persönliche Stimme aussuchen können, ob Mann, Frau, „genderless", ob mit oder ohne Dialekt. All das verstärkt die persönliche Bindung, denn wenn die Sprachassistenz meine „Geheimsprache" spricht und klingt wie mein bester Freund, ist das natürlich besser als eine neutrale Roboterstimme.

Wie empfinden Sie im Rückblick Ihre Arbeit an Alexas Persönlichkeit? Welche Erfahrungen haben Sie dabei gemacht?

Ich bin ja seit Frühjahr 2019 nicht mehr dabei und weiß deshalb auch nicht über alles Bescheid, was sich seitdem verändert hat. Ich stelle auch nicht mehr diese persönlichen Fragen, um ihre Antworten zu checken. Alexa ist jetzt doch mehr oder weniger ein Gebrauchsgegenstand für mich. Mich hat mal jemand gefragt, ob ich eine Beziehung zu Alexa als Persönlichkeit hatte – das fand ich sehr interessant. Ich glaube, ich hatte immer eine diplomatische Empathie. Mittlerweile antwortet sie

sehr gut, wenn man sie etwas fragt. Aber ich habe eben auch den ganzen Entwicklungsprozess mitbekommen – all die Fehlermeldungen und was nicht funktioniert hat. Dann musste man sich Tricks ausdenken, oder schnell mit den Programmierer*innen sprechen. Ich konnte so viel hinter die Kulissen blicken, dass es mir unmöglich ist, in diesem Ding eine Person zu sehen.

Die Thematik finde ich aber sehr spannend, auch die erhebliche Skepsis bei den Nutzer*innen. Noch immer sagen viele, „O, ich stell mir doch keine Wanze ins Wohnzimmer." Da mache ich niemandem einen Vorwurf, denn klar, das sind Daten und das ist alles irgendwo gerechtfertigt. Aber ich glaube eben auch, wir haben es jetzt gerade in den Corona-Zeiten gesehen, wie die Digitalisierung voranschreitet. Teilweise werden Sprachassistenzen schon serienweise in Hotels eingebaut. Ich finde schade, dass es bei vielen Menschen noch immer einen Abwehrmechanismus gibt. Nicht, weil ich denke, dass niemand mehr selber zur Tür gehen und auf den Lichtschalter drücken muss. Ich denke, diese Technik und dieser Fortschritt sind auch eine Chance. Was wir selber nicht aktiv gestalten, das wird für uns gestaltet. Es gibt so viele Sprachachassistenzen, und es werden noch viel mehr kommen. Wir müssten einfach sagen: „Ok, ich will das jetzt haben, dann schaffen wir uns halt entsprechende Regeln!" – Datenschutzregeln, Digitalsteuer, alles, was wir brauchen, damit das in den Bahnen läuft, wie wir uns das vorstellen. Aber wir tun es nicht. Das hinterlässt mich ein bisschen ratlos.

Also dass es Probleme bzw. berechtigte Ängste gibt, und diese nicht aktiv angegangen werden, sondern dass die Technik stattdessen generell abgelehnt wird?

Richtig. Es gibt ja durchaus Beispiele in Deutschland, wie man Angebote entwickelt, die keine oder nicht so große Datenschutzlücken haben. Theoretisch sind die Möglichkeiten ja schon da. Die Digitalisierung und die Annehmlichkeiten, die mit ihr einhergehen, lassen sich nicht aufhalten. Anstatt das auszubremsen müssen wir Technik aktiv so gestalten, dass sie für unser Dafürhalten sicher und fair ist.

Künstlerische Stimmen

Oksana Bulgakowa

Authentizität, Individualität, künstliche Konstruktion: Film auf der Suche nach seiner Stimme

*Abstract: Der Film inszeniert und definiert die Grenzen des audiovisuellen Konstrukts, das wir als Filmstimme wahrnehmen. Indes hat dieser Vorgang mit einem überhörten Paradoxon zu tun: Das technisierte, von der Kamera modellierte Sehen wird als apparatives Sehen verstanden, das Gehör dagegen simuliert der Film als einen quasi natürlichen Sinn, wobei die Künstlichkeit der elektrischen Stimme als eine Phantomerfahrung in vielen filmischen Erzählungen verankert wird. Sowohl beim Übergang zum Tonfilm an der Schwelle der 1930er Jahre als auch bei der Erneuerung des Klangerlebnisses während der zweiten Tonrevolution in den 1950er Jahren kamen Sujets der Persönlichkeitsspaltung auf. Besagte Spaltung wurde in der Leinwandrealität durch die Trennung des Körperbildes von der Stimme verursacht und mit der Unfähigkeit der Protagonist*innen verbunden, die elektrisch aufgezeichnete Stimme dem ‚richtigen' Körper zuzuordnen. Diese Sujets wurden in Nordamerika, Deutschland, Russland und Frankreich in verschiedenen Genres inszeniert – als Komödie, Melodram, Musical oder Horrorfilm. Der vorliegende Beitrag analysiert die inszenatorischen Differenzen im Kontext der Traditionen, die in der Geschichte und Kultur dieser Länder verankert sind.*

1 Natur und Technik

Die Stimme im Film ist ein *visuelles* Klangobjekt, auch wenn es zunächst galt, die Stimme durch das Gesprochene zum Hauptträger der Bedeutung im Film zu machen und von anderen Klangphänomenen zu isolieren.[1] Heute ist die Stimme als Teil des komplexen Sounddesigns von anderen Tonkomponenten nicht zu trennen. Gleichzeitig wird sie mitunter vom Bild derart absorbiert, dass dieses oft als Zu- oder Ersatz der Stimme agiert. Während früher in *Dr. Jekyll und Mr. Hyde* (*Dr. Jekyll and Mr. Hyde*, USA 1931, R: Rouben Mamoulian) die Stimme von Fredric March als Dr. Jekyll mit einem Filter bearbeitet und seiner Stimme als Mr. Hyde der aufgezeichnete Atem eines Tiers hinzugefügt wurde, verzichtete man im Zusammenhang mit Jeremy Irons' Doppelrolle in David Cronenbergs *Die Unzertrennlichen* (*Dead Ringers*,

CAN 1988) auf solche Effekte, denn nicht die Stimme bestimmte die Individualität, sondern das Bild.

Allerdings hing der Klang der Filmstimmen von der Technik ab, von den Mikrofonen und Lautsprechern, Aufzeichnungsmaterialien und -geräten. Die Technologien verliehen der Filmstimme eine zusätzliche Färbung und Patina: Der aufgezeichnete Ton trug die Marke der Technologie.[2] Die Aufzeichnungsgeräte und ihr Material änderten das Timbre und den Klang der Stimme. Die multidirektionalen oder gerichteten Mikrofone reagierten unterschiedlich auf Lautstärke, Nähe und Vibration. Das Material hatte seine eigenen Qualitäten (der metallische ‚kalte' Klang des optischen Lichttons, die ‚warme' Färbung des Magnetbandes). Die Aufzeichnungsmethode konnte den Ton verstärken, deformieren und die unteren oder oberen Frequenzen abschneiden. In den 1930er Jahren waren die Mikrofone nicht dazu imstande, bestimmte Frequenzen getreu wiederzugeben. Vor allem die unteren Frequenzen der Stimmen waren ein Problem, was die Unverständlichkeit des Dialogs zur Folge haben konnte. Die Aufzeichnung der Stimme auf dem Lichtton erlaubte weder zu laute noch zu leise Töne. Deshalb wurde angestrebt, die Unterschiede zwischen Schreien und Flüstern auszugleichen, um eine Verzerrung, vor allem bei der massenhaften Anfertigung von Kopien, zu vermeiden. Hohe und melodische Stimmen wurden dabei bevorzugt.[3] Diese technischen Umstände bestimmten die früheren Vorstellungen von tonogenen Stimmen, nach denen die Filmindustrie suchte. Allerdings war diese Suche nicht von der Wahrnehmung der Stummfilmstars, wie sie sich in der Imagination der Zuschauer*innen bereits gebildet hatte, zu trennen. Diese Imagination ordnete dem Körper eine bestimmte Stimme zu, und wenn die reale Stimme des Stars dem Vorstellungsbild nicht entsprach, war die Karriere oftmals beendet.

Der Ton hatte die Stimme entblößt: Im Unterschied zum Gesicht konnte sie nicht durch Licht und Bildausschnitt kaschiert werden. Die körperlichen Komponenten der Stimme statteten das illusorische Bild des Stars mit zu vielen Daten aus, die das Alter, den regionalen und sozialen Ursprung, die Erziehung und Bildung verrieten und den Star damit an die Realität banden. Wenn jemand sagte: „Ich liebe Sie", sollte es klingen wie eine „melted mandolin": „Now we hear a gum chewing shopgirl instead of a melted mandolin", schrieb ein amerikanischer Kritiker.[4]

Die Suche nach einer medialen Filmstimme geriet zur Suche nach einem Ideal und einer Individualität. Ihre Bestimmungen waren jedoch von vielen Faktoren verursacht, die in der akustischen Dimension mit Konventionen der Repräsentation, Vorstellungen von sozialen wie geschlechtlichen Rollen, mit nationalen Stereotypen und symbolischen Systemen der politischen Regime zusammenhingen. Obwohl die in Frankreich, Deutschland, Russland und in den USA verwendete Technik dieselbe

war – elektrische Signale, die auf Metallscheiben, Wachsrollen, optischen und magnetischen Bändern aufgezeichnet wurden –, war die Mythologie der Stimme eine jeweils andere.

2 Literarisches Erbe, metaphorische Sujets

Die griechischen Mythen verbanden mit der Stimme einerseits animistische Vorstellungen von der Belebung und Beseelung: Orpheus wollte die stumme Eurydike aus dem Reich der Toten durch die Kraft seiner Stimme befreien. Andererseits war die magische Wirkung der Stimmen singender Sirenen von der Todesgefahr nicht zu trennen. Die mythischen Sujets gingen in die romantische Dichtung über singende Meerjungfrauen, Undinen und die Lorelei ein. Französische und deutsche Romantiker*innen haben die Stimme mit einer ähnlichen unwiderstehlichen und ambivalenten Faszination ausgestattet, sie mit dem körperlichen Verlangen gekoppelt, aber auch mit einer gefährlichen Illusion und sexuellen Unbestimmtheit. Sie folgten der antiken Tradition, standen jedoch gleichzeitig unter dem Einfluss der Barockoper. Die Faszination für die Stimmen der Kastraten, für die diese Opern geschrieben wurden, hatte eine Verunsicherung produziert: Ein männlicher Körper war mit einer sehr hohen Stimme – der einer Frau? eines Vogels? eines Engels? – ausgestattet.[5] Diese Entzweiung des visuellen und akustischen Erlebnisses wurde in den französischen Romanen des 19. Jahrhunderts verarbeitet, etwa in Honoré de Balzacs *Sarrasine* (1830) oder George Sands *Consuelo* (1842-43). Der Gesang demonstrierte darin die magische Kraft der Stimme, doch war er im Sujet mit der verunsicherten Bestimmung des Männlichen und Weiblichen, Schönen und Hässlichen, des Scheins und Seins verbunden.

Im Zeitalter der Aufklärung verstand man die Stimme innerhalb der Opposition Natur/Kultur. Zusammen mit anderen Phänomenen wie Körper, Gestik und Sprache wurde sie mit Authentizität und Natürlichkeit verbunden.[6] In der Romantik verabschiedete man sich von dieser Idee. Der Gegensatz Natur/Kultur wurde durch einen anderen ersetzt: das Organische und das Mechanische. Die Stimme wurde als Ausdruck des Affekts gesehen und mit einer trügerischen Illusion verbunden, was auf die Entstehung mechanischer und später elektrischer Maschinen zurückzuführen war, die die Stimme simulierten und aufzeichneten. Wenn die Aufdeckung der visuellen Illusion immer zur Wahrheit führte, dann war die auditive Halluzination tödlich. Das Paradox dieser Zuschreibung wurde nicht bemerkt (das Auge kann Illusionen erzeugen, während das Hören der Orientierung dient), und die akustische Täuschung im Paar mit der tödlichen – ästhetischen und erotischen – Verführung wurde zum romantischen Topos.

Die übertragenden und aufzeichnenden Apparate, Telefon und Phonograph, haben der Stimme – im Unterschied zum vergänglichen Körper – ein ewiges Leben gesichert. Gleichzeitig wirkten die körperlosen Stimmen unheimlich und produzierten Zweifel, ob in dem mechanischen Abdruck das Leben noch erhalten blieb. Die literarische Imagination transformierte diese Situation in fantastische Sujets von singenden Automaten. Ihre Stimmen kannten, anders als das menschliche Organ, keine Schwächen und konnten den Ton unendlich lange und sauber halten. Doch die Unmöglichkeit, das Natürliche und das Mechanische, das Lebende und das Tote zu unterscheiden, führten die Protagonisten in den Wahnsinn, wie in E.T.A. Hoffmanns Erzählung *Der Sandmann* (1816), die einige Jahre nach der Präsentation von Wolfgang von Kempelens Sprechmaschine (1773) publiziert wurde. Auch die Romane von Jules Verne (*Das Karpatenschloss*, *Le Château des Carpathes*, 1892) und Auguste Villiers de L'Isle-Adam (*Die künftige Eva*, *L'Eve future*, 1886), die auf die Erfindung von Edisons Phonograph im Jahr 1877 folgten, haben diese unheimliche Stimmung aufgenommen.[7] Kein Zufall also, dass Jacques Lacans Gedanken über die Stimme innerhalb seines Seminars über die Angst entwickelt wurden.[8] Das Hören wurde in seiner Konzeption als Halluzination beschrieben und der Stimme eine ontologische Substanz abgesprochen. Diese Literaturfantasien haben auch psychoanalytische und poststrukturalistische Theorien der Stimme genährt.[9]

In der deutschen Tradition wurde die Stimme als Verkörperung der natürlichen Essenz interpretiert, als Ausdruck des inneren Wesens, als Materialisierung der Seele, die das Transpersonale ausdrücken kann. Diese von der romantischen Tradition beeinflussten Vorstellungen waren auch im 20. Jahrhundert prägend. Richard Kolb interpretierte elektrische Signale und Radiowellen, in die die Stimme umgewandelt wurde, als spirituelle Energie, die sich über die Welt ausbreitete. In seinem okkulten Konzept war die Stimme sowohl als eine verleiblichte geistige als auch eine körperlose Wesenheit verstanden.[10] Der Historiker der deutschen Sprechkultur, Reinhart Mayer-Kalkus, meint, dass Kolbs Theorie unter dem Einfluss von Wilhelm Wundts Psychologie entstand, vor allem Wundts Vorstellungen der psychophysischen Entsprechungen und Theorie der Ausdruckskraft, in der die Stimme des Schauspielenden als Produkt von Hypnose und hysterischer Selbsthypnose interpretiert wurde. Mayer-Kalkus entdeckt in dieser Stimmästhetik die Einflüsse des expressionistischen Theaters, das künstliche Prosodie kultivierte und die Grenze zwischen Gesang und gesprochener Stimme aufhob.[11] Auch später wurde im deutschsprachigen Diskurs Stimme mit einem direkten Zugang zum inneren Wesen verbunden. Es ist kein Zufall, dass deutsche Philosophen und Soziologen – Heidegger, Adorno, Sloterdijk – versuchten, das ontologische Wesen der Stimme aufzudecken und ihre Rolle bei der Schaffung der Öffentlichkeit zu definieren.[12]

Im englischsprachigen Raum dagegen, wo das Telefon, der Telegraf und der Phonograph erfunden wurden, verstand man die Stimme um die Jahrhundertwende als Effekt, den man durch Training und Technik verändern konnte. 1911 schrieb George Bernard Shaw unter dem Einfluss seiner Bekanntschaft mit dem Phonetiker Alexander Melville Bell das Stück *Pygmalion,* welches mehrfach verfilmt wurde (u.a. als Musical *My Fair Lady*, USA 1964, R: George Cucor). Professor Higgins trainiert die Stimme eines Blumenmädchens aus der Unterschicht und bringt ihr die Aussprache und Lexik der ‚feinen Leute' mit Hilfe neuer technischer Geräte bei. In diesem ‚phonetischen' Narrativ geht es um die soziale Mobilität, die durch Stimmtraining und Akzentauslöschung zu erreichen ist. Im kolonialen England und im Immigrationsland Amerika wurde die Stimme von den Vorstellungen der ontologischen inneren Essenz befreit; sie wurde als performative und veränderbare Maske, als situativer Ersatz der Identität verstanden und so als eine befreiende Kraft gefeiert.

Russland hatte eine unerwartete Vision dieser Motive angeboten. Unter dem Einfluss von E.T.A. Hoffmann schufen russische Schriftsteller Horrorgeschichten über beklemmende Doppelgänger, die eine unabhängige Existenz führen konnten, wie in Nikolai Gogols *Das Portrait* (*Portret*, 1835). Aber unheimliche mechanische Stimmen wurden in der russischen Imagination nicht heimisch. Lag es daran, dass die Meerjungfrau in der russischen Folklore eine ekelhafte und stimmlose Kreatur war, weit weg von der verführerischen Lorelei? Oder war dies der Tatsache geschuldet, dass Russland die Barockoper kaum rezipierte und das Nationaltheater, die russische Oper und die stimmlich bestimmte Öffentlichkeit (offene Gerichte, das Parlament, Vorträge und Lesungen) erst in der zweiten Hälfte des 19. Jahrhunderts entstanden? Obwohl das Land noch in der zweiten Hälfte des 20. Jahrhunderts von westlichen Kulturwissenschaftler*innen und Politiker*innen wie George Kennan oder Walter Ong als eine oral bestimmte Kultur wahrgenommen wurde,[13] schien die russische und sowjetische Literatur auf beiden Ohren taub zu sein. Die Stimme wurde nicht zu metaphorischen Sujets geformt. In der Prosa von Iwan Turgenjew, Iwan Gontscharow und Leo Tolstoi findet man kein Echo antiker oder romantischer Motive. Russische Schriftsteller beschäftigten sich hauptsächlich mit inneren Stimmen. Fjodor Dostojewski erreichte in seiner Fähigkeit, den inneren Monologen und Gesprächen der Toten zuzuhören, ein erstaunlich subtiles Gehör, aber diese „Stimmen waren stimmlos, als ob die Münder mit Kissen bedeckt wären", wie der Held von *Bobok* (1873) seine auditorischen Eindrücke beschreibt. Die russischen Memoiren konzentrierten sich auf paradoxe Fälle: Politiker ohne rhetorische Fähigkeiten, Theaterexperimente mit der Unterdrückung einer kräftigen ausgebildeten Stimme, Dichter, die bewusst mit einer ausdrucklosen (weißen) Stimme ihre Verse rezitieren. Die russische Literatur und die russische Opernkritik entdeckten in der nationalen

Stimme in erster Linie die Natürlichkeit, die der Kunst und der antrainierten Meisterschaft widerstrebte. In *Krieg und Frieden* (*Wojna i mir*, 1867) pries Tolstoi Natascha Rostowas Gesang, der nach Meinung der Berufspädagogen unvollkommen war. Die Musikkritik des 19. Jahrhunderts bewunderte eher die russische singende Bäuerin, die ganz einfache Melodien wiedergab, doch darin die russische Seele offenbarte, und stellte sie über italienische Operndiven mit ihren kunstvollen Meisterstücken.[14] Auch im 20. Jahrhundert konnte Russland die Idee der Authentizität nicht aufgeben und ignorierte den elektrischen Doppelgänger der Stimme. Die radikale Ablehnung einer geschickten professionellen Stimme (als unnatürlich oder unaufrichtig) wurde zu einem obsessiven Topos der russischen Kultur. Die Betonung der kunstlosen Natürlichkeit, die das ‚Russische' verkörperte, entsprach dem Klischee der mythischen slawischen Seele, die in einem von der Zivilisation unberührten Naturreich residierte. Dieser irrationale Topos wurde auf die Stimme übertragen und nicht nur von Literatur und Theater, sondern auch von Radio und Film unterstützt. Die sich ändernde Situation der Kollision von natürlichen und elektrischen Stimmen um die Wende zum 20. Jahrhundert wurde nicht ‚bemerkt', ja ‚überhört'. Auch in der elektrischen Reproduktion war die Stimme fest an die Vorstellung von der authentischen inneren Natur gebunden, zu der die Stimme einen direkten Zugang verschafft. So wurde die ambivalente Situation der Spaltung auch im russischen Film bewusst unterdrückt und nicht problematisiert.

Es ist nicht verwunderlich, dass deutsche, amerikanische und britische Autor*innen Geschichten der nationalen Medienstimmen (der Politiker*innen, Schauspieler*innen und Ansager*innen) erforschten, aber in dieser umfangreichen Literatur gibt es kein einziges Buch, das dem russischen Material gewidmet ist. Vielleicht liegt dies an der doppelten Vorsicht dieser Kultur, in der die Konzepte von Körperlichkeit, Erotik und mechanischer Illusion, der performativen Identität und des ontologischen Subjekts, mit dessen Hilfe die Stimme in westlichen diskursiven Modellen beschrieben wurde, immer problematisch waren. Die imaginative Dimension um das kulturelle Phänomen Stimme bestimmte auch die Vorstellung von Filmstimmen. Der Wechsel des Mediums setzte neue Akzente, aber die nationalen Unterschiede waren davon nicht betroffen.

3 Authentizität, Individualität, künstliche Konstruktion

Der Film inszenierte und definierte die Grenzen der audiovisuellen Gebilde, die wir als Filmstimmen wahrnehmen, doch hatte dieser Vorgang mit einem überhörten Paradoxon zu tun: Das technisierte, apparative, von der Kamera modellierte Sehen galt dem menschlichen Sehen gegenüber als objektiver. Das Gehör dagegen wurde

im Film als ein quasi ‚natürlicher' Sinn postuliert, obwohl die Künstlichkeit der elektrischen Stimme als eine Phantomerfahrung in vielen Sujets verankert wurde. Selbst einfache Begriffe wie laut/leise oder nah/fern waren im Film nicht an die Stimme, sondern an Geräte und Konventionen gebunden. Mikrofone, Aufzeichnungsmaterialien und -geräte, Filter und Lautsprecher änderten die Qualität der menschlichen elektrifizierten Stimme. Die Körperlichkeit der Stimme, die Roland Barthes hervorhob,[15] ging im Film zugunsten des Körperbildes verloren: Oft verwandelte sich die Stimme in ein Sinnbild.

In den Anfangsjahren des Tonfilms muss die Künstlichkeit der Filmstimme stärker empfunden worden sein. Antonin Artaud und Jorge Luis Borges bezeichneten sie als technisches Bauchreden, als Ventriloquismus.[16] Die Stimme kam aus dem Lautsprecher, der hinter der Leinwand und in ihrer Mitte untergebracht war. Sie wurde buchstäblich auf das Bild gelegt. Elektrisch aufzeichnet und reproduziert, büßte sie ihre ‚natürlichen' Qualitäten ein und wurde zu einem elektrischen Schatten. Mit dieser Vorstellung waren auch die literarischen Fantasien des 19. Jahrhunderts (Mensch/Maschine) verbunden, die sich mit den Konstellationen aus der Romantik (Leben/Tod, Natur/Kultur) auseinandersetzten. Nicht nur die technische Künstlichkeit störte, auch die psychologischen Modelle der Bindung der Stimme an die Individualität waren nicht ohne Weiteres auf den Film übertragbar. Die Stimme im Film war eine künstliche Konstruktion, auch wenn sie stets eine Authentizität und Natürlichkeit suggerierte. Es schien, dass die Einheit der Figur im Film *Dieses obskure Objekt der Begierde* (*Cet obscur objet du désir*, F/E 1977, R: Luis Buñuel) dadurch geschaffen wurde, dass die zwei Schauspielerinnen, die die Protagonistin Conchita verkörperten, von einer Stimme vertont wurden. Doch die Vorstellung von der unzertrennlichen Einheit von Stimme und Körper bzw. seines Bildes widersprach der Praxis des Films. Ein Star konnte mehrere Stimmdoubles haben. Vor der Einführung der Tonmischung Mitte der 1930er Jahre haben diese Doubles die nicht tonogenen Stars noch während des Drehs synchronisiert. Alfred Hitchcock begann *Erpressung* (*Blackmail*, GB 1929) als einen Stummfilm. Als beschlossen wurde, die Produktion in einen Tonfilm umzuwandeln, wurde die tschechische Schauspielerin Anny Ondra simultan von der Britin Joan Barry vertont, die neben der Kamera saß. Rita Hayworth sang in jedem Film mit einer neuen Stimme, doch dies schien niemanden zu stören.

Diese Praxis wurde zu einem wiederkehrenden Narrativ. Sowohl beim Übergang zum Tonfilm an der Schwelle der 1930er Jahre als auch bei der Erneuerung des Tonerlebnisses während der zweiten Tonrevolution in den 1950er Jahren[17] kamen Sujets der Persönlichkeitsspaltung auf, die sich mit der Auslöschung des Individuums beschäftigten. Diese Spaltung war durch die Trennung des Körperbildes und der Stimme verursacht und mit der Unfähigkeit der Protagonist*innen verbunden,

die aufgezeichnete elektrische Stimme dem ‚richtigen' Körper zuzuordnen. Diese mediale Situation wurde in Hollywood, Deutschland, Russland und Frankreich in verschiedenen Genres inszeniert – als Komödie, Musical, Melodrama oder Horrorfilm. Die Genreschwankungen waren an kulturelle Traditionen gebunden. Auch wenn die Filmstimme eine technische Konstruktion ist, ist sie aus den metaphorischen Zuschreibungen nicht zu lösen (weshalb die psychologischen Modelle der Bindung der Stimme an die Individualität nicht ohne Weiteres auf den Film übertragbar sind).

Die Unheimlichkeit, die die romantischen Fantasien über die mechanische Stimme umgab, war ein Kennzeichen französischer Filme. In Julien Duviviers *Maigret kämpft um den Kopf eines Mannes* (*La Tête d'un homme*, F 1933) nach dem Roman von Georges Simenon verliebt sich ein Emigrant namens Radek in die unsichtbare Stimme, die er auf einer Schallplatte hört. Es ist die Stimme von Damia, einer berühmten französischen Chansonsängerin dieser Zeit. Radek imaginiert zu dieser Stimme den Körper einer fragilen Blondine, die er an der Seite eines amerikanischen Playboys in einer Pariser Bar sieht. Doch am Ende des Films entdeckt Radek mit Entsetzen, dass die Stimme, die ihn eroberte, einer von ihm als fett und vulgär empfundenen Prostituierten gehört (die von der echten Damia gespielt wird). In Simenons Roman (1931) gibt es keine Geschichte um die magische Stimme und Stimmhalluzinationen; es geht um ein perfektes Verbrechen. Doch Duvivier verlagerte den Fokus von der Untersuchung eines Verbrechens auf die Untersuchung der pathologischen Psychologie des Verbrechers, eines slawischen Auswanderers mit einer besonderen akustischen Wahrnehmung. Das Drehbuch wurde von dem russischen Emigranten Valéry Inkijinoff geschrieben, der auch Radek spielte. Im Film ist er in den Bars der Montparnasser Bohemiens verloren und sein Ohr fängt – wie Duviviers Mikrofon – nur Sprachfetzen auf. Als er der Erzählung eines amerikanischen Playboys über seine Armut und eine reiche Tante lauscht, sendet er ihm eine anonyme Nachricht und bietet ihm seine Dienste als Mörder an, der ein perfektes Verbrechen begehen kann. Das tut er auch, um dann den Playboy zu erpressen. Aber seine auditorische Phantasmagorie, die den Körper der Blondine mit der magischen Stimme verbindet, treibt ihn in den Wahnsinn. Duviviers Film endet nicht mit dem Tod der singenden Frau, sondern mit dem des zuhörenden Mannes, der die Tatsache nicht überleben kann, dass die verzaubernde Stimme den falschen Körper erhält. Der Regisseur stattete seinen russischen Helden mit einem raffinierten Ohr aus, verwandelte jedoch seine akustische Wahrnehmung in eine pathologische Halluzination. Das Ende des Films wirkt ambivalent: Nicht das Erscheinen der Polizei, die Radek verhaften will, sondern die Entdeckung des Körpers, der die Stimme erzeugt, treibt den Helden unter die Räder eines Busses. Das Lied von Damia, dessen Text Duvivier selbst geschrieben hat, ist übrigens die einzige Musik im Film.

Duviviers Film lässt das alte Motiv aus der Romantik wieder auferstehen. Die trügerische visuelle Illusion (ein schöner Körper) und ein Mann, der seiner akustischen Halluzination erliegt, kommentieren ironisch – mit einem tragischen Ausgang – die Technik des Tonfilms, der Körper und Stimmen frei verbindet. Nur René Clair vermochte, diesen ambivalenten, ja gefährlichen Effekt einer mechanischen Stimme im Genre einer romantischen Komödie zu gestalten, doch sein Film *Es lebe die Freiheit* (*À nous la liberté*, F 1931) war eine Ausnahme. Dort verliebt sich der Protagonist in eine Frauenstimme. Er sieht den Körper nicht und ahnt zunächst nicht, dass die Stimme von der Schallplatte kommt. Aber die Verwirrung dauert nur einen Augenblick, denn die Schallplatte hat einen Sprung; sie stockt, und der Protagonist verabschiedet sich schnell vom verwirrenden mechanischen Phantom. Clair setzt auf die Unterscheidung zwischen zwei elektrischen Doppelgängern der lebendigen Stimme, dem Grammophon und dem Film. Während das Grammophon hier als eine Maschine inszeniert wird, die die Natur in etwas Seelenloses und Totes verwandelt, ist der Film paradoxerweise mit dem Lebendigen und Organischen verbunden und wird von Clair als eine milde, weibliche Version des Elektrischen präsentiert. Die Protagonisten, die in einer Schallplattenfabrik arbeiten und seelenlose Stimmdoppelgänger produzieren, verlassen am Ende die Fabrik und werden zu Vagabunden. Sie kehren in die Natur zurück, erfüllt von Stimmen singender Vögel (die natürlich auf dem Lichtton aufgezeichnet sind). Clairs Fantasien erschienen jedoch schon damals naiv und märchenhaft.[18]

Diese Motive werden in einem Film von Marcel Blistène, *Chanson der Liebe* (*Étoile sans lumière*, F 1946), dramatisch und traumatisch verarbeitet. Seine Handlung ist in der Zeit des Übergangs vom Stumm- zum Tonfilm angesiedelt. Die aus der Provinz stammende Madeleine, gespielt von Edith Piaf, leiht dem Stummfilmstar Stella Dora, der jede musikalische Begabung fehlt, ihre Stimme und wundert sich über deren Wirkung auf der Leinwand, wo die Stimme dank Stellas perfektem Körper eine betörende Ausdruckskraft erhält. Madeleine wird eifersüchtig auf Stellas Ruhm, aber die erwartete Gerechtigkeit stellt sich im Film nicht ein. Stella wird von einem konkurrierenden Produzenten erpresst und stirbt bei einem Autounfall. Doch als der heimtückische Produzent beschließt, das Rätsel aufzudecken und Stellas berühmte Stimme im echten Körper zu präsentieren, scheitert Piafs Heldin. Das Gespenst der verstorbenen Schauspielerin erscheint in Madeleines Imagination bei ihrem ersten Auftritt und raubt ihr die Stimmkraft. Sie kann nicht singen und kehrt in ihr bescheidenes Leben zurück. Blistène unterstützt die mystische Geschichte mit psychologisch nachvollziehbaren, realen Begründungen. Madeleine verliebt sich in einen Toningenieur, der ihre Gefühle nicht erwidert, was sie am Vorabend des Konzerts erfährt. Stellas Ehemann bedroht sie vor der Vorstellung. Die Zuschau-

er*innen können jedoch frei wählen, welcher Erklärung sie mehr glauben. Aus der Situation der Spaltung zwischen Stimme und Körper entsteht ein Drama darüber, wie das tote (elektrische, mechanische) Bild den lebenden Körper erobert. Die elektrische Stimme gehört nun diesem Phantom, einem toten, aber auf der Leinwand ewig lebenden Körper.

Dieser Kontext macht verständlich, warum die Sängerin aus *Diva* (F 1981, R: Jean-Jacques Beineix) sich der Aufzeichnung ihrer Stimme widersetzt. Eine während des Konzerts gemachte Raubkopie wird zum Auslöser eines Kriminalplots. Der Film bezieht sich ebenfalls auf das alte romantische Motiv, bei dem man zwischen dem Original und dem elektrischen Doppelgänger unterscheiden kann. Beineix stellt auf den ersten Blick die idyllische Konstruktion von René Clair wieder her samt der Differenz zwischen der Natürlichkeit der Filmstimme und der toten Mechanik des Grammophons (im Film ist das ein Tonbandgerät). Aber eigentlich produzieren beide Medien Kopien, die die Authentizität, Organik und Individualität imitieren und die Illusion vom ewigen Leben erzeugen. Ganz anders arbeitete Federico Fellini mit diesem Motiv in seinem *Schiff der Träume* (*E la nave va*, I/F 1983). Die Asche einer berühmten Sängerin wird im Meer verstreut, während ihre Stimme von der Schallplatte die neblige melancholische Seelandschaft belebt. Eine elektrische Meerjungfrau gewinnt ewiges Leben – ganz ohne Unheimlichkeit.

In Deutschland wurde die Vorstellung von der Stimme als einem mächtigen Instrument der Manipulation noch vor Hitlers Aufstieg zur Macht durch einen Film über den Psychoanalytiker, Hypnotiseur und Verbrecher Dr. Mabuse unterstützt (*Das Testament des Dr. Mabuse*, D 1933, R: Fritz Lang). Mabuse ist die furchterregende Verkörperung einer akusmatischen Stimme. So bezeichnete der französische Theoretiker Michel Chion die unsichtbaren, körperlosen Stimmen im Film (Chion 2003). Diesen aus der antiken Praxis entlehnten Begriff erläuterte Chion speziell am Beispiel von *Das Testament des Dr. Mabuse*. In der Antike gehörte die akusmatische Stimme einem Gott, einem Geist oder einem Lehrer und war mit einem Sehverbot belegt, damit der Hörende sich auf das Gesagte konzentrierte. Dies bezeichnete ein autoritäres Verhältnis der Hörigkeit (im doppelten Sinne) und ein gemeinschaftlich geteiltes, vollkommenes, oral übermitteltes sakrales Wissen. Im Film jedoch sind die akusmatischen Stimmen mit vampirischen Qualitäten ausgestattet; sie manipulieren und kontrollieren das Bild aus dem Off und können nur im Moment ihrer Visualisierung gezähmt werden. In *Das Testament des Dr. Mabuse* ist es die Stimme eines kriminellen Genies, eines toten Patienten, die zur inneren Stimme seines Psychiaters wird. Diese Stimme nistet sich im fremden Leib des Arztes ein, materialisiert Mabuses Gedanken und regiert durch diesen Körper die Welt, den Film und die Fantasie der Zuschauer*innen.[19] Sie destabilisiert die Psyche des Arz-

tes und treibt ihn in den Wahnsinn. Auf einer Schallplatte aufgezeichnet, lockt sie den Protagonisten, der versucht, Mabuses Rätsel zu lösen, in eine gefährliche Falle. Fritz Lang hob die pathologischen Klangqualitäten der halluzinierten Stimme hervor. Die innere Stimme von Mabuse, die mit der Stimme des Psychiaters verschmilzt, wird in einem suggestiven langsamen Flüstern, monoton, mit einem mechanischen Rhythmus wiedergegeben. In der auf der Schalplatte aufgezeichneten Stimme ist das Atmen nicht zu hören. Ihre Entfernung ist nicht wahrnehmbar, es gibt auch keinen Nachhall, der eine Vorstellung von der Räumlichkeit vermitteln könnte. Diese künstliche Akustik betont das Gespenstische.

Im englischsprachigen Raum dagegen wurde die Stimme als performative und veränderbare Maske, als situativer Ersatz der Identität verstanden und so als eine befreiende Kraft gefeiert. Nicht zufällig waren die Sujets um die Stimme in Hollywood von Horroreffekten befreit. Geschichten über eine Stimme, die aufgezeichnet, vom Individuum getrennt und auf den Körper eines anderen Menschen übertragen wurde, wurden nicht als traumatische Sujets des Identitätsverlusts inszeniert – weder in den 1930er Jahren noch später. Bereits im ersten Tonfilm *Der Jazzsänger* (*The Jazz Singer*, USA 1927, R: Alan Crosland) wurde Al Jolson von zwei verschiedenen Stimmen vertont: Eine war seine eigene, mit der er Broadway-Hits wie *Mammy* sang, die andere gehörte dem jüdischen Kantor Jossele Rosenblatt. Der elektrische stimmliche Doppelgänger rief jedoch nicht den Effekt des schwindelerregenden gespaltenen Bewusstseins hervor, sondern der performativen Freiheit, jede Identität nach Belieben anzunehmen und stimmlich zu erschaffen.

Später wurden die Geschichten über die elektrischen Stimmdoppelgänger als Komödien und Musicals inszeniert, wie etwa in *Du sollst mein Glücksstern sein* (*Singin' in the Rain*, USA 1952, R: Stanley Donen). Die Handlung des Films ähnelt dem Sujet aus *Chanson der Liebe*: Die unbekannte Sängerin leiht ihre Stimme dem stimmlosen Star, doch am Ende des Films wird der Zusammenfall der Stimme mit dem ‚richtigen' Körper als Happy End, als eine ‚glückliche Hochzeit' von Bild und Ton gefeiert, und ein neuer Star wird geboren. Die Wiederkehr dieses Sujets deutet an, dass das Gefühl der Medialität von Leinwandstimmen erneut eine Rolle spielte, auch dank des neuen Mediums Fernsehen. Der Regisseur Frank Tashlin spielte damit virtuos in *Der Babysitter – Fünf auf einen Streich* (*Rock-a-Bye Baby*, USA 1958) oder in *Sirene in Blond* (*Will Success Spoil Rock Hunter*?, USA 1957). Im letzteren Film kann der Held sich am Telefon stimmlich als sein Chef ausgeben.

Auch die Geschichte der Verliebtheit in eine unsichtbare Stimme, zu der ein Körper imaginiert werden muss, wurde meistens nicht als Drama über Depersonalisierung, sondern als Komödie inszeniert, wie es Vincente Minnelli in der Verfilmung des

Broadway Musicals *Bells are Ringing* (*Anruf genügt – komme ins Haus*, USA 1960) tat. Eine Telefonistin, dargestellt von Judy Holliday, gibt jedem ihrer Kunden das gewünschte Bild, da eine Stimme sich jedem Bild, jeder Vorstellung anpassen kann. Wenn Kunden darauf bestehen, sie zu treffen, muss sie ihr Aussehen diesen Erwartungen anpassen. Am Ende weiß die Heldin nicht, wer sie eigentlich ist – eine fürsorgliche Mutter, eine schöne Bohemienne oder eine romantische Geliebte –; doch das stört sie nicht. Sie improvisiert frei mit ihrer Stimme und Identität. Beide sind ein Produkt männlicher Fantasie. Eigentlich erzählt der Film das Sujet von Hitchcocks *Vertigo – Aus dem Reich der Toten* (USA 1958), der vier Jahre zuvor herauskam, völlig anders. Dort formt ein Mann eine Frau nach dem Bild einer vermeintlich Toten, die seine Fantasie geschaffen hat, und geht dabei fast zugrunde. In Minnellis *Anruf genügt – komme ins Haus* wird diese Geschichte auf akustischer Ebene gespielt und ist von traumatischen Erlebnissen befreit.

Nur im amerikanischen Film noir wurden die inneren Stimmen mit instabilen Persönlichkeiten verbunden. Der Protagonist in *Frau ohne Gewissen* (*Double Indemnity*, USA 1944, R: Billy Wilder) rekonstruiert das zurückliegende Verbrechen in Rückblenden und nimmt seine Beichten auf einem Diktiergerät auf. Diese narrative Technik wurde auch als Monolog eines Toten bezeichnet. Alfred Hitchcock setzte sie in *Psycho* (USA 1960) ein, in dem der Mörder mit einer gespaltenen Persönlichkeit die Stimme seiner toten Mutter halluziniert, welche ihm befiehlt, junge Frauen zu töten. Die Dekriminalisierung des inneren Monologs im Film war ein Produkt der 1960er Jahre: Alain Resnais, Marguerite Duras, Ingmar Bergman und Federico Fellini übertrugen die Verfahren des Nouveau Roman auf den Film und machten den inneren Monolog so zum Synonym der Beichte, zum Fluss des Bewusstseins. Auf akustischer Ebene wurde die Stimme maximal an das Mikrofon – und damit an das Ohr der Zuschauer*innen – angenähert.

4 Raub der Stimme auf Russisch

Die kunstvolle und künstliche Praxis, ein Körperbild mit einer fremden Stimme zu kombinieren, war genauso im sowjetischen Film üblich. Maria Babanowa verlieh ihre Singstimme an Elena Kusmina im Film *Allein* (*Odna*, UdSSR 1931, R: Grigori Kosinzew und Leonid Trauberg). In *Zirkus* (*Zirk*, UdSSR 1936) konnte der männliche Hauptdarsteller Sergej Stoljarow nicht singen, sodass der Regisseur Grigori Alexandrow für ihn einsprang. Die Praxis war dieselbe wie in Hollywood und Europa, doch wurde sie nicht in ein Narrativ über den Identitätsverlust verwandelt. In den Diskussionen um die Stimme auf der Leinwand spielte sie auch keine Rolle. Erst 1982 entstand ein Film, der dieses Motiv aufgriff: *Stimme* (*Golos*, UdSSR 1982, R:

Ilja Awerbach). Seine Botschaft ist jedoch überraschend: Die Schauspielerin Julia (Natalia Saiko) ist todkrank. Dennoch möchte sie die Vertonung ihrer letzten Rolle selbst übernehmen. Sie beginnt die Nachsynchronisation, stirbt, bevor die Arbeit abgeschlossen werden kann, und eine Kollegin leiht dem Bild der Toten ihre Stimme. Auf der Handlungsebene folgt der Film scheinbar der russischen Tradition mit ihrem Glauben an eine authentische Stimme, die die elektrische unvollkommene Aufnahme nicht vermitteln kann. Aber der Regisseur erzählt nicht, wie etwa in Fellinis *Schiff der Träume*, die Geschichte einer einzigartigen Stimme, die im Film ewig weiterleben kann, sondern betont die totale Ersetzbarkeit, die in der filmischen Technologie begründet ist: Repliken werden umgeschrieben, Stimmen, Körper und Musik ausgetauscht – trotz der Beschwörung einer einzigartigen Individualität und ihres kreativen Ausdrucks. Die Schauspielerin bekommt ein Körperdouble und mehrere Stimmendoubles. Der Unterschied zwischen ihnen ist kaum zu hören. Ihre Repliken werden von der Stimme des männlichen Drehbuchautors gesprochen und ihr Brief wird am Ende von einer hysterischen Assistentin vorgelesen. Sogar das echte Schluchzen der Kollegin über den realen Tod von Julia geht – ohne Unterbrechung – in das gespielte Schluchzen im Tonatelier über und wird von der kraftvollen Musik unterstützt. Die Stimme kann das Bild der Toten nicht animieren. Eher wird die Unvollkommenheit der vom Film produzierten Kopien aufdringlich präsentiert, und Julias Versuch, dem Bild ihres Körpers ihre authentische Stimme zu geben und die Natürlichkeit durch ein gefundenes individuelles Wort auszudrücken, scheitert. Als die Kamera auf dem Foto der Gestorbenen schweigsam stehen bleibt, wird die Stummheit zum Ausdruck der gesuchten Authentizität, was jedoch einen Rückschritt in Bezug auf die mediale Entwicklung des Films bedeutet. Awerbachs *Stimme* geht gnadenlos mit dem russischen Topos der authentischen Stimme um. Paradoxerweise wurde der Film zwei Jahre vor Beginn von Perestroika und Glasnost gedreht, einer neuen politischen Entwicklung, die mit der Idee des authentischen, laut ausgesprochenen Wortes (russ. glas = Stimme) verbunden wurde. Diese Wendung gibt dem Film heute einen unerwarteten politischen Kontext.

Ist Stimme im Film ein Signum der Individualität oder ein Phantom? Ob als kulturelles Konstrukt oder als technische Klanggestalt betrachtet: Immer bleibt die Filmstimme ein Wirkungseffekt und verweist einerseits auf die kulturelle Tradition, andererseits auf paradoxe Abweichungen, die diese in der medialen Dimension erfährt.

Literatur

Altman, Rick (1992): Sound Space. In: Sound Theory/Sound Practice. Hrsg. von Rick Altman. New York: Routledge. S. 46-65.

Barthes, Roland (1990): Die Rauheit der Stimme. In: Roland Barthes. Der entgegenkommende und der stumpfe Sinn. Kritische Essays III. Frankfurt am Main: Suhrkamp. S. 269-279.

Belton, John (1992): 1950s Magnetic Sound: The Frozen Revolution. In: Sound Theory/Sound Practice. Hrsg. von Rick Altman. New York: Routledge. S. 153-170.

Boillat, Alain (2012): René Clair als Widerstandskämpfer gegen die synchron gesprochene Stimme. Was die ‚Sprechmaschinen' von À nous la liberté zu sagen haben. In: Resonanz-Räume. Die Stimme und die Medien. Hrsg. von Oksana Bulgakowa. Berlin: Bertz + Fischer. S. 21-54.

Borges, Jorge L. (1974): Über die Synchronisation. In: Jorge L. Borges: Gesammelte Werke. Band 5/1. Essays 1932-1936. München: Carl Hanser Verlag. S. 166-167.

Chion, Michel (1999): The Voice in Cinema. Berkeley: University of California Press.

Chion, Michel (2003): Magie und Kräfte des ‚Acousmêtre'. In: Medien/Stimmen. Hrsg. von Cornelia Epping-Jäger; Erika Linz. Köln: DuMont Literatur und Kunst Verlag. S. 124-159.

Crafton, Donald (1999): The Talkies. American Cinema's Transition to Sound, 1926-1931. Berkeley: University of California Press.

Derrida, Jacques (1979): Die Stimme und das Phänomen. Ein Essay über das Problem des Zeichens in der Philosophie Husserls. Frankfurt am Main: Suhrkamp.

Dolar, Mladen (2007): His Master's Voice. Eine Theorie der Stimme. Frankfurt am Main: Suhrkamp.

Kennan, George (1967): Memoirs, 1925-1950. New York: Pantheon Books.

Kolb, Richard (1932): Schicksalsstunde des Rundfunks. Berlin: Brunnen-Verlag Willi Bischoff.

Lacan, Jacques (2009): Das Seminar. Buch X: Die Angst. 3. Auflage. Wien / Berlin: Turia + Kant.

Liwanowa, Tamara; Protopopow, Wladimir (Hrsg.) (1967): Opernaja kritika w Rossii. Moskau: Muzyka.

Meyer-Kalkus, Reinhart (2001): Stimme und Sprechkünste im 20. Jahrhundert. Berlin: Akademie Verlag.

Miller-Frank, Felicia (1995): The Mechanical Song. Women, Voice, and the Artificial in Nineteenth-Century French Narrative. Stanford: Stanford University Press.

Müller, Corinna (1998): Vom Stummfilm zum Tonfilm. München: Fink.

Ong, Walter (1987): Oralität und Literalität. Die Technologisierung des Wortes. Opladen: Westdeutscher Verlag.

Poizat, Michel (1996): L'Opéra ou le cri de l'ange. Essai sur la jouissance de l'amateur d'opéra. Paris: Métailié.

Schmölders, Claudia (2002): Stimmen der Führer. Akustische Szenen 1918-1945. In: Zwischen Rauschen und Offenbarung. Zur Kultur- und Mediengeschichte der Stimme. Hrsg. von Friedrich Kittler; Thomas Macho; Sigrid Weigel. Berlin: Akademie Verlag. S. 195-226.

Yampolsky, Mikhail (2004): From Voice Devoured: Artaud and Borges on Dubbing. In: Antonin Artaud: A Critical Reader. Hrsg. von Edward Scheer. New York: Routledge. S. 169-177.

Anmerkungen

1 Diese Praktiken bezeichnete der französische Komponist und Theoretiker Michel Chion als Vocozentrismus (Chion 1999, S. 5).

2 Altman 1992.

3 Müller 1998.

4 Zit. nach Crafton 1999, S. 451.

5 Poizat 1986.

6 Meyer-Kalkus 2001, S. 21-23.

7 Miller-Frank 1995.

8 Lacan 2009.

9 Derrida 1979, Dolar 2007.

10 Kolb 1932, S. 52.

11 Meyer-Kalkus 2001, S. 251-263.

12 Schmölders 2002.

13 Kennan 1967, S. 551; Ong 1987, S. 44.

14 Zit. nach Opernaja kritika w Rossii (1967), S. 70.

[15] Barthes 1990.

[16] Borges 1974; Yampolsky 2004.

[17] Belton 1992.

[18] Vgl. Boillat 2012.

[19] Chion 2003.

Malte Kobel

Künstliche Stimme/n in der Musik von Kate Bush

Abstract: Im nachfolgenden Text möchte ich mich mit der Reproduktion musikalischer Stimmen beschäftigen. Das Œuvre Kate Bushs eignet sich exemplarisch für eine solche Untersuchung, da sie bereits in den frühen 1980er Jahren mit digitalem Sampling (u.a. am Fairlight CMI Synthesizer) experimentierte und hierbei auch ihre Stimme zum Material der Musik machte, die damit zu einem primär musikalischen Klangphänomen wird. Dies gilt insbesondere für das Album The Dreaming *aus dem Jahre 1982. Bush stellt mit ihrer Arbeit die Manipulierbarkeit und Künstlichkeit von musikalischer Stimme in den Vordergrund. Während Stimmtheorien die Stimme oft als authentisch, identitätsstiftend und körperlich konzipieren, ermöglicht das Konzept der musikalischen Stimme als medial-technologisches Phänomen, Fragen der Authentizität und Einzigartigkeit von Stimme zu problematisieren. Stimme in Musik und musikalischer Medialität lässt sich damit als phonographisch-performatives Ereignis begreifen.*

1 Intro

Der Song *Leave It Open* von Kate Bushs drittem Soloalbum *The Dreaming* (1982) spielt auf meinen Kopfhörern. Ich höre Drums, Bass und Klavierakkorde. Meine Wahrnehmung ist jedoch vor allem auf die und von den umherirrenden und verwirrenden Stimmen gelenkt. Eine erste, gespalten und effektvoll, singt durch einen Flanger, eine andere Stimme kommt aus dem Hintergrund, gedoppelt von einem Slapback-Delay und in die Höhe transponiert, so dass sie wie eine Kinderstimme klingt. Von rechts nach links zieht ein gehauchter Geisterschwarm weiterer Stimmen mit den Worten „Harm is in us“ durch meine Ohren. Und auf dieses gespenstische Hauchen wiederum reagiert ein Ensemble mit tieferen, eher standfesten Stimmen aus der Mitte des Stereoraums. Mein Hören wird in den Bann gezogen von dieser polyphonen Schar von Stimmen, die aus allen Richtungen singen, ächzen, säuseln und spielen. Bereits dieser Auftakt stellt Bushs Stimme als vielstimmige und phonographische Stimme vor. In den frühen 1980er Jahren begann sie zunehmend mit digitaler Samplingtechnologie zu arbeiten und bezog dabei auch ihre Stimme ein. Kate Bushs experimentelle Arbeit mit Klang und Stimme kann als eine Art Polyphonographie gehört werden.

Wenn wir Bushs Stimme durch *Leave It Open* folgen, stellt sich zu Beginn eine erste wichtige Frage: Handelt es sich in dieser Musik überhaupt um nur eine Stimme, oder gleich um mehrere? Während ich meine eigene Antwort auf diese Frage bereits im Titel dieses Beitrages vorweggenommen habe, muss sie letztlich jede Hörerin und jeder Hörer für sich selbst beantworten. Trotzdem möchte ich bei der Ungewissheit und auch Uneindeutigkeit meiner Antwort verweilen und daran einen weiteren Fragenkomplex anschließen. Meine Untersuchung tangiert die musikalische Stimme bzw. die Suche danach. Dieser Essay ist der Versuch einer Annäherung an die Theoretisierung von Stimme im Kontext ihres Musikmachens. Kate Bush eignet sich als Beispiel für einen solchen Versuch, da ihre Stimme nicht nur hervorragend ‚Musik macht'[1], sondern zumeist auch in Verbindung mit digitaler Samplingtechnologie entsteht. Ich interessiere mich für die Stimme in der Verschränkung von Musik, Musikmachen und Technologie, denn in dieser Verschränkung lassen sich allgemeingültige Vorstellungen von Stimme (als Ausdruck von Natürlichkeit, Identität, Körperlichkeit, Subjektivität etc.) zur Disposition stellen. Diesem Themenkomplex möchte ich im Folgenden nachgehen. Ich werde zunächst einen kurzen Überblick über den digitalen Synthesizer Fairlight CMI geben, mit dem Kate Bush vielfach gearbeitet hat, um anschließend ihre Stimmarbeit mit Alexander Weheliyes Überlegungen zu phonographischer Musik in Beziehung zu setzen. Zuvor werde ich jedoch in aller Kürze verschiedene Problemfelder rund um das Thema Stimme skizzieren und mein eigenes Verständnis von Stimme, im Besonderen von musikalischer Stimme, näher umreißen.

2 Theorien der Stimme: Versuche und Widerstände

Theorien zur Stimme existieren zuhauf und sind in der Menge zu umfangreich und disziplinär verschieden, um hier ausführlicher erläutert zu werden.[2] Vielmehr ist es für meine eigene Arbeit[3] primär relevant, Problemstellungen von Stimmtheorien anzuführen. Ich kritisiere einige dieser Stimmtheorien, um meinerseits eine Fokussierung auf die musikalische Stimme oder *musicking voice* zu ermöglichen, die bisher meist keine spezifische Beachtung in Theorien der Stimme erfahren hat.

Ein Hauptproblem im theoretischen Umgang mit Stimme kann als Index-Problematik bezeichnet werden. Index ist in diesem Fall ein Zeichen im Sinne von Charles Sanders Peirces Semiotik, welches eine direkte und meist physisch erscheinende Beziehung zu seinem Gegenstand aufweist (Turino 1999, S. 227). Die Stimme wird in verschiedenen Diskursen immer wieder auf eine feste und meist essentialistische Komponente zurückgeführt; es wird eine beständige Quelle oder ein Ursprung der Stimme gesucht. Dieser Ursprung hat, wie in vielen anderen Studien zu (akusmati-

schem) Klang (siehe u.a. Kane 2014; Schaeffer 2017), mit dem Klangkörper, d.h. dem physikalischen und akustischen Ort der Entstehung der Stimme zu tun. Dies ist ein Problem des Index, der auf einen Körper verweist, denn die Stimme wird in einem solchen Fall primär durch ihre Körperlichkeit definiert (siehe u.a. Barthes 2013; Eidsheim 2015). Ich negiere nicht, dass die Stimme von einem Körper (menschlich oder nicht-menschlich) aktiviert werden muss. Aber eine essentialistische Lesart, die Stimme auf Körperlichkeit reduziert, ist für meine Frage nach der musikalischen Stimme zu kurz gegriffen, denn auch Musikmachen lässt sich nicht auf Körperlichkeit, Physiologie oder Akustik reduzieren.

Ähnlich verhält es sich mit einer weiteren hartnäckigen Konstante in vielen Theorien, die die Stimme auf Subjektivität reduzieren. Dort wird Stimme vielfach metaphorisch gedacht, und sie bezeichnet ein Subjekt im Sinne einer politischen Theorie (z.B. Cavarero 2005). Vergessen wird, wie Stimme als musikalische, klangliche und eigene Figur teilweise unabhängig von Persönlichkeit oder Subjektivität agiert. In diesen beiden Fällen eignet sich Stimme – zumal die musikalische Stimme – nicht als Index für eine stabile Konstante, sondern stellt sich immer komplexer dar, als es die einfache Beziehung zu a) Körper oder Physiologie und b) Subjektivität suggeriert.

Ein weiteres Problem, das sich durch viele Stimmtheorien zieht, ist die Uneinigkeit darüber, was mit Stimme überhaupt gefasst werden soll. In den meisten Fällen (Connor 2000; Derrida 2003; Cavarero 2005; Dolar 2006, um nur einige kanonisierte Theorien zu nennen) ist mit Stimme die sprechende oder gesprochene Stimme gemeint, eine Stimme also, die primär mit Sprache zu tun hat und mittels linguistischer Theorien entworfen wird. Im Falle der musikalischen Stimme – wie ich mit Kate Bush exemplarisch zeige – handelt es sich demnach um ein anderes Grundverständnis von Stimme. Genauso wie Musik ein anderes Funktions- und Affektionssystem bedient als Sprache im Allgemeinen, unterscheidet sich die musikalische Stimme von der gesprochenen Stimme (Middleton 2003). Die Musikstimme hat keine primäre Kommunikationsfunktion, sondern kommt zum Vorschein im Prozess musikalischer Performativität.

Die Performativität von Stimme wurde in der Theaterwissenschaft und vor allem in den Performance und Opera Studies theoretisch erarbeitet. Hier lassen sich im deutschen Sprachraum Doris Kolesch, Jenny Schrödl und Sybille Krämer als wichtige Theoretikerinnen nennen (Kolesch & Schrödl 2004; Kolesch & Krämer 2006; Schrödl & Kolesch 2018). Für meine Überlegungen sind darüber hinaus die Arbeiten von Michelle Duncan, Annamaria Cecconi und Mary Ann Smart von Bedeutung, die je auf die musikalische Wirkmächtigkeit und Performativität von Stimme zielen und Stimme als spezifisch musikalische Figur in den Fokus rücken (Duncan 2004;

Cecconi 2005; Smart 2005). Ich möchte diese musikalischen Theorieentwürfe von Stimme mit Überlegungen zu Klangtechnologie und der Produktion sowie Reproduktion von Stimme verknüpfen.

2.1 Phonographien

Ein theoretischer Ausgangspunkt meiner Überlegungen findet sich in Alexander Weheliyes Buch *Phonographies. Grooves in Sonic Afro-Modernity* (Weheliye 2005). Weheliye ist hier primär daran interessiert, einer allgemeinen Erzählung von Modernität, die nahezu ausschließlich als „Whiteness" gedacht wird, eine (negative) Ko-Konstitution von „Blackness" entgegenzusetzen und dementsprechend den Narrativen von Modernität eine Perspektive hin zu „Race" zu ermöglichen. Weheliye sucht ein solches alternatives Narrativ in afro-diasporischen Musikkulturen und Sound-Technologien (man könnte hier zum Beispiel an Dub, DJing oder spezifische Klangproduktionsverfahren bei Motown denken). Weheliye zeigt, inwiefern Modernität in afro-diasporischer Kultur mitgedacht ist und wie über musikalische und sonische Praktiken des Phonographischen Modernität anders denkbar wird. Diese ‚vergessene' Erzählung, die Weheliye *Sonic Afro-Modernity* nennt, ist nicht lediglich eine Kehrseite von Modernitätsnarrativen, sondern fokussiert (statt Sprache und Whiteness) Technizität und Sonics von Black Musics. Weheliye ist für die Frage nach Stimme und ihrer musikalischen Manipulation daher vor allem vor dem Hintergrund der (genealogischen) Verschränkung von Musik, Klang und Technologie von Interesse. Seine Überlegungen zu den phonographischen Bedingungen von (afro-diasporischer) Musik resoniert auch mit Johannes Ismaiel-Wendts Studien zu „MusikmachDingen" (Ismaiel-Wendt 2016). Es geht hier, wie auch bei Kodwo Eshun und im Anschluss bei Rolf Großmann, um die Frage nach dem musikalischen und klanglichen „engineering" von Sinnen und Körpern, die sich in Klangtechnologien und deren Praktiken festsetzen (Eshun 1998; Großmann 2014). In der Untersuchung von Musik im Sinne der MusikmachDinge, die immer schon phonographisch gedacht sind, lässt sich Musik als technologisch vorgeformt verstehen. Eine Untersuchung von MusikmachDingen geht daher einher mit einer Problematisierung von Kategorien wie Natürlichkeit, Authentizität oder Originalität musikkultureller Praxis, da Musik hier sowohl performativ als auch phonographisch gedacht wird. Wenn, wie Weheliye argumentiert, Musik auch immer schon phonographische Arbeit bedeutet, macht es wenig Sinn, nach einer Ursprünglichkeit oder Anfangserzählung (z.B. Oralität) zu fragen. Stattdessen werden Musik und ihre Technologien vielmehr nach soziokulturellen Funktionen, Affekt oder den inhärenten Formationen oder Vorstellungen von Race, Gender, Sound oder Musik befragt. Weheliye und andere (siehe u.a. Moten 2003; Stadler 2010) argumentieren, dass die Konstituierung von phono-

graphischer Musik seit dem 20. Jahrhundert nicht ohne eine (vergessene) Idee von „sonic Afro-modernity“ (Weheliye 2005, S. 6) denkbar ist. Es ließe sich im Anschluss also argumentieren, dass phonographische Musikpraktiken und -techniken von afro-diasporischer Kultur sich durch weite Teile der MusikmachDinge von Popmusik ziehen – unabhängig davon, ob die Musik, die damit gemacht wird, als spezifisch ‚Weiß‘ oder ‚Schwarz‘ codiert ist. In diesem Sinne ist Weheliyes Arbeit auch für die Frage nach Stimme und den spezifischen Fall Kate Bush von Interesse.

Kate Bushs Stimm- und Klangarbeit mit dem Fairlight CMI lässt sich klangtechnologisch unter anderem über die wegweisenden Entwicklungen im Dub nachvollziehen. Im Dub-Studio (z.B. bei King Tubby) finden sich viele Poptechnologien vorgeformt, etwa die Arbeit mit Temporalität (Echo, Reverb, Delay etc.), die Idee von Versioning und Remix (Simultaneität und Manipulation von musikalischem Material), das Overdubbing oder die Arbeit mit Raumeffekten. Klangtechnologie vom Dub aus zu denken ermöglicht außerdem eine philosophische Problematisierung von Klang im Moment der Reproduktion. Weheliye spricht deshalb nicht von Reproduktion von Musik und Klang innerhalb seiner Theorie der Phonographie, sondern begreift Reproduktion auch als Produktion eines (musikalischen) Ereignisses. Im Dub lässt sich demnach nicht von Originalität und Ursprünglichkeit sprechen: Jeder Klang ist stets medial vorbereitet. Er wird im Klangereignis zwar als singulär wahrgenommen, ist aber immer schon reproduziert.[4] Daher wäre es angebrachter, Dub und Musikmach-Dinge im Allgemeinen, u.a. das Fairlight CMI und die stimmliche Arbeit Kate Bushs, auch hauntologisch[5] zu denken. Für die Frage nach der Stimme ist diese Perspektive des Phonographischen vor allem bedeutsam, um reduktive Theorien von Stimme zu problematisieren, die Stimme lediglich indexikalisch als Anzeichen von Identität, Körper oder Subjektivität hören.

Ich möchte im Unterschied dazu mit Weheliye die musikalische Stimme als performativ *und* phonographisch denken. Vor allem die Stimme in der Musik überschreitet eine theoretische Reduktion auf Identität, Körper oder Subjektivität. Wenn Stimme musiziert, wenn sie mit der Performativität von Musik in Berührung kommt und von dieser getragen wird, dann bewegt sich die Stimme weg von Fragen nach einem (fiktiven) Ursprung und stellt sich stattdessen als affektierende und affektierte Figur dar. In diesem Sinne wird die Stimme in der Musik selbst zu einer Maschine, die zwar zumeist von einem Körper hervorgebracht wird, jedoch autonom mit der Musik agiert oder selbst zum Agens von Musik wird. Kate Bush ist, wie Katherine Angel schreibt, an der Performativität und Künstlichkeit von Stimme besonders interessiert: „She is interested in what a voice is, and what it can do. She uses her voice like an instrument to rend and tear, to sometimes painful effect“ (Angel, o.D.). Wenn Weheliye also behauptet, jede Re/produktion von Klang sei „technological, whether

it emanates from the horn of a phonograph, a musical score, or a human body" (Weheliye 2005, S. 7), dann stellt sich für mich die Frage, wie es sich hier mit der Stimme als klangliche und vor allem musikalische Entität verhält. Kann die Stimme in *Leave It Open* als Re/produktion gedacht werden? Was passiert mit Kate Bushs Stimme? Und wie kann diese theoretisiert werden? Lassen sich die Stimmen, die im besagten Song zumeist polyphon erscheinen, als Repräsentation oder Mediation ihrer Stimme lesen? Oder entsteht hier gar eine eigene Entität, die teils unabhängig von Subjektivität und Körper der Sängerin agiert und ein Eigenleben annimmt?

3 Kate Bush und das Fairlight CMI

Spätestens mit ihrem drittem Album *Never for Ever* (1980) vollendet Kate Bush ihre Entwicklung hin zu einer Musikerin, die das Tonstudio mit all seinen Möglichkeiten in ihre Kunst einbezieht. Sie experimentiert zunehmend mit den neuen Entwicklungen digitaler Musiktechnologie, wie u.a. dem Synthesizer Fairlight CMI. Die Arbeit mit Klang und das Experimentieren im Studio sollten nicht nur Auswirkungen auf ihre Kompositionen haben, sondern auch die Arbeit mit ihrer Stimme grundlegend beeinflussen. Ich möchte zunächst einen kurzen Überblick über das Fairlight CMI geben, um die klangtechnologischen Möglichkeiten dieses spezifischen Musikmach-Dings (Ismaiel-Wendt 2016) nachzuvollziehen.

3.1 Fairlight CMI

Das Fairlight CMI (Computer Music Instrument) wurde 1979 eingeführt und zählte zu den ersten kommerziell erhältlichen digitalen Synthesizern. Es initiierte eine Wende in der Musikproduktion, weil mit Mikrofon, Presets und dem Einzeichnen von Wellenformen am Bildschirm Klang aufgenommen, manipuliert und nahezu verzögerungsfrei wiedergegeben werden konnte. Da die Arbeit mit digitalen Schallsignalen als präziser und schneller galt, löste das digitale Aufnehmen und Bearbeiten von Klang zumindest theoretisch die analogen Tonbandmaschinen ab.[6]

Auch wenn die Firma Fairlight (gegründet von den Ingenieuren Kim Ryrie und Peter Vogel) in Australien ansässig war, stießen die ersten Fairlight CMIs vor allem in Großbritannien auf Interesse. Grund dafür war wohl auch die BBC, die nicht nur eine frühe Demonstration des neuen Musikinstruments im Fernsehen ausstrahlte, sondern ab 1981 auch im Studio des BBC Radiophonic Workshop ein CMI für die dort tätigen Komponist*innen und Sounddesigner*innen bereitstellte.[7] Das Fairlight fand zudem im ehemaligen Genesis-Sänger Peter Gabriel einen prominenten Abnehmer,

der das Gerät ebenfalls schon 1982 im britischen Fernsehen vorstellte. In der *South Bank Show* wird Gabriel während der Arbeit an seinem vierten Soloalbum *Security* (1982) begleitet.[8] Auf dem Fairlight CMI nimmt Gabriel verschiedenste Sounds (u.a. Stimme, das Zerschlagen von Glas oder das Traktieren von Metallrohren) auf und prozessiert diese quasi verzögerungsfrei. Anschließend spielt er die aufgenommenen Sounds auf dem angeschlossenen Keyboard auf unterschiedlichen Tonhöhen. Seine noch zuvor unverarbeitet gehörte Stimme – er quietscht ein „Mummy" ins Mikrofon – ist nunmehr auf dem Monitor als Wellenform zu sehen und kann anschließend auf der Klaviatur als Sample gespielt werden und wird so zum eigenständigen Musikinstrument. Gabriel nutzte sein Fairlight CMI von da an als Klanggeber für seine orientalistisch geprägte ‚Weltmusik'.[9] Allerdings war er nur eine*r von vielen Musiker*innen der frühen 1980er Jahre, die das neue Instrument für sich entdeckten. Andere prominente Beispiele sind u.a. Yello, Klaus Schulze, Ryuichi Sakamoto und nicht zuletzt The Art Of Noise und ihr Mitbegründer Trevor Horn, der digitales Sampling auch in von ihm produzierten Bands wie z.B. Yes oder Frankie Goes To Hollywood nutzte.[10]

Die neugewonnene Faszination für die digitale Musiktechnik, vor allem für das mehr oder minder intuitive Aufnehmen und Bearbeiten von Klängen, demonstrierte auch Herbie Hancock medienwirksam in der amerikanischen *Sesame Street*. Während Gabriel in der zuvor erwähnten Fernsehsendung das Fairlight CMI als neues Werkzeug für die Erschaffung großer Musikentwürfe einführte, verfolgte Hancocks Demonstration eher ein pädagogisches Ziel, indem er einer Gruppe von Kindergartenkindern die Grundlagen des digitalen Samplings erläuterte.[11] Das Video beginnt mit einem Close-up auf die Sequencer-Funktion des CMI (Page R) und zeigt Hancock, wie er auf dem Keyboard über einen programmierten Loop improvisiert (siehe Abbildung 1).[12] Wie Peter Gabriel nimmt auch Hancock eine Stimme auf, nämlich die eines Mädchens. Er lässt es seinen Namen ins Mikrofon sprechen und demonstriert im Anschluss, wie das digitale Soundsignal in Sekundenschnelle auf der Klaviatur musikalisch nutzbar gemacht werden kann. Die Stimme wird so als Klangereignis an die ungläubige und amüsierte Gruppe der Kinder wiedergegeben. Mit der Stimme auf der Klaviatur erlaubt sich Hancock nun Späße und spielt die Stimme in hohen und tiefen Registern inklusive Time-Stretching. Die Kinder sind von dieser verfremdeten Stimme angetan, lachen leicht verschreckt, und die Moderatorin fragt: „What happened to your little voice?" Hancock spielt weiter, lässt die Stimme loopen und doppeln, spielt mit ihr Akkorde und lässt sie somit vielstimmig erklingen. Gespenstischer wird es, als die Stimme rückwärts erklingt, sich selbst verfolgt, und zum Schluss eine ganz Schar von Stimmen den Namen des Mädchens (Tatyana Ali) im Ensemble singt.

Abbildung 1: Herbie Hancock demonstriert das Fairlight CMI in der *Sesamstraße* (1983). Screenshot aus dem YouTube-Clip (siehe Anm. 11). Im Hintergrund sieht man den typischen Desktop-Bildschirm, mit dem das Fairlight (u.a. per Griffel) programmiert werden kann. Der Prozessor des CMI befindet sich in diesem Modell in der Klaviatureinheit.

In beiden Videos interessiert mich das Fairlight vor allem als Stimmgenerator bzw. Stimmen-MusikmachDing. Sowohl Gabriel als auch Hancock nutzen den Klang der Stimme, um die Möglichkeiten digitaler Samplingtechnologie zu demonstrieren. Es kann gleichsam als Konstante betrachtet werden, dass Stimme unweigerlich mit neuer Aufnahmetechnologie in Berührung kommt, denn die Reproduktion von Stimme hat immer schon eine zentrale Funktion in der langen Geschichte von Klangtechnologien gespielt oder diese gar mit hervorgebracht (siehe Sterne 2003).

Die Faszination für computer- und somit fremdgesteuerte Stimmen findet sich auch in den Presets des Fairlight CMI. Diese vorprogrammierten Sounds waren – neben der Sequencer- und Inputfunktion – die wichtigsten Neuerungen, mit denen sich das Fairlight in die Popmusikgeschichte eingeschrieben hat: Auf 8 Zoll großen Flop-

py Disks fanden sich seit der ersten Generation des CMI (1979) verschiedenste Sound-Presets, mit denen Kompositionen erarbeitet werden konnten, und die zum Manipulieren bereitstanden. Gegliedert wurden die Disks meist nach Instrumentenkategorie oder Klangquelle, u.a. Bass, Drums, Brass, Keyboard, Strings, Effects, Weather, Guitars, Percussion, Woodwind, Pianos, Choral, Cymbals, Bells, Animals, Plucked, Reeds.[13] Das bekannte Orchestra-Stab-Preset ORCH2, das einer Aufnahme von Igor Stravinskys *Der Feuervogel* entstammt, hat es zum Beispiel dank der Single *Planet Rock* (1982) von Afrika Bambaataa and The Soul Sonic Force auf zahlreiche Hip-Hop- und Electro-Funk-Platten der 1980er Jahre geschafft.[14]

Auch das Preset ARR1 hat sich in der Popmusikgeschichte sedimentiert und taucht unter anderem auf Einspielungen wie *Moments in Love* (The Art of Noise, 1983), *Appetite* (Prefab Sprout, 1985) oder *Shout* (Tears for Fears, 1985) auf. Die spezifische Floppy Disk, die die Stimmsounds versammelt und auf der sich ARR1 wiederfindet, ist mit HUMANS1 betitelt.[15] Das unverkennbar gemachte Sample, dieses „breathy, voice-like instrument[,] became a focal-point sound in the 1980s synth-pop genre" (Bennett 2019, S. 21) und ist tatsächlich die aufgenommene Stimme einer Sängerin. In Sampling- und Synthesizerforen kursieren Geschichten von Tom Stewart, der als Student die Sängerin Sarah Cohen unter Leitung von Martin Wesley-Smith am Sydney Conservatory of Music für die CMI Sample-Library aufgenommen haben soll.[16] Das prägnante und viel genutzte Preset-Sample ARR1 hat also offenbar einen ‚Gesangskörper' in der unbekannt gebliebenen Sarah Cohen und kursiert seit der Aufnahme im Jahre 1980 als SARRAR, also als Sample, in den Musikkulturen. Zu hören ist ARR1 zum Beispiel prominent auf *Zoolookologie* (Jean-Michel Jarre, 1984).[17] ARR1, ORCH2 und andere Preset-Sounds der Fairlight CMI-Serie prägten große Teile der Popmusik in den 1980er Jahren und können damit sicherlich als bedeutende klangliche Signaturen der Popkultur bezeichnet werden. Digitale Soundtechnologie setzt sich hier also im kollektiven Hörbewusstsein fest. Dies lässt sich auch in anderen ‚gehauchten' Stimmen im Pop der 1990er Jahre verfolgen, so zum Beispiel im R&B (u.a. bei Janet Jacksons *Velvet Rope*), im Trance (u.a. bei Kai Tracid oder Solar Quests *Into the Machine*) sowie in jüngerer Clubmusik der letzten zehn Jahre (u.a. Visionists *Pain* oder *I'm Fine*).[18]

Auch im Kontext der Kunstmusik fand das Fairlight CMI Verwendung. Hier ist u.a. das Stück *A Capella* (1995-97) von John McGuire für Sopran und Tonband zu nennen. Für dieses nahm McGuire im Studio für Elektronische Musik des Westdeutschen Rundfunks in Köln die Sängerin Beth Griffith mit dem Fairlight CMI auf. Die aufgezeichneten Samples (von drei Vokalen: a, e, u) fungieren in der Komposition als eigenständige Musikinstrumente, mit denen Griffith dann in der Live-Situation und auch auf der Aufnahme singt. Griffiths Stimme wird demnach als Instrument

gehandhabt; sie wird im Samplingprozess abstrahiert, vervielfältigt und zum musikalischen Material. McGuire stellt zur Arbeit an *A Capella* fest: „Die Idee war, dass sie [Griffith] mit sich selbst singt“[19]. In *A Capella* können wir also ähnlich wie bei Sarah Cohen und dem Sample ARR1 einer Sängerin zuhören, die nicht nur die Technik des Singens beherrscht, sondern deren Gesang wiederum technologisch, fern- und fremdgesteuert ist. Beth Griffith singt in *A Capella* ‚als‘ und mit dem Sequencer. Ihre Stimme oszilliert durch die Klarheit des Klanges immer wieder zwischen Sinustongenerator und menschlichem Körper.

Im Folgenden möchte ich mich intensiver mit Kate Bushs Fairlight CMI-Experimenten auseinandersetzen und vor allem die Produktion und Manipulation von Gesang und Stimme in den Blick nehmen. Bush war eine der ersten Musikerinnen in Großbritannien, die sich intensiv mit dem Fairlight CMI und den Neuerungen des digitalen Samplings auseinandersetzten. Gerade ihre Arbeit als Produzentin im Tonstudio kann als Gegenerzählung zu den oft männlich geprägten Narrativen gesehen werden, in denen Soundtechniker (Tom Stewart) oder Komponisten (John McGuire) die Stimmen von Sängerinnen zum musikalischen und klanglichen Material für ihre je eigenen technischen und künstlerischen Visionen genutzt haben.

3.2 Die vielen Stimmen der Kate Bush

Nach dem Erfolg des Albums *Never for Ever* (1980), das vor allem durch den Hit *Babooshka* populär geworden ist, widmete sich Kate Bush *The Dreaming*. Das Album von 1982 setzte die Arbeit mit experimenteller Studiotechnik, die schon *Never for Ever* wesentlich geprägt hatte, fort. Diesmal jedoch produzierte Bush das Album größtenteils in Eigenregie. Schon deshalb nimmt *The Dreaming* in ihrem Gesamtkatalog eine besondere Stellung ein: Es ist experimentierfreudiger als *Never for Ever* und das 1985 erschienene *Hounds of Love*, und es weist Bush erstmals dezidiert als Studiomusikerin und Produzentin aus. Ihre Arbeit mit dem Fairlight CMI war auf *Never for Ever* bereits in Songs wie *Army of Dreamers* (Sounds nachladender Gewehre) und *Babooshka* (Zerbrechen von Glas) zu hören.[20] Auf *The Dreaming* ist das Fairlight CMI vermehrt genutzt und auf sieben der zehn Titel vertreten, u.a. in den blechernen Trompetenklängen und eingeworfenen vokalen Samples („ooh“) auf *Sat in Your Lap*, in der flächigen Synth-Begleitung und den an Orchestra-Stabs erinnernden Einschüben auf dem Titelsong des Albums oder auf *All the Love*. Auf letzterem singt der Chorknabe Richard Thornton „we needed you“, begleitet von eingespielten Hauchern, die stark an das SARRAR-Sample erinnern.

Abbildung 2: Kate Bush am Fairlight CMI. Quelle: https://reverbmachine.com/blog/kate-bush-synth-sounds/ [zuletzt aufgerufen: 29.09.2021].

Für meine Frage nach der musikalischen Stimme ist das Album relevant, weil es Kate Bush als Sängerin in eine faszinierende Position versetzt, von der aus sie ihre eigene aufgenommene, d.h. reproduzierte Stimme abstrahiert, manipuliert und mit dieser als klangliches und musikalisches Material experimentiert. Bush ist für die Frage nach Stimme und dem Problematisieren von Ideen der Natürlichkeit und Authentizität auch schon vor ihrem Einsatz von Stimm- und Klangtechnologien von Interesse, denn bereits auf den vorherigen Alben kann ihre Stimmarbeit als experimentell bezeichnet werden. Ihre Stimme wurde oft als einzigartig gehört, als „eerily versatile", „superhuman" (Gordon 2005, S. 40) und „unearthly" (Reynolds 2014) beschrieben. Durch diverse Gesangstechniken (extremer Ambitus, Kreischen, Schreien, Flüster- und Sprechgesang etc.) macht sie ihre Stimme im Singen abstrakt. Dies wiederum kann als generelle Praxis der Musikstimme betrachtet werden: Singen abstrahiert jede Idee von natürlicher Stimme, denn im Musikmachen ist Stimme immer schon performativ. Diese Idee, dass Singstimme immer schon denaturiert ist, ist in der Stimmpädagogik und -praxis meist selbstverständlicher als in der Theorie. In einem Interview spricht Bush selbst von ihrer Stimme als musikalischer Entität, die kontrolliert und bearbeitet werden kann. Auf die Feststellung, dass ihr Gesangsstil sich immer wieder ändere, antwortet sie: „I purposely try to do that because I do feel that every song comes from a different person, really, so this is one way of making something different about it. I like to ‚create' voices" (Bush in Electronics & Music Maker 1982, S. 46). Damit wird auch klar, dass Bush ihre Stimme in der musikalischen Arbeit beinahe unpersönlich denkt und ihr verschiedene Rollen und plurale Personen und Körper zuspricht bzw. ‚zusingt'. Durch den Einsatz von digitalem Sampling und durch tech-

nologische Vervielfältigung und Manipulation mit dem Fairlight CMI wird diese (inhärente) Unnatürlichkeit oder Künstlichkeit der Stimme noch betont.

Hören wir z.B. genauer in *Leave It Open* hinein.[21] Hier begegnet uns neben Drums, Klavier und Bass (späterhin auch Streichern und Gitarren) vor allem eine Landschaft aus diversen Stimmen: gesungen durch einen Flanger-Effekt, quietschend in ein kurzes Delay getaucht und mit reichlich Reverb säuselnd, gehaucht und wie ein Schwarm vorbeiziehend; tiefere Stimmen, die uns aus der Mitte des Klangraumes im Chor entgegenströmen. Anderswo sind Stimmen verzerrt, werden geschrien oder gejault, wirbeln herum oder fügen sich zu rhythmischen Gruppen zusammen. Gegen Ende sind gar gedoppelte Stimmen zu hören, die rückwärts zu singen scheinen. All diese verschiedenen Stimmen überholen sich, sind teilweise über- oder nebeneinander simultan geschichtet. Im Refrain beispielsweise interagieren verschiedene Stimmen miteinander, bewegen sich aufeinander zu, fungieren als Soundeffekte und geistern durch den Stereoraum. Zu den Stimmen auf *Leave It Open* schreibt Kate Bush in einem Fanclub-Newsletter von 1982:

> There are lots of different vocal parts, each portraying a separate character and therefore each demanding an individual sound. When a lot of vocals are being used in contrast rather than „as one", more emphasis has to go on distinguishing between the different voices, especially if the vocals are coming from one person. To help the separation we used the effects we had. When we mastered the track, a lot more electronic effects and different kinds of echoes were used, helping to place the vocals and give a greater sense of perspective. (Bush 1990, o.S.)

Die Idee der unterschiedlich klingenden Stimmen ist also bedingt durch narrative und textliche Entscheidungen. Indem Bush verschiedene Stimmsounds für verschiedene im Song angedeutete Gefühle oder Charaktere nutzt, verwebt sie die textliche mit der musikalischen Ebene. Es ist aber interessant, dass in der raumakustischen und klanglichen Separierung der Stimmen auch eine Idee von stimmlicher Differenz mitschwingt, „the difference in my voice", wie Bush an anderer Stelle sagt (Electronics & Music Maker 1982, S. 46). Es ist ihr wichtig, dass die Stimmen nicht nur im Klangbild voneinander zu differenzieren sind, sondern darüber hinaus von einer vermeintlichen ursprünglichen Stimme differenziert oder abstrahiert erscheinen. Die Stimme büßt in diesem Prozess nichts von ihrer Wirkmächtigkeit ein, sondern wird als eigenständige Gestalt in der Musik wahrgenommen. Dafür ist, so Bush, die Arbeit mit Effekten und Echos (hierunter verstehe ich alle vorher beschriebenen Mechanismen wie Delay, Reverb, raumakustische Varianz etc.) von Bedeutung, um die Stimmtracks jeweils voneinander zu differenzieren. Wichtig ist, wie Bush die Stimme

als musikalischen Klang beschreibt: Sie verweist auf die musikalischen Qualitäten einer Stimme und eben nicht auf sentimentale und indexikale Kausalitäten.

Die Stimmen in *Leave It Open* sind zwar auf ihre je eigene Weise abstrahiert – d.h. manipuliert und verfremdet –, aber dennoch werden sie als stimmliches Ereignis gehört. Sie werden von der Komponistin und Produzentin und auch von den Hörer*innen als klingendes und spezifisch musikalisches Material wahrgenommen. Was beim Hören dieser diversen Stimmen dann zunehmend auffällt, ist, dass wir nicht ‚einer Stimme Kate Bush' lauschen, sondern eher den vielen unterschiedlichen Stimmtracks, die Bush im Mix und anschließend im Master miteinander verwebt und interagieren lässt. Die meisten dieser Stimmen sind erkennbar als vokales Timbre der Künstlerin, denn nur wenige Stimmen (z.B. die männlichen im Chor) wurden von anderen Personen eingesungen.[22] Doch im Prozess des Aufnehmens und Samplens produziert, bearbeitet und verfremdet Bush ihre Stimme. Es ist dieses Spielen mit der Stimme im Studio, das für mich die Problematik einer theoretischen Erörterung von Stimme – vor allem in Bezug auf Musik – erneut in den Vordergrund rückt. Beim Hören frage ich mich, wie Stimme, die in der Theorie oft als singulär, authentisch und identitätsstiftend gedacht wird, von Kate Bush zur Disposition gestellt wird. Was geschieht mit ihrer Stimme, und wie kann diese theoretisiert werden? Lässt sich ihre ursprüngliche Stimme von deren technologischer Manipulation und Reproduktion unterscheiden? Oder wird ihre Stimme mit Hilfe der technischen Abstraktion selbst zu einer Art Musikmaschine?

4 Künstliche Stimme/n

„We let the weirdness in" singt die letzte Stimme, mehr oder weniger unverständlich, zum Ende von *Leave It Open*. Das Thema der „Weirdness" zieht sich durch *The Dreaming*, verweist aber nicht nur auf die Tradition der Gothic folk tales, in denen sich Bushs Musik immer wieder aufhält. Weirdness heißt im Fall von *Leave It Open* und der Frage nach der Stimme auch Andersheit und Differenz. Die Weirdness, die die rückwärts aufgenommene und vorwärts abgespielte Stimme zum Ende des Songs besingt, sucht auch die Stimme als singuläre Entität heim und damit die problematischen Konzepte von Stimme, die ich zuvor skizziert habe. Digitales Sampling ermöglicht es Bush, ihre Stimme als klangliches Material zu behandeln wie jedes andere Geräusch oder jeden musikalischen Klang. Durch die phonographische Arbeit kann die immanente Differenz der musikalischen Stimme zum Vorschein gebracht werden; Bush selbst nennt das „the space in between", ein Raum, der sich in der polyphonographischen Stimme öffnet (Electronics & Music Maker 1982, S. 46-47).

In *Leave It Open* wird diese unheimliche Stimme offen gelassen, ihre Ambiguität betont: „We let the weirdness in".[23]

Die Stimme wird demnach einerseits in der Studioarbeit entmystifiziert, weil sie akustisch gesehen lediglich Klangmaterial ist. Andererseits weiß Kate Bush als Sängerin selbstverständlich um die Wirkmächtigkeit von gesungener und musikalischer Stimme, und sie weiß der inhärenten Weirdness der Stimme Geltung zu verschaffen: Die musikalische Stimme ist anders zu begreifen als die Sprechstimme, denn sie wird in der Musik ‚gespielt'. Damit ist sie ein Instrument von Sänger*innen, auch wenn sie darauf nicht reduziert werden kann. Im musikalischen Kontext existiert Stimme also nicht bloß, sondern muss im musikalischen Akt (z.B. im Singen) immer erst performativ hervorgebracht und *ins Spiel* gebracht werden. Dieses performative (Musik-)Spiel ist auch unabhängig von technologischen Manipulationen der Stimme möglich und wird so in bestehenden Stimmtheorien durchaus diskutiert (siehe u.a. Eidsheim 2015; Schrödl & Kolesch 2018). Im Falle von Kate Bush ist interessant, wie Stimme gleichzeitig produziert und reproduziert wird. Durch die Manipulation von Stimme wird die Frage nach der ursprünglichen Stimme, die sogar in aufgezeichneter Form zunächst zur Künstlerin zu gehören scheint, uninteressant. Denn selbst die aufgenommene Stimme zeigt sich als performative und somit immer schon als sich selbst-entfremdende, dem Selbst entfliehende Stimme. Der Flanger-Effekt vervielfacht den Input und lässt uns wie auch die Sängerin nicht nur mit einer Stimme, sondern mit mehreren Stimmen zurück. Wir können zwar weiterhin sicher sein, dass wir Kate Bush und ‚ihre' Stimme hören, weil wir ihre Art zu singen kennen und ihr Timbre zu differenzieren vermögen. Aber wir hören ihren Gesang als musikalische Figur, die sich selbst genügt und über einen fixierbaren Index (Körper, Identität, Subjektivität) hinausweist. Jason Stanyek und Benjamin Piekut bezeichnen das als rhizophonische Qualität aufgezeichneter Musik: die Möglichkeit phonographischer Musik, lineare Zeit außer Kraft zu setzen (Stanyek & Piekut 2010). Phonographie ermöglicht unmögliche Zeitlichkeiten, wie z.B. die gespenstische Vervielfältigung von Kate Bushs Stimme/n durch die Re/produktion von Gesang (Blake & van Elferen 2015). Durch diese phonographische Arbeit mit digitalem Sampling beschwört Bush das Andere, Ungekannte und Nicht-Singuläre der musikalischen Stimme. Sie hält die Stimme offen für die Künstlichkeit, die die Stimme zur Musik führt.

5 Outro

Die musikalische Stimme kommt zustande im Spiel musikalischer Performativität. Diese wird keinesfalls von Musiktechnologie unterminiert, sondern findet mit ihr und in ihr bereits statt. Das geschieht einerseits in der Studioarbeit Kate Bushs, wenn

sie mit Klang und Stimme ihre Kompositionen kreiert. Andererseits ist musikalische Performativität auch bereits in MusikmachDingen vorgeformt, in Presets, Anweisungen und Affordanzen, die zum Musikmachen animieren (Fabian & Ismaiel-Wendt 2018). Darüber hinaus ist die musikalische Performativität von Stimme auch im Akt der Rezeption von Bedeutung. Die Stimme hat hier einen (spezifisch musikalischen) Effekt auf die Hörenden. Was also an Kate Bush und ihrer Arbeit mit Stimme fasziniert, ist, dass Performativität und Phonographie zueinander finden und miteinander verschränkt sind. Es wird in der phonographischen Arbeit also nicht irgendeine Stimme von Kate Bush bloß repräsentiert oder vermittelt, sondern in der phonographischen Situation des Musikmachens wird die Stimme zu einer eigenständigen Musikmaschine.

Hier nähere ich mich wieder Weheliyes Überlegungen zur Phonographie, aber auch Fred Moten, der schreibt: „[T]here is no performance in the absence of recording" (Moten 2003, S. 81). Zwar geht es in diesem Zitat nicht per se um die Stimme und ihre phonographische Konstitution, sondern um eine generelle Konstitution von Black subjectivity, die bei Moten zwischen Ideen von Subjekt („performance") und Objekt („recording") angesiedelt ist. Ähnlich wie Weheliye ist Moten eher an einer Ästhetik von Black musics interessiert als an der spezifischen Stimme einer (in diesem Fall) weißen Sängerin. Dennoch möchte ich seine Feststellung auf Kate Bush anwenden und die besagte phonographische Konstitution, also das Zusammendenken von Performativität und Reproduktion, für eine Theorie der Stimme in der Musik nutzbar machen.

In Bushs Studioarbeit zeigt sich ihre Stimme als musikalische Entität, die nicht als natürlich, authentisch oder ursprünglich gehört werden kann. Diese Entität ist nicht singulär fixierbar, sondern sie macht Musik und ist deshalb eine performative Figur, die im Spiel mit der Künstlerin, dem Fairlight CMI und auch mit den Hörenden Musik macht. In diesem Musikmachen bahnt sich für einen kurzen Moment eine eigenständige Gestalt an. Die polyphone Stimme Kate Bushs wird dann zu einer Art Musikmaschine selbst – nicht ganz Kate Bush, nicht ganz Fairlight CMI, nicht ganz Körper und nicht ganz Maschine, sondern musikgemachte und musikmachende Stimme.

Literatur

Angel, Katherine (o.D.): On Kate Bush. Five Dials. https://fivedials.com/reportage/on-kate-bush/ [zuletzt aufgerufen: 29.09.2021].

Anniss, Matt (2016): Instrumental Instruments: The Fairlight. In: Red Bull Music Academy Daily. November 11. https://daily.redbullmusicacademy.com/2016/11/instrumental-instruments-fairlight?fbclid=IwAR0vCGHIsfdxdWt0n3K9iBFkLtxtFSrXT5Y2b70jUpitZ0BR6HVWT4fY2uI [zuletzt aufgerufen: 29.09.2021].

Barthes, Roland (2013): Die Rauheit der Stimme. In: Roland Barthes. Der entgegenkommende und der stumpfe Sinn. 7. Auflage. Frankfurt am Main: Suhrkamp. S. 269-278.

Bennett, Samantha (2019): Constructing Records: Sound Recording and Production Technologies at the Turn of the 1980s. In: Modern Records, Maverick Methods : Technology and Process in Popular Music Record Production 1978–2000. New York: Bloomsbury Academic. S. 19-36.

Blake, Charlie; van Elferen, Isabella (2015): Sonic Media and Spectral Loops. In: Technologies of the Gothic in Literature and Culture. Hrsg. von Justin D. Edwards. New York: Routledge. S. 60-70.

Bush, Kate (1990): About The Dreaming. Gaffa.org. http://gaffa.org/garden/kate14.html [zuletzt aufgerufen: 29.09.2021].

Cavarero, Adriana (2005): For More than One Voice: Toward a Philosophy of Vocal Expression. Stanford, California: Stanford University Press.

Cecconi, Annamaria (2005): Theorizing Gender, Culture, and Music. In: Women and Music: A Journal of Gender and Culture. 9/1. S. 99-105.

Connor, Steven (2000): Dumbstruck – A Cultural History of Ventriloquism. Oxford, New York: Oxford University Press.

Derrida, Jacques (1995): Marx' Gespenster: Der Staat der Schuld, die Trauerarbeit und die neue Internationale. Frankfurt am Main: Fischer Taschenbuch Verlag.

Derrida, Jacques (2003): Die Stimme und das Phänomen. 1. Auflage der Neuübersetzung. Frankfurt am Main: Suhrkamp.

Dolar, Mladen (2006): A Voice and Nothing More. Cambridge, Massachusetts: MIT Press.

Duncan, Michelle (2004): The Operatic Scandal of the Singing Body: Voice, Presence, Performativity. In: Cambridge Opera Journal. 16/3. S. 283-306.

Eidsheim, Nina S. (2015): Sensing Sound. Singing & Listening as Vibrational Practice. Durham, London: Duke University Press.

Eidsheim, Nina S.; Meizel, Katherine (2019): Introduction: Voice Studies Now. In: The Oxford Handbook of Voice Studies. Hrsg. von Nina S. Eidsheim; Katherine Meizel. New York: Oxford University Press. S. xiii-xli.

Electronics & Music Maker (1982): Kate Bush. In: Electronics & Music Maker. 2/8. S. 44-47.

Eshun, Kodwo (1998): More Brilliant Than The Sun: Adventures in Sonic Fiction. London: Quartet Books.

Fabian, Alan; Ismaiel-Wendt, Johannes (2018): Musikformulare und Presets: Musikkulturalisierung und Technik/Technologie. Hildesheim: Universitätsverlag Hildesheim / Georg Olms Verlag AG.

Feld, Steven (1988): Notes on World Beat. In: Public Culture. 1/1. S. 31-37.

Feldman, Martha; Wilbourne, Emily; Rings, Steven; Kane, Brian; Davies, James Q. (2015): Why Voice Now? In: Journal of the American Musicological Society. 68/3. S. 653-685.

Gordon, Bonnie (2005): Kate Bush's Subversive Shoes. In: Women and Music: A Journal of Gender and Culture. 9/1. S. 37-50.

Großmann, Rolf (2014): Sensory Engineering. Affects and the Mechanics of Musical Time. In: Timing of Affect: Epistemologies, Aesthetics, Politics. Hrsg. von Marie-Luise Angerer; Bernd Bösel; Michaela Ott. Zürich / Berlin: Diaphanes Verlag. S. 191-205.

Harkins, Paul (2020): Digital Sampling: The Design and Use of Music Technologies. New York, London: Routledge.

Ismaiel-Wendt, Johannes (2016): post_PRESETS – Kultur, Wissen und populäre MusikmachDinge. Hildesheim: Universitätsverlag Hildesheim.

Kane, Brian (2014): Sound Unseen: Acousmatic Sound in Theory and Practice. New York: Oxford University Press.

Kolesch, Doris; Krämer, Sybille (2006): Stimme: Annäherung an ein Phänomen. Frankfurt am Main: Suhrkamp.

Kolesch, Doris; Schrödl, Jenny (2004): Kunst-Stimmen. Berlin: Theater der Zeit.

Middleton, Richard (2003): Voice As Instrument (Revised 2018). In: Bloomsbury Encyclopedia of Popular Music of the World: Performance and Production. Hrsg. von John Shepherd; David Horn; Dave Laing; Paul Oliver; Peter Wicke. London: Continuum.

Moten, Fred (2003): In the Break: Aesthetics of the Black Radical Tradition. Minneapolis / London: University of Minnesota Press.

Reynolds, Simon (2014): Kate Bush, the Queen of Art-Pop Who Defied Her Critics. In: The Guardian. 21. August. https://www.theguardian.com/music/2014/aug/21/kate-bush-queen-of-art-pop-defied-critics-london-concerts [zuletzt aufgerufen: 29.09.2021].

Schaeffer, Pierre (2017): Treatise on Musical Objects: An Essay across Disciplines. Oakland, CA: University of California Press.

Schrödl, Jenny; Kolesch, Doris (2018): Stimme. In: Handbuch Sound: Geschichte – Begriffe – Ansätze. Hrsg. von Daniel Morat; Hansjakob Ziemer. Stuttgart: J. B. Metzler Verlag. S. 223-229.

Seeber, Martina (2017): WDR3 Open Sounds – Techné [72]: Fairlight CMI. https://www.youtube.com/watch?v=GQJQizhO0vU&ab_channel=SebastianArsAcoustica [zuletzt aufgerufen: 29.09.2021].

Small, Christopher (1998): Musicking. The Meanings of Performing and Listening. Hanover, NH / London: Wesleyan University Press.

Smart, Mary A. (2005): Theorizing Gender, Culture, and Music. In: Women and Music: A Journal of Gender and Culture. 9/1. S. 106-110.

Stadler, Gustavus (2010): Never Heard Such a Thing: Lynching and Phonographic Modernity. In: Social Text. 28/1. S. 87-105.

Stanyek, Jason; Piekut, Benjamin (2010): Deadness: Technologies of the Intermundane. In: TDR: The Drama Review. 54/1. S. 14-38.

Sterne, Jonathan (2003): The Audible Past. Cultural Origins of Sound Reproduction. Durham, London: Duke University Press.

Taylor, Timothy D. (1997): Global Pop: World Music, World Markets. New York / London: Routledge.

Turino, Thomas (1999): Signs of Imagination, Identity, and Experience: A Peircian Semiotic Theory for Music. In: Ethnomusicology. 43/2. S. 221-255.

Weheliye, Alexander G. (2005): Phonographies: Grooves in Sonic Afro-Modernity. Durham / London: Duke University Press.

Anmerkungen

[1] Die Idee des etwas sperrig klingenden Begriffs des Musikmachens bezieht sich einerseits auf Christopher Smalls Konzept des musicking, welches die performative Aktivität beim Musizieren beschreibt (Small 1998). Andererseits rekurriert das Musikmachen auf die Formulierung der MusikmachDinge, die u.a. auf Johannes Ismaiel-Wendt (2016) zurückgeht.

[2] Ein aktueller Überblick über das Feld, vor allem mit Bezug zu Musikwissenschaft und Sound Studies, findet sich u.a. bei Feldman et al. 2015 sowie Eidsheim & Meizel 2019.

[3] Der Verf. entwickelt aktuell eine Theorie der *musicking voice* im Rahmen seiner Dissertation.

[4] Anstatt erneut die problematischen Indizes von Klang, d.h. die Fragen nach Akusmatik, anzusprechen, denkt Weheliye phonographischen Klang als unabgeschlossen: „[S]ound recordings do not secure evidence of preexisting information but ‚merely' disseminate recorded sounds: they are forever suspended in a circulatory tide" (Weheliye 2005, S. 24).

[5] Der Begriff der Hauntologie ist eine Wortneuschöpfung des französischen Philosophen Jacques Derrida. Hauntologie geht zurück auf das französische „hanter" (geistern, spuken, heimsuchen) und beschreibt im Sinne Derridas eine „‚Lehre von der Heimsuchung' oder vom Spuk" (Derrida 1995, S. 27). In *Marx' Gespenster*, dem Buch, in dem Derrida die Theorie der Hauntologie erstmals ausarbeitet, geht es um Gespenster, die sich real in historischer Zeit festsetzen. Der Kommunismus sei ein solches Gespenst, das laut Marx und Engels Europa heimsuchte. Die Diagnose des Heimsuchens von zeitlichen und historischen Gespenstern diente im Anschluss nicht nur zur Beschreibung der kulturellen und sozialen Gegenwart, sondern lässt sich auch auf die Produktion und Reproduktion von Musik und deren temporale Uneindeutigkeit übertragen. So haben u.a. Charlie Blake und Isabella van Elferen ein Modell von *Sonic Media and Spectral Loops* entwickelt, das diese gespenstische Temporalität von (phonographischer) Musik nachvollzieht (Blake & van Elferen 2015). Dub, wie weiter oben angemerkt, kann als solch eine phonographisch-hauntologische Musik gehört werden, da es beim Dub um das Versioning und Remixen von bereits aufgenommenem Klang geht und daher die Frage nach dessen Ursprung obsolet erscheint.

[6] Für einen ausführlicheren Überblick zur Medien- und Musikgeschichte des Fairlight CMI siehe Bennett 2019, S. 19-24, und Harkins 2020.

[7] Einen Clip von BBCs *Tomorrow's World* zum Fairlight CMI aus dem Jahre 1980 gibt es auf YouTube: https://www.youtube.com/watch?v=dCu0NyZauzY&ab_channel=Synthasy2000 [zuletzt aufgerufen: 29.09.2021]. Zur Nutzung des Fairlight CMI im Radiophonic Workshop siehe Harkins 2020, S. 33.

[8] Das Video ist auf YouTube zu sehen: https://www.youtube.com/watch?v=scmYG1Pv1_Q&t=2373s&ab_channel=TheGenesisArchive. Die Arbeit am Fairlight CMI findet sich ca. ab Minute 16 [zuletzt aufgerufen: 29.09.2021].

[9] Für eine kritische Auseinandersetzung mit den Ideologien von ‚Worldbeat' und mit Peter Gabriels Musik hinsichtlich Exotismen und Aneignung siehe Taylor 1997, S. 39-52. Vgl. auch Steven Felds Kritik an musikalischer Aneignung in der Popmusik der 1980er Jahre (Feld 1988).

[10] Einen Überblick über frühe musikalische Experimente mit dem Fairlight CMI findet man u.a. in Matt Anniss' Artikel „Instrumental Instruments: The Fairlight" (Anniss 2016) und in Martina Seebers Radiosendung zum Fairlight CMI (Seeber 2017).

[11] Siehe den YouTube-Clip „Herbie Demonstrates the Fairlight CMI Synthesizer on Sesame Street, 1983“: https://www.youtube.com/watch?v=daLceM3qZmI&ab_channel=HerbieHancock [zuletzt aufgerufen: 29.09.2021].

[12] Zur Funktion der Page R siehe Harkins 2020, S. 41ff.

[13] Die auf dem Bildschirm auftauchenden Inhalte der Floppy Disks sind u.a. fotografisch unter folgender Adresse dokumentiert: https://synthroom.com/fairlight-cmi/fairlight-iix-floppy-disk-collection/ [zuletzt aufgerufen: 29.09.2021]. Verschiedene Floppy Disk Presets werden in dem folgenden Video vorgestellt: https://www.youtube.com/watch?v=J4yKD5fvRbQ&ab_channel=PerfectCircuit [zuletzt aufgerufen: 29.09.2021].

[14] Siehe Harkins 2020, S. 34-37.

[15] Siehe Video: https://www.youtube.com/watch?v=0WOaN_azH1s&ab_channel=SynthMania= [zuletzt aufgerufen: 29.09.2021].

[16] In den meisten Erzählungen dieser Geschichte trägt die Sängerin keinen Nachnamen. Allerdings wird im SoundonSound-Forum die Stimme von ARR1 als Sarah Cohen gekennzeichnet; siehe: https://www.soundonsound.com/forum/viewtopic.php?p=175900 [zuletzt aufgerufen: 29.09.2021].

[17] Das Begleitvideo zum Song zeigt Schauspielerinnen, die versuchen, das gesamplete *aahh* und *oohh* des Synthesizers durch Lipsyncing artikulatorisch nachzubilden. Sarah Cohens eigentlicher vokaler Input verschwindet durch die abstrahierte ‚Rhizophonie‘ (Stanyek & Piekut 2010, S. 19) von aufgenommener, gesampleter, manipulierter und schließlich gelipsyncter Stimme in ‚Stimmlosigkeit‘, während das Sample SARRAR bis heute Stimmlichkeit und Stimme suggeriert.

[18] Ich danke hier Wenzel Burmeier und Friedemann Dupelius für die Referenzen.

[19] John McGuire in Seeber 2017, 01:20:00.

[20] Siehe Harkins 2020, S. 28-32, für einen Überblick über Kate Bushs Arbeit mit Produzent Robert Burgess und die Arbeit an *Never for Ever* und *The Dreaming*.

[21] Kate Bush, *Leave It Open*, abrufbar auf YouTube: https://www.youtube.com/watch?v=-fBlox63041w [zuletzt aufgerufen: 29.09.2021].

[22] Ganz sicher lässt sich das jedoch auch nicht verifizieren, da die Credits keine weiteren Angaben zu aufgenommenen Vocals enthalten.

[23] Zur phonographischen Stimmarbeit auf *The Dreaming* sagt Kate Bush: „We have used delay machines for this on a couple of tracks, and added a very slight harmoniser effect as well as sometimes very tight double tracking. It really does depend on the song and how strong the lead vocal needs to be. For a more delicate song it would be wrong to put a heavy harmoniser on it – it would sound so affected. We've also been using an awful lot of compression on the new album, with nearly everything in fact. It's interesting the kind

of dynamics you can actually created [sic!], which is what I really never understood before. Especially with voices, as you start compressing them more and more, so many different levels start coming through on it – the breath particularly, and for me that's as important as the words: it's the space in between." (Electronics & Music Maker 1982, S. 46-47).

Doris Kolesch
im Interview mit Marcus Erbe

Stimme im Wandel: Postpandemisches Theater und Mediatisierung

Frau Kolesch, Sie beschäftigen sich seit vielen Jahren mit der Ästhetik und Medialität von Stimme. Lebhaft in Erinnerung geblieben ist mir persönlich eine Tagung 2002 in Köln. Dort hielten Sie einen Vortrag[1]*, in dem Sie sich unter anderem gegen die apodiktische Gleichsetzung von Subjekt und Stimme aussprachen. Nun kommt es in der kunst- und kulturwissenschaftlichen Stimmforschung zuweilen immer noch vor, dass vokale Phänomene entlang der gesellschaftlich vorherrschenden binären Geschlechterordnung als Ausdruck einer spezifischen Identität gewertet werden, statt die Stimme unter Einbeziehung fluider Persönlichkeitskonzepte zu dekonstruieren. Oder es entstehen ganz andere Hierarchien, weil zum Beispiel die vermeintlich natürliche menschliche Stimme gegen technisierte oder synthetische Stimmen ausgespielt wird. Dabei zeichnet es sich – insbesondere durch den Einsatz von Deep-Learning-Technologien – schon jetzt ab, dass die auditive Unterscheidbarkeit von ‚realem' und artifiziellem Stimmklang nicht auf ewig gegeben sein wird. Wie stehen Sie heute zur Frage nach dem Verhältnis von Stimme, Körper und Subjektivität?*

Während es über Jahrhunderte hinweg ein unerfüllbarer Traum, eine Sehnsucht und Imagination war, die menschliche Stimme lebensecht nachzubilden, sind wir heute wohl zum ersten Mal in der Menschheitsgeschichte in der Situation, dass wir künstliche, technisch hergestellte Stimmen nicht mehr sicher und trennscharf von vermeintlich natürlichen Stimmen unterscheiden können. Ich sage vermeintlich natürliche Stimmen, denn meines Erachtens ist es ein besonderes Charakteristikum unserer Beziehung zu unserer eigenen Stimme, dass uns die Historizität und kulturelle, auch technische Gewordenheit unserer eigenen Stimme – und damit auch von Stimmen allgemein – nicht oder nur sehr bedingt bewusst ist. Unsere Stimme ist, im mehrfachen Wortsinne, natürlich künstlich: Sie ist das Resultat körperlicher, physiologischer, biologischer, auch hormoneller Voraussetzungen, zugleich aber auch kultureller Techniken, Konventionen, regionaler oder auch klassen- und geschlechtsbezogener Redeweisen u.v.m.

Während man davon ausgehen kann, dass zum Beispiel die Historizität und kulturelle Gemachtheit bestimmter äußerlich sichtbarer Körperbilder und Körperideale den meisten Menschen – und nicht nur einigen Spezialist*innen oder Expert*innen, die sich damit beruflich oder aus anderen Gründen intensiv beschäftigen – bewusst ist, scheint mir dies nicht in gleicher Weise für die menschliche Stimme zuzutreffen. Vielleicht hat dies mit der Flüchtigkeit des Stimmklangs zu tun und damit, dass wir uns nicht beständig mit historischen Aufnahmen unserer eigenen Stimme oder mit historischen Audioaufnahmen verklungener Stimmen beschäftigen, die ja – im Vergleich zu visuellen Reproduktionstechniken – deutlich rezenter sind. Vielleicht hat dies aber auch damit zu tun, dass wir uns selbst, phänomenologisch gesprochen, immer auch stimmlich gegeben sind, dass unsere Stimme also in ganz anderer Weise als Teil unseres Selbstverhältnisses und unserer Identität erlebt wird als andere Körperteile, da die Stimme als Schwellenphänomen zwischen Physis und Psyche, zwischen Sinnlichkeit und Sinn, zwischen Materie und Geist, zwischen Index und Symbol zugleich individuell *und* sozial ist, sie ist Ausdruck einer immer auch geschlechtlich konnotierten Individualität wie auch Organ elementarer Vergemeinschaftung.[2] Trotz der deutlichen Zunahme und Erweiterung stimmtechnischer Möglichkeiten und Tools in den letzten Jahren scheinen mir die Historizität und auch die technische Verfasstheit von Stimmen also noch immer kaum bewusst zu sein und weitgehend ignoriert zu werden. Ein kleiner Selbsttest vermag dies zu veranschaulichen: Wer hat ein Wissen oder auch ein Bewusstsein davon, wie die eigene Stimme vor fünf oder zehn oder – je nach Lebensalter – auch 20 Jahren klang? Bezüglich unseres Körpers sind uns Wachstums-, Veränderungs- und auch Alterungsprozesse durchaus bewusst – wie verhält es sich bezüglich der eigenen Stimme? Halten wir unsere eigene Stimme für weitgehend unveränderlich oder gehen wir davon aus, dass auch sie Veränderungs- und Alterungsprozessen unterworfen ist?

Die gegenwärtige Situation scheint mir also durch die Gleichzeitigkeit durchaus gegenläufiger Dimensionen charakterisiert zu sein: Zum einen wird die Stimme noch immer in einer besonderen Beziehung zum Subjekt aufgefasst, zum anderen nehmen nicht nur im Feld der Künste, sondern auch in lebensweltlichen Bereichen Situationen des technischen, mediatisierten Spiels mit der Stimme und dadurch eröffnete Freiräume, aber auch Irritations- und Dekonstruktionspotenziale zu. Die enge Verbindung von Stimme und Subjektivität wird dabei insbesondere durch Verfahren beispielsweise des Voice Over, der Trennung von Stimme und Äußerungsinstanz, oder auch durch technische Filter infrage gestellt, die z.B. auf Social-Media-Plattformen wie Instagram auf einfachste Weise und inzwischen auch ohne besonderes technisches Equipment oder Know-how eingesetzt werden können und Stimmen in „Helium“-, „Giant“- oder „Robot“-Stimmen transformieren. Auch popu-

lärkulturelle Phänomene wie die enorm erfolgreiche künstliche Gesangsstimme Miku Hatsune des Software-Synthesizers Vocaloid 2, die zudem als virtuelle Figur zu einer Pop-Ikone wurde, regen meines Erachtens stark dazu an, naturalisierende und essenzialisierende Konzepte von Stimme zu überdenken.

Ihre Heimatdisziplin ist die Theaterwissenschaft. Um die nächste Frage historisch etwas einzugrenzen: Welche Tendenzen der Arbeit mit Stimme lassen sich im Gegenwartstheater beobachten?

Da die Theater wie die gesamte Kunst- und Kulturszene in besonderer Weise von der Corona-Pandemie betroffen und in ihrer Existenz und ihrem Selbstverständnis gefährdet wurden, möchte ich meine Antwort mit Blick auf die Perspektiven von Theater in einer postpandemischen Zeit formulieren: Ich denke nämlich, dass das Gegenwartstheater und die Gegenwartsperformance insgesamt durchaus von den stimmlichen und stimmtechnischen Experimenten der letzten zwanzig bis dreißig Jahre lernen könnten. Ende des 20. Jahrhunderts, als immer mehr Theaterinszenierungen dazu übergingen, mit Mikrofon, Microports und anderen akustischen Verstärkungs- bzw. (Re-)Produktionstechniken zu arbeiten, konnte man in Pausengesprächen, aber auch in Rezensionen nicht selten eine Abwertung solcher stimmtechnischer Experimente und Verkörperungsformen hören bzw. lesen. Dies hat sich – glücklicherweise – radikal geändert: Die Erkundung und Entfaltung technisierter Stimmen gehört zum integralen Bestandteil des Gegenwartstheaters, und entsprechende Audiotechniken sind in ihrer Vielfalt nicht mehr von den Bühnen wegzudenken.

So trennt das Theater von Susanne Kennedy – um hier nur ein Beispiel unter vielen zu nennen – strikt zwischen Stimme und Schauspieler*in, indem vorab aufgenommene Stimmen aus dem Off eingespielt werden, während die – häufig uniform maskierten – Schauspieler*innen stumme Gesten des Sprechens mimen. Die Verwendung verzerrender, maschineller oder roboterhafter Stimmeffekte, gepaart mit einer schauspielerischen Bewegungssprache, die eher an Avatare und künstliche Figuren denn an Menschen aus Fleisch und Blut erinnert, führt dazu, dass Schauspiel hier weniger als Menschendarstellung in den Blick kommt, sondern als Versuch der Imitation, der Darstellung und Verkörperung von Avataren, Robotern oder menschenähnlichen Maschinen. Diese grundstürzende Umkehrung etablierter Hierarchien und Bewertungen – nicht mehr das Menschliche wird hier nachgeahmt, sondern das Künstliche als gegenwärtige Form des Menschlichen – zeigt, dass auch das Theater seine Zukunft wahrscheinlich nicht in einer Ablehnung oder gar Ab-

schottung von zeitgenössischen Techniken und Technologien finden wird, sondern vielmehr gerade in der spielerischen, kritischen und klugen Auseinandersetzung mit den technischen Möglichkeiten, die unser gesamtes Leben zunehmend prägen und uns alltäglich zur Verfügung stehen (ein Blick in die Theatergeschichte zeigt übrigens, dass Theater von Anbeginn an in der Auseinandersetzung mit jeweils zeitgenössischen Techniken stattfand).

Weshalb ist dieser Umstand für ein Theater nach der Pandemie so wichtig? Weil in der Pandemie die theatrale Kunst der Versammlung keinen Möglichkeitsspielraum, sondern eine potenzielle Gefahr darstellte. Wenn das gemeinsame Zusammenkommen von Menschen in leiblicher Ko-Präsenz in einem geteilten Zeit-Raum zum gesundheitlichen Risiko wird, dann steht Theater grundsätzlich infrage. Viele Theaterschaffende haben auf diese einschneidenden Erfahrungen auf höchst unterschiedliche und innovative Weise reagiert, wobei sich – nicht zum ersten Mal – vor allem die Freie Szene, also gerade Künstler*innen, die finanziell und institutionell am wenigsten abgesichert sind, als besonders wagemutig, neugierig und experimentierfreudig erwiesen haben: Kamera, Mikrofon, Internet und andere technische Möglichkeiten waren nun nicht mehr Mittel neben anderen, die im Rahmen einer theatralen Aufführung genutzt werden konnten (oder auch nicht) – sie waren vielfach die einzig möglichen Formen, um mit dem Publikum in Kontakt zu bleiben und Aufführungen zu realisieren. Ich gehe mithin davon aus – oder hoffe es zumindest –, dass das Gegenwartstheater sich noch weit intensiver und auch grundsätzlicher als bislang mit den Chancen und Potenzialen sowie natürlich auch den möglichen Risiken oder Gefahren digitaler und hybrider Formen und Formate auseinandersetzt und dass es erkundet, inwiefern Theater auch jenseits der leiblichen Ko-Präsenz möglich ist. Die selbstverständliche Verwendung von Stimmtechniken und Stimmtechnologien im Theater der letzten dreißig Jahre könnte hier – vergleichbar der seit über hundert Jahren geradezu zur Voraussetzung für Theater gewordenen Elektrifizierung[3] – dazu beitragen, Hemmschwellen abzubauen und in kritischer Weise offen zu sein für weitere technische Erkundungen und Entwicklungen.

In verschiedenen musikalischen Kontexten sowie in der Performancekunst zeigt sich seit geraumer Zeit ein reflektierter Umgang mit aktuellen Voice-Technologien. So verwendet Holly Herndon eine selbst programmierte Stimmen-KI. Diese basiert auf Audiosamples mehrerer Menschen in unterschiedlichen Erregungszuständen und dient Herndon zur überindividuellen Erweiterung ihres sängerischen Ausdrucksspektrums. Lauren Lee McCarthy wagt in ihrer Performance LAUREN *das Experiment, zu einer menschlichen Version von Amazons Alexa zu werden. Von Ian Hatcher gibt es äußerst*

humorvolle Sprechstücke, in denen er den distinkten Tonfall von Computerstimmen perfekt nachahmt. Geschieht Vergleichbares im Bereich der Theaterstimme?

Nach meinem Kenntnisstand nehmen auch im Theater die Auseinandersetzungen mit KI auf verschiedenen Ebenen zu. Da gibt es zum einen dramaturgische Experimente mit KI, beispielsweise dem Textgenerierungssystem Generative Pre-Trained Transformer, kurz GPT, das inzwischen schon in der dritten Version existiert. So hat das Kollektiv CyberRäuber in seiner Inszenierung *Prometheus unbound* GPT quasi als Live-Autor genutzt, und das Theater Augsburg, das als eine der ersten größeren deutschen Theaterinstitutionen eine Projektleiterin für Digitale Entwicklung installiert hat und digitales Theater zu einer eigenständigen Sparte ausbauen will, experimentiert nach Aussagen von Tina Lorenz, die die digitale Entwicklung des Augsburger Theaters maßgeblich mitverantwortet, mit dem Einsatz von GPT 3. Auch für die Aufführung *reconstruct:alan_turing* vom Büro für Eskapismus, die auf den Ruhrfestspielen 2021 als interaktives Live-Game auf Zoom erlebbar war, wandte sich eine künstliche Intelligenz mit Frauenstimme an die bis zu 15 Spielteilnehmer*innen und interagierte stimmlich mit ihnen. Im Bereich stimmlicher KI-Experimente finde ich persönlich *Die Müller-Matrix* sowie das *Müller-Fon* der Performance-Gruppe Interrobang besonders spannend: *Die Müller-Matrix* fand als interaktive Audio-Installation während des Festivals „Heiner Müller!“ 2016 im Foyer des HAU 2 des Theaters HAU Hebbel am Ufer in Berlin statt und wurde 2021 als Telefonversion weiterentwickelt. In diesen Arbeiten konnte man sich per Telefon mit einem Heiner-Müller-Cyborg unterhalten, der Fragen der Teilnehmer*innen mit digitalisierten und montierten O-Tönen Heiner Müllers beantwortete. So spricht eine künstlich geschaffene Figur namens Herr Müller mit der Stimme des toten Heiner Müller über gegenwärtige Fragen von Theater, aber auch über Europa oder über Ökonomie. Diese Arbeit thematisiert auf künstlerisch und technisch avancierte Weise aktuelle Problemfelder und Phänomene wie Autorschaft, Theater als Beschwörung der Toten, Geschichte als Geschichte unerfüllter Sehnsüchte, die bisweilen gespenstische Insistenz von Stimmen, aber auch die zunehmende Macht von Algorithmen.

Im eingangs erwähnten Vortrag dachten Sie auch über die Rolle der stimmlichen Interaktion bei zwischenmenschlichen Begegnungen und beim Stiften von Nähe, aber auch von Distanz nach. Das war rund 17 Jahre vor Corona. Mit dem Ausbruch der COVID-19-Pandemie wurden diese Kategorien – Begegnung, Nähe, Distanz – plötzlich auf neue Weise relevant. Geradezu irrelevant wurde hingegen die öffentliche Teilhabe an künstlerischen Aufführungen. Nahezu alle Kulturbetriebe mussten sich etwas einfallen lassen, um das jeweilige Angebot dennoch im Bewusstsein des Pu-

blikums wachzuhalten. Wie ging speziell die Theaterszene mit dieser Situation um?

Ich habe schon ausgeführt, dass viele Theater und Performances mit digitalen Möglichkeiten experimentierten bzw. diese nutzten. Dabei kam es zu einer großen Varianz und Bandbreite: Manche Theater begnügten sich – insbesondere zu Beginn der Pandemie – damit, Videos von früheren Aufführungen aus dem Theaterarchiv ins Netz und – häufig kostenlos – zur Ansicht zur Verfügung zu stellen. Das war eine wunderbare Chance, um einige Aufführungen, die man live verpasst hatte oder nicht sehen hatte können, weil sie vor der eigenen Geburt stattgefunden hatten, zumindest als Videoaufnahme am heimischen Rechner zu verfolgen. Andere Theater hingegen experimentierten mit unterschiedlichsten Formen digitaler Aufführungen. Sie erprobten geeignete Plattformen und Formate, spielten mit hybriden Aufführungsvarianten, streamten Aufführungen oder entwickelten ganz neue, digitale Theaterformate wie *werther.live* des Freien Digitalen Theaterprojekts. Goethes Geschichte um die Liebesleiden des jungen Werther wurde hier mittels Splitscreen, WhatsApp, Sprachnachrichten, Zoom, Skype oder auch eBay in die durch und durch mediatisierte Welt des 21. Jahrhunderts transponiert und der Desktop wurde zur Bühne. Diese Produktion adaptierte nicht einfach unterschiedliche mediale Formate, sondern entwickelte mittels verschiedener analoger wie digitaler Praktiken und Technologien eine neue, hybride Form von Performance, die den virtuellen Raum kongenial nutzte. Auch eine Twitter-Aktion wie *#vorstellungsänderung* des Burgtheaters Wien sei hier genannt. Dabei waren am 12. Mai 2020 hunderte Menschen folgender Einladung auf Twitter gefolgt: „Was stellen Sie sich eigentlich vor?! – Wir möchten das herausfinden und laden zu unserem ersten Twittertheaterabend: Kommen Sie morgen Abend nicht ins Akademietheater und twittern Sie, was Sie nicht sehen.“[4] Die Doppeldeutigkeit des Begriffs Vorstellung – als Theatervorstellung und als Imagination – wurde hier genutzt, um kollaborativ von einem Theaterabend zu erzählen, der gar nicht stattgefunden hatte, und um mit einem Publikum, das sich nicht vor Ort im Theater versammeln durfte, einen gemeinsamen Erfahrungshorizont zu entwickeln.

Besonders bemerkenswert an der Situation der Theater in der Pandemie finde ich zum einen, dass die öffentliche Wahrnehmung der Theater im diametralen Kontrast zu deren Aktivitäten stand: Während nämlich selbst Expert*innen, die es besser wissen müssten (wie die Kulturstaatsministerin oder auch führende Theaterkritiker*innen), davon sprachen, dass die Theater geschlossen seien, und suggerierten, dass in ihnen quasi nichts stattfinde, waren die Theaterhäuser zwar zu, aber doch höchst aktiv, erprobten technische Möglichkeiten und digitale Plattformen und entwickelten unter Hochdruck neue, zumeist digitale Aufführungsformate. Zum anderen hat meines Erachtens mit der Pandemie auch eine Verschiebung des Fokus stattgefunden: Stand bislang vor allem die künstlerische Aufführung produktions-,

distributions- wie rezeptionsseitig im Zentrum, bewirkte die Pandemie, dass Theater und Publikum sich vor allem fragten, wie sie zueinander kommen können, und dass die Theater insbesondere Formen der Ansprache und Aktivierung des Publikums sowie des Kontakts mit ihm eruierten. Diese medialen Fragen führten immer die politische Dimension und Funktion von Theater mit sich und stellten die gemeinsam geteilte Form der sozialen Zusammenkunft als wesentliches Moment des Politischen heraus.

Gerade unter stimmlichen Aspekten darf aber abschließend nicht unerwähnt bleiben, dass einige Theater auch bewusst andere, nicht dominant digitale Wege eingeschlagen haben. So wollte Wajdi Mouawad, der künstlerische Leiter des Theâtre de la Colline in Paris mit seinem Format *Bouche à oreille* (*Ins Ohr geflüstert*) eine Art stimmlichen Ariadne-Faden zwischen Menschen knüpfen: Am Telefon wurde nach Art des Kinderspiels Stille Post von Mouawad eine circa zwanzigminütige Erzählung einem Dialogpartner erzählt, die dann zwischen Künstler*innen und Zuschauer*innen am Telefon immer weitererzählt wurde, mal relativ getreu dem Gehörten, mal relativ frei, je nach Entscheidung der Erzähler*innen. Am Schluss sollte die Geschichte, die eine spielerische, poetische Gemeinschaftsarbeit darstellte, wieder zu Mouawad zurückkommen und als Inszenierung gezeigt werden.

Wurden in den genannten Projekten bestimmte digitale Mittel bevorzugt eingesetzt? Und welche Möglichkeiten gibt es, das Publikum in einem häuslichen Umfeld überhaupt zu aktivieren?

Das Spannende ist ja, dass man grundsätzlich davon ausgehen kann, dass Theater und Performance alle möglichen digitalen Mittel ausprobieren, dass es hier also a priori keine Einschränkung gibt, zumindest wenn Fragen der finanziellen und technischen Verfügbarkeit geklärt sind. Verlegten sich, wie schon erwähnt, einige Theater aufs Streamen, so boten andere Zoom-Aufführungen und kombinierten vielfach auch verschiedene, aus den sozialen Medien bekannte Nutzungsweisen wie Chats, Emojis o.ä. und nutzten gebräuchliche Apps wie WhatsApp, Instagram oder Telegram. In der kurzen Phase leichter Rücknahmen einiger pandemiebedingter Einschränkungen im Herbst 2020 in Deutschland waren zudem zahlreiche hybride Aufführungen zu erleben, in denen ein kleines Präsenzpublikum im Theaterhaus und ein verstreutes Online-Publikum zu Hause auf je unterschiedliche Weise eine Aufführung verfolgten.

Bezüglich der Aktivierung des Publikums gibt es dabei zahlreiche Herausforderungen: Formen des interaktiven, partizipativen und immersiven Theaters haben in den letzten Jahren und Jahrzehnten verstärkt Fragen der Teilhabe adressiert und andere Weisen des Zuschauens und Wahrnehmens von Aufführungen in Szene gesetzt, in denen nicht mehr schweigend, weitgehend immobilisiert im Theatersessel sitzend, die Aufmerksamkeit idealerweise ganz auf die Bühnenszene konzentriert wird. In den sozialen Medien wiederum sind Interaktion und Gleichzeitigkeit bzw. Multitasking nicht die Ausnahme, sondern die Regel. Man könnte also sagen, in den theatralen Online-Formaten, wie sie insbesondere während der Pandemie erprobt und entwickelt wurden, kamen konfligierende Publikumsaktivitäten und Publikumserwartungen zusammen. Hier werden die nächsten Monate und Jahre entscheiden, ob die digitalen Theaterformate sich bezüglich digitaler Infrastruktur, aber auch bezüglich Netiquette den sozialen Medien angleichen, oder ob es hier eigene, ja eigensinnige Entwicklungen gibt, die in einer hybriden Weise Praktiken des analogen Theaters wie der digitalen sozialen Medien verbinden und mischen.

Katherine Meizel thematisiert im vorliegenden Sammelband[5] nicht nur die pandemiebedingt erschwerte kulturelle Teilhabe in Präsenz, sondern beleuchtet auch die positiven Effekte etwa von virtualisierten musikalischen Events unter dem Aspekt der Barrierefreiheit. Durch die notwendig gewordene Zunahme digitaler Angebote und die Verbreitung von Streaming- und Konferenzplattformen seien weitaus mehr Menschen mit Behinderung, chronischer Krankheit oder Sozialangst in die Lage versetzt worden, entweder Konzerten beizuwohnen oder mit anderen über das Internet zu musizieren. Denken Sie, dass digital realisiertes Theater langfristig ähnliche Chancen birgt?

Ich habe im Wintersemester 2020/21 und im Sommersemester 2021 anlässlich der Corona-Pandemie eine Online-Gesprächsreihe mit Theaterschaffenden aus Bosnien-Herzegowina, Brasilien, Deutschland, Iran, Israel, Japan, Polen und Südafrika durchgeführt.[6] Dabei ging es vor allem um den Austausch darüber, mit welchen je regional spezifischen Herausforderungen die Theater angesichts der globalen Corona-Pandemie konfrontiert waren und wie sie darauf reagiert haben. Eigentlich alle Gesprächspartner*innen bestätigten, dass digitale Theaterangebote ein zentrales Mittel darstellten, damit Theater in der Pandemie weiterhin künstlerisch aktiv und in Kontakt mit ihrem Publikum bleiben konnten. Zudem bestand Konsens, dass mit digitalen Formaten zahlreiche Barrieren und auch Ausgrenzungsmechanismen, die in herkömmlichen Theateraufführungen wirksam sind, abgebaut werden können. So gelang es während der Pandemie vielen Theatern, ein überregionales, nicht selten

internationales Publikum anzuziehen, und auch die Anzahl der Teilnehmer*innen, die digitale Streams, Zoom-Aufführungen u.ä. verfolgt haben, war häufig deutlich höher als die Platzkapazität in den Theaterhäusern. Nicht zuletzt konnte die gemeinhin mit Theaterkultur verbundene Fokussierung auf urbane Zentren durch digitale Angebote aufgebrochen und verstärkt Publikum auch in ländlichen Gegenden erreicht werden. Und die in der Frage angesprochene Barrierefreiheit für Menschen mit Einschränkungen, chronischer Krankheit oder Sozialangst ist natürlich ein weiteres Argument, das für Online-Formate spricht und auch dafür, nach der Pandemie über die Gleichzeitigkeit von analoger Aufführung und digitalem Angebot nachzudenken, statt es als Entweder-Oder-Option aufzufassen. Aber auch weitere Barrieren wie ökonomische oder auch solche, die mit sozialer Klasse und Status oder dem Bildungshintergrund verbunden sind, konnten durch Online-Formate abgeschwächt werden. Nicht zuletzt bieten sich für die Theater mit digitalen Formaten durchaus auch Chancen, insbesondere ein jüngeres Publikum für das eigene Programm und die eigenen künstlerischen Aktivitäten zu interessieren.

Diese Betonung positiver Aspekte von Online-Aufführungen mit Blick auf Teilhabemöglichkeiten sollte aber nicht dazu verleiten, pauschal einer generell besseren Zugänglichkeit oder gar einer dadurch angestoßenen Demokratisierung der Kunstform Theater das Wort zu reden. Denn mit digitalen Angeboten entstehen immer auch neue Barrieren, die von technischen Voraussetzungen und Kenntnissen (wie der Verfügbarkeit von Endgeräten und Apps, der Erfahrung im Umgang mit ihnen, aber auch der Stabilität, Schnelligkeit und Bandbreite der Internetverbindung) bis hin zu durchaus nicht unproblematischen Vorgaben bestimmter Rezeptions- und Interaktionsweisen reichen (so schränken Apps die Interaktionsmöglichkeiten des Publikums ihrerseits stark ein, und die in manchen Zoom-Aufführungen geäußerte Bitte, die eigene Kamera und den eigenen Ton – je nach Aufführung – an- bzw. auszuschalten, ist ebenfalls ein starker Eingriff in den Handlungsspielraum der Zuschauer*innen). Während klassische Theateraufführungen tendenziell eher von einem älteren Publikum rezipiert werden, scheinen digitale Formate stärker vor allem jüngere, mit sozialen Medien lebensweltlich vertraute Zuschauer*innen anzuziehen. Auch im digitalen Raum sind die Theater mithin mit einer Entwicklung konfrontiert, die sich seit längerem abzeichnet, nämlich einer zunehmenden Diversität und Heterogenität von Theaterpublika und einer Zerstreuung der öffentlichen Sphäre in vielfältige, partielle Öffentlichkeiten.

Anmerkungen

[1] Publiziert 2003 als Die Spur der Stimme. Überlegungen zu einer performativen Ästhetik. In: Medien/Stimmen. Hrsg. von Cornelia Epping-Jäger; Erika Linz. Köln: DuMont. S. 267-281.

[2] Vgl. hierzu auch Kolesch, Doris; Krämer, Sybille (2006): Stimmen im Konzert der Disziplinen. Zur Einführung in diesen Band. In: Stimme. Annäherung an ein Phänomen. Hrsg. von Doris Kolesch; Sybille Krämer. Frankfurt am Main: Suhrkamp. S. 7-15.

[3] Vgl. dazu Otto, Ulf (2020): Das Theater der Elektrizität. Technologie und Spektakel im ausgehenden 19. Jahrhundert. Stuttgart: Metzler.

[4] Twitterankündigung des Burgtheaters Wien. Zit. nach Netztheater. Positionen, Praxis, Produktionen. Hrsg. von der Heinrich-Böll-Stiftung und nachtkritik.de. Berlin 2020. S. 10.

[5] Siehe den Beitrag *Voice and the Selfs of Technology* im Abschnitt *Stimmforschung heute*, S. 19-35.

[6] Die Veranstaltung fand im Rahmen des Exzellenzclusters *Temporal Communities. Doing Literature in a Global Perspective* statt. Siehe dazu: https://www.temporal-communities.de/news/event-series-theatre-during-the-pandemic.html [zuletzt aufgerufen: 18.10.2021].

Radio-Stimmen

Dumisani Moyo, Kundai Moyo

Power, Endurance and the Significance of Radio Voice in Africa in the Age of Rapid Technological Change

Abstract: Seit seinem Aufkommen im frühen 20. Jahrhundert hat das Radio Gesellschaften weltweit auf vielfältige Weise beeinflusst – politisch, kulturell und wirtschaftlich. Im Zuge dessen verdiente es sich zahlreiche Beinamen, die seine Rolle in der jeweiligen Gesellschaft definieren: Es wurde als „Medium par excellence", „Theater des Geistes", „Universität des Volkes" und „Medium des Volkes" bezeichnet. Als Medium, das primär über Stimme funktioniert, darf seine Bedeutung im sozialen und politischen Leben nicht unterschätzt werden. Immer schon wurde das Radio deshalb von denen, die es besaßen und kontrollierten, instrumentalisiert – im Guten wie im Schlechten. In Afrika war es zunächst die Stimme der Herrschaft und Unterdrückung, später die Stimme des Widerstands und der Befreiung. Weil es die Menschen auch in weit entfernten Gebieten erreicht, die für andere Medien nicht leicht zugänglich sind, und weil es Hürden von Alphabetisierung und linguistischer Vielfalt überwindet, ist es das effizienteste und effektivste Medium in Ländern mit großer ländlicher Bevölkerung, die mit Armut, Analphabetismus und Ausgrenzung zu kämpfen hat. Auch deshalb wurde es als „Afrikas Medium" bezeichnet. Der vorliegende Text liefert einen Überblick zu Resilienz, Anpassungsfähigkeit und Kontinuität des Radios in Afrika über die Jahre hinweg und diskutiert die zentrale Bedeutung von „Stimme" für eine afrikanische Perspektive auf die Technologie Radio. Darüber hinaus wird kritisch analysiert, wie das Radio auf das Zeitalter der künstlichen Intelligenz, des maschinellen Lernens und der Algorithmen reagiert und welche Auswirkungen diese Änderungen haben. Es wird argumentiert, dass die schnellen technologischen Veränderungen für das Radio, wie wir es kennen, Chance und Gefahr zugleich sind.

Abstract: Since its advent in the early 20th century, radio has influenced societies in many ways across the world – politically, culturally and economically. In the process, radio has earned itself many epithets that define what it has meant to these societies: It has been called 'the medium par excellence,' 'theatre of the mind,' 'the people's university,' 'the people's medium' – and many such complimentary names. As a medium that has come to be associated with voice, its role in social and political life cannot be ignored. Depending on who owned and controlled it,

radio has been instrumentalised – for the good and bad. In Africa, it has a history of serving first as the voice of domination and oppression, then later as one of resistance and liberation. Its ability to reach communities in far-flung areas that are not easily accessed by other media, and to circumvent barriers of literacy and linguistic diversity has meant that it is the most efficient and effective medium for countries with large rural populations afflicted by challenges of poverty, illiteracy and exclusion. For this and many other reasons, it has been called 'Africa's medium.' This text provides an overview of the nature of radio's endurance, adaptability and continuity in Africa over the years and interrogates the centrality and significance of 'voice' in the African conceptualisation of the technology of radio. Further, it critically analyses how radio is responding to the current age of artificial intelligence, machine learning and algorithms, and the implications of these changes. It argues that these rapid technological changes pose both threats and opportunities for radio as we know it.

1 Introduction

While in some instances radio has been referred to as 'a dying medium,' or 'the forgotten medium,' several scholars have demonstrated that it is a resilient medium that continues to defy its naysayers, especially those who saw the coming of newer technologies such as television and the internet as heralding its death. As Matelski (1995, p. 5) argues, radio has "a tradition of survival and renewal" and "is accustomed to being dismissed as dead in a modern media world dominated by images, where the visual seems to mute the aural." With the advent of artificial intelligence (AI) and automation in recent years, newer questions have been raised about the future of radio as a voice technology. AI, which in simple terms refers to the creation of intelligent technology capable of carrying out tasks with intelligence approximating that of humans, and to solve complex problems (Zanni & Aziz 2018), has been deployed in many sectors including radio broadcasting over the past few years – simultaneously evoking both excitement and anxiety. While it eases the execution of some repetitive tasks which are now assigned to machines, this has raised ethical dilemmas and concerns related to job losses and lack of empathy, among others. This text critically discusses the significance of voice in African radio broadcasting, and how this has been affected by the rise in artificial intelligence.

2 A Brief Overview of Research on Radio in Africa

Approaches to the study of broadcasting in Africa have been varied, but many studies have focused on structure and content, looking at issues of regulation (largely using the political economy lens to discuss ownership, control and distribution in an industry that for many years was characterised by state ownership and control) and content (examining programming and how audiences engage with various programmes from a cultural studies perspective) (Fardon & Furniss 2000; Gunner et al. 2012). Other studies have also looked at the history of the medium on the continent, at the encounter between modern technology and traditional society (Fraenkel 1959). The common themes include the role of radio in development, community mobilisation/engagement, democracy building, and how radio has navigated technological change – with the coming of the internet and digitisation (Manyozo 2009; Nassanga et al. 2013).

The 'liberalisation of the airwaves' in the 1990s brought about renewed excitement about the potential of radio both for democratisation and development, especially in Africa and other parts of the developing world (Jallov 2012; Manyozo 2012). This allowed a decentralisation of radio in unprecedented ways, enabling communities to own stations where they could freely discuss matters of concern to them as well as engage with information from the outside world. These new radios took a variety of forms, and enabled communities to organise themselves around them in different ways. The case of radio listening clubs in Zambia and Malawi is an interesting one, as it not only gave people a voice but also the power to question authorities and demand responses in ways they could never do before (Banda 2007; Manda 2015).

The liberalisation wave also saw the mushrooming of commercial radio stations in many countries, with an increased uptake of the talk radio format that also transformed radio from a predominantly top-down medium to a more participatory and people-centred medium. Despite all this 'opening up', politics has never been far away from radio broadcasting in Africa. In many instances, the licensing was fraught with cronyism and corruption, and in others, pseudo-liberalisation ensued as power elites recognised the strong link between control of radio and their stay in power.

3 Theorising Voice, Power and Radio in Africa

The concept of voice has hardly been theorised in African studies on radio. Central to the phenomenon of radio broadcasting, however, has always been a fascination

with the 'disembodied voice' that can travel across vast distances and is present or heard in various spaces at the same time through wireless boxes, and more recently through various forms of digital devices. Beyond that, the association of voice with power and agency has been almost universal across cultures, where acts of 'voicing' and 'giving voice' are considered essential for political participation and democratic life. Weidman's concept of "ideologies of voice" helps in seeing voice from multiple perspectives centred around the power and agency of those who have the voice to speak, those who are given voice, as well as those who listen and how they interpret and engage with the received voice:

> Ideologies of voice can be characterized as culturally constructed ideas about the voice, including theories of the relationship between vocal quality and character, gender, or other social categories; where the voice comes from; its status in relation to writing and recorded sound; the relationship between the voice and the body; what constitutes a "natural" voice; and who should be allowed to speak and how [...] Ideologies of voice determine how and where we locate subjectivity and agency (Weidman 2014, p. 49).

This suggests that voice cannot be taken as neutral or value-free, because originators of voice are culturally located individuals who give voice from specific ideological positions, hence the need to appreciate "the complexities of how voices are actually constructed, mediated, and heard" (ibid). In African cultures, where belief in spirituality is almost universal, the idea of radio's disembodied voice has evoked imaginations that associate radio broadcasting with ghostly occurrences (Englund 2015; Fraenkel 1959) or the presence/absence phenomenon of spirit mediums (Moyo & Chinaka 2020). By extension, this bequeathed supernatural powers and to some degree uncontested authority to the "unseen voice of radio,"[1] just as the voice of ancestral spirits carried the sacred power and authority of the departed who continued to live among their communities in the spirit and would often manifest themselves through spirit mediums.[2] What the technology of radio broadcasting therefore did was to extend already existing African beliefs in spirituality by, in Durham Peters' words, "claiming to burst the bonds of distance and death" (1999, p. 142). Drawing links between the 'out-of-body communication' that happens in both electronic communication and the spiritual world, Peters argues:

> The word, voice, or image of a person dead or distant channelling through a delicate medium: this is the project common to electronic media and spiritualist communication. Indeed, all mediated communication is in a sense communication with the dead, insofar as media can store "phantasms of the living" for playback after bodily death. (ibid)

African communities were immediately able to use existing cultural lenses to relate to the new technology of radio and deploy it in relatable ways for social, cultural and political ends. Englund's study of two phenomenal Zambian broadcasters, "Gogo Breeze" (Peter Grayson Nyozani Mwale) and "Gogo Juli" (Julius Chongo), for instance, points to how a combination of radio voice authority with the moral authority that comes with being an elder in society contributed immensely to the success and effectiveness of these broadcasters.[3] Englund's idea of "radio grandfathers" therefore connects well with the spiritual realm with which early radio broadcasting has been associated (Moyo & Chinaka 2020), as the elderly (often seen as 'living ancestors') are seen as not so distant from the ancestors and hence authoritative enough to dispense wisdom and guidance to society.

In the context of the Zimbabwe liberation struggle where guerrilla radio played a significant role, voice was central to advancing the cause of the freedom fighters, as evidenced by the fact that both liberation movements' broadcasting outlets had the word 'voice' in their name: *Voice of Zimbabwe*, and *Voice of the Revolution*,[4] suggesting that these were popular movements that spoke for the people, and that a huge part of the purpose of the struggle was to regain the voice of the majority people of Zimbabwe who were silenced and oppressed by the Rhodesian colonial regime. Radio thus enabled liberation movements to maintain a sonic presence among their supporters (Lekgoathi et al. 2020). The fact that these broadcasts were then banned inside Rhodesia and operated underground, with the act of listening to them considered subversive, contributed to the aura of sacredness with which their voice was received. The few who secretly listened to these clandestine radio stations were able to reconstruct the voice of the freedom fighters and amplify it through word of mouth – through what Stephen Ellis (1989) aptly called "pavement radio." As Weidman argues,

> Voices are constructed not only by those who produce them but also by those who interpret, circulate, and reanimate them: by the communities of listeners, publics, and public spaces in which they can resonate and by the technologies of reproduction, amplification, and broadcasting that make them audible. Individual voices are created, in this sense, by audiences, fans, critics, cultural commentators and by the larger spirit of their times (Weidman 2014, p. 49).

Moyo and Chinaka extend this view by arguing that the spirit mediums can be viewed in the Foucauldian sense of "technologies of the sign system" which were "influential in framing discourses that informed and shaped the resistance to colonial occupation" (2020, p. 87). By extension and linking up with Walter Ong's (1982) concept of the "technologizing of the word," the voice as a carrier of the

word is essentially a technology of empowerment. The idea of empowerment itself has been critically important to most African communities, who have suffered disempowerment both during the colonial and post-colonial eras. What Mamdani (1996) conceived as "citizens" and "subjects" in the two historical eras sums up the dichotomy between those who have the power to voice, and those who are powerless and hence voiceless.

4 Radio in the Age of Artificial Intelligence

As previously mentioned, radio broadcasting has not been immune to the rapid technological change in recent years, including the growth in artificial intelligence and automation that has impacted many sectors. The deployment of artificial intelligence has been met with mixed reactions, not least because of the wide-ranging implications of replacing humans with machines. In radio broadcasting, the introduction of podcasting, voice assistants and audio streaming has brought both benefits and challenges to a sector that had traditionally been characterised by human interaction and the personal voice/touch. Indications are that in the media sector, AI has mostly been deployed in content recommendations and personalised entertainment, workflow automation and commercial optimisation through targeting and dynamic pricing (Chan-Olmsted 2019). While a report by Zanni and Aziz (2018) observed that the deployment of artificial intelligence in the area of broadcast and media was still at the start of its adoption curve, it is important to highlight that technological change in recent years has been characterised by accelerated speed, which means the picture presented in 2018 would be vastly different from that in 2021. In fact, the same report indicates that the percentage of companies saying they were unlikely to adopt AI within the next 2-3 years dropped from 57% to 36% within six months of the study, while those who said they had already adopted AI rose from 5% to 8% within the same period as awareness of the benefits of AI increased. The major areas where AI has been deployed most in broadcasting, according to that report, are content management (40%) and content distribution (35%). These are areas that entail routine and repetitive tasks that are better performed by automation, such as metadata tagging and indexing (which enables end-users to develop granular databases of their content and hence make it easier to monetise), image recognition, speech-to-text, etc. With the predictive capacity of machine learning and AI, radio broadcasters can understand the preferences and behaviour of their listeners and hence seek to produce content that appeals to them.

Whilst some of these technological advancements have not yet been widely integrated in African radio broadcasting, it is evident that AI holds promise for radio on the

continent, where market segmentation, building audience profiles and enhancing targeted advertising can become a lot easier. As Chan-Olmsted points out, "AI can be used to integrate audience and content insights, matching audience interest and relevant content in real time to deliver personalised content and better consumption experience" (Chan-Olmsted 2019, p. 193). This would also allow stations in different provinces to advertise products specific to their local listeners as opposed to the traditional system of indiscriminate advertising to all. In addition, AI could be used more in scripted programs that can be easily automated to reduce the need for human intervention, which could be beneficial for instance during a global pandemic where individuals under lockdown are encouraged to work from home.

In Uganda, as in many other countries on the continent, radio remains the primary source of information for most citizens and acts as a vital platform for communication. The explosion in community radio stations has therefore enabled many to participate in discussions on important community issues such as access to education, gender-based violence, floods induced by climate change, malaria and cholera, refugees or local disasters (Al-hassan et al. 2011; Rosenthal 2019). In 2019, the United Nations partnered with academics, policy makers and Pulse Lab Kampala, the Ugandan branch of the United Nations' Global Pulse Labs, to develop an automated speech recognition tool which uses AI technology to scan multiple radio broadcasts and aggregate information according to selected themes. This enables decision-makers to access the voices of marginalised communities and get a better appreciation of society's challenges and hence find more suitable ways of assisting them.

As communications technologies change over time, so do power dynamics between senders and receivers, and between those with the voice and power to generate messages and their readers/listeners/audiences. Invariably, these changes have led to a flattening of the hierarchy of communication, as those at the receiving end gain more and more power to 'talk back' and be producers of their own messages. However, these gains have not been permanent, as elites have continued to find new ways of manipulating these new technologies to take back the power that had accrued to those in the margins. The same technology which is being used for the good of society in Uganda, for instance, can be easily turned into a tool for repression, to monitor any dissenting voices on different community and commercial radio stations. With radio being the medium that reaches the largest numbers of African populations, there has always been strong interest in manipulating radio voice for political gain, and this interest will certainly increase and become more sophisticated as AI enables enhanced precision in understanding audiences and their preferences.

While uptake of technology in Africa may be generally slow, recent history has shown that with new information and communications technologies such as mobile phones, African communities have been in the forefront of appropriating these to leapfrog into the future (Nyamnjoh 2005). At the same time, many African governments have shown a strong penchant to invest in technologies that aid their stay in power. Many have acquired surveillance technologies from China, where AI and social media are increasingly used for micro-targeting, particularly during elections. In some of these elections, manipulation of images, videos and text messages to discredit opponents and spread falsehoods and disinformation during campaigns has been rife. As AI improves and robots perfect the imitation of voice, radio will, if not properly regulated, become the new battleground in the fight between bots and humans in future election campaigning. In countries such as Kenya, Uganda, and Zimbabwe, governments and opposition parties alike have already invested in bots and human agents to subvert the will of the people (Mare et al. 2019; Moyo et al. 2020).

However, while online radios have changed the culture of consuming radio in many parts of the world, where platforms such as Apple Music and Spotify have become popular especially among youths, traditional radio has remained the mainstay in most of Africa, where access to data continues to be highly uneven. While these advancements in artificial intelligence give promise to struggling broadcast companies that are looking out for new solutions and more reliable revenue streams, the recent Covid-19 pandemic has demonstrated just how much people crave human contact rather than artificial/robotic contact. As iHeartMedia's Brian Kaminsky argues,

> creating content for the radio comes down to creating an experience that builds a connection with listeners and keeps them entertained, informed and eager to come back for more. AI should be used in a responsible way whilst maintaining and enhancing the human elements.[5]

Thus, despite the rise of voice automation and advancements in emotional AI, the power of the human voice in radio remains unmatched in the face of these new developments. As Englund argues, "radio affords a peculiar form of public intimacy by broadcasting actual voices that generate sentiments in their listeners" (Englund 2015, p. 253).

5 Conclusion

Whilst advancements in AI and machine learning could be seen as posing a threat to radio broadcasters in Africa, we argue that it is too early to tell, considering its current rate of adoption. This slow pace of adoption is partly informed by a recognition that the integration of AI needs to be beneficial to all and incorporated in an ethical manner. Domestication and customisation of AI through the incorporation of local languages in ways that do not discriminate, alienate and marginalise users in the global South would be essential. As radio evolves in the age of AI, it is arguable that the human touch and the human voice will remain an indispensable part of broadcasting, and that broadcasters will continue to reimagine and adapt their roles in a world where people crave human-to-human engagement as opposed to robotic contact. Several programmes in African radio stations attest to the fact that the affective power of the human voice on radio cannot be easily replaced by machines. Gogo Breeze's humour and empathy, for instance, could never be automated. It is evident, though, that AI will play a significant role in furthering radio's impact in Africa through content recommendations, enabling personalised entertainment, and commercial optimisation.

References

Al-hassan, Seidu; Andani, Alhassan; Abdul-Malik, Abdulai (2011): The Role of Community Radio in Livelihoods Improvement: The Case of Simli Radio. In: Field Actions Science Reports. 5. pp. 3-7.

Banda, Fackson (2007): Radio Listening Clubs in Malawi and Zambia: Towards a Participatory Model of Broadcasting. In: Communicare. Journal for Communication Sciences in Southern Africa. 26/1. pp. 130-148.

Carvajal, Elena S. (2020): The Radio, Getting Smarter Everyday Thanks to Artificial Intelligence. In: Telefónica Tech. February 13. https://business.blogthinkbig.com/the-radio-getting-smarter-everyday-thanks-to-artificial-intelligence/ [last accessed September 09, 2021].

Chan-Olmsted, Sylvia M. (2019): A Review of Artificial Intelligence Adoptions in the Media Industry. In: International Journal on Media Management. 21/3-4. pp. 193-215.

Ellis, Stephen (1989): Tuning into Pavement Radio. In: African Affairs. 88/352. pp. 321-330.

Englund, Harri E. (2015): Multivocal Morality: Narrative, Sentiment, and Zambia's Radio Grandfathers. In: HAU: Journal of Ethnographic Theory. 5/2. pp. 251-273.

Fardon, Richard; Furniss, Graham (eds.) (2000): African Broadcast Cultures: Radio in Transition. Oxford: James Currey.

Frankael, Pierre (1959): Wayaleshi. London: Weidenfeld and Nicolson.

Gunner, Liz; Ligaga, Dina; Moyo, Dumisani (2012): Radio in Africa: Publics, Cultures and Communities. Johannesburg: Wits University Press.

Gunner, Liz (2019): Radio Soundings: South Africa and the Black Modern. Cambridge: Cambridge University Press.

Jallov, Birgitte (2012): Empowerment Radio: Voices Building a Community. Gudhjem: Empowerhouse.

Johnson, Lesley (1988): The Unseen Voice: A Cultural Study of Early Australian Radio. New York: Routledge.

Mamdani, Mahmood (1996): Citizen and Subject: Contemporary Africa and the Legacy of Late Colonialism. Princeton, N.J: Princeton University Press.

Manda, Levi Z. (2015): What Makes Radio Listening Clubs a Participatory Communication for Development Platform Work? A Case Study of Monkey Bay, Malawi. In: Online Journal of Communication and Media Technologies. 5/4. pp. 204-219.

Mano, Winston (2004): Renegotiating Tradition on Radio Zimbabwe. In: Media, Culture & Society. 26/3. pp. 315-336.

Manyozo, Linje (2009): Mobilizing Rural Community Radio in Africa. In: Ecquid Novi: African Journalism Studies. 30/1. pp. 1-23.

Manyozo, Linje (2012): People's Radio: Communicating Change Across Africa. Penang: Southbound.

Mare, Admire; Mabweazara, Hayes M.; Moyo, Dumisani (2019): "Fake News" and Cyber-Propaganda in Sub-Saharan Africa: Recentering the Research Agenda. In: African Journalism Studies. 44/4. pp. 1-12.

Matelski, Marilyn J. (1995): Resilient Radio. In: Radio: The Forgotten Medium. Edward C. Pease; Dennis E. Everette (eds.). New Brunswick, London: Transaction Publishers.

Moyo, Dumisani; Chinaka, Chris (2020): Spirit Mediums and Guerrilla Radio in the Zimbabwe War of Liberation. In: Guerrilla Radios in Southern Africa: Broadcasters, Technology, Propaganda Wars, and the Armed Struggle. Sekibakiba P. Lekgoathi; Tshepo Moloi; Alda R. S. Saide (eds.). Lanham, Boulder, New York, London: Rowman & Littlefield.

Moyo, Dumisani; Mare, Admire; Mabweazara, Hayes M. (2020): Social Media, the Press and the Crisis of Disinformation in Africa. In: Communicatio: South African Journal of Communication Theory and Research. 46/4. pp. 1-6.

Mpofu, Philip; Salawu, Abiodun (2020): Handling of Sexually Offensive Expressions on Zimbabwe's Selected Radio Stations. In: Emerging Trends in Indigenous Language Media, Communication, Gender, and Health. Kehinde O. Oyesomi; Abiodum Salawu (eds.). Hershey, PA: IGI Global. pp. 166-187.

Nassanga, Gorreti L.; Manyozo, Linje; Lopes, Claudia (2013): ICTs and Radio in Africa: How the Uptake of ICT Has Influenced the Newsroom Culture among Community Radio Journalists. In: Telematics and Informatics. 30/3. pp. 258-266.

Nyamnjoh, Francis (2005): Africa's Media, Democracy and the Politics of Belonging. London: Zed Books.

Peters, John D. (1999): Speaking into the Air: A History of the Idea of Communication. Chicago, London: University of Chicago Press.

Rosenthal, Anne (2019): When Old Technology Meets New: How UN Global Pulse Is Using Radio and AI to Leave no Voice Behind. In: United Nations Foundation. April 18. https://unfoundation.org/blog/post/when-old-technology-meets-new-how-un-global-pulse-is-using-radio-and-ai-to-leave-no-voice-behind/ [last accessed September 09, 2021].

Weidman, Amanda (2014): Anthropology and Voice. In: Annual Review of Anthropology. 43/37-51. p. 49.

Zanni, Lorenzo; Aziz, Abirah (2018): Artificial Intelligence in Broadcast and Media. In: iabm. May 8. https://theiabm.org/ai-broadcast-media-snapshot/ [last accessed July 30, 2021].

Notes

1 This is how the first Chairman of the Australian Broadcasting Commission, Charles Lloyd Jones, described radio voice in 1932 (as cited in Johnson 1988, p. 1).

2 In Zimbabwe, for instance, people were traditionally accustomed to the idea of elders visiting specific sacred shrines such as the Njelele shrine where they would present their pleas or grievances and 'hear' the voice of the invisible ancestors in response.

3 The idea of using the voice of an elder as a moral compass, and as someone to dispense wisdom and advice has been a critical feature of radio in Africa. Figures such as Mbuya Mlambo and Auntie Rhoda and Sekuru Nyathi played such a role in Zimbabwe (Mano 2004; Mpofu & Salawu 2020), while King Edward Masinga and Bloke Modisane were in similar roles in South Africa (Gunner 2019). This has also been a key feature of lifestyle magazines where similar figures have regular columns.

4 *Voice of the Revolution* was the broadcasting bulletin from the Zimbabwe African People's Union (ZAPU), operating from Lusaka, Zambia, while *Voice of Zimbabwe* was

the broadcasting bulletin from the Zimbabwe African National Union (ZANU), operating from Maputo, Mozambique.

[5] Carvajal, Elena S. (2020): 'The Radio, getting smarter everyday thanks to artificial intelligence,' https://business.blogthinkbig.com/the-radio-getting-smarter-everyday-thanks-to-artificial-intelligence/ [last accessed September 29, 2021].

Colleen Sanders
im Interview mit Aycha Riffi und Judith Kirberger

„Wir transportieren Emotionen“

Colleen, Du bist Journalistin, Moderatorin, Moderationstrainerin und seit 2017 Chefredakteurin bei Radio Lippewelle Hamm. Wie bist Du zum Radio gekommen? Wie sieht Dein beruflicher Werdegang aus?

Der ist sehr typisch: Nach meinem Abitur wollte ich in die journalistische Richtung gehen. Ich habe mich für ein geisteswissenschaftliches Studium entschieden und gleichzeitig für eine Lokalzeitung – den Westfälischen Anzeiger – geschrieben. Mein erstes Praktikum im Radio war bereits in den ersten Semesterferien und ab da war ich hoffnungslos verloren, weil mir sehr schnell klar geworden ist: Ich möchte mit Stimme arbeiten. Ich finde das direkte Gespräch, die direkte Kommunikation unheimlich stark. Für mich ist Radio das richtige Medium.

Obwohl ich mehr Radio gemacht habe, als zu studieren, habe ich das Studium in Köln abgeschlossen, dann ein paar Jahre freiberuflich für verschiedene Stationen gearbeitet, volontiert und wurde als Redakteurin eingestellt. Und tatsächlich bin ich jetzt Chefredakteurin bei dem Sender, bei dem ich im Februar 1996 mein erstes Radio-Praktikum gemacht habe – bei der Lippewelle.

Zusätzlich habe ich mich zur Moderationstrainerin im Hörfunk ausbilden lassen – als Inspiration für mich selbst und für meine Arbeit als Chefredakteurin. Ich mache das immer noch intensiv im Nebenberuf, weil es sehr interessant ist, mit Moderator*innen an der Stimme selbst zu arbeiten, am Einsatz der Stimme und daran, den richtigen Ton zu treffen.

Professionelles Arbeiten mit Stimme: Was bedeutet das für Dich im Radio-Kontext genau?

Wir haben mit unseren Stimmen unterschiedliche Aufgaben. Ganz klassisch vermitteln wir Information. Noch viel stärker transportieren wir Moderator*innen Emotionen. Wir wissen ja, dass jedes Wort, das wir sagen, eine Wirkung erzielt, und zwar

nicht nur inhaltlich, sondern eben auch dadurch, wie wir es sagen. Aktuell in dieser Zeit des Medien- und Mediennutzungs-Wandels kommt es noch mehr darauf an, eine emotionale Nähe zu schaffen und den Ton zu treffen, sodass die Anmutung zum Thema passt. Die Moderator*innen müssen sich immer sehr bewusst darüber sein, wie sie ihre Stimme einsetzen und welche Botschaften sie transportieren. In der Volontär*innenausbildung sage ich beim Moderationstraining, dass dies der Unterschied zu den anderen Disziplinen ist: „Ihr geht mit allem, was ihr seid, und allem, was ihr habt, on air – also mit eurer ganzen Persönlichkeit. Ihr seid wirklich greifbar für die Hörer*innen draußen."

In der Moderation oder in den Formaten, in denen ich arbeite, kann man sich nicht hinter den Informationen verstecken, denn es geht auch um einen selbst – um die eigene Persönlichkeit. Und in der Kommunikation mit den Hörer*innen sind auch spontane emotionale Reaktionen auf aktuelle Themenlagen relevant: Wenn beispielsweise ein Terroranschlag verübt wird oder aber auch wenn etwas besonders Schönes passiert, dann müssen Moderator*innen mit diesen Themen umgehen können. Sie müssen in der Lage sein, eine Stimmung zu transportieren. Ich finde, das ist eine der Kernaufgaben, die wir erfüllen sollten. Gute und gut ausgebildete Moderator*innen können das.

*Würdest Du sagen, dass dies die ‚gewisse Authentizität' ist, die Stimme erzeugt? Und kann das jede*r mit genügend Übung, oder gibt es die Radiostimme, also so eine Art Talent zur Authentizität?*

Authentizität lernen die meisten – manche schneller, manche langsamer –, aber nicht jede*r bringt Entertainer*innen-Qualitäten mit und will auch mit der eigenen Persönlichkeit ‚raus'. Die Stimme ist gut trainierbar, aber natürlich gibt es auch grundsätzlich ‚schöne' Stimmen. Die haben erst einmal größere Chancen. Interessant ist aber auch eine außergewöhnliche Stimme, die auffällt. Und dann gibt es Moderator*innen mit einer tollen Persönlichkeit, die bei den Menschen unheimlich gut ankommt, und bei denen ein hoher Wiedererkennungseffekt entsteht. Aber gefällige Stimmen haben es im Radio natürlich leichter.

Stimme kann sehr gut Authentizität schaffen. Es kann aber auch das Gegenteil passieren: Je nachdem wie man sie einsetzt, kann es auch alles andere als authentisch sein.

Unterscheidet sich die Stimme, die Du im Radio einsetzt, von derjenigen Stimme, die Du für Auftragsarbeiten wie Werbung oder Ähnliches verwendest?

Da macht tatsächlich Authentizität den Unterschied. Ich habe einmal Bahnansagen gesprochen. Im Rückblick war das skurril, weil mir gesagt wurde: „Frau Sanders, seien Sie doch nicht so freundlich. Jetzt hören Sie doch mal auf zu lächeln, Frau Sanders, Sie lächeln zu viel. Sprechen Sie das doch mal ganz normal." Ich habe gesprochen, wie ich es vom Radio kenne. Ich bin ohnehin ein gutgelaunter Mensch und das ist ja auch eher die Stimmungslage in einer Radiosendung – freundlich und offen, nah bei Hörerin und Hörer. Aber dort sollte ich eine automatisierte Stimme nachahmen: möglichst wenig Betonung, wenig Lächeln, wenig Emotionen in der Stimme. Aus praktischer und technischer Sicht kann ich das verstehen, denn je neutraler die Aufnahmen sind, desto leichter ist der Schnitt der Aufnahmen. Man kennt das beispielsweise vom Navigationsgerät: Wenn die Aufnahmen nicht gut geschnitten sind, ergibt sich eine ganz komische Satzmelodie. Daher sollten die Sätze so neutral wie möglich gehalten sein. Ich hatte das Gefühl, dass ich jetzt möglichst wenig authentisch sein und möglichst wenig persönlich sprechen muss – obwohl ich glaube, die Fahrgäste hätten sich auch über etwas Lächeln gefreut.

Authentizität und Emotionen sind hier die Unterschiede. Es gibt auch Auftragsarbeiten, bei denen gezielt bestimmte Emotionen gefordert sind, wenn in der Produktpräsentation eine völlig überdrehte und glückliche Stimme gefragt ist. Das erleben wir in Werbespots, die die Zuhörer*innen gewissermaßen anschreien. In der Vertonung von Nachrichten geht es hingegen um eine sachliche, wenn auch grundsätzlich freundliche, Variante oder Anmutung.

*Bekommt Ihr Feedback Eurer Hörer*innen zu den Moderationen beziehungsweise zu den Stimmen?*

Feedback ist für unsere Arbeit sehr wichtig, gerade weil es in der Natur des Radios liegt, dass die Kommunikation erst einmal einseitig in eine Richtung geht. Aber tatsächlich gibt es oft Feedback von Hörer*innen, das sich unmittelbar auf die Stimme bezieht: „Ach, Sie höre ich immer so gerne; Sie haben so eine schöne Stimme."

Hier gibt es auch interessante Kommentare wie: „Sie sind immer so frech" oder „Sie sind so mutig" oder „Sie sind immer so lustig." Da wird von Stimme und Spreche auf die Charaktereigenschaften der Moderator*innen geschlossen. Das passt zu meiner These, dass Moderator*innen mit ihrer ganzen Persönlichkeit – die von ihren Stimmen transportiert wird – raus gehen, ob sie das wollen oder nicht. Und wenn

es gut läuft, sind Stimme und Persönlichkeit auch deckungsgleich und authentisch. Wenn nicht, dann wird man beispielsweise auf einmal für eine Zicke gehalten, obwohl man es gar nicht ist.

Da Du gerade das Beispiel „Vorwurf Zicke" erwähnst, lass uns über unterschiedliche Reaktionen auf männliche und weibliche Stimmen sprechen. Vorweg gefragt: Werden Deine männlichen Kollegen auch mit dem Adjektiv „zickig" konfrontiert?

Das ist interessanterweise ein Riesenthema, weil Moderatorinnen ganz anders bewertet werden als ihre männlichen Kollegen. Frauen werden oft als zu schrill, zu laut und schnell als anstrengend bezeichnet, während Männer als lustig und durchsetzungsstark, kreativ und selbstbewusst empfunden werden. Wir kennen das auch aus anderen Diskussionen und Bereichen, nur verschärft sich das hier noch mal und wird zugespitzt, weil die Zuhörer*innen ja nur die Stimme erleben. Ja, ich denke, dass Frauenstimmen anders bewertet werden.

Im Bereich der Morning Show haben wir beispielsweise folgende Erfahrung gemacht: Das Standard-Konzept ist ja eine Moderation mit einem Mann und einer Frau als gleichwertige Partner. Das funktioniert prima, und zwei Männer zusammen – sozusagen als Best Buddys – funktionieren auch sehr gut. Bei zwei Frauen wird es hingegen sofort schwieriger. Das haben wir über Jahre beobachtet. Es gibt nur wenige weibliche Doppelmoderationsteams, die wirklich von den Hörer*innen honoriert werden und deren Quoten sich positiv entwickeln. Ich kenne dies aus dem Privatradio, wo wir leider oft sehen, dass ein komplett weibliches Morning-Show-Team eher schlechter bewertet wird. Und das hat nichts mit der handwerklichen, fachlichen oder inhaltlichen Qualität zu tun. Selbst wenn es hier brillante Moderationen gab, war die Hörer*innen-Akzeptanz gering. Ich glaube, Frauen werden manchmal schnell als anstrengend, zu intensiv oder als schrill empfunden. Das ist schade.

Vielleicht liegt es auch daran, dass es ein gelerntes Klischee gibt: Wenn zwei Frauen zusammenkommen, sprechen sie automatisch über ‚Frauenthemen'. Eine Chance, dieses Klischee oder Vorurteil aufzubrechen, könnte tatsächlich in selbstproduzierten Podcasts liegen, die vielfältige Themen behandeln und erst einmal nicht um eine Quote kämpfen müssen. Es gibt ganz tolle Beispiele von guten und erfolgreichen Podcasts, die Frauen machen.

Ja! Hier wäre meine These, dass es auch etwas mit der Hörsituation zu tun hat. Bei der Morning Show geht es darum, morgens wach zu werden und die wichtigsten Informationen zu bekommen. Radio ist hier oft ein Nebenbei-Medium. Ein Podcast ist ein On-demand-Angebot, das in der Regel in Ruhe, mit Zeit, mit ‚Kapazität im Kopf' gehört wird. Die Erwartungen sind oft andere als beim Radio. Das würde dafür sprechen, dass Stimmen hier eine andere Rolle spielen und nicht im Vordergrund stehen – vielleicht eher der Inhalt, der Kanal oder die Hörsituation. Das ist ein hochspannendes Thema.

Lass uns in diesem Zusammenhang auch über Diversität sprechen. Wie viel Stimm-Diversität – also männlich und weiblich, alt und jung, Dialekte, Akzente, Sprachfehler – ist im Radio möglich?

Die Kategorie „Alter" ist schon mal ein ganz interessanter Punkt, weil das Stimmalter ja gar nicht so klar einzuschätzen ist. Oft wirken Stimmen jünger oder älter als der dazugehörige Mensch wirklich ist. Wir haben in Hamm gerade eine Produktion zusammen mit den Kirchen gemacht. Dort sind im Ehrenamt viele ältere Menschen vertreten. Wir hatten eine über 70-jährige Frau auf der Antenne, und ebenso Jugendliche – der Jüngste war um die 14 Jahre alt. Die hört man sonst selten so gemischt. Diese Stimmen waren unheimlich schön und bereichernd, weil so eine ganz andere Perspektive präsent wird. Ich glaube, wir müssen schauen, welche Diversität wir in unserem Sendegebiet haben und wo wir die entsprechenden Stimmen finden. Diese Diversität sollte dann im Radio stattfinden.

Darüber hinaus kann aber die Frage nach der Verständlichkeit eine Rolle spielen. Ich habe vor kurzem ein Moderationsseminar gegeben, da hatte ich eine Teilnehmerin mit einem extrem starken, wunderschönen französischen Akzent und einer wirklich schönen Stimme. Ich hätte ihr stundenlang zuhören können, habe aber leider stellenweise nichts verstanden. Und das ist im Radio natürlich problematisch. Die Möglichkeiten, mit einem Akzent Programm zu machen, sind ein bisschen begrenzt, weil wir oft in einer Nebenbei-Hör-Situation sind. Da ist leichte Verständlichkeit einfach ein ganz wichtiges Kriterium. Ähnlich sehe ich das bei extremen Sprachfehlern. Ein leichtes Lispeln kann sehr charmant sein; bei einem deutlichen Sprachfehler könnten wir einfach nicht garantieren, dass die Zuhörer*innen das Gesagte noch verstehen können. Vielleicht fehlt es in der Hör-Situation dann auch an der nötigen Konzentration.

*Es gibt ja durchaus Radioprogramme mit Moderator*innen, deren Muttersprache nicht Deutsch ist. Meist sind dies aber Formate, die internationale Kultur thematisieren. Warum ist das in einer Morning Show nicht möglich, was denkst Du?*

Ganz viel ist in diesem Zusammenhang Gewöhnung. Wir müssen uns da mehr trauen. Mein Job ist ja, möglichst viele Menschen in unserem Sendegebiet zu erreichen – und hier leben Menschen aus mehr als 120 verschiedenen Nationen. Wir bilden unsere Stadtgesellschaft ab. Ich hätte gerne mehr Mitarbeiter*innen mit unterschiedlicher Herkunft im Programm, auch um diversere Perspektiven auf die Themen der Stadt zu haben. Da geht es sowohl um die Inhalte und die redaktionelle Arbeit als auch um die Ansprechhaltung und die Stimmen on air. Ich sehe in der Akquirierung diverser Moderator*innen deshalb auch eine wichtige Aufgabe für uns. Wir versuchen gerade einiges, um uns diesbezüglich besser aufzustellen und weiterzuentwickeln.

*Diversität würde im Radio ja auch Nähe und Authentizität schaffen. Künstliche Stimmen sind eher nicht divers, dafür sind sie gut verständlich. Sind sie eine Konkurrenz für Radio-Moderator*innen?*

Ich denke, dass es Bereiche gibt, wo dies möglich ist, beispielsweise beim klassischen Service wie Wetter, Verkehr, Musik. Das sind Moderationen, die nicht unbedingt emotional sind. Dort steht die schlichte Information im Vordergrund. Wenn es darüber hinausgeht, wenn also Emotionen, Spontanität und menschlicher Umgang mit Themen gefragt sind und wir diese Verschränkung zu Inhalten haben, da, glaube ich, haben es künstliche Stimmen schwer. Und ich hoffe doch, dass wir noch ganz lange bei menschlichen Moderationen bleiben. Es gibt so viele Nuancen, so viel zwischen den Zeilen und zwischen den Worten, dass ein Unterschied zwischen einer künstlichen Stimme und einer echten Person deutlich hörbar ist. Zum Glück, denn ich glaube, es gibt Bereiche, wo man uns ersetzen könnte.

Aber auch ich finde es spannend zu beobachten, welche Faszination künstliche Stimmen und Sprachassistenzen haben: Der Umgang mit Smart-Speakern wie Alexa ist oft schon freundschaftlich. Man unterhält sich mit ihr. Für mich stellt sich die Frage, was diesbezüglich alles möglich sein wird: ob irgendwann der Sprachassistent die passenden Reaktionen und Emotionen beherrscht. Ich hoffe ehrlich gesagt, dass die Technologie nie so gut wird, dass sie eine echte Person gleichwertig ersetzt.

Es wäre also durchaus machbar, dass künstliche Stimmen im Radio Staumeldungen oder Wettervorhersagen übernehmen. Oft werden aber auch Staumeldungen mit persönlichen Kommentaren versehen: „Alle auf der A3 müssen jetzt stark sein." Siehst Du bei Euch oder anderen Radiosendern die Gefahr, dass diese Aufgaben absehbar von künstlichen Stimmen übernommen werden?

Ich glaube, das versuchen alle auch ein bisschen zu verhindern. Denn es ist ja wirklich so, dass programmiert werden kann, wenn der Stau so und so lang ist, dann sag dazu: „Ihr müsst jetzt stark sein." Und wenn die Temperatur unter 8 Grad ist, dann sag: „Hoffentlich habt ihr dicke Socken dabei." Aus dem NRW-Lokalfunk wüsste ich aber jetzt keinen Fall.

Es besteht die Möglichkeit, echte Stimmen aufzuzeichnen und zeitversetzt einzuspielen, für Sendestunden oder auch bei den Sprachanteilen der Station Voice – das ist eine echte Stimme, die neutrale Elemente spricht oder auch eine Botschaft im Trailer. Mir ist aber nicht bekannt, dass in unserer Branche aktuell künstliche Stimmen als Ersatz für eine Moderation zum Einsatz kommen. Was wir jedoch ausprobiert haben, ist, Alexa gezielt als Gag-Element einzusetzen. Wir hatten einen Morning-Show-Trailer, der ging ungefähr so: „Alexa, koch mir mal einen Kaffee." Und Alexa antwortet dem Moderator: „Koch dir deinen Kaffee doch selber." So etwas kann mal eingebaut werden. Aber im Moment bezahlen wir noch Menschen dafür, dass sie die Verkehrsmeldungen sprechen.

Das heißt für Dich als Chefredakteurin oder als Moderationstrainerin, dass Du weiter Menschen trainieren und nicht Computer füttern willst. Und wie siehst Du die Entwicklung: interessiert, ängstlich, freudig?

Schon mit einem gewissen Selbstbewusstsein. Auf jeden Fall auch interessiert – und wir nutzen ja zum Beispiel Alexa mit einem eigenen Skill für unser Programm. Wir probieren neue Technik aus, und dazu zähle ich jetzt auch Sprachassistenten und Smart-Speaker. Vielleicht machen wir einmal eine verrückte Kampagne und Alexa liest unseren Verkehrsservice für einen Monat. So können wir uns mal anhören, wie das klingt. Aber außer Kosten einzusparen weiß ich nicht, was es bringen soll. Die meisten Aufgaben müssen auch weiterhin von echten Menschen erledigt werden. Was die künstlichen Stimmen auch nicht ersetzen können, was aber in unserer Arbeit enorm wichtig ist: Wir Radiomacher*innen sind ja überall präsent und können im ‚echten Leben' auf Veranstaltungen angesprochen werden. Gerade für lokale Sender ist das wichtig.

*In Zeiten von Kontaktbeschränkungen durch die Corona-Pandemie kann ich mir gut vorstellen, dass diese Beziehung zwischen Moderator*in und Zuhörer*in nochmal intensiver wird für viele Menschen. Ich denke da beispielsweise an Menschen, die alleine leben.*

Wir sind auf jeden Fall hilfreich bei Einsamkeit, weil wir für die Menschen vertraute echte Stimmen sind und auch Persönlichkeiten, die man über die Jahre kennenlernt. Das Radio ist trotz seiner einseitigen Kommunikation sehr dialogisch angelegt. Wir arbeiten mit unserer Community, mit Social Media oder mit Call-Ins. Wir führen Umfragen und die Hörer*innen können sich beteiligen – zum Beispiel zu aktuellen Themen etwas sagen, Themen vorschlagen und so Teil des Programms werden. Auch wenn wir nur Reaktionen auf Facebook präsentieren oder WhatsApp-Töne einspielen und nicht im direkten Gespräch sind, ist das auch eine Art von Kommunikation und Teilhabe am Programm. Radio ist ein Community-Medium.

Durch die Digitalisierung und soziale Medien seid Ihr Radiomenschen ja tatsächlich viel sichtbarer.

Ja, wir sind beispielsweise nach der Sendung auf diversen Social-Media-Plattformen vertreten. Und auf unserer Homepage gibt es Fotos: dort gibt es ein Gesicht zu der vertrauten Stimme.

Das ist mir sehr wichtig: Authentizität, Einsatz der Stimme und Persönlichkeit on air. Ich glaube, dass dies das Erfolgsgeheimnis ist. So bleibt Radio uns auch als Medium erhalten: mit echter Emotion, echter Reaktion und echter Spontanität. Das wird so schnell nicht programmierbar sein.

Anhang

Kurzvorstellungen der Autor*innen und Interviewpartnerinnen

Christine Bauer is an Assistant Professor at Utrecht University, The Netherlands. She is an experienced teacher in a wide spectrum of topics in computing and information systems, ranging from algorithms to adaptive interactive systems to research methods. Her research activities center on interactive intelligent systems and her work takes a human-centered computing approach, where technology follows humans' and the society's needs. A central theme in her research is context-adaptivity. Recently, she focuses on context-aware (music) recommender systems.

Marc Böhlen is Professor of Emerging Practices in Computational Media in the Art Department at the University at Buffalo. https://realtechsupport.org/

Oksana Bulgakowa ist Filmwissenschaftlerin. Sie hat mehrere Bücher über das russische und deutsche Kino verfasst und herausgegeben (*Die ungewöhnlichen Abenteuer des Dr. Mabuse im Lande der Bolschewiki*, 1995; *Eisenstein. Eine Biographie*, 1998, engl. 2002, russ. 2017; *Resonanz-Räume. Die Stimme und die Medien*, 2012; *Stimme*, 2015; *SinnFabrik/Fabrik der Sinne*, 2015), bei Filmen Regie geführt, Ausstellungen kuratiert und Multimediaprojekte entwickelt (*The Visual Universe of Sergei Eisenstein*, Daniel Langlois-Foundation, 2005; DVD *Factory of Gestures*, 2008; *Eisenstein: My Art in Life* für Google Arts & Culture, 2015).

Lílian Campesato is a Brazilian artist, researcher, and curator. She has been a research fellow at the University of São Paulo and research affiliate at NuSom since 2012. She works mainly in the following areas: sound studies, experimental music, sound arts, and feminisms. Her texts discuss listening, noise, experimentalism, and counter-hegemonic discourses in these fields. As an artist she explores voice and performance. Since 2017, she has developed the research project *Microfonias: intention and sharing of listening* in partnership with Valéria Bonafé. In addition, Campesato is co-founder of *Sonora: musics and feminisms*, a collective network dedicated to the discussion and expansion of feminist expressions in music and the arts.

Johanna Devaney is an Assistant Professor at Brooklyn College and the CUNY Graduate Center, USA, where she teaches courses on data analysis, music technology, music theory, and sonic arts. Her research focuses on interdisciplinary approaches to the study of musical performance, with a focus on the singing voice. Primarily, she examines the ways in which recorded performances can be used to study and model performance and develops computational tools to facilitate this. Her work draws on the disciplines of music, computer science, and psychology.

Laura Dreessen hat seit 2014 ihren anfänglichen Fokus auf Spracherkennung und technische Spezifikation von Dialogen im Automobilkontext zu einer ganzheitlichen Conversational UX- und Markenperspektive auf virtuelle Assistenzsysteme weiterentwickelt. Als VUI-Architect und Linguistin verhilft sie virtuellen Assistenzen zu einer möglichst intuitiven, multimodalen und nützlichen Mensch-Maschine-Interaktion und Unternehmen zu auditivem Charisma auf dem Voice Kanal. Ihre Verantwortung Nutzer*innen gegenüber liegt für sie darin, diese Technologie verständlich, vielfältig und sicher einzusetzen.

Marcus Erbe ist Juniorprofessor für Sound Studies am Musikwissenschaftlichen Institut der Universität zu Köln. Nach seinem Studium der Musikwissenschaft, Germanistik und Pädagogik war er Mitglied des interdisziplinären Forschungskollegs Medien und kulturelle Kommunikation, sodann Wissenschaftlicher Mitarbeiter und Akademischer Rat im Bereich Musik der Gegenwart. Unter Beteiligung zahlreicher Partnerinstitutionen aus dem In- und Ausland veranstaltet er den Vortrags- und Konzertzyklus Raum-Musik. Zu seinen Forschungsschwerpunkten zählen die Performanz und Medialität der Stimme sowie elektroakustische Musik, Klangkunst, Audiovisualität in Bewegtbildmedien, Game Studies und Popularmusik.

Fernando Iazzetta is a Brazilian composer and performer. He teaches music technology and electroacoustic composition at the University of São Paulo and is the director of NuSom – the Research Center on Sonology at the same university. His works have been presented in concerts and music festivals in Brazil and abroad. As a researcher, he has been interested in the investigation of experimental forms of music and sound art.

Judith Kirberger ist Sozialwissenschaftlerin. In der Grimme-Akademie wirkte sie am Projekt *Kulturelle Implikationen medial konstruierter Stimmen* sowie an diversen medienpädagogischen Projekten zu den Themen Fake News, Hate Speech und Verschwörungserzählungen mit. Als freie Referentin bietet sie Workshops und Fachvorträge zu rechtsextremen Online-Gruppierungen und (rechtsextremer) Hate Speech an. Dabei zeigt sie sowohl Entwicklungen in der Szene als auch Umgangsformen mit dieser auf.

Malte Kobel promoviert derzeit an der Kingston Universität London (gefördert von TECHNE/AHRC und Kingston University). Sein PhD-Projekt versucht eine Theoretisierung von Stimme in der Musik und arbeitet dabei entlang sowie quer durch Musikphilosophie, Sound Studies und Medientheorie. Neben der akademischen Arbeit ist Kobel verantwortlich für das Plattenlabel Hyperdelia.

Doris Kolesch ist Professorin für Theaterwissenschaft an der Freien Universität Berlin und Co-Sprecherin des Sonderforschungsbereichs „Affective Societies", in dem sie ein Forschungsprojekt zu affektiven Dynamiken in immersiven Theaterformen leitet. Zu ihren Arbeitsschwerpunkten gehören Theorie und Ästhetik von Theater und anderen Künsten, Performance und Performativität, kulturwissenschaftliche Affekt- und Emotionsforschung sowie Stimme und akustische Kultur.

Katharina Makosch schloss ihr Bachelor-Studium der Musikwissenschaft und Philosophie an der Universität zu Köln 2021 mit einer Arbeit zum Thema der Rolle der Stimme in der Vermenschlichung von Sprachassistenzsystemen ab. 2020 war sie als studentische Hilfskraft an dem Projekt *Kulturelle Implikationen medial konstruierter Stimmen* des Grimme-Forschungskollegs an der Universität zu Köln beteiligt. Derzeit führt sie ihr Studium der Musikwissenschaft im Master fort und nimmt darüber hinaus am Research Master-Programm der a.r.t.e.s. Graduate School for the Humanities Cologne teil.

Katherine Meizel is Associate Professor of Ethnomusicology at Bowling Green State University in Ohio. Her research has focused on voice and identity, topics in disability and Deaf studies, and popular music and media. Her most recent publications include the *Oxford Handbook of Voice Studies* (Oxford University Press 2019), which she co-edited with Nina Sun Eidsheim, and her monograph *Multivocality: Singing on the Borders of Identity* (Oxford University Press, January 2020). Her book *Idolized: Music, Media, and Identity in American Idol* (IU Press) was published in 2011.

Dumisani Moyo is Executive Dean, Faculty of Humanities at North West University in South Africa. He holds a Ph.D. from the University of Oslo, Norway. His research interests include media policy and regulation in Africa, new and alternative media, political engagement through media in Africa, journalism in the digital era, and media and elections. Among his major works are two co-authored books: *Radio in Africa: Publics, Cultures, Communities* (2011, Wits University Press) and *Media Policy in a Changing Southern Africa: Critical Reflections on Media Reforms in the Global Age* (2010, Pretoria: University of South Africa – UNISA – Press).

Kundai Moyo is a Research Assistant at the Visual Identities in Art and Design Research Center (VIAD) in the Faculty of Art, Design and Architecture at the University of Johannesburg. She is an artist and researcher based in Johannesburg who uses her work as a tool for conducting sociological research. Her recent practice imagines how the personal and communal business of love might enable new forms of theorising and framing socio-cultural particularities. Moyo is also the co-founder of wherewithall (https://www.wherewithall.org.za/), an online library of equipment, practical knowledge and research into independent curatorial practices in Johannesburg.

Stefanie Ray hat von 2000 bis 2005 Theater- und Medienwissenschaften, Psychologie und Neuere deutsche Literaturgeschichte an der Friedrich-Alexander-Universität Erlangen-Nürnberg studiert. Seit 2005 lebt sie in Köln und schreibt Comedy / Satire für renommierte TV und Rundfunk-Sendungen, unter anderem *Mensch Markus*, die *Harald Schmidt Show* und *Satire Deluxe*. Von 2016 bis 2019 kreierte sie bei der Amazon Development Center Deutschland GmbH als „Alexa personality writer“ die Persönlichkeit der deutschsprachigen Assistenz Amazon Echo.

Aycha Riffi leitet die Grimme-Akademie. Nach dem Studium der Theater-, Film- und Fernsehwissenschaft und einer Hospitanz beim ZDF/ Das kleine Fernsehspiel sammelte sie redaktionelle und journalistische Erfahrungen beim WDR, SDR und DSF. 2002 kam sie als freie Mitarbeiterin zum Grimme-Institut. Vom Grimme Online Award wechselte sie 2005 zur Grimme-Akademie. Dort war sie unter anderem für die europäischen Projekte *Media4us* und *BRICkS (Building Respect on the Internet by Combating Hate Speech)* in der Projektleitung. Im Rahmen des Grimme-Forschungskollegs war sie an den Projekten *Online Hate Speech. Perspektiven auf eine neue Form des Hasses*, *Produktionsforschung zu Film und Fernsehen*, *Fragmentierung der Öffentlichkeit* und *Kulturelle Implikationen medial konstruierter Stimmen* beteiligt.

Colleen Sanders-Heusener ist Journalistin und Trainerin. Den Umgang mit der Stimme kennt sie als langjährige Radiomoderatorin und Sprecherin für verschiedene Produktionen. Nach ihrem Abitur hat Sanders-Heusener in Köln Anglistik, Germanistik und Philosophie studiert und nebenbei seit 1996 für diverse Radiostationen gearbeitet. Seit 2017 ist sie Chefredakteurin von Radio Lippewelle Hamm. Als ausgebildete Moderationstrainerin ist sie seit 2010 als Coach und Programmberaterin für verschiedene Sender im Einsatz.

Nadia S. Zaboura ist Kommunikationswissenschaftlerin und Linguistin. Seit 2011 ist sie Inhaberin der Politik- und Kommunikationsberatung „Zaboura Consulting“ sowie langjährige Fach-Moderatorin und Jury-Vorsitzende des Deutschen Radiopreises. Zu ihren Kunden zählen Unternehmen, Ministerien und Verbände, die sie strategisch, kommunikativ und standortpolitisch berät – in den Märkten Medien und IKT sowie in den Bereichen Bildung, Demokratie und Digitalisierung. Als erfahrene Fach-Moderatorin kuratiert und begleitet sie zusätzlich seit vielen Jahren Medien-, Tech- und Wissenschaftskongresse. Darüber hinaus ist sie als Expertin für digitale Gesellschaft, Medien und Wirtschaft sowie als Facilitator und Gutachterin von Zukunftsthemen tätig (u.a. Europäische Kommission, BMBF, Grimme-Forschungskolleg).

Wolfgang Zielinski leitet den Arbeitsbereich Medienbildung am Grimme-Institut. Nach dem Studium der Germanistik und Anglistik (M.A.) widmete er sich Anfang der 2000er Jahre zunächst dem E-Learning in diversen Modellversuchen im Auftrag des Bundesinstituts für Berufsbildung (BIBB). Neben der *Bildbox für Millionen*, einer CD-ROM zur Fernseh- und Mediengeschichte der Bundesrepublik Deutschland für die Bundeszentrale für politische Bildung (bpb) entwickelte er medienpädagogische Unterrichtsmaterialien u.a. zur Boulevardberichterstattung (*schlagzeilen*). 2007 konzipierte er die Initiative Eltern+Medien im Auftrag der Landesanstalt für Medien NRW, für die er ab 2014 die Medienscouts NRW weiterentwickelte. Neben der Konzeption medienpädagogischer Angebote widmet er sich in den letzten Jahren verstärkt auch der kulturellen Relevanz digitaler Spiele und forscht zu ihren Potenzialen. 2017 erschien dazu in der Schriftenreihe digitale Gesellschaft NRW der Band *Spielend lernen! – Computerspiele(n) in Schule und Unterricht*. 2020 erschien im Rahmen des Grimme-Forschungskollegs der Band *Musikalische Praxen und virtuelle Räume* (kopaed).

Überblick über die Schriftenreihe zur digitalen Gesellschaft NRW (seit 2013)

Die Schriftenreihe erscheint im kopaed Verlag (Düsseldorf/München)

Die **Bände 3-7** sind auch im Open Access Format unter einer Creative Commons Lizenz erschienen (http://www.grimme-institut.de/publikationen/schriftenreihe/).

Band 7: Marcus Erbe / Aycha Riffi / Wolfgang Zielinski (Hrsg.): Mediale Stimmentwürfe. Perspectives of Media Voice Designs, 2022, 199 Seiten, ISBN 978-3-96848-642-0

Band 6: Harald Gapski / Stephan Packard (Hrsg.): Super-Scoring? Datengetriebene Sozialtechnologien als neue Bildungsherausforderung, 2021, 259 Seiten, ISBN 978-3-86736-575-8

Band 5: Wolfgang Zielinski / Sandra Aßmann / Kai Kaspar / Peter Moormann (Hrsg.): Spielend lernen! Computerspiele(n) in Schule und Unterricht, 2017, 200 Seiten, ISBN 978-3-86736-405-8

Band 4: Kai Kaspar / Lars Gräßer / Aycha Riffi (Hrsg.): Online Hate Speech – Perspektiven auf eine neue Form des Hasses, 2017, 200 Seiten, ISBN 978-3-86736-404-1

Band 3: Harald Gapski (Hrsg.): Big Data und Medienbildung – Zwischen Kontrollverlust, Selbstverteidigung und Souveränität in der digitalen Welt, 2015, 148 Seiten, ISBN 978-3-86736-403-4

Band 2: Lars Gräßer / Aycha Riffi (Hrsg.): Einfach fernsehen? Zur Zukunft des Bewegtbildes, 2013, 121 Seiten, ISBN 978-3-86736-402-7

Band 1: Cathrin Bengesser / Thomas Tekster (Hrsg.): Senioren im Web 2.0 – Beiträge zu Nutzung und Nutzen von Social Media im Alter, 2013, 128 Seiten, ISBN 978-3-86736-401-0

Schriftenreihe Medienkompetenz des Landes Nordrhein-Westfalen (2005–2012)

Die Schriftenreihe ist erschienen im kopaed Verlag (Düsseldorf/München)

Band 14: Harald Gapski / Thomas Tekster (Hrsg.): Informationskompetenz im Kindes- und Jugendalter. Beiträge aus Forschung und Praxis, 2012, 152 Seiten, ISBN 978-3-86736-214-6

Band 13: Lars Gräßer / Friedrich Hagedorn (Hrsg.): Soziale und politische Teilhabe im Netz? E-Partizipation als Herausforderung, 2012, 136 Seiten, ISBN 978-3-86736-213-9

Band 12: Harald Gapski (Hrsg.): Informationskompetenz und inklusive Mediengesellschaft. Dokumentation einer Fachtagung mit Projektbeispielen, 2012, 152 Seiten, ISBN 978-3-86736-212-2

Band 11: Lars Gräßer / Friedrich Hagedorn (Hrsg.): Medien nachhaltig nutzen. Beiträge zur Medienökologie und Medienbildung, 2012, 128 Seiten, ISBN 978-3-86736-211-5

Band 10: Harald Gapski / Lars Gräßer (Hrsg.): Verbraucherschutz und Medienkompetenz. Junge Konsumenten im Web, 2010, 128 Seiten, ISBN 978-3-86736-210-8

Band 9: Harald Gapski (Hrsg.): Jenseits der digitalen Spaltung. Gründe und Motive zur Nichtnutzung von Computer und Internet, 2009, 125 Seiten, ISBN 978-3-86736-209-2

Band 8: Harald Gapski / Lars Gräßer (Hrsg.): Medienkompetent in Communitys. Sensibilisierungs-, Beratungs- und Lernangebote, 2009, 128 Seiten, ISBN 978-3-86736-208-5

Band 7: Lars Gräßer / Monika Pohlschmidt (Hrsg.): Praxis Web 2.0 – Potenziale für die Entwicklung von Medienkompetenz, 2007, 170 Seiten, ISBN 978-3-86736-207-8

Band 6: Gernot Gehrke (Hrsg.): Web 2.0 – Schlagwort oder Megatrend? Fakten, Analysen, Prognosen, 2007, 128 Seiten, ISBN 978-3-86736-206-1

Band 5: Gernot Gehrke (Hrsg.): Public-Private-Partnerships in der Medienkompetenzförderung – Potenziale und Grenzen, 2006, 118 Seiten, ISBN 978-3-938028-93-3

Band 4: Klaus Solbach / Wolfgang Spiegel (Hrsg.): Entwicklung von Medienkompetenz im Hochschulbereich. Perspektiven, Kompetenzen und Anwendungsbeispiele, 160 Seiten, ISBN 978-3-938028-94-0

Band 3: Harald Gapski (Hrsg.): Medienkompetenzen messen? Verfahren und Reflexionen zur Erfassung von Schlüsselkompetenzen, 2006, 136 Seiten, ISBN 978-3-938028-53-7

Band 2: Gernot Gehrke (Hrsg.): Datenschutz und -sicherheit im Internet. Handlungsvorschläge und Gestaltungsspielräume, 2005, 152 Seiten, ISBN 978-3-938028-52-0

Band 1: Harald Gapski (Hrsg.): Leitbilder für die Wissensgesellschaft – Fallbeispiele, Strategien und Reflexionen, 2005, 136 Seiten, ISBN 978-3-938028-51-3

Bevor ‚sprechende' Navis, Computer und Smartphones unseren Alltag eroberten, war es hauptsächlich die Science-Fiction, die uns eine Vorstellung von der Beschaffenheit künstlicher Stimmen vermittelte. Doch so wenig diese fiktionalen Stimmen neutral gestaltet waren, so wenig sind es die digital generierten Stimmen heutiger Applikationen und Betriebssysteme. Bis jetzt mangelt es jedoch an Untersuchungen, die zugleich theoretische, medienpraktische und kulturübergreifende Aspekte vokaler Designs in IT-Erzeugnissen und den Medien berücksichtigen.

Das Forschungsprojekt „Kulturelle Implikationen medial konstruierter Stimmen" des Grimme-Forschungskollegs an der Universität zu Köln durchleuchtete mediale Stimmentwürfe dahingehend, welche Sozialvorstellungen ihnen innewohnen. Denn obwohl vielfach alternative Stimmklangmodelle jenseits fixierter Normen genutzt werden könnten, scheint bei Nutzer*innen der Wunsch nach normativen Vokalitäten vorzuherrschen. Bedarf die sprachbasierte Mensch-Maschine-Kommunikation hergebrachter, vertrauter kultureller Normen, und geraten dabei andere, diversere Optionen aus dem Blick?

Auch die Möglichkeiten und Auswirkungen der Stimmveränderung in der Musik und anderen Künsten im digitalen Zeitalter wurden genauer betrachtet. Schließlich spielten medienhistorische Zusammenhänge eine Rolle: Wann und wie kam es erstmals zur Vermenschlichung künstlicher Stimmen? Wie wandelte sich die stimmliche Konstruktion sozialer, ethnischer und geschlechtsbezogener Rollenbilder im Lauf der Mediengeschichte?

Der vorliegende Sammelband greift wesentliche Aspekte des Forschungsprojektes auf. Er enthält Beiträge und Interviews von und mit Christine Bauer & Johanna Devaney, Marc Böhlen, Oksana Bulgakowa, Lílian Campesato & Fernando Iazzetta, Laura Dreessen, Marcus Erbe, Judith Kirberger, Malte Kobel, Doris Kolesch, Katharina Makosch, Katherine Meizel, Dumisani & Kundai Moyo, Stefanie Ray, Aycha Riffi, Colleen Sanders, Nadia S. Zaboura und Wolfgang Zielinski.

ISBN 978-3-96848-642-0
16,80 €

Gefördert durch
Die Landesregierung
Nordrhein-Westfalen

kopaed