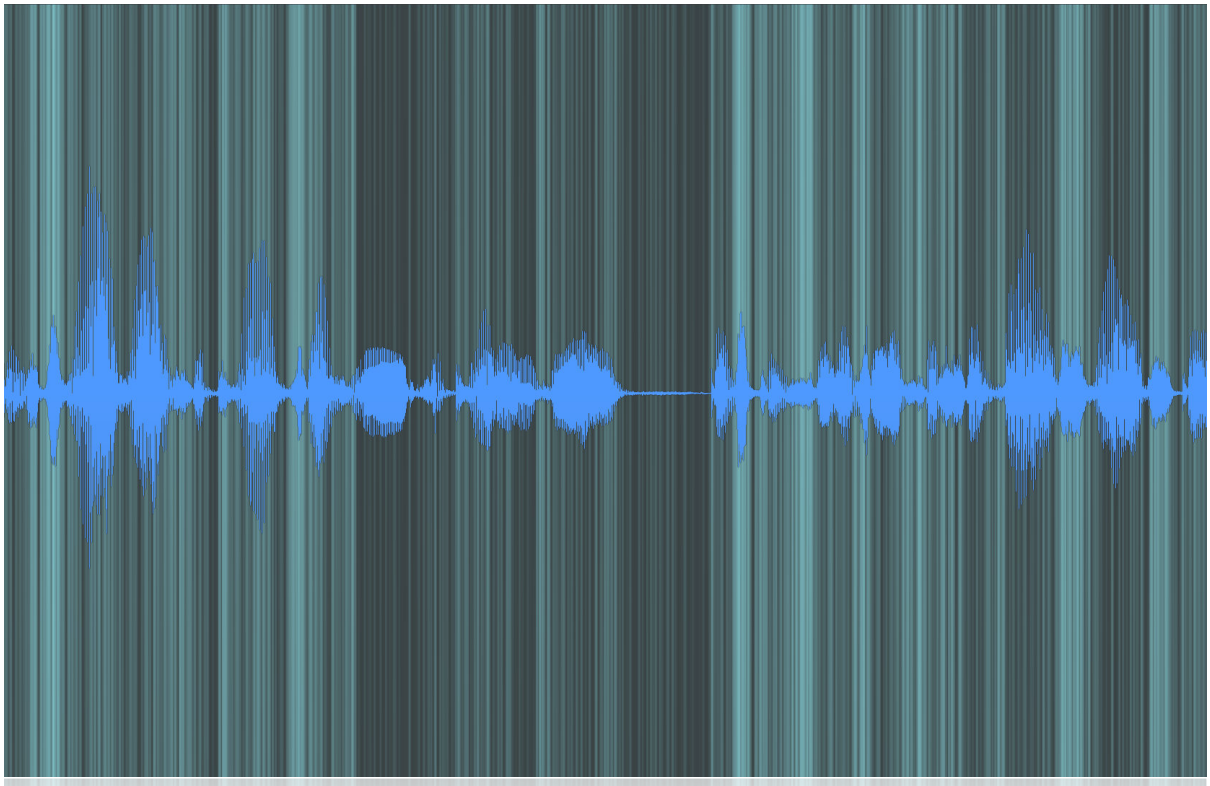
Aktuelle Themen, AM3
SS 2020
Benjamin Mütsch und Jan Bieletzki

KÜNSTLICHE INTELLIGENZ

in der Audioerkennung



Introduction

Was ist *KI*? Der sprachliche Ausdruck KI leitet sich aus dem Begriff *Künstliche Intelligenz* (eng. *Artificial Intelligence*) ab. Dahinter verbirgt sich im Wesentlichen ein Teilgebiet der Informatik, welches sich mit der Entwicklung effizienter, intelligenter Automatisierungs- und Problemlösungsprozesse sowie maschinellem Lernen (eng. *Machine Learning*) befasst.[1]

Bereits seit den 50er Jahren beschäftigt man sich mit den Möglichkeiten maschineller Logiksysteme, die den Entscheidungsfindungen des Menschen ähneln sollen. Auch durch die technische Weiterentwicklung der Computersysteme und die globale Vernetzung von Daten - Stichwort *Big Data* - gewinnt das Thema zunehmend an Relevanz. Der Aufbau von sogenannten *Neuronalen Netzen* und der Einsatz von *Deep Learning-T*echnologien sind heutzutage schon nicht mehr wegzudenken.[2] Dabei sind die

[1] Vgl.: https://de.wikipedia.org/w/index.php?title=Künstliche_Intelligenz&oldid=198129163
[2] Vgl.: https://www.gruenderszene.de/lexikon/begriffe/kuenstliche-intelligenz?interstitial

Anwendungsmöglichkeiten noch lange nicht vollends ausgereizt.

Forschungsergebnisse, sowie neuartige zum Teil geradezu spektakuläre Entwicklungen aus diesem Bereich sind immer wieder Teil großer medialer Aufmerksamkeit. So berichtete kürzlich ZEIT ONLINE in dem Artikel „Mit künstlicher Intelligenz gegen das Coronavirus“ über die Möglichkeiten von Algorithmen, welche in der medizinischen Diagnose helfen und gezielt Wirkstoffe finden könnten. Auch bei der Entwicklung von besseren Vorhersagemodellen über die Verbreitung des Corona-Virus könnten eigenständige, intelligente Analyse-Systeme helfen.[3]

Die Einsatzbereiche von künstlicher Intelligenz sind breit gefächert. Ein besonders spannender Bereich findet sich dabei in der Erkennung, Auswertung und Restauration von Audiodaten. Neben den grundlegenden technischen Hintergründen werden im Folgenden auch diverse Tools und Anwendungsbeispiele aus der Praxis vorgestellt.

Sound Detection - how does it work?

Dieser Abschnitt widmet sich der Frage, wie die Tonerkennung überhaupt funktioniert und in welchen Formen und Varianten diese erfolgen kann. Es soll zudem ein Überblick über die unterschiedlichen Technologien gegeben werden, welche dazu dienen können, Töne als Informationen zu erfassen. Dabei wird zunächst kurz erläutert, wie Schall, beziehungsweise ein Schallereignis oder auch der Schalldruckpegel definiert ist.

- *Schall* ist eine Schwingung mechanischer Art, die ein flüssiges oder gasförmiges Übertragungsmedium benötigt.

- *Schalldruckpegel* bezeichnet die Intensität dieser mechanischen Schwingung und wird in dB SPL (Sound Pressure Level) angegeben.

- *Schallereignis* darunter versteht man einen physikalisch-akustischen Vorgang der durch physikalische Parameter bestimmt ist. Ein Schallereignis wird z.B. über Schallfeldgrößen wie die Entfernung und Richtung der Schallquelle oder auch deren Frequenzspektrum definiert.[4]

Soll nun ein Schallereignis als Datengrundlage für eine Tonerkennung dienen, muss dieses zunächst in ein quantifizierbares und interpretierbares Signal gewandelt werden. Aus akustischen Schwingungen wird deshalb ein digitaler Datensatz generiert mit dem Algorithmen und intelligente Systeme letztendlich arbeiten können. Mikrofone fungieren an dieser Stelle häufig als Schallwandler und übertragen die Schwingungen in eine elektrische Ausgangsspannung. Hier setzt dann die *Analog-Digital-Wandlung* (kurz: AD-Wandlung) an.

[3] Vgl.: https://www.zeit.de/digital/internet/2020-03/covid-19-kuenstliche-intelligenz-coronavirus-diagnose-technik
[4] Vgl.: https://de.wikipedia.org/wiki/Schallereignis

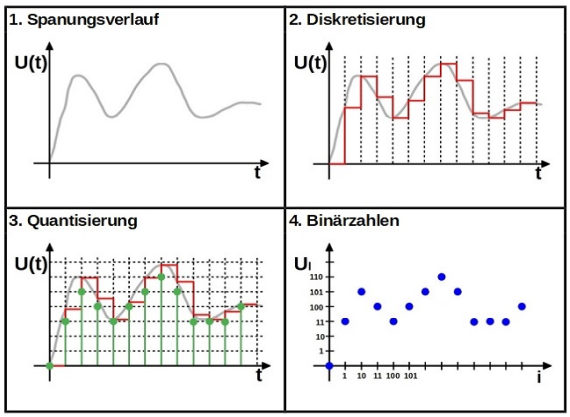


Abb. 1: Analog-Digital-Wandlung[5]

Der Spannungsverlauf über die Zeit in Abbildung 1 steht exemplarisch für die variierende Ausgangsspannung eines Mikrofons, bei dem die Membranauslenkung die Höhe der Spannung bestimmt. Dieses analoge Signal ist sowohl zeit- als auch wertekontinuierlich. Um dieses in die „digitale Welt“ zu übertragen, werden diskrete Werte benötigt. Daher wird das Signal zeitlich abgetastet. Die Abtastrate, auch Samplerate genannt, bestimmt in wie viele Samples pro Sekunde das eingehende Signal unterteilt wird. Je höher die Samplerate desto genauer ist auch die zeitliche Auflösung. Diese zeitliche Rasterung wird in einem weiteren Schritt, der sog. *Quantisierung*, auch auf die Werteebene (y-Achse) übertragen. Für jedes einzelne Sample wird der zugehörige Spannungswert dem nächstgelegenen Wert auf dem Raster zugeordnet. Die vertikale Auflösung wird durch die Bittiefe bestimmt. Zuletzt erfolgt die Umsetzung in Binärzahlen.

Liegt ein digitaler Datensatz vor, können Algorithmen und Methoden zur Erkennung spezifischer Merkmale angewendet werden. Die Anwendungsgebiete sind dabei sehr vielfältig: Sprachsteuerung, Sprachsynthese, Spracherkennung, impulshafte Gefahrengeräusche sowie Alarmsignale. Auch die Bestimmung von zusätzlichen Eigenschaften diverser Signale, wie die Tonhöhe oder Aussprache können beispielsweise Auskunft über die emotionale Stimmung eines Sprechers geben.

Interessant hierbei ist, dass in modernen Geräten wie *Smart Home Devices* und persönlichen Assistenten nicht einfach eine kleine Mikrofonkapsel zum Einsatz kommt. Es werden an dieser Stelle oft aufwendige Arrays aus mehreren Mikrofonen verwendet um Umgebungslärm zu reduzieren oder die Reichweite zu erhöhen.[6] Diese Technik bewirkt auch, dass zwischen Umgebungsgeräuschen und Haushaltslärm eine effektive Stimmerkennung möglich ist und sich der Sprechende nicht in unmittelbarer Nähe zum Gerät befinden muss.

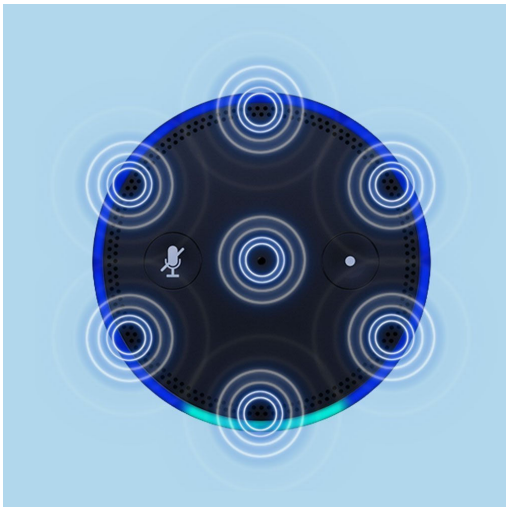


Abb. 2: Mikrofonarray für Voice-Control[7]

[5] https://www.technik-unterrichten.de/Robotik/Schallsensor/Bilder/Signalwandler.jpg
[6] Vgl.: https://electronics.howstuffworks.com/gadgets/high-tech-gadgets/amazon-echo1.htm
[7] https://www.techproviderzone.com/sites/techproviderzone/files/Amazon%20Echo’s%207-microphone%20far-field%20array.png

Auch im Bereich des öffentlichen Lebens findet Tonerkennung statt um die allgemeine Sicherheit zu erhöhen und eine zusätzliche Dimension zur Videoüberwachung zu ermöglichen. Dort sollen Gefahren anhand von spezifischen Geräuschmustern in Echtzeit erkannt werden, welche sich auch anderen Überwachungsmöglichkeiten wie Rauch- oder Bewegungsmeldern entziehen. Hierzu werden vor allem impulshafte Signale analysiert, wie zum Beispiel Explosionen, Unfälle oder Schüsse.[8]

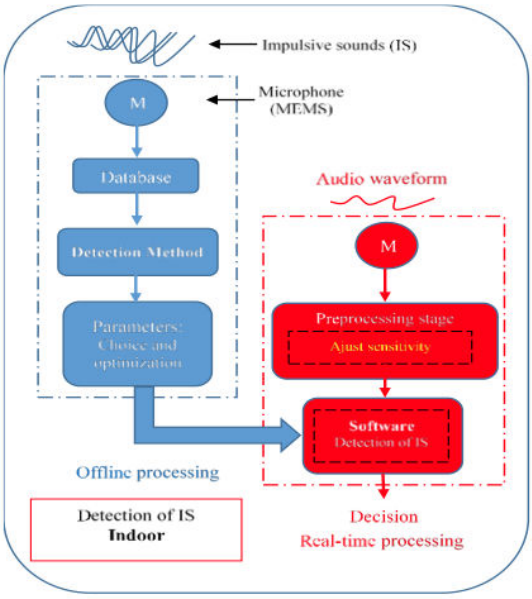


Abb. 3: Analyse akustischer Ereignisse[9]

Hier muss ebenfalls wieder zwischen irrelevanten und relevanten Schallereignissen entschieden werden, was in Echtzeit oder auch „offline“ geschehen kann. Künstliche Intelligenz und Machine Learning werden dabei bevorzugt eingesetzt, um Erkennungsmethoden zu optimieren und zu automatisieren.

Sound Intelligence for Health Care

„KI ist über die automatische Erkennung gesprochener Inhalte weit hinaus (…) Dank Verfahren des maschinellen Lernens in Kombination mit entsprechenden Daten und Modellen können aus dem Audiosignal neben klassischen demographischen Merkmalen (…) auch menschliche Emotionen, Persönlichkeitsmerkmale und Gesundheitszustände erkannt werden.“[10] Hierbei spielt das sogenannte *Affective Computing* eine wesentliche Rolle. Dabei handelt es sich um ein interdisziplinäres Gebiet der KI-Forschung, bei der Systeme entwickelt werden, die Emotionen verstehen können. Damit sind Maschinen beispielsweise in medizinischen Bereichen in der Lage, „den individuellen Zustand eines Menschen (…) zu bewerten und ‚intelligent‘ zu reagieren.“[11] Gerade im Gesundheitswesen besitzen Technologien mit Künstlicher Intelligenz daher ein enormes Potenzial.

Im Falle eines Parkinson Patienten können Veränderungen in der Stimme frühzeitig erkannt werden, was bei neurokognitiven/-degenerativen Krankheiten die Diagnostik beschleunigen und somit die Therapie verbessern kann. Parkinson als Nervenkrankheit ist an dieser Stelle nur ein Beispiel. Mit den Sprachanalysetools von audEERING soll es möglich sein, motorische Defizite in der Artikulationsmuskulatur zu erkennen noch bevor das krankheitsbedingte Zittern ausbricht.

[8] Vgl.: http://ceur-ws.org/Vol-2351/paper_49.pdf
[9] http://ceur-ws.org/Vol-2351/paper_49.pdf
[10] https://www.bitkom.org/sites/default/files/2019-10/20191014_sof2_healthai_1.pdf
[11] https://www.bitkom.org/sites/default/files/2019-10/20191014_sof2_healthai_1.pdf

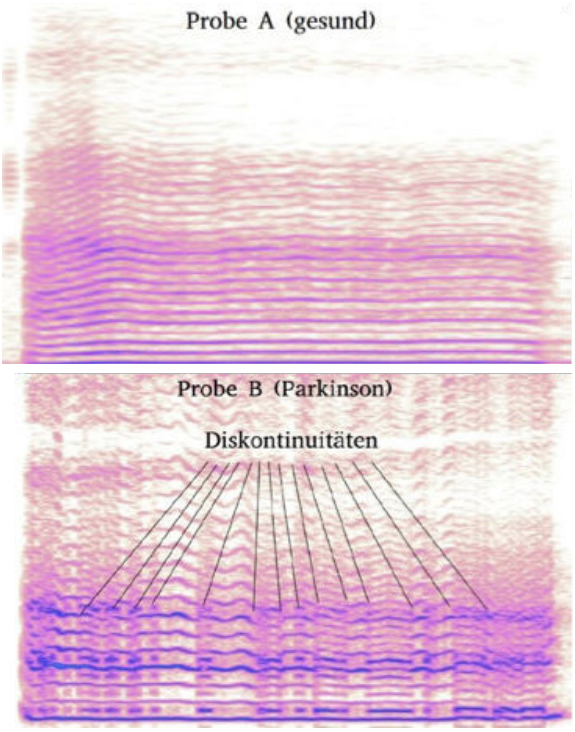


Abb. 4: Vorher-/ Nachher-Spektrogrammdarstellung bei einem Parkinson-Patienten[12]

Für eine intelligente Sprachanalyse muss der Algorithmus zuvor mit „annotierten Sprechdaten", also Audioaufnahmen welche einem eindeutigen Krankheitsbild zugewiesen werden können, trainiert werden.[13]

Die intelligente Audioerkennung kann auch beim Monitoring im Gesundheitswesen zur Verbesserung der Betriebseffizienz eingesetzt werden. Derartige Sensoren helfen überall, wo regelmäßige Kontrollen der Patienten vonnöten sind. Beispielsweise in Krankenhäusern und Pflegeheimen: Dort stehen regelmäßige Kontrollen nicht nur an der Tagesordnung, sondern auch während der Nacht. Diese Kontrollen gewährleisten keine sofortige Hilfestellung im Falle eines Notfalls. Zudem stören sie auch die Privatsphäre eines Patienten und erfordern einen hohen Personalaufwand, wodurch andere Aufgaben zwischenzeitlich unterbrochen werden müssen. In einem Szenario mit akustischem Überwachungssystem werden Notfälle hingegen unmittelbar detektiert.

Überschreitet ein bestimmtes Klangprofil einen Schwellwert, wird ein Alarm an die zentrale Überwachungsstation und/oder direkt an ein verfügbares Mitglied des Pflegepersonals gesendet. Diese hochmoderne Sensorik ermöglicht ein schnelleres Eingreifen und eine effizientere Auslastung des Personals. Zugleich wird eine bessere Privatsphäre für die Patienten gewährleistet[14]

Gerade bei primitiveren akustischen Überwachungssystemen können Fehlalarme durch irrelevante Umgebungsgeräusche ausgelöst werden. Um unnötige akustische Trigger zu vermeiden, hat sich die Erkennungstechnologie von rein pegelbasierten Systemen hin zu einer intelligenten Mustererkennungssoftware weiterentwickelt bei der unter Berücksichtigung des Verhaltens im Frequenzbereich einzelne Geräusche unterschieden werden können. Gerade hier bieten sich die Ansatzpunkte für weitere Optimierungen mithilfe von KI.

[12] https://www.bitkom.org/sites/default/files/2019-10/20191014_sof2_healthai_1.pdf
[13] Vgl. https://www.bitkom.org/sites/default/files/2019-10/20191014_sof2_healthai_1.pdf
[14] Vgl.: https://global.clb.nl

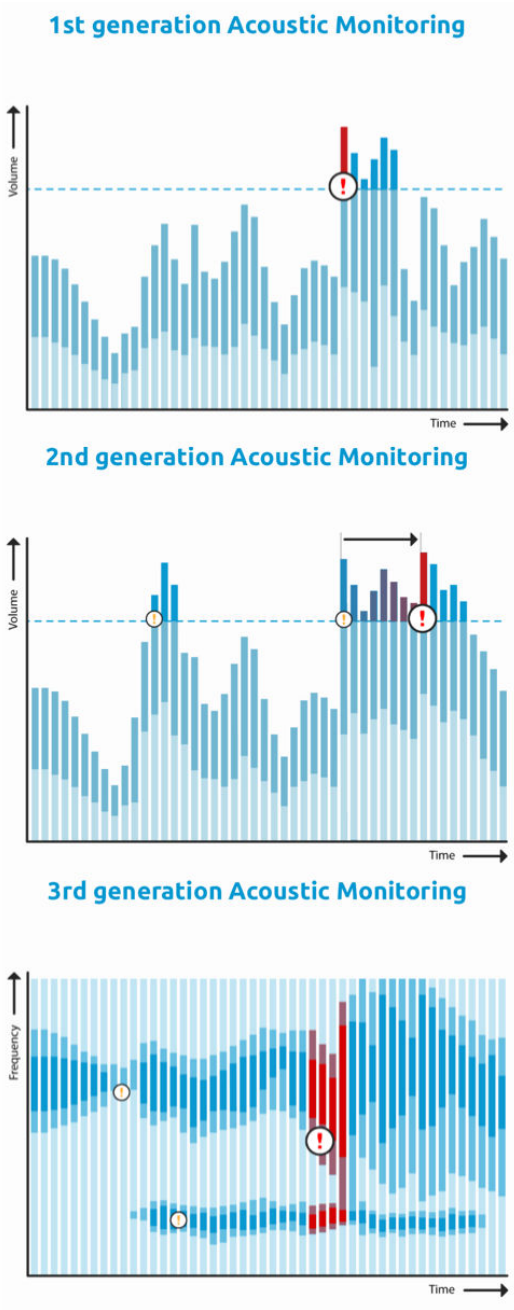


Abb. 5: Entwicklung des akustischen Monitoring[15]

Aus der nachfolgenden Abbildung gehen die Vorteile eines akustischen Überwachungssystems im Vergleich zu herkömmlichen Notfall-Systemen im Gesundheitswesen hervor. Dabei soll hervorgehoben werden, dass intelligente akustische Monitoringsysteme nicht alleinstehend sondern gerade als Ergänzung zu bestehenden Systemen als sinnvoll erachtet werden.

	Nurse Call	Acoustic Monitoring	Video Monitoring
Ability to raise alarm	=	+	−
Effect on resident's sleeping pattern	−	+	+
Effect on resident's privacy	−	+	−
Effect on operational costs	−	+	−
Provide care when and where needed	=	+	+

Abb. 6: Vorteile des akustischen Monitorings gegenüber herkömmlichen Methoden[16]

Sound Detection for Security

Ein besonderer Markt für intelligente Audiolösungen findet sich im Bereich neuartiger Sicherheits- und Alarmsysteme. Die niederländische Firma *Sound Intelligence*, gegründet im Jahr 2000, hat sich dabei auf die Entwicklung fortschrittlicher Audioanalysetechnologien spezialisiert. Zu ihrem Produktportfolio zählen Aggressionserkennung, Schusswaffenerkennung, Einbruchserkennung und Autoalarmsysteme.[17]

Die Erkennung von menschlichen Emotionen wie Angst und Aggression in der Stimme einer Person kann bei softwarebasierten Frühwarnsystemen helfen und so zur Abschreckung und Verhinderung physikalischer Gewalt beitragen. Bei einem Vorfall wird automatisch das Sicherheitspersonal

[15] https://global.clb.nl/solution/#acousticmonitoring
[16] https://global.clb.nl/solution/#acousticmonitoring
[17] Vgl.: http://www.soundintel.com/about-si/who-we-are/

benachrichtigt, welches dadurch schneller reagieren kann. Die Vorfälle sollen zudem in Echtzeit dokumentiert werden. Dies ermöglicht auch eine Überprüfung des akustischen Ereignisses im Nachhinein. Derartige Systeme lassen sich mit bereits vorhandenen Video-Überwachungslösungen kombinieren und an die jeweilige Umgebung adaptieren.[18]

Die Erkennung und Klassifizierung von Schusswaffen in Audio-Überwachungssystemen funktioniert ganz ähnlich. Das Vorfallmanagement in einem solchen Szenario übernimmt auch hier wieder das Sicherheitspersonal, welches von einem sicheren Standort aus agieren kann. Wird ein Alarm ausgelöst, werden diese unmittelbar benachrichtigt. Auch hier kann das Signal zur Verifikation nochmals überprüft und die Bedrohung besser abgeschätzt werden.[19]

Um bei Einbrüchen besser geschützt zu sein, kann auch ein Glasbruchdetektor helfen. Banken, Büros und Geschäfte mit Ladenfronten und bereits vorhandenen Sicherheitssystemen werden um einen Schallsensor erweitert, der akustische Glasbrüche erkennt und alarmiert. Die Empfindlichkeit lässt sich auf die jeweilige Umgebung einstellen.[20]

Eine ähnliche Sensorik lässt sich beispielsweise in Autohäuser und Parkplätze integrieren um gängige Autoalarme akustisch in Echtzeit zu erkennen und den Standort zu melden. Dafür muss der Detektor das spezifische Klangmuster genau identifizieren können. Die Software erkennt laut Herstellerseite Autoalarm-Tonmuster in einer Entfernung von bis zu 90 Metern. Dieser Detektor ermöglicht es dem Sicherheitspersonal so die Reaktionszeit auf Vorfälle zu verbessern.[21]

Die in diesem Abschnitt beschriebenen Sound-Erkennungssysteme lassen sich vielfältig einsetzen. Sound Intelligence schlägt hierfür beispielsweise ganz allgemein öffentliche Plätze und Fortbewegungsmittel vor, aber auch Schulen, Empfangsschalter, Verkaufshäuser, Gefängnisse, Banken sowie der Bereich Stadtüberwachung seien denkbar.[22]

Sound Detection for Offline Audio Production

Künstliche Intelligenz und Machine Learning haben mittlerweile auch in die Welt der Audioproduktion Einzug erhalten und finden sich zum Teil hinter vielen Algorithmen verschiedener Anwendungen und Plugins. Grundsätzlich muss hierbei erneut unterschieden werden, ob es sich um Echtzeit-Anwendungen handelt oder eine Tonerkennung und -interpretation im „Offlinemodus“ vorgenommen wird. Für beide Fälle gibt es mittlerweile gängige Anwendungsbeispiele.

An dieser Stelle sollen im Bereich der offline arbeitenden Audioerkennungstechnologien die Produkte der Firma i*Zotope, Inc.* genauer vorgestellt werden. Hierbei wird die moderne Audioerkennung für die Restauration von Tonmaterial eingesetzt oder auch das vorhandene Audiomaterial

[18] Vgl.: https://www.soundintel.com/products/overview/aggression/
[19] Vgl.: https://www.soundintel.com/products/overview/gun-shot/
[20] Vgl.: https://www.soundintel.com/products/overview/breaking-glass/
[21] Vgl.: https://www.soundintel.com/products/overview/car-alarm/
[22] Vgl.: https://www.soundintel.com/markets/

analysiert, um mittels KI sowie Machine Learning den Mischungsprozess zu unterstützen. Je größer die Datengrundlage, desto besser wird dabei der Algorithmus, weshalb iZotope, aber auch Apple (Siri) und Amazon (Alexa) auf diese Lernmechanismen zurückgreifen.[23]

Des weiteren existieren inzwischen verschiedene Verfahren für die Audioerkennung, welche aus den Bereichen Musikproduktion und Medientechnik bekannt sind. Hierbei handelt es sich zumeist um Algorithmen zur Erkennung impulshafter Signale, zur Tonhöhendetektion sowie deren Korrektur oder Aufzeichnung. Die Firma iZotope ist dafür bekannt, dass sie künstliche Intelligenz und Machine Learning zur Entwicklung ihrer Software einsetzt. Insbesondere deren Audiorestaurationssuite *RX* - aktuell in Version 7 am Markt verfügbar[24] - hat dabei große Bekanntheit erlangt. Zwei Bestandteile dieses Softwarepakets spielen dahingehend eine wesentliche Rolle: *RX7 Music Rebalance* und *RX7 Repair Assistant*.

Music Rebalance ist ein Programm, welches Musikstücke analysiert und die Möglichkeit bietet, eine fertige Gesamtmischung im Nachhinein zu verändern. Die Technologie beruht dabei auf tiefen neuronalen Netzwerken, wofür explizit hunderttausende Musikstücke mittels Machine Learning analysiert und in einzelne Bestandteile aufgeschlüsselt wurden, sodass diese nachträglich erfasst und verändert werden können.[25]

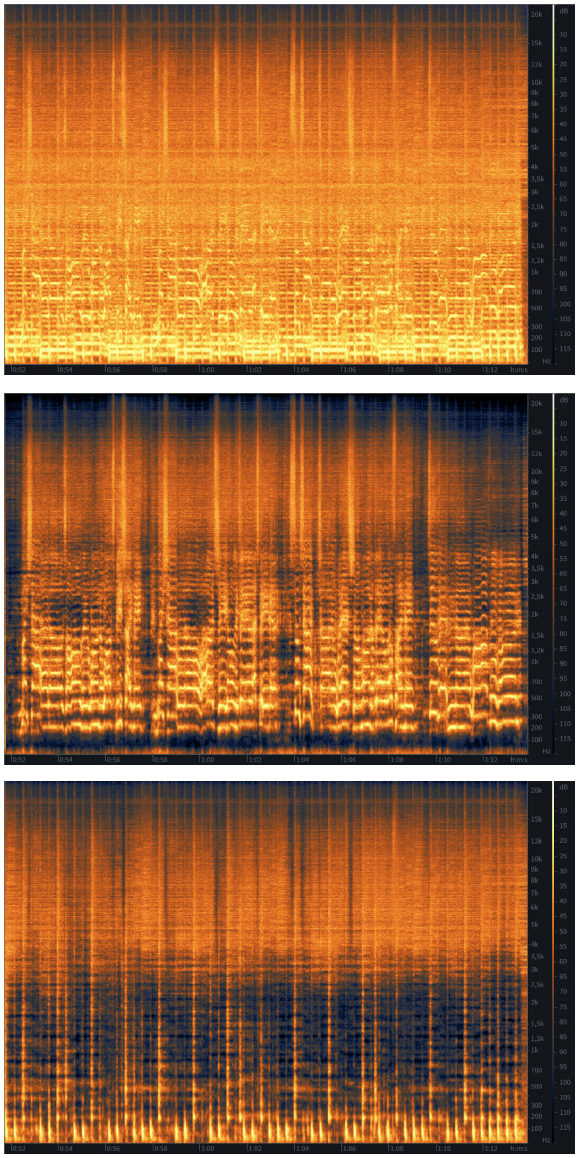


Abb. 7: Original (oben), extrahierte Vocals (mitte) und extrahierte perkussive Instrumente (unten) mit Music Rebalance

Der *Repair Assistant* bietet verschiedene Reparaturmöglichkeiten wie zum Beispiel die Entfernung unerwünschter Nebengeräusche, Klicks oder Wind auf Mikrofonaufnahmen. Das Tool arbeitet auf der Ebene des Frequenzspektrums über der Zeit. Es nutzt dafür iZotopes *Assistive Audio Technology*, welche auch hier wieder

[23] Vgl.: https://www.izotope.com/en/learn/what-the-machine-learning-in-rx-6-advanced-means-for-the-future-of-audio- repair-technology.html
[24] Stand: Juli 2020.
[25] Vgl.: https://www.izotope.com/en/learn/exploring-the-technology-that-makes-rx-7-music-rebalance-possible.html

auf neuronalen Netzen basiert, die durch Machine Learning trainiert wurden.[26] Der Hersteller selbst beschreibt diese Technologie wie folgt: „Assistive audio technology intelligently analyzes your audio and provides custom-built presets that are tailored to the sound you're trying to achieve. It combines years of intelligent digital signal processing (DSP) algorithm development with modern machine learning techniques to analyze your audio signal and make the best possible suggestions.“[27]

Auch in anderen Produkten der Firma, wie der Masteringsuite *Ozone*, oder dem Channelstrip *Neutron* kommt diese Technologie zum Einsatz. Die Idee besteht jeweils darin, durch Analyse und Eingabe diverser Parameter ein Profil oder Preset zu generieren, das den klanglichen Vorstellungen des Nutzers möglichst nahe kommt. Dadurch kann beispielsweise beim Mixing- oder Mastering einiges an Arbeit und vor allem viele Einzelschritte eingespart werden.

Sound Detection for Realtime Audio Production

Noise Cancelation muss mittlerweile nicht mehr zwangsläufig offline berechnet werden. Im August 2018 stellte der kalifornische Hersteller für Grafikprozessoren Nvidia eine neue Generation seiner „GeForce“ genannten Grafikkarten vor, die GeForce 20 Serie, deren Architektur unter dem Codenamen „Turing“ (nach Alan Turing) entwickelt wurde.[28]

Als Hauptfeatures der neuen Generation werden Raytracing in Echtzeit und ein neuronales Netz namens DLSS 2.0 (Deep Learning Super Sampling) genannt. Letzteres soll höhere Frameraten bei besserer Bildqualität ermöglichen.[29] Neben weiteren neuen Funktionen führten diese Eigenschaften dazu, dass Nvidia ein neues Kürzel für die Hochleistungskarten einführte, aus GTX wurde RTX, was für Raytracing steht.

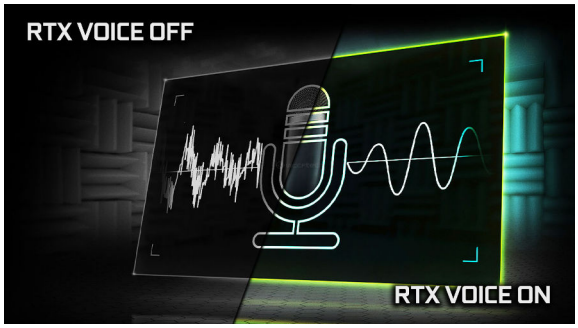


Abb. 8: RTX Voice[30]

2020 stellte Nvidia ein Plugin für seine RTX Grafikkarten vor, welches dazu dient, mittels KI per Mikrofon am Computer aufgenommene Sprache von Nebengeräuschen zu befreien. Initiale Idee war es, Tastaturklappern beim Spielen zu unterdrücken. Wie festgestellt wurde, funktioniert das System aber auch bei vielen andere Nebengeräuschen, wie z.B. Kindergeschrei, Staubsaugen, Verkehrslärm. Dabei wird das digitale Audiosignal durch die Grafikkarte geleitet und mittels AI prozessiert. Das Plugin befindet sich aktuell in der Betaphase, wird aber bereits von vielen Herstellern unterstützt (u.a. Skype, Discord, Zoom). In aktuellen Zeiten, den damit verbundenen Videokonferenzen und Meetings eine

[26] Vgl.: https://www.izotope.com/en/learn/speed-up-your-workflow-with-assistive-audio-technology.html
[27] https://www.izotope.com/en/learn/speed-up-your-workflow-with-assistive-audio-technology.html
[28] Vgl.: https://de.wikipedia.org/wiki/Nvidia-GeForce-20-Serie
[29] Vgl.: https://www.nvidia.com/de-de/geforce/news/nvidia-dlss-2-0-a-big-leap-in-ai-rendering/
[30] https://cdn.wccftech.com/wp-content/uploads/2020/04/How-To-Turn-NVIDIA-RTX-Voice-On-Featured-Image-Guide.jpg

gelungene Lösung, um die allgemeine Sprachverständlichkeit bei Auftreten von Nebengeräuschen zu verbessern.[31]

Tonerkennung bzw. Tonhöhenanalyse spielen heutzutage auch eine wichtige Rolle in der Musikproduktion, bei der Postproduktion von Gesangs- und Instrumentenaufnahmen, ebenso wie im Live-Bereich. Die entweder in Echtzeit oder nachträglich angewandte Tonhöhenkorrektur wird dabei nicht nur zur technisch unauffälligen Korrektur unsauber aufgenommener oder eingespielter Töne genutzt, sondern gerne auch als Stilmittel für Gesang in verschiedenen Genres verwendet. Auch mit Einstellungen, die nicht mehr unauffällig, sondern gezielt synthetisch klingen.

Die bekannteste Software hierfür ist *Antares Auto-Tune*, die bereits 1997 auf den Mark kam und mittlerweile auf unzähligen Aufnahmen zu hören ist. Interessanterweise wurde diese von einem Unternehmen erfunden, dessen Gründer Geowissenschaftler war, der allerdings Kenntnisse über digitales Signalprocessing und ein Interesse an Musikproduktion hatte.[32] Die Technik der digitalen Tonhöhenmanipulation ist mit einer der größten Meilensteine der jüngeren Recordinggeschichte.

Eine Weiterentwicklung davon findet sich eher im Kreativbereich: *Voice-to-MIDI*, welches hier am Beispiel von Vochlea vorgestellt werden soll. Es handelt sich dabei um ein Tool, welches gesungene oder gepfiffene Melodien direkt in Midi-Signale konvertiert und dabei die Mechanismen der Tonhöhenerkennung ausnutzt. Ebenso kann die Software auch impulshafte Signale erkennen und damit beispielsweise einen Drumcomputer auf bis zu acht Spuren triggern. So können per Beatboxing oder Plopp, Klick, und Zischlauten ganze Schlagzeugsequenzen aufgenommen werden. Der Ansatz dahinter ist, dass die eigene Stimme selbst bei Nichtinstrumentalisten das am besten trainierte Instrument ist und damit intuitiv am einfachsten zu handhaben ist.

Rein technisch betrachtet können analoge kontinuierliche Signale in ihrer Frequenz analysiert werden, wie das Beispiel eines Gitarren-Stimmgerätes zeigt. Manipulationen im analogen Bereich sind allerdings deutlich komplizierter zu realisieren, als wenn das Signal in der digitalen Domäne vorliegt. So klingt beispielsweise eine Schallplatte oder eine Tonbandaufnahme, welche schneller abgespielt wird, meist unnatürlich was die Tonhöhe angeht. Eine Veränderung der Frequenz eines Signals ohne Auswirkung auf die Tonhöhe und die Signalqualität ist nicht unproblematisch.

Während aus der eingehenden Frequenz direkt eine Information extrahiert und beispielsweise in eine MIDI-Note umgewandelt werden kann, muss das Signal für eine Tonhöhenkorrektur weiter manipuliert werden. Die dabei zugrundeliegende Technik ist ein sogenannter *Phase Vocoder* - ein Algorithmus, der auf der Fouriertransformation beruht.[33]

Das Signal wird in kleinere Segmente unterteilt und jeweils einer FFT-Analyse (Fast Fourier Transformation) unterzogen.

---

[31] Vgl.: https://www.nvidia.com/en-us/geforce/guides/nvidia-rtx-voice-setup-guide/
[32] Vgl.: https://www.mixonline.com/technology/1997-antares-auto-tune-383728
[33] Vgl.: https://sethares.engr.wisc.edu/vocoders/phasevocoder.html

Als Abbildung kann dieses Signal in seinem Spektrum manipuliert werden, was zahlreiche Audioeffekte ermöglicht.

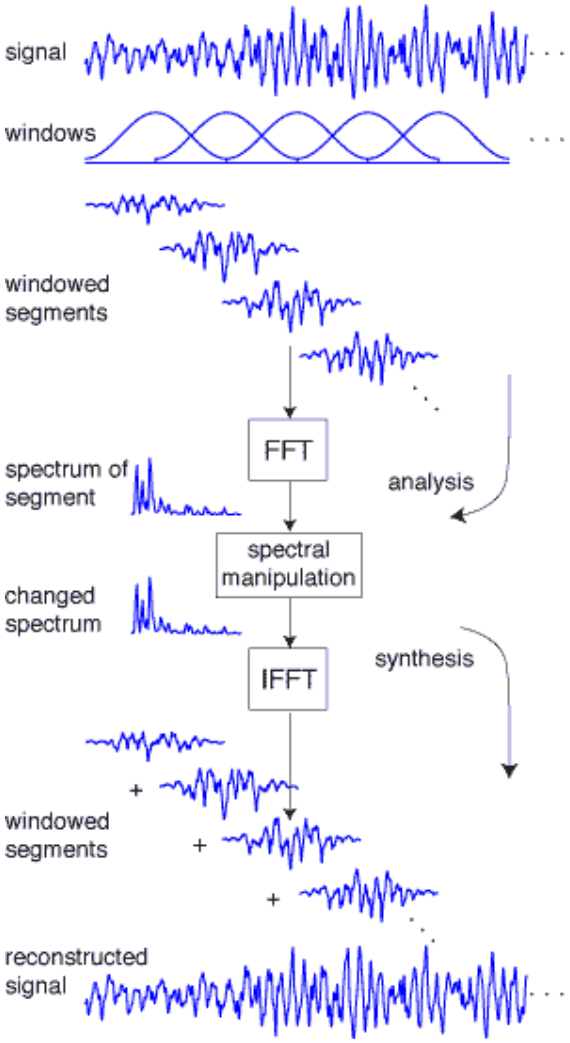


Abb. 9: Funktionsprinzip Phase Vocoder[34]

Um etwa die Tonhöhe zu manipulieren, wird das Originalsignal möglichst exakt rekonstruiert und nur in der Frequenz in gewünschtem Maße verändert. Dazu wird zunächst die Dauer des Signalfensters verändert ohne die Tonhöhe zu beeinträchtigen. Im Anschluss wird die Frequenz so angepasst, dass ein möglichst originalgetreues Signal entsteht, welches die originale Länge bei einer geänderten Frequenz aufweist.[35]

Conclusion

Im Rahmen der Lehrveranstaltung *Aktuelle Themen* unter Leitung von Prof. Dr. Andreas Koch werden jedes Semester die neuesten Forschungsergebnisse und Entwicklungen aus dem Bereich der künstlichen Intelligenz zusammengetragen. Das vorliegende Paper thematisierte in diesem Zusammenhang unterschiedliche Aspekte aus dem Bereich Audio, da moderne Machine-Learning-Technologien auch in der Tonerkennung eingesetzt werden können. Dies betrifft neben der klassischen Audioproduktion beispielsweise auch das Gesundheitswesen, sowie Sicherheits- und Überwachungssysteme. Neben den technischen Hintergründen wurden auch die Vorteile dieser intelligenten Systeme erläutert. Klar ist: Künstliche Intelligenz besitzt hier noch viel Potenzial. Hinsichtlich zukünftiger Entwicklungen darf man also zurecht gespannt sein.

[34] https://sethares.engr.wisc.edu/vocoders/phasevocoder.html
[35] http://www.physics.org/article-questions.asp?id=75

Quellen

https://de.wikipedia.org/w/index.php?title=Künstliche_Intelligenz&oldid=198129163

https://www.gruenderszene.de/lexikon/begriffe/kuenstliche-intelligenz?interstitial

https://www.zeit.de/digital/internet/2020-03/covid-19-kuenstliche-intelligenz-coronavirus-diagnose-technik

https://de.wikipedia.org/wiki/Schallereignis

https://electronics.howstuffworks.com/gadgets/high-tech-gadgets/amazon-echo1.htm

http://ceur-ws.org/Vol-2351/paper_49.pdf

https://www.bitkom.org/sites/default/files/2019-10/20191014_sof2_healthai_1.pdf

https://global.clb.nl

http://www.soundintel.com/about-si/who-we-are/

https://www.soundintel.com/products/overview/aggression/

https://www.soundintel.com/products/overview/gun-shot/

https://www.soundintel.com/products/overview/breaking-glass/

https://www.soundintel.com/products/overview/car-alarm/

https://www.soundintel.com/markets/

https://www.izotope.com/en/learn/what-the-machine-learning-in-rx-6-advanced-means-for-the-future-of-audio-repair-technology.html

https://www.izotope.com/en/learn/exploring-the-technology-that-makes-rx-7-music-rebalance-possible.html

https://www.izotope.com/en/learn/speed-up-your-workflow-with-assistive-audio-technology.html

https://de.wikipedia.org/wiki/Nvidia-GeForce-20-Serie

https://www.nvidia.com/de-de/geforce/news/nvidia-dlss-2-0-a-big-leap-in-ai-rendering/

https://www.nvidia.com/en-us/geforce/guides/nvidia-rtx-voice-setup-guide/

https://www.mixonline.com/technology/1997-antares-auto-tune-383728

https://sethares.engr.wisc.edu/vocoders/phasevocoder.html

Abbildungen

Titelbild: Eigene Abbildung

Abb. 1: https://www.technik-unterrichten.de/Robotik/Schallsensor/Bilder/Signalwandler.jpg

Abb. 2: https://www.techproviderzone.com/sites/techproviderzone/files/Amazon%20Echo's%207-microphone%20far-field%20array.png

Abb. 3: http://ceur-ws.org/Vol-2351/paper_49.pdf

Abb. 4: https://www.bitkom.org/sites/default/files/2019-10/20191014_sof2_healthai_1.pdf

Abb. 5: https://global.clb.nl/solution/#acousticmonitoring

Abb. 6: https://global.clb.nl/solution/#acousticmonitoring

Abb. 7: Eigene Abbildung

Abb. 8: https://cdn.wccftech.com/wp-content/uploads/2020/04/How-To-Turn-NVIDIA-RTX-Voice-On-Featured-Image-Guide.jpg

Abb. 9: https://sethares.engr.wisc.edu/vocoders/phasevocoder.html