SPRACHSYNTHESE

mithilfe von K.I. zur effektiven Vertonung eines Videospielcharakters

---

MASTERARBEIT

zur Erlangung des akademischen Grades
Master of Arts (M.A.)
im Masterstudium Sounddesign
an der FH JOANNEUM Graz
vorgelegt von Lukas Matthias Robausch

Betreuer: Prof. Dr. Josef Gründler
Graz, 2019

EIDESSTATTLICHE

Erklärung

„Ich erkläre ehrenwörtlich, dass ich die vorliegende Masterarbeit selbstständig angefertigt und die mit ihr verbundenen Tätigkeiten selbst erbracht habe. Ich erkläre weiters, dass ich keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle aus gedruckten, ungedruckten oder dem Internet im Wortlaut oder im wesentlichen Inhalt übernommenen Formulierungen und Konzepte sind gemäß den Regeln für gutes wissenschaftliches Arbeiten zitiert und durch Fußnoten bzw. durch andere genaue Quellenangaben gekennzeichnet.

Die vorliegende Originalarbeit ist in dieser Form zur Erreichung eines akademischen Grades noch keiner anderen Hochschule vorgelegt worden. Diese Arbeit wurde in gedruckter und elektronischer Form abgegeben. Ich bestätige, dass der Inhalt der digitalen Version vollständig mit der gedruckten Version übereinstimmt. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben kann.“

ABSTRACT

ENG

Artificial intelligence is on the run, whether it evaluates automatic calls and forwards them or hides in smartphones like Siri, Google or Alexa. These technologies try to imitate original human speech and intonation getting audible by speech synthesis. They handle complex questionnaires as well as colloquial words. They have their specific kind of emphasis and fluency. In this thesis, I will explore their properties and whether this type of speech synthesis is effectively suitable for replacing a voice actor. The resulting material and its work process are then evaluated for quality and effectiveness. Finally, it is tested whether one of these methods is suitable for creating voice lines and sound effects for a video game character which defines my practical part.

GER

Die künstlichen Intelligenzen sind auf dem Vormarsch, ob es nun diese sind die Fragen in einer Anrufschleife auswerten und weiterleiten können oder welche die sich in unseren Smartphones verstecken wie Siri, Google oder Alexa. Diese versuchen den normalen Sprachfluss und Intonation der normalen menschlichen Sprache zu imitieren, welche durch die Sprachsynthese hörbar gemacht wird. Sei es eine kompliziertere Fragestellung oder umgangssprachliche Wörter. Sie weisen somit eine eigene Art von Betonung und Sprachfluss auf. In dieser Arbeit werde ich deren Eigenschaften erforschen und ob diese Art der Sprachsynthese sich effektiv dazu eignet einen Synchronsprecher zu ersetzen. Das daraus resultierende Material und dessen Arbeitsprozess wird danach auf Qualität und Effektivität evaluiert. Zuletzt wird getestet ob sich eine dieser Methoden für die Vertonung eines Videospielcharakters meines praktischen Teils eignet.

DANKSAGUNG

Mein besonderer Dank gilt meiner Familie, insbesondere meinen Eltern, die mir mein Studium ermöglicht, die Arbeit korrekturgelesen und mich in all meinen Entscheidungen unterstützt haben.

Herzlich bedanken möchte ich mich auch bei meiner Freundin, der mich immer wieder ermutigte und mit vielen nützlichen Tipps einen wesentlichen Teil zur Masterarbeit beigetragen hat.

Schließlich danke ich meinen Freunden während der Studienzeit für die schöne Zeit im auf der Fachhochschule Joanneum und meinem Betreuer Josef Gründler für die wunderbare Betreuung und die mit mir vermittelte Expertise.

INHALTS VERZEICHNIS

1. EINLEITUNG

Das Gebiet der KI-Forschung wurde 1956 in einem Workshop am Dartmouth College in New Hampshire geboren. Es wird eine Methode geschaffen die es einem Menschen gleichtut, und die komplexen Abläufe vom Handeln, Denken und Entscheiden nachzustellen und diese zu verstehen. Allen Newell, Herbert Simon, John McCarthy, Marvin Minsky und Arthur Samuel (IBM) waren die Gründer und Leiter der KI-Forschung. Sie lernten Schachstrategien und spielten Berichten zufolge 1959 besser als der Durchschnittsmensch[1]. Probleme in der Algebra, Beweisen von logischen Theoremen und Sprechen von Englisch, dies sind Beispiele für Anwendungsgebiete für. künstliche Intelligenz.

Das US-Verteidigungsministerium begann Mitte der 1960er Jahre großes Interesse zu zeigen und richtete weltweit Laboratorien ein. Sie blicken optimistisch in die Zukunft: Herbert Simon sagte voraus, „Maschinen werden in zwanzig Jahren in der Lage sein, jede Arbeit zu verrichten, die der Mensch leisten kann“. Marvin Minsky stimmte zu und schrieb: „Innerhalb einer Generation ... wird das Problem der Schaffung ‚künstlicher Intelligenz‘ im Wesentlichen gelöst sein.“[4]

Heutzutage haben 57,6% der Deutschen[2] zumindest einmal von künstlicher Intelligenz wie dem Sprachassistenten „Alexa“ von Amazon oder Google „Assistant“ gehört[3]. Wenn Sprache so effektiv wiedergegeben werden kann, ist dies ein Ansatzpunkt für einschlägige Experimente. Den Synchronsprecher zu ersetzen oder die Sprachausgabe geregelt zu manipulieren indem man gewisse Regeln voranstellt ist eine Methode. Diese Masterarbeit behandelt im ersten Teil eine Übersicht und diverse Experimente zu gängigen Methoden mit Sprachausgabe und Training eines neuralen Netzwerkes. Ebenso werden in der Schlussbetrachtung die Ergebnisse diskutiert. Im zweiten, praktischen, Teil wird aufgrund der Erkenntnisse des ersten Teils, eine Sprache für einen Protagonisten eines Videospiels erschaffen. Die Arbeitsweise für diese Vertonung zielt auf das bestmögliche Ergebnis ab, welche aus den Forschungsergebnissen abgeleitet wird.

1) Vgl. Dreyfus, Hubert L.: Die Grenzen künstlicher Intelligenz. Was Computer nicht können. Athenäum, Königstein 1985
2) Statista, und Norstat. „In immer mehr Bereichen des Lebens spielen digitale Sprachassistenten eine Rolle. Welche dieser Sprachassistenten kennen Sie?.“ Chart. 29. März, 2017 https://de.statista.com/statistik/daten/studie/739040/umfrage/umfrage-zur-bekanntheit-ausgewaehlter-sprachassistenten-in-deutschland/ (zuletzt aufgerufen am 09.7.2019)
3) Google Assistant: What it can do. In: https://assistant.google.com/learn/ (zuletzt aufgerufen am 09.7.2019)
4) Vgl. Wikipedia. Die freie Enzyklopädie (14.11.2012), s.v. Bibliothek, https://en.wikipedia.org/wiki/Artificial_intelligence

2 THEORIE

In diesem Teil werden die Basiselemente der nachstehenden Forschung erläutert um die Teilfragen und Hypothesen aufzustellen. Im ersten Abschnitt wird die Künstliche Intelligenz an sich behandelt und wie diese in unserer heutigen Zeit definiert wird. Im zweiten Abschnitt wird die Sprache beziehungsweise Sprachsynthese genauer erklärt und welche Parameter vorhanden sein müssen um damit effektiv kommunizieren zu können. Unabhängig davon, ob es sich um einen Menschen oder eine Maschine handelt, welche der Sender oder Empfänger ist.

2...1 KÜNSTLICHE Intelligenz

Künstliche Intelligenz wird auch artifizielle Intelligenz genannt oder „K.I." abgekürzt und kommt aus dem Gebiet der Informatik. Im Englischen auch Artificial Intelligence („AI") genannt. Dieser Teilbereich konzentriert sich auf die Automatisierung von intelligentem Verhalten. Maschinelles Lernen ist ein wichtiger Teil davon, um menschliches Verhalten in irgendeiner Weise zu imitieren[4].

Laut Hubert L. Dreyfus[1] wird zwischen einer schwachen und einer starken künstlichen Intelligenz unterschieden. Diese „starke" Intelligenz wäre eine der kognitive Wahrnehmung vorangestellte. Ein System welches jedoch eine nur menschenähnliche Denkweise besitzt. Dieses System entscheidet aufgrund von Erfahrungen und Eindrücken. Es soll Gefühle wie Liebe, Angst und Hass nicht besitzen, sie jedoch simulieren können. Diese Art zu erschaffen ist aber bis jetzt erfolglos geblieben.
Im Gegenzug dazu ist die schwache künstliche Intelligenz die Art die sich in den letzten Jahren stark weiterentwickelt hat. Hierbei wird Fokus auf der Unterstützung des menschlichen Denkens gelegt, es wird ein intelligentes Verhalten simuliert welches sich auf Mathematik und Informatik stützt und auf schon vorhandene Fälle die zu einem bestimmten Ergebnis führen[5].

Eine Unterordnung im Gebiet der K.I. ist Machine Learning welche eine Sammlung von mathematischen Methoden der Mustererkennung. darstellt. Diese Methoden erkennen Muster beispielsweise durch bestmögliche Zerlegung von Datenbeständen in hierarchische Strukturen[4]. Ein Teilgebiet von Machine Learning ist wiederrum „Deep Learning", eine Disziplin des maschinellen Lernes unter Einsatz von künstlichen neuronalen Netzen. Während die Ideen für Entscheidungsoptionen aus einer gewissen mathematischen Logik heraus entwickelt wurden, gibt es für künstliche neuronale Netze ein Vorbild aus der Natur: Biologische neuronale Netze[35].

1) Vgl. Dreyfus, Hubert L.: Die Grenzen künstlicher Intelligenz. Was Computer nicht können. Athenäum, Königstein 1985
4) Vgl. Wikipedia. Die freie Enzyklopädie (14.11.2012), s.v. Bibliothek, https://de.wikipedia.org/wiki/Künstliche_Intelligenz (zuletzt aufgerufen am 09.7.2019)
5) Vgl. Bostrom, Nick: Superintelligenz. Szenarien einer kommenden Revolution. Suhrkamp, Frankfurt am Main. 2016
35) Vgl. Aunkofer, Benjamin: Data Science Blog: Machine Learning vs. Deep Learning, https://data-science-blog.com/blog/2018/05/14/machine-learning-vs-deep-learning-wo-liegt-der-unterschied/ (zuletzt aufgerufen am 18.8.2019)

2.2 SPRACHE

Im Allgemeinen

Sprache
gilt als die Summe aller Elemente von Lauten und Schriftzeichen. Es gibt konstruierte Sprachen wie diverse Programmiersprachen zum Beispiel C, C++ oder Python, oder Sprachen natürlichen Ursprungs wie Deutsch oder Englisch. Was Letztere ausmacht sind zusätzliche Gestik, Mimik oder auch der Tonfall zur zusätzlichen Modulation der Kommunikation[6]. Ebenso gibt es hier eine Abgrenzung zwischen jenen Sprachen, die durch natürliche Weise entstanden sind und sich über Jahrhunderte entwickelt haben und denen, die bewusst ausgearbeitet und konstruiert wurden, wie Elbisch oder Klingonisch.

Sprache dient nicht nur dazu um Informationen zu überliefern sondern auch als Schlüssel zum Selbst- und Weltverständnis. Edward Sapir bezeichnete 1921 die Sprache als eine menscheneigene und nicht im Instinkt verankerte Methode zur Nachrichtenübermittlung. Im Tierreich wird diese zum Beispiel durch Körpersprache, Laute, Farbgebung oder Düfte ersetzt[7]. Es besteht aber im Tierreich keine neuronales Defizit welches eine Voraussetzung für eine Sprache nicht erlauben würde. Jedoch wurden beim Menschen ein Satz von vier sprachbezogenen Genen identifiziert welcher für die Stimmkontrolle eine Rolle spielt[8].

Die Sprachsynthese
welche für die Sprachausgabe in unseren Smartphones oder Smart-Home-Assistenten verantwortlich ist, können wir nur verstehen weil wir die Sprache die sie sprechen erlernt haben. Diese wurde ihr wiederum von uns angelernt. Sie bedienen sich großen Datenmengen an gesprochen Texten, welche verknüpft mit dem dazugehörigen Texten analysiert werden. Der Klang bestimmter Buchstabenkombinationen und Wörtern welche in einer bestimmten Reihenfolge gesprochen[9], bildet eine erweiterte Variation eines schwachen künstlichen Systems wie es bereits in 2.1.1 erklärt wurde.

6) Vgl. Lewandowski, Theodor: Linguistisches Wörterbuch. 4., neu bearbeitete Aufl. Quelle & Meyer, Heidelberg 1985
7) Vgl. Sapir, Edward: Language. An Introduction to the Study of Speech. Harcourt Brace, New York, 1921
8) Vgl. C. Lalueza-Fox u.a: The derived FOXP2 variant of modern humans was shared with Neandertals. In: Curr Biol. 17(21), 2007
9) Vgl. Grüner, Sebastian: Tacotron 2: Googles Sprachsynthese erreicht fast menschliche Qualität - Golem.de. In: golem.de. 21. Dezember 2017 (zuletzt aufgerufen am 09.7.2019)

2.3 TEILFRAGEN

Die Teilfragen die sich stellen:

- Kann mit K.I. ein Charakter in einem Videospiel oder Film vertonent werden?
- Reichen die technischen Mittel eines Sounddesigners dazu aus?
- Kann die eigene Stimme als Ausgangsmaterial verwendet werden?
- Inwieweit muss das Ergebnis noch bearbeitet werden um ein befriedigenden Ergebnis zu erhalten?
- Welche/s Wissen/Software muss erlangt werden um effektiv arbeiten zu können?
- Kann ein solches System in ein Computerspiel eingebunden werden um effektiv Charaktere anhand derer Sprach-Skripts zu vertonen?

2.4 HYPOTHESEN

Resultierende Annahmen:

- Es wird eine Menge Zeit in Anspruch nehmen sich das Know-How über neuronale Netzwerke und die damit in Verbindung stehende Software anzueignen.
- Würde ein Charakter mit robotischen Spracheigenschaften vertont werden, wäre ein befriedigendes Ergebnis schneller zu erreichen.
- Es wird eine Variation an Ergebnissen, auch künstlerischer Natur, mit dem erlangten Wissen erhalten werden.
- Text-to-Speech-Systeme können schnell eine Vielzahl an Lösungen anbieten.

3 METHODIK

Durch Testen an verschiedenen gängigen künstlichen Intelligenzen werden Sprachversuche durchgeführt um mit diversen Ausgangsdaten eine Vielzahl an Ergebnissen zu erhalten. Mit der Fütterung von neuralen Netzwerken sollen Vergleiche zu schon existierenden Sprachsynthesesystemen geschaffen werden um dann das beste Ergebnis zu küren und zu verwenden. Allen Methoden wurde ein Testskript zugrunde gelegt um die Sprache zu synthetisieren. Ebenfalls wird das Ergebnis mit der von Google veröffentlichten Grafik (Abb. 0) verglichen.

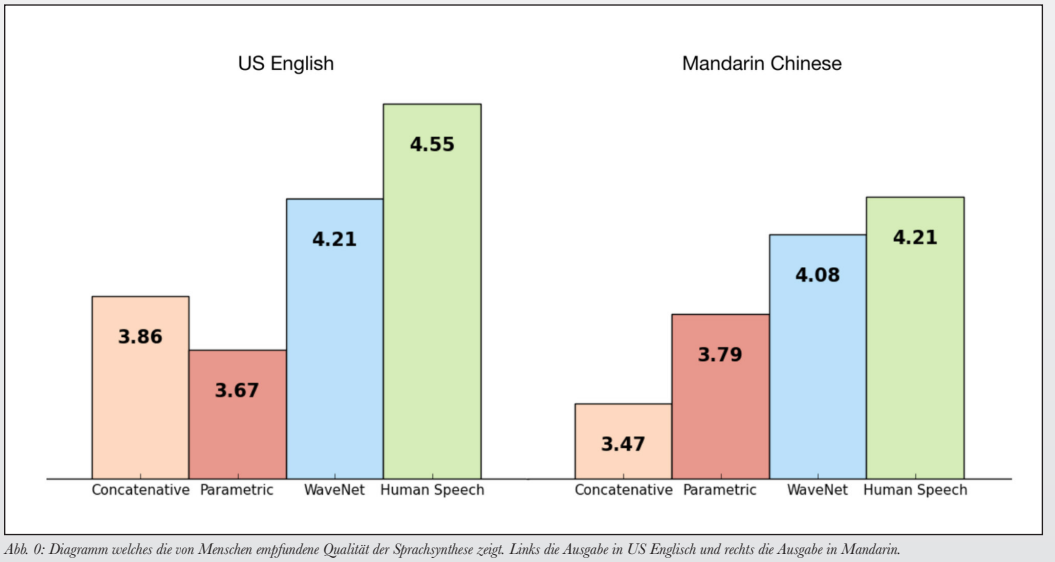

Abb. 0: Diagramm welches die von Menschen empfundene Qualität der Sprachsynthese zeigt. Links die Ausgabe in US Englisch und rechts die Ausgabe in Mandarin.

3.1 CONCATENATIVE

Text-to-Speech

Ein Großteil der Sprache die künstlich erzeugt wird, entsteht immer noch durch sogenanntes „concatenative TTS“, welches frei übersetzt „verketteter Text zu Sprache“ bedeutet. Hier wird eine große Datenbank von einem einzigen Sprecher angelegt in welcher kurze Sprach-Fragmente zu finden sind. Diese sind ebenso einer bestimmten Textbasis zugeordnet. Es entsteht damit sehr wenig Manipulation der Signale bei ihrer Verknüpfung und es ist somit eine höherer Grad an Natürlichkeit der Synthese gegeben[10].

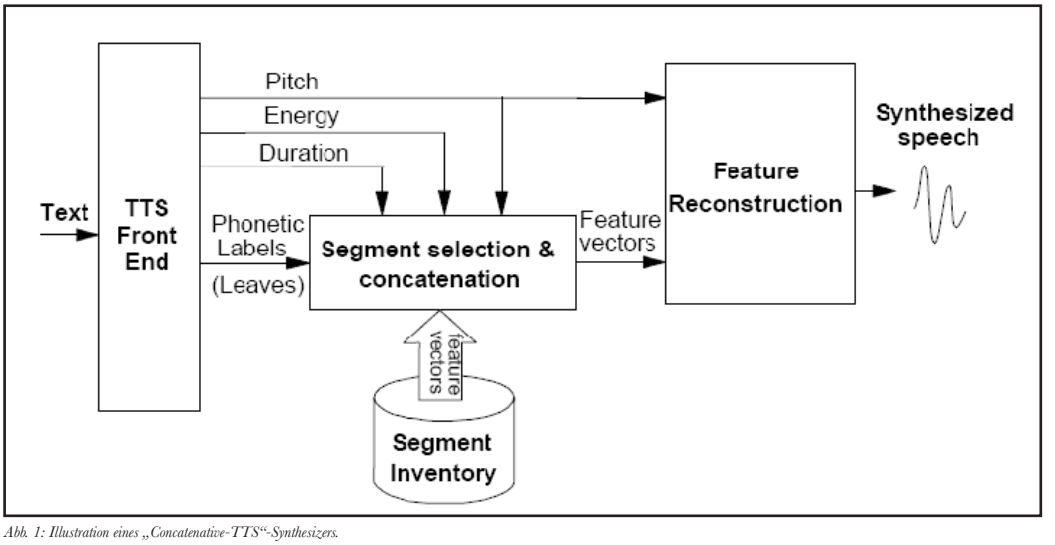

Abb. 1: Illustration eines „Concatenative-TTS“-Synthesizers.

10) Reichel, Uwe: Unit-Selection-Synthese. Datengetriebenes Vorgehen vs. Signalmanipulation. Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München, München 2017

3.2 PARAMETRIC

Text-to-Speech

Im Gegensatz dazu gibt es den neueren Ansatz namens „parametric TTS“, also parametrische Text-zu-Sprache-Methode. Hier werden die ganzen Informationen um Sprache zu generieren in den jeweiligen Parametern des Modells gespeichert[10]. Voraussetzung ist, dass die aus einer Textvorverarbeitung erhaltenen, symbolischen Phonemfolgen eine statistische Modellierung durchlaufen, indem sie zuerst in Einzelsegmente zerlegt und jedem dieser Segmente sodann ein bestimmtes Modell aus einer bestehenden Datenbank zugeordnet wird. Somit können auch die einzelnen Aspekte und Feinheiten der Sprache über die Eingabe kontrolliert werden.

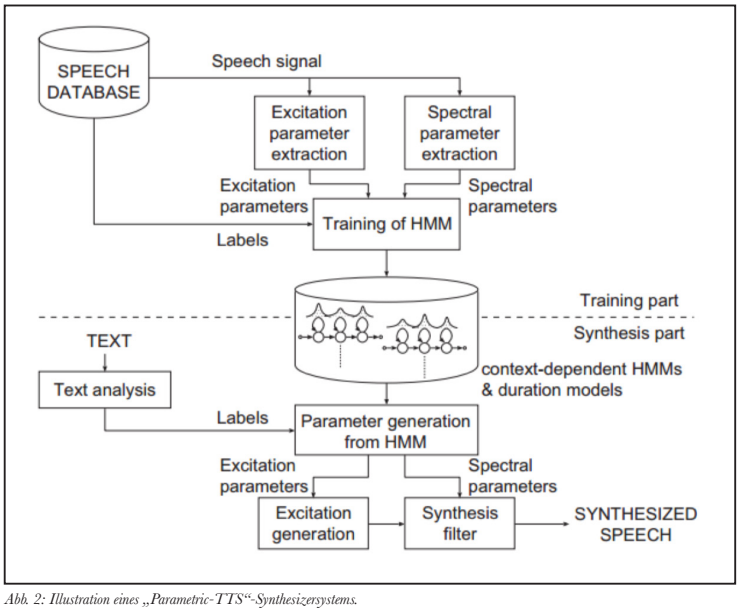

Abb. 2: Illustration eines „Parametric-TTS“-Synthesizersystems.

10) Reichel, Uwe: Unit-Selection-Synthese. Datengetriebenes Vorgehen vs. Signalmanipulation. Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München, München 2017

3.3 WAVENET

Neural Network (Google)

Google´s WaveNet hat hier einen anderen Ansatz, indem diese Methode direkt die unbearbeitete Wellenform des Audiosignal modelliert[11], jedes Sample wird einzeln analysiert und mit seinem Vorgänger und Nachfolger verglichen. Damit dieser Prozess aber funktioniert müssen ebenso zusätzlich noch Textdateien zur Sprache eingearbeitet werden. Dies passiert indem der Text in eine Reihe phonetischer und linguistischer Eigenschaften eingeteilt wird. Somit errechnet WaveNet nicht nur den nächsten logischsten Schritt bei der Berechnung sondern bezieht es auch auf den Text den wir ausgegeben haben möchten. Zusätzlich zur Ausgabe von natürlich klingendere Sprachsynthese kann diese ebenso auf alle Arten von Audio angewandt werden, wie zuvor schon erwähnt auf Musik und Samples.

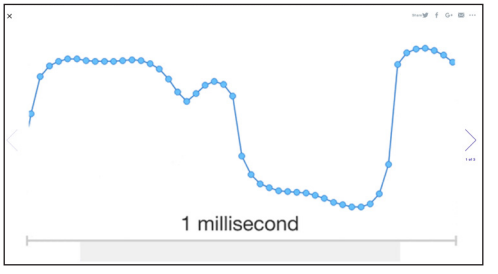

Abb. 3: Detailansicht der zu berechnbaren Punkte eines Signals.

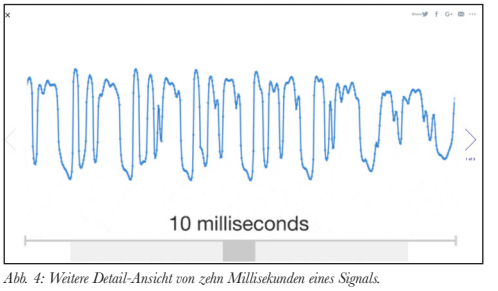

Abb. 4: Weitere Detail-Ansicht von zehn Millisekunden eines Signals.

11) Vgl. Örnek, Evin Pinar: medium. 31.1.2018, https://medium.com/@evinpinar/wavenet-implementation-and-experiments-2d2ee57105d5 (zuletzt aufgerufen am 17.7.2019)

3.4 WAVENET

Neural Network (Customized)

Hier wird mit einem unveränderten neuralen Netzwerk wie Google es benützt Versuche durchgeführt. Das WaveNet verwendet ein autoregressives Lernen mit Hilfe von Faltungsnetzwerken. Grundsätzlich gibt es ein Faltungsfenster, das auf den Audiodaten gleitet und versucht, bei jedem Schritt, den nächsten Abtastwert vorherzusagen den es noch nicht gesehen hat. Mit anderen Worten, es wird ein Netzwerk aufgebaut, das die kausalen Beziehungen zwischen aufeinander folgenden Zeitschritten lernt. Ohne zusätzliche Regeln und Algorithmen wie bei Googles WaveNet, wird hier nicht Wert auf vordefinierte Parameter oder Anhaltspunkte gelegt sondern mit den Standardeinstellungen des Netzwerkes ein eigenes Ergebnis errechnet. Dieses variiert je nach Datenmenge und Genauigkeit der Ausgabe. Allein für ein paar Sekunden an Sprachaufnahmen mit einer Samplerate von 22k oder 16k gibt es für eine einzelne Datei über 100.000 verschiedene Werte[11].

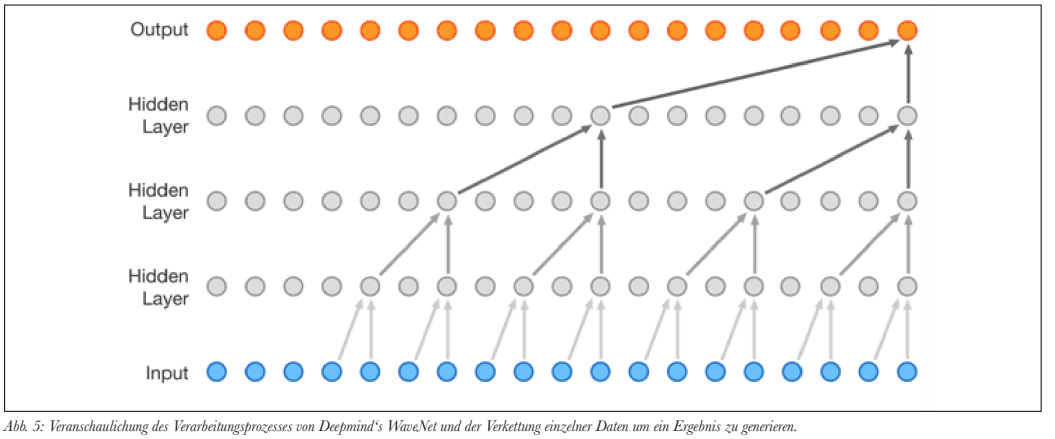

Abb. 5: Veranschaulichung des Verarbeitungsprozesses von Deepmind's WaveNet und der Verkettung einzelner Daten um ein Ergebnis zu generieren.

11) Vgl. Örnek, Evin Pinar: medium. 31.1.2018, https://medium.com/@evinpinar/wavenet-implementation-and-experiments-2d2ee57105d5 (zuletzt aufgerufen am 17.7.2019)

3.4.1 LERNARTEN

„Überwachtes“ gegen „Freies“ Lernen

Um die Arbeitsweise von neuralen Netzwerken besser zu verstehen um so auch den Punkt 3.4 besser Testen zu können muss die „Lernart“ des benutzen Netzwerks verstanden werden. Im Bereich des maschinellen Lernens gibt es zwei Haupttypen von Methoden: „Überwacht“ und „Frei“. Der entscheidende Unterschied zwischen den beiden Typen besteht darin, dass wir vorher wissen, wie die Ausgabewerte für unsere Proben sein sollten. Daher ist das Ziel des überwachten Lernens eine Funktion, die unter Berücksichtigung einer Stichprobe von Daten und gewünschten Ergebnissen die in den Daten beobachtbare Beziehung zwischen Eingabe und Ausgabe am besten erreicht. Unüberwachtes Lernen hat hingegen keine gekennzeichneten Ergebnisse. Ziel ist es daher, auf die natürliche Struktur innerhalb eines Satzes von Datenpunkten zu schließen[12].

Überwachtes Lernen wird normalerweise im Kontext der Klassifizierung durchgeführt, wenn Eingaben auf intelligent kategorisiert, oder wenn Regressionen auf eine kontinuierliche Ausgabe abgebildet werden sollen. Zu den gängigen Algorithmen beim überwachten Lernen gehören der Bayes-Klassifikator, Vektorunterstützungsmaschinen und künstliche neuronale Netze. Sowohl bei der Regression als auch bei der Klassifizierung besteht das Ziel darin, bestimmte Beziehungen oder Strukturen in den Eingabedaten zu finden, die es ermöglichen, korrekte Ausgabedaten zu erzeugen. Es muss beachtet werden, dass die „korrekte“ Ausgabe vollständig aus den Trainingsdaten bestimmt wird, es existiert also eine fundamentale „Wahrheit“, von der unser Modell annimmt, dass sie das perfekte Ergebnis ist.

12) Vgl. Soni, Devin: Towards: Data Science. 22.3.2018, https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d (zuletzt aufgerufen am 19.8.2019, Übersetzung des Verfassers)

Freies Lernen

ist am häufigsten präsent bei Clustering, Repräsentationen und Dichteschätzung. In all diesen Fällen wird die Struktur unserer Daten erlernt, ohne explizite Vorgaben zu verwenden. Einige gebräuchliche Algorithmen umfassen k-Means-Clustering, Hauptkomponentenanalyse und Autoencoder. Da keine Vorgaben bereitgestellt werden, gibt es bei den meisten unbeaufsichtigten Lernmethoden keine spezifische Möglichkeit, die Modellleistung zu vergleichen. Zwei häufige Anwendungsfälle für freies Lernen sind explorative Analysen und Dimensionsreduktion.

Unbeaufsichtigtes Lernen ist bei der explorativen Analyse sehr nützlich, da es automatisch die Struktur in den Daten erkennen kann. Wenn ein Analyst beispielsweise versuchen würde, Verbraucher zu segmentieren, wären freie Clustering-Methoden ein guter Ausgangspunkt für ihre Analyse. In Situationen, in denen das Ergebnis nicht zufriedenstellend ist, kann jedoch nicht gezielt in den Prozess eingegriffen werden.

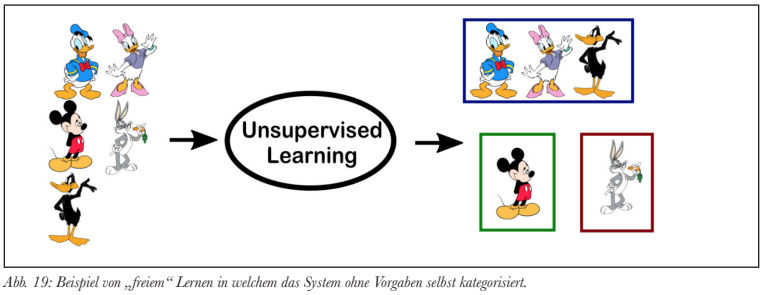

Abb. 19: Beispiel von „freiem" Lernen in welchem das System ohne Vorgaben selbst kategorisiert.

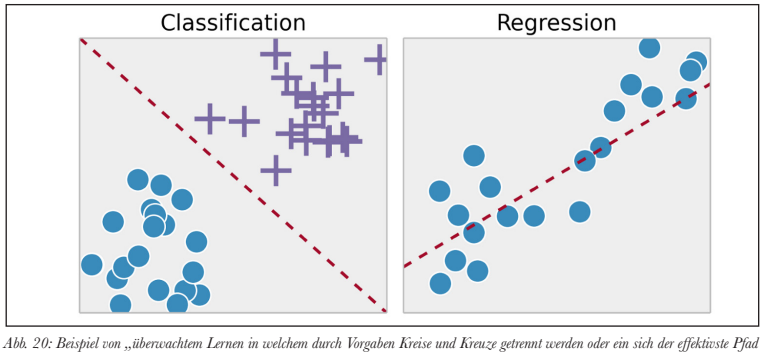

Abb. 20: Beispiel von „überwachtem Lernen in welchem durch Vorgaben Kreise und Kreuze getrennt werden oder ein sich der effektivste Pfad abbilden soll.

4 ERGEBNISSE

der Testreihen

In diesem Abschnitt erläutert der Verfasser dieser Arbeit die einzelnen Ergebnisse der erarbeiteten Methoden. Es wird die Sprachsynthese von diversen Text-to-Speech Systemen behandelt als auch der Umgang mit neuralen Netzwerken. Zuletzt werden noch die vorangegangenen Teilfragen und Hypothesen beantwortet und ausgewertet. Es wird der Arbeit ein Datenträger mit aufgezeichneten Audiodateien beigelegt welche diese Dokumentation fundieren und mit den Methoden erarbeitet wurden.

4.1 CONCATENATIVE

Text-to-Speech

Die Ergebnisse dieser Methode hören sich zu einem hohen Grad sehr natürlich an.
Es wird hier ein eingegebener Text von einer Sprachsynthesesoftware vom klassischen Google-Übersetzer ohne adjustierbare Parameter ausgegeben. Jedoch kann man in den einzelnen Wörtern, Silben und Phonemen Ungereimtheiten ausmachen, die diese sofort als Sprachsynthese entlarven. Ebenso hat man hier keinen Einfluss auf die Modulation der Sprache um Emotionen, Sprachgeschwindigkeit oder den Tonfall zum Ausdruck zu bringen. Folglich ist der Aufwand um akustisches Ausgangs-Material zu sammeln und aufzubereiten sehr hoch.

Abb. 6: Das Interface zur Textein- und Ausgabe vom klassischen Google-Übersetzer-Tool welches sich der verketteten Sprachsynthese bedient.

4.2 PARAMETRIC

Text-to-Speech

Die Sprachsynthese dieser Version hört sich meist metallischer und synthetischer an als die der ersteren „verketteten“ Methode. Von der Seite des Sprachflusses und der Charakteristika des Ausgangsmaterials des Sprechers kann diese Arbeitsweise dennoch das beste Ergebnis aufweisen. Nach 20 gesprochenen Beispielen war eine deutliche Verbesserung der Sprachsynthese zu hören. In diesem Test wurden manuell Textabschnitte eingelesen und diese mit dem dazugehörigen Text verknüpft. Nach über 130 gesprochenen Sätzen verbessert sich die Qualität der Synthese jedoch nur mehr in kleinen und feineren Schritten. Der abnehmende Ertrag der Ergebnisse spricht nicht für den noch unbekannten zu leistenden Arbeitsaufwand um ein eventuell akzeptables Ergebnis zu erzielen.

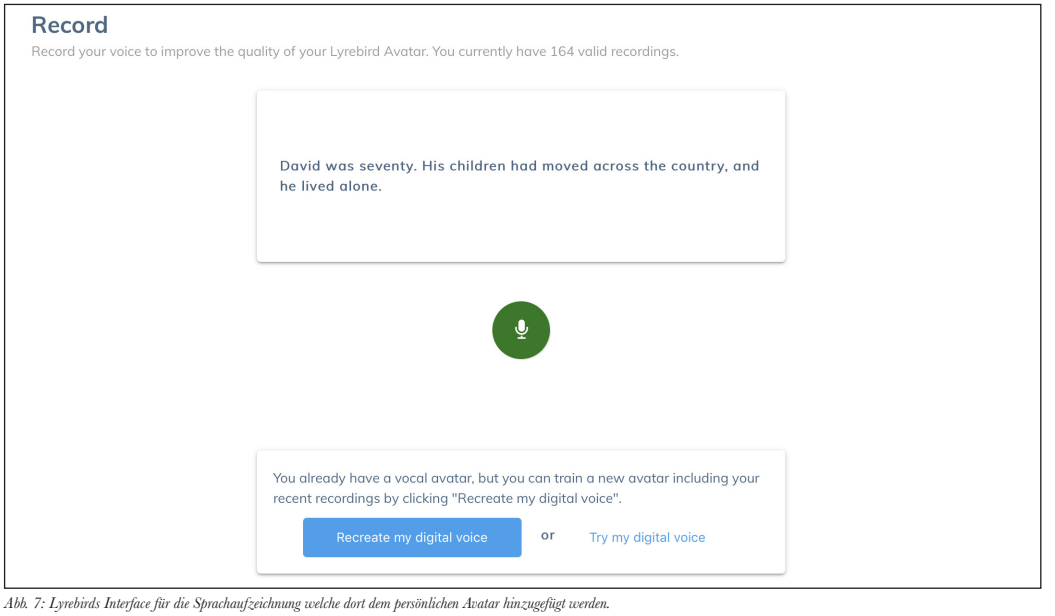

Abb. 7: Lyrebirds Interface für die Sprachaufzeichnung welche dort dem persönlichen Avatar hinzugefügt werden.

4.3 WAVENET

Neural Network (Google)

Hier wird wie im Beispiel 4.1.1 ein Text in eine Textfeld eingesetzt welches im Anschluss mithilfe von Sprachsynthese ausgegeben wird. Es gibt hier noch zusätzliche Parameter wie die Auswahl aus vier verschiedenen englischen Sprechern, Tempo und Tonhöhe. Die einzelnen Parameter für Tempo oder Tonhöhe können nur minimal verändert werden ohne eine zu hohe digitale Verzerrung oder Artefakte hervorzurufen welche das Resultat unbrauchbar machen. Die Synthese an sich hört sich wie eine Mischung aus den Tests von 4.1.1 und 4.1.2 an. Es wird eine minimierte Anzahl an Ungereimtheiten zwischen den einzelnen Silben hörbar und eine Reduktion des metallischen Effekts, welcher sich über die Ausgabe legt. Hier bleibt auch wie schon in den vorangegangenen Tests, die nicht vorhandene Kontrolle über Intonation, Sprachtempo und Emotionen zu bemängeln, welche die Ergebnisse für den zu vertonenden Charakter brauchbar machen würden.

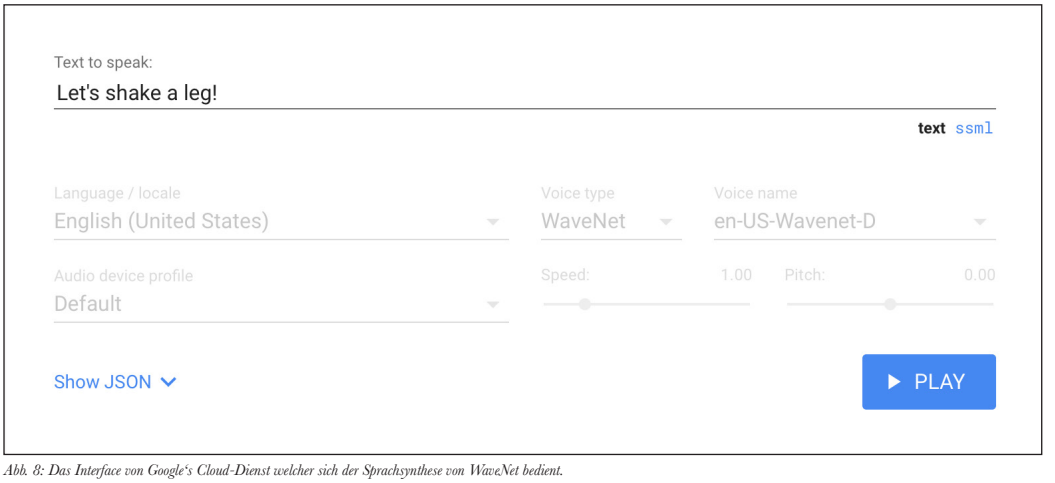

Abb. 8: Das Interface von Google's Cloud-Dienst welcher sich der Sprachsynthese von WaveNet bedient.

4.4 WAVENET

Neural Network (Customized)

Hier bestätigt sich nun die Hypothese wie in 2.3 beschrieben. Das Fachwissen um ein solches Netzwerk mit Daten zu füttern und die Ausgabe zu kontrollieren benötigt ein fundiertes Wissen in Programmiersprachen wie Python. Eine vertrautheit mit einschlägigen Codezeilen im Bereich neuraler Netzwerke trägt stark zum Verständnis bei. Die Aneignung dieses Wissens übersteigt die zur Verfügung stehenden zeitlichen Ressourcen für diese Masterarbeit.

Die Differenz

zwischen „überwachtem“ und „unüberwachtem“ Lernen miteinzubeziehen ist essentiell, welches im nächsten Abschnitt erklärt wird. Des Weiteren ist ein neurales Netzwerk ohne Vorkenntnisse unberechenbar und die Ausgabe oder die Ähnlichkeiten die dieses erkennt können stark von dem Erwarteten abweichen. Das unberechenbare Ergebnis rechtfertigt nicht den dafür notwendigen Arbeitsaufwand und ist somit für die Vertonung eines Videospielcharakters ungeeignet.

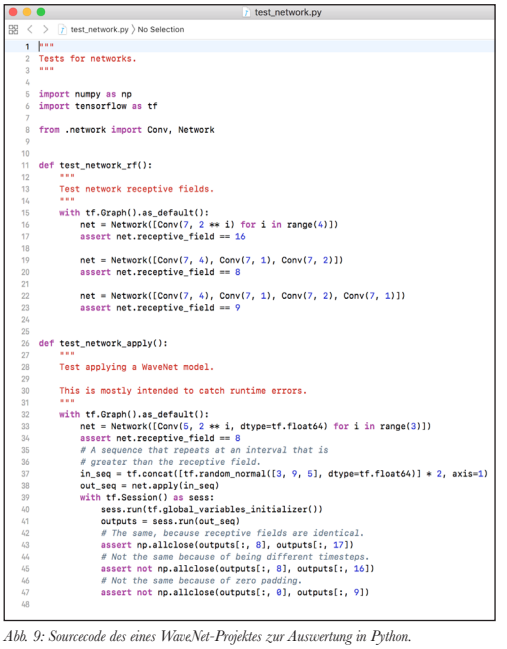

Abb. 9: Sourcecode des eines WaveNet-Projektes zur Auswertung in Python.

4.5 ZUSAMMEN

Fassung

Der zu vertonende Charakter ist ein organisches Wesen welches sich nicht mit einer künstlichen oder synthetischen Lebensform identifiziert. Der Fokus liegt hierbei auf einer humanoiden Sprachausgabe welche die Emotionen und Gestik des Protagonisten verstärken soll. Sämtliche Ergebnisse der durchgeführten Experimente und Analysen weisen ähnliche Defizite auf. Der Mangel und Möglichkeiten die Sprache effektiv zu modulieren um die Emotionen und Tonfall des Gesprochenen hervorzuheben ist keine Ausgangsbasis auf der man aufbauen kann, es klingt an diversen Stellen künstlich, unnatürlich oder wird mit einem metallischen Effekt versehen. Texte selbst zu sprechen oder diese einem Synchronsprecher zu übergeben ist bei weitem die kosten- und zeiteffizienteste Lösung. Ebenso um innerhalb weniger Sekunden verschiedene Emotionen und Tonfälle zu dem gewünschten gesprochenen Text zu erhalten. Die Testergebnisse spiegeln die untenstehende Grafik wieder.

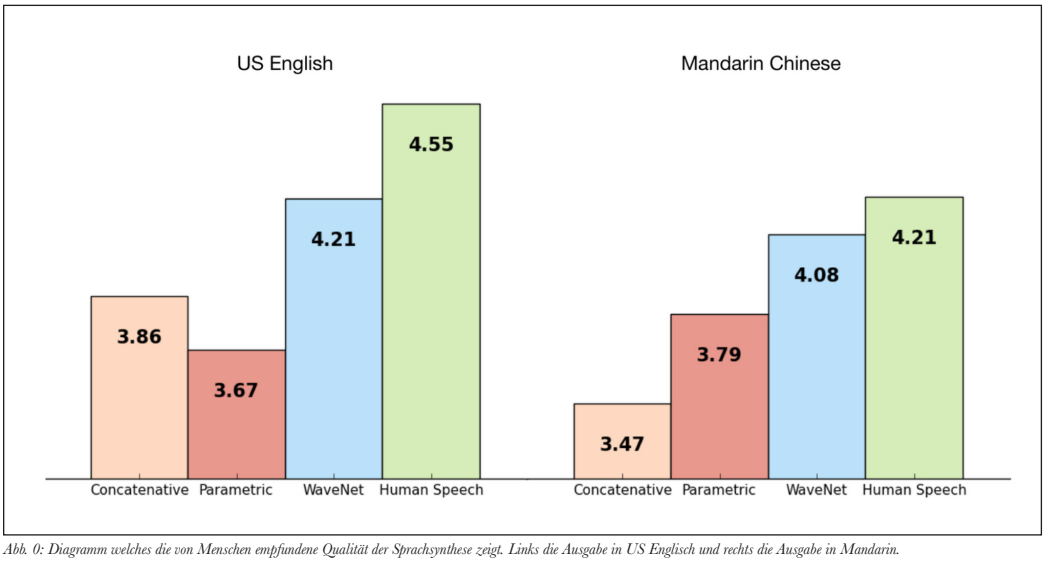

Abb. 0: Diagramm welches die von Menschen empfundene Qualität der Sprachsynthese zeigt. Links die Ausgabe in US Englisch und rechts die Ausgabe in Mandarin.

5 DISKUSSION

Nach der Durchführung der zuvor beschriebenen Tests ist ein Muster erkennbar in welchem die ausgeführte Methoden immer zu ähnlichen Endergebnissen führen. Die größten Auffälligkeiten der Resultate in dieser Testreihe sind die Qualitätsunterschiede der Sprachsynthese. In diesem Abschnitt werde ich Erklärungen anführen und auf neue Erkenntnisse hinweisen, welche Ansatzpunkte aufzeigen, um zukünftige Forschungen weiterzuführen.

5.1 ERKLÄRUNG

der Ergebnisse

Durch das Füttern und das Erstellen eines Avatars welcher Sprachsynthese man sich bedient hat, erhält man zu guter Letzt immer eine Art unbefriedigendes und unflexibles Durchschnittsergebnis. Dieses Resultat ist geprägt von spezifischen Lauten, Buchstaben oder deren Kombinationen, für jedes „s“ im Text wird immer auf denselben errechneten Ton zurückgegriffen. Variationen davon gibt es durch bestimmte Kombinationen. Zum Beispiel wird ein „st“ anders synthetisiert als ein „sk“. Ein weiterer Teil welcher das Ergebnis beeinflusst ist die Qualität und die Menge an Ausgangsmaterial welches für die Berechnungen herangezogen wird. Das Ergebnis im Test von 4.1.2 unterliegt hinsichtlich des Ausgangsmaterial ganz klar, einer von Google erstellten Version (4.1.3) und der des klassischen „Übersetzers“ in dem Test 4.1.

5.2 ERKENNTNISSE

Die Endergebnisse können zwar alle die gewünschte Information in Bezug auf Wörter und Buchstaben enthalten, sind jedoch nicht beeinflussbar hinsichtlich Emotionen oder Tonalität. Ebenso ist die Qualität der Ergebnisse stark abhängig vom Arbeitsaufwand in Bezug auf die Tests 4.1, 4.2 und 4.3 in welchen die Systeme aktiv mit Sprachaufnahmen versehen werden. Mit den ersten Daten steigt Ausgabequalität stark an, jedoch flacht diese Steigerung schnell ab. Ab einem mediokren Grad an akzeptabler Sprachsynthese werden die benötigten Datenmengen immer größer, was allerdings den Arbeitsaufwand für eine Vertonung eines Videospielcharakters nicht rechtfertigen würde. Denn das Endergebnis, selbst mit den offiziell zugänglichen und ausgereiften Sprachavataren von Google und Co. ist nicht ausreichend um einen Menschen als Synchronsprecher zu ersetzen. Die Flexibilität, Originalität und der geringen Arbeitsaufwand den ein humanoides Individuum mit sich bringt überwiegt stark in einer Gegenüberstellung.

5.3 EMPFEHLUNGEN

für zukünftige Forschungen

Um an den hier beschriebenen Testergebnissen anzusetzen und um neue Ergebnisse zu erhalten muss der neueste Stand der Technik berücksichtigt werden. Künstliche Intelligenz wird immer weiter verfeinert und intelligentere Ergebnisse liefern. Die Sprachsynthese ist zum Teil schon sehr gut, ihr fehlt es jedoch an den richtigen Werkzeugen um Emotionen, Sprachtempi oder Gestiken zu modulieren. Eine Automationsspur in welcher Gefühle oder Betonungen festgelegt werden können wäre sehr hilfreich. Weiters kann ein definierter Workflow oder ein geschaffener Standard, mit der Sprachaufnahmen gemacht werden, sehr hilfreich sein, um am effektivsten neurale Netzwerke mit Daten zu versorgen, ebenso um Zeit einzusparen und einen brauchbaren Avatar in kürzester Zeit zu erstellen.

6.1 FAZIT

der Teilfragen

Kann mit K.I. ein Charakter in einem Videospiel oder Film vertonent werden?

Ja und nein. Auf der einen Seite stehen die Ansprüche an Qualität und Arbeitsaufwand um an ein zufriedenstellendes Ergebnis zu kommen. Auf der anderen Seite steht der Wunsch um in möglichst viele Arbeitsvorgänge eingreifen und zu zentralisieren zu können. Ich konnte den Methoden mithilfe der künstlichen Intelligenz für mein Projekt keine verwertbaren Ergebnisse entlocken. Der Arbeitsaufwand rechnet sich nicht für das Resultat einer mangelhafte und unbrauchbare Vertonung

Reichen die technischen Mittel eines Sounddesigners dazu aus?

Ja die technischen Mittel reichen aus. Jedoch beansprucht es demnach auch einen bestimmten Zeitrahmen für jede der Methoden. Die einen, wie im Test 4.1 und 4.3 können innerhalb von Sekunden benützt werden, für andere, wie in 4.2 und 4.4, kann die Vorarbeit mit mehreren Tagen datiert sein.

- Kann die eigene Stimme als Ausgangsmaterial verwendet werden?

Ja die eigene Stimme ist hinsichtlich des Workflows sogar am brauchbarsten und kann jederzeit ohne andere Ressourcen hinzuzuziehen verwendet werden.

- Inwieweit muss das Ergebnis noch bearbeitet werden um ein befriedigenden Ergebnis zu erhalten?

Es müssen wie bei Sprachaufnahmen immer noch diverse Postproduktionsschritte durchgeführt werden um die Sprachsynthese an das Endprodukt anzupassen.

- Welche/s Wissen/Software muss erlangt werden um effektiv arbeiten zu können?

Ein Grundwissen in „Python“ wäre mir gerade in der Hinsicht auf neurale Netzwerke sehr hilfreich gewesen. Jedoch ist ein Grundstock in Signalverarbeitung und das Wissen über Frequenzgänge der menschlichen Sprache von Vorteil um die Ergebnisse effektiv beurteilen und Fehlerquellen isolieren zu können

- Kann ein solches System in ein Computerspiel eingebunden werden um effektiv Charaktere anhand derer Sprach-Skripts zu vertonen?

Ja dies ist möglich, „Lyrebird“ bietet zum Beispiel schon einen Bot an welcher die Daten vom persönlichen Sprachavatar dazu benützt um diversen sozialen Medien Sprachnachrichten zu versenden. Somit ist die Grundvoraussetzung für eine effektive Verwendungsmethode gegeben.

6.2 FAZIT

der Hypothesen

- Es wird eine Menge Zeit in Anspruch nehmen sich das Know-How über neuronale Netzwerke und die damit in Verbindung stehende Software anzueignen.

  Diese Hypothese hat sich bestätigt, sich diverses Fachwissen in Softwareentwicklung ging jedoch bei weitem über die zeitlich zur Verfügung gestellten Ressourcen dieser Arbeit hinaus.

- Würde ein Charakter mit robotischen Spracheigenschaften vertont werden, wäre ein befriedigendes Ergebnis schneller zu erreichen.

  Diese Hypothese bestätigt sich nur teilweise. Ein synthetischer Charakter hätte definitiv ein passenderes Medium für die Sprachsynthese dargestellt. Jedoch wären mir über andere Workflows bessere Lösungswege bekannt mit denen ich die Eigenschaften des Protagonisten besser und effektiver hervorheben hätte können.

- Es wird eine Variation an Ergebnissen, auch künstlerischer Natur, mit dem erlangten Wissen erhalten werden.

  Hat sich im Abschnitt 4.1 - 4.5 bestätigt.

- Text-to-Speech-Systeme können schnell eine Vielzahl an Lösungen anbieten.

  Schnell ja, aber die Vielzahl unterscheidet sich nur in einzelnen Stilelementen welche mir nicht das erhoffte breite Spektrum an Ergebnissen dargebracht haben. Da sich alle Methoden als unbrauchbar herausgestellt haben kann ich diese jedoch nicht als solche Lösungen ansehen.

7 PRAXIS

Vertonung eines Videospielcharakters

Die Vertonung des Protagonisten eines Videospiels ist die Aufgabe die ich hier nun gründlich dokumentiert habe bildet den praktischen Teil dieses Schriftstücks. Aufgrund gründlicher Recherche und den oben angeführten Testreihen wurden sämtliche Methoden der Sprachsynthese außen vor gelassen. Diese Entscheidung wurde nicht nur durch den steigenden Arbeitsaufwand der Sprachsynthese gefällt, sondern ebenso wegen des ungeeigneten Sprachstils der Ergebnisse.

Abb. 10: Erste Skizzen des Hauptcharakters.

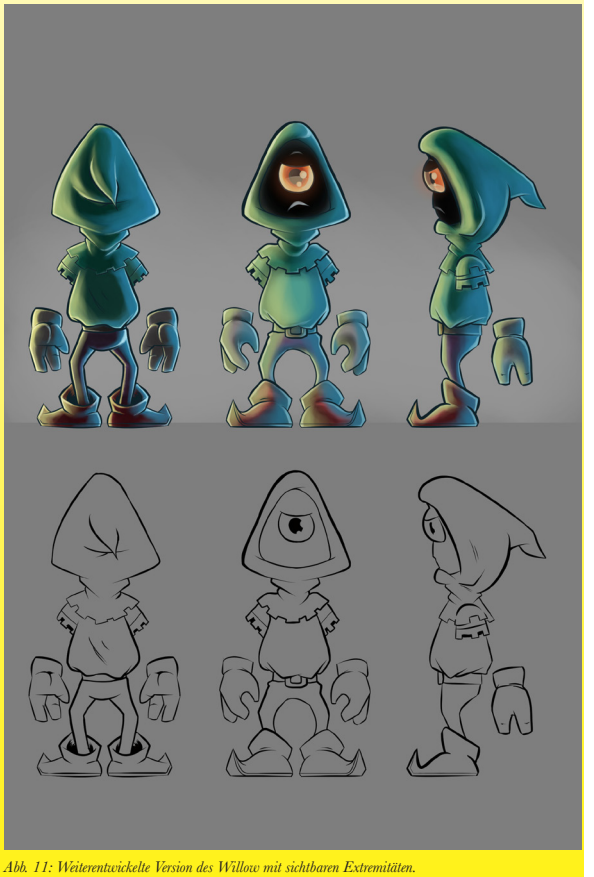

Abb. 11: Weiterentwickelte Version des Willow mit sichtbaren Extremitäten.

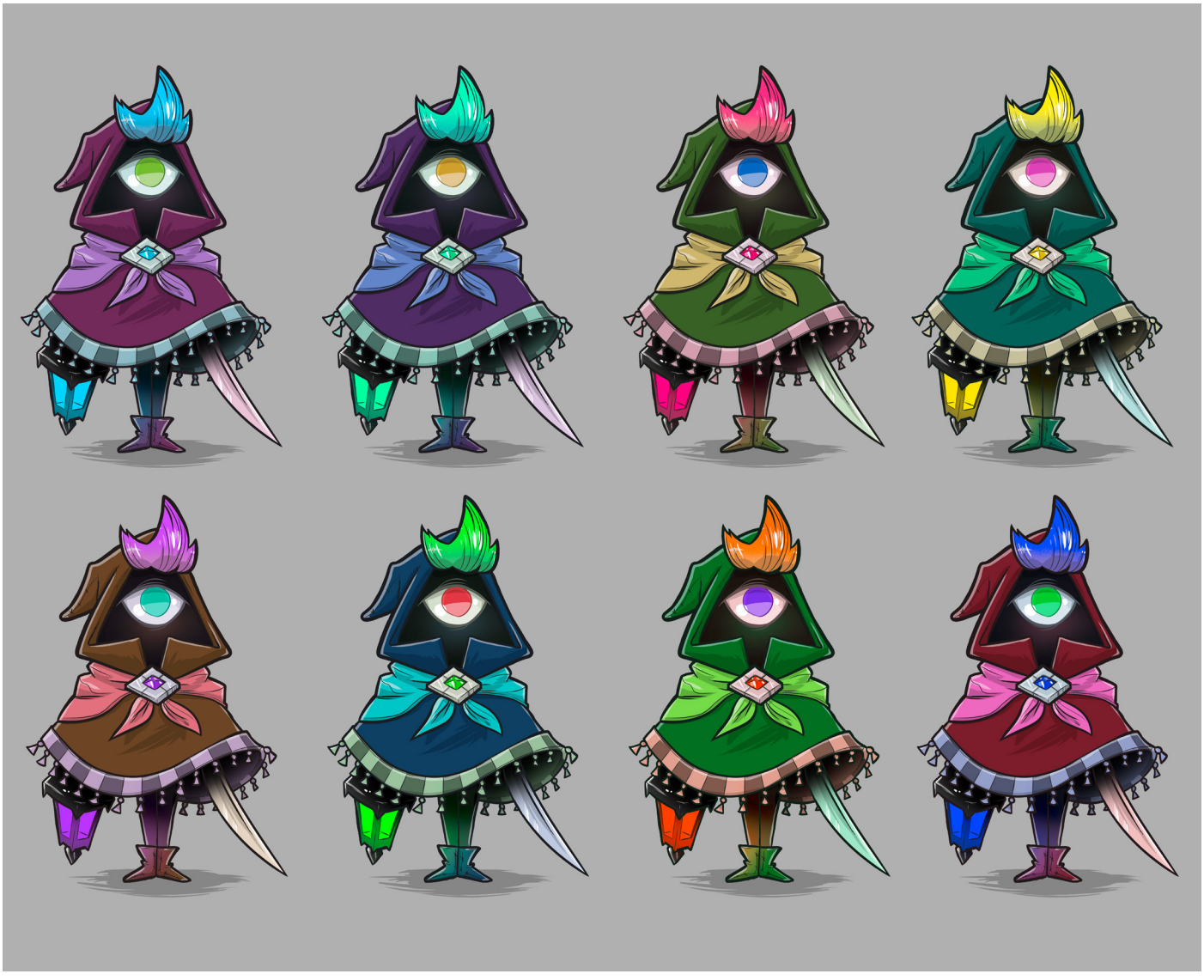

Abb. 12: Finale Konzeptzeichnung des Videospielcharakter in verschiedenen Farbvarianten.

8.1 SOUND

Konzept der Spielwelt

Das Ambiente der Welt setzt sich aus einer Mixtur von Fantasy, Mystery mit einer Prise komödiantischen Anspielungen zusammen. Finstere Höhlen, von Ranken umklammerte Türen und Steine die eigentlich lebendig sein könnten. Die Farbwelt ist ebenso düster wie auch durch bunte, hell erleuchteten Flecken durchbrochen. In der auditiven Version dieses Universums werden diese Bilder immer als Ausgangsidee für die Produktion sämtlicher produzierter Soundeffekte und musikalischer Begleitung dienen.

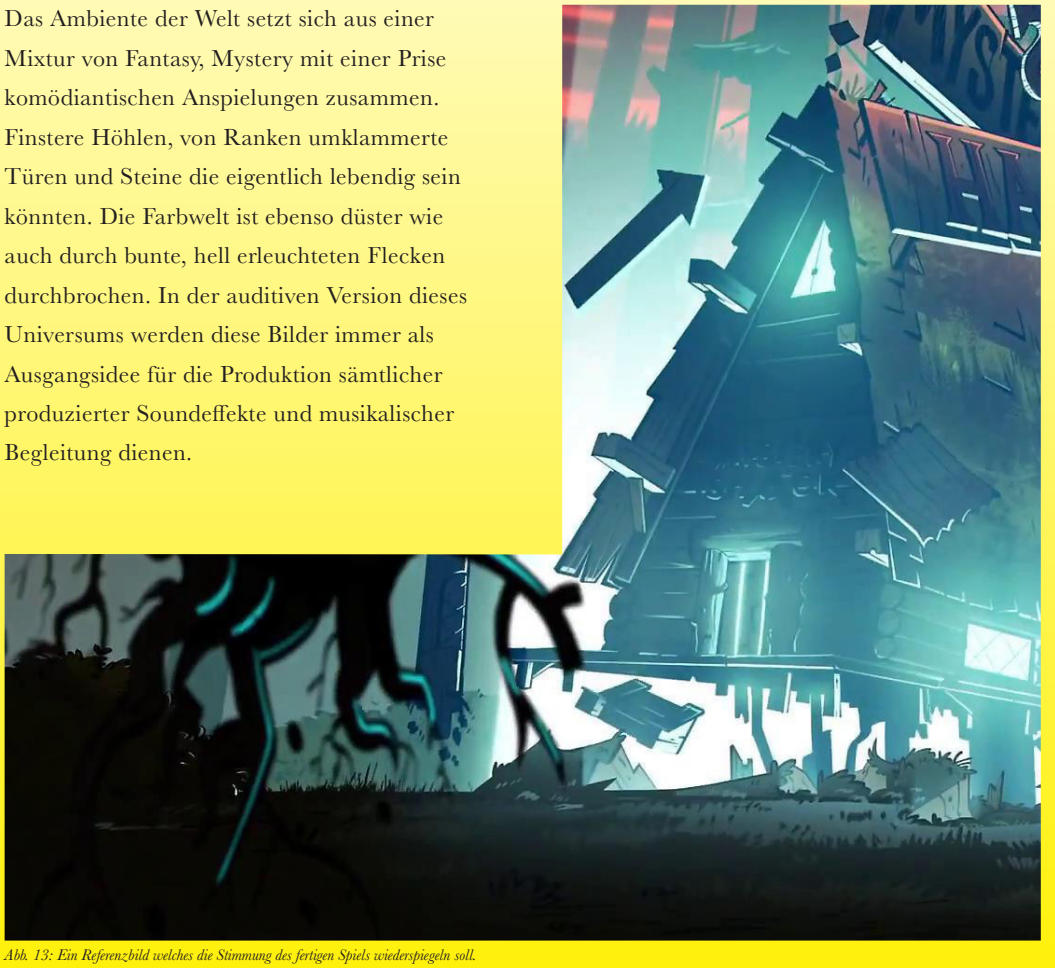

Abb. 13: Ein Referenzbild welches die Stimmung des fertigen Spiels wiederspiegeln soll.

8.2 SOFTWARE

Reason and Wwise

Die Komponenten

für die Produktion des Werkstückes bestehen einerseits aus einer „DAW“ und einer „Middleware“. Als Produktions- und Postproduktionsumgebung habe ich „Reason“ aus dem Hause Propellerhead gewählt. Einerseits wegen der intuitiven Handhabung des Programms, sowohl als auch diverser Möglichkeiten von CV-Verknüpfungen im Rack-Modus wegen. Zu guter Letzt habe ich mit dieser Software auch am meisten Erfahrung und kann die gewünschten Ergebnisse im abschätzbaren Rahmen erzielen.

Die Software

die benützt wird um die finalen Audioeffekte und Musikstücke dynamisch in das Spiel einzubinden nennt sich „Wwise“. Das kanadische Softwareunternehmen „Audiokinetic“ welches diese entwickelt hat beschreibt ihr Produkt selbst als die am weitesten entwickelte und komponentenreichste Lösung für Videospiele. Sei es eine Multimillionen-Dollar-Produktion oder ein kleines Indie-Spiel[12]. Integriert in die „GoDot-Game-Engine“ zentralisiert und verwaltet diese Middleware sämtliche benötigte Audiodatenbanken und vereinheitlicht den Workflow welchen man in anderen Projekten wiederaufnehmen kann. Weiters ist der Einstieg in die Software sehr benutzerfreundlich und wird durch hauseigene Videotutorials und frei verfügbare Testprojekte noch zugänglicher gemacht.

Viele „AAA-Titel“

wie „Overwatch[13]“ oder „The Witcher[14]“ setzen auf Wwise als Audio-Engine, zuallererst wegen der Stabilität der Anwendung, welche unabdingbar für die effektive Produktion von Titeln jeder Art ist. Folgend ist die Anzahl der Möglichkeiten der interaktiven Einbindungen einzelner Sounds und ganzer Arrangements ausschlaggebend.

13) Vgl. Audiokinetic: Wwise, The Enige Powering Interactive Audio. In: https://www.audiokinetic.com/products/wwise/ (zuletzt aufgerufen am 27.7.2019)
14) Vgl. Audiokinetic: Audiokinetic Video Channel, The interactive audio channel. In: https://www.audiokinetic.com/products/wwise/ (zuletzt aufgerufen am 27.7.2019)
15) Vgl. Audiokinetic: Audiokinetic Video Channel, The interactive audio channel. In: https://www.audiokinetic.com/learn/videos/-XZO9PgHmFA/ (zuletzt aufgerufen am 27.7.2019)

9. PROTAGONIST

Jemand namens Willow

Der Protagonist

namens Willow ist das Aushängeschild des Spiels. Sein visuelles Äußeres erinnert an einen Kobold mit einem Poncho und einer Kapuze, welch Gesicht sich jedoch im Schatten dieser versteckt, bleibt dem Spieler verborgen. Nur ein Haarschopf und das Auge sticht ganz klar heraus welches als Ankerpunkt für sämtliche gezeigte Emotionen dient. Unter der Kutte ragt eine leuchtende Laterne und ein Schwert hervor welche wichtige Spielelemente bilden. Die Laterne einerseits spendet Licht im Dunkeln und das Schwert dient zur Selbstverteidigung und als Grundlage für das Kampfsystem.

Das Verhalten

und die Persönlichkeit des Willow beruht auf einer aufsässigen aber dennoch akzeptierenden unterwürfigen Art. Er ist ein Diener welcher vom Spieler, dem transzendenten Magier heraufbeschworen wurde. Wie schon in Abschnitt 4.2 beschrieben ist er ein Wesen aus Fleisch und Blut, welches geisterhaft erscheinen mag, dennoch hat er einen zähen Körper unter den Textilien. Diese Information ist Grundlegend für die nachfolgende Dokumentation der einzelnen Sprachaufnahmen und deren Postproduktion sowie die Evaluation und Eignung der Testergebnisse.

Abb. 14: Finale Farbauswahl des Protagonisten.

9.1 SPRACHE

Aufnahmen

Die Sprachaufnahmen wurden von mir persönlich gesprochen, aufgenommen und postproduziert. Zuallererst wurde die Persönlichkeit Willow's mit diversen anderen fiktiven Charakteren verglichen, die in etwa eine gleiche Anmutung haben. Emotionen, Betonungen und Aussprache waren hier essenziell um die einzelnen Aspekte hervorzuheben. Weiters ist es essentiell für jede Zeile mehrere Alternativen zu produzieren um der Repetitivität um Spiel vorzubeugen. Wenn exemplarisch Aktion „A" ausgeführt wird gibt es Sound „X", dennoch muss bei wiederholter Aktion „A" Sound „X.1", „X.2" oder „X.3" abgespielt werden um für mehr Immersion und Abwechslung innerhalb der Spielwelt zu sorgen.

9.1.1 WORKFLOW

der Sprachaufnahmen

Ein Skript,
das zuerst in den einzelnen zu sprechenden Texten vorliegt, wird in weiterer Folge, basierend auf einer Gegenstands- und Spielmechanik-Liste, bearbeitet. Eine funktionierende, spielbare Demo-Version des Spiels liegt bis zu dem Zeitpunkt des Verfassens dieser Masterarbeit noch nicht vor. Jedoch wird hier eine Bibliothek vorproduziert mit der schon jetzt in Wwise gearbeitet werden kann.

Das Einsprechen
der einzelnen Zeilen erfolgt mithilfe eines Rode NT1-A Mikrofons in einer schallarmen Sprecherkabine. Die Signale werden direkt in die Audiosoftware „Reason“ als Monosignal gespeist und können von dort aus weiterbearbeitet werden. Die Aufnahmen wurden in einer Abtastrate von 48kHz und einer Tiefe von 24Bit getätigt um ein hochqualitatives Rohmaterial zu erhalten. Diese Voreinstellungen wurden ebenso bei weiteren Foley-Aufnahmen angewendet, auf welche ich in dieser Arbeit noch zu einem späteren Zeitpunkt eingehe. Es wurden von jeder Textzeile bis zu 15 verschiedene Aufnahmen gesprochen. Ein paar wenige davon wurden jedoch durch Spannungsgeräusche und Klicks bei der Aufnahme unbrauchbar gemacht. Während des Auswahlprozesses wurden von jeder Voiceline durchschnittlich 3-4 Versionen für die Weiterbearbeitung ausgewählt.

Nach dem Auswahlverfahren
werden diese in der Postproduktion geschnitten und von Unregelmäßigkeiten bereinigt. Einige Aufnahmen wurden nachträglich noch mit Pitch-Anpassungen versehen um besser ins Gesamtbild zu passen.. Um die Klarheit der Aussprache zu maximieren wurden hauptsächlich „T“-Laute am Ende von Wörtern zusätzlich in der Lautstärke angepasst.

Die Vereinheitlichung
des Klanges der Sprachaufnahmen wurde mit einem MClass Equalizer erzielt. Die Frequenzen um 2.3kHz wurden um knapp 6db angehoben und mit einem MClass-Stereo-Imager

die Stereobreite erhöht. Für die Sprachaufnahmen habe ich speziell Stereo anstatt Mono gewählt um etwaigen Mono-Soundeffekten im Endarrangement mehr Platz bieten zu können. Danach wurden noch leichte Anhebungen im Bereich über 5kHz vorgenommen sowohl als auch speziell im tieferen Bereich um 300Hz um der Stimme mehr Präsenz zu geben. Es folgt ein Compressor um das Gesamtbild ein wenig zu verdichten welcher den Bereich über -22,4dB mit einem Ratio von 14:1 komprimiert. Danach folgt ein „Scream 4“ welcher den Aufnahmen mehr Struktur in Form einer Bandmaschine verleiht. Dieser Effekt wurde jedoch nur mit einer 18%-igen Stärke angewandt. Zu guter Letzt wird noch ein MClass-Maximizer geschaltet welcher die Lautstärke in einen Bereich um 9db anhebt.

Zuletzt wird jeder Sound in einer Abtastrate von 44.1kHz exportiert, in Kombination mit einer Bitrate von 16 ergeben diese Voreinstellungen kleinere Datenmengen. Diese sind essentiell für die Performance der Software des Endprodukts. Gespeichert wird dies im Containerformat „WAV“.

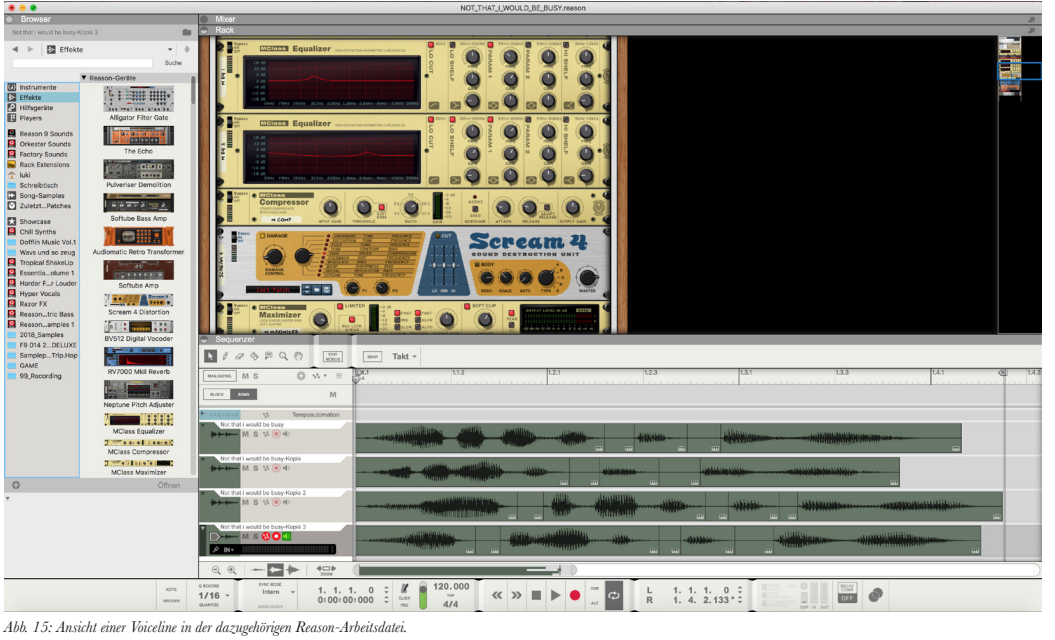

Abb. 15: Ansicht einer Voiceline in der dazugehörigen Reason-Arbeitsdatei.

10. ITEMS

Waffen und Verbrauchbares

Folgende Gegenstände, welcher der Spieler benützen oder in einer Art interagieren kann, wurden für die erste Demo-Version des Spiels entwickelt:

- Ein magischer Feuerring, welcher mit einer Flammenbrunst seine Opfer über große Distanz verbrennen kann.
- Ein Heiltrank, welcher verlorenen Lebenspunkte nach einem Kampf wieder auffüllt.
  Eine Stoppuhr, welche die Zeit für sämtliche feindselige Einheiten einfriert, um selbst im Vorteil zu sein.
- Ein aktivierbares Magieschild um Projektile oder andere eintreffende Angriffe zu annullieren und zurückzuwerfen.
- Das klassische Schwert des Willow, welches schon im Abschnitt 9.1 schon erklärt wurde und als Grundausstattung dient.

Dieses Arsenal an Waffen und Unterstützungsgegenständen wird im Stil der Bildwelt erschaffen. Als Ausgangsmaterial dienen großteils Klänge organischer Natur wie Metall, Holz und Stein.

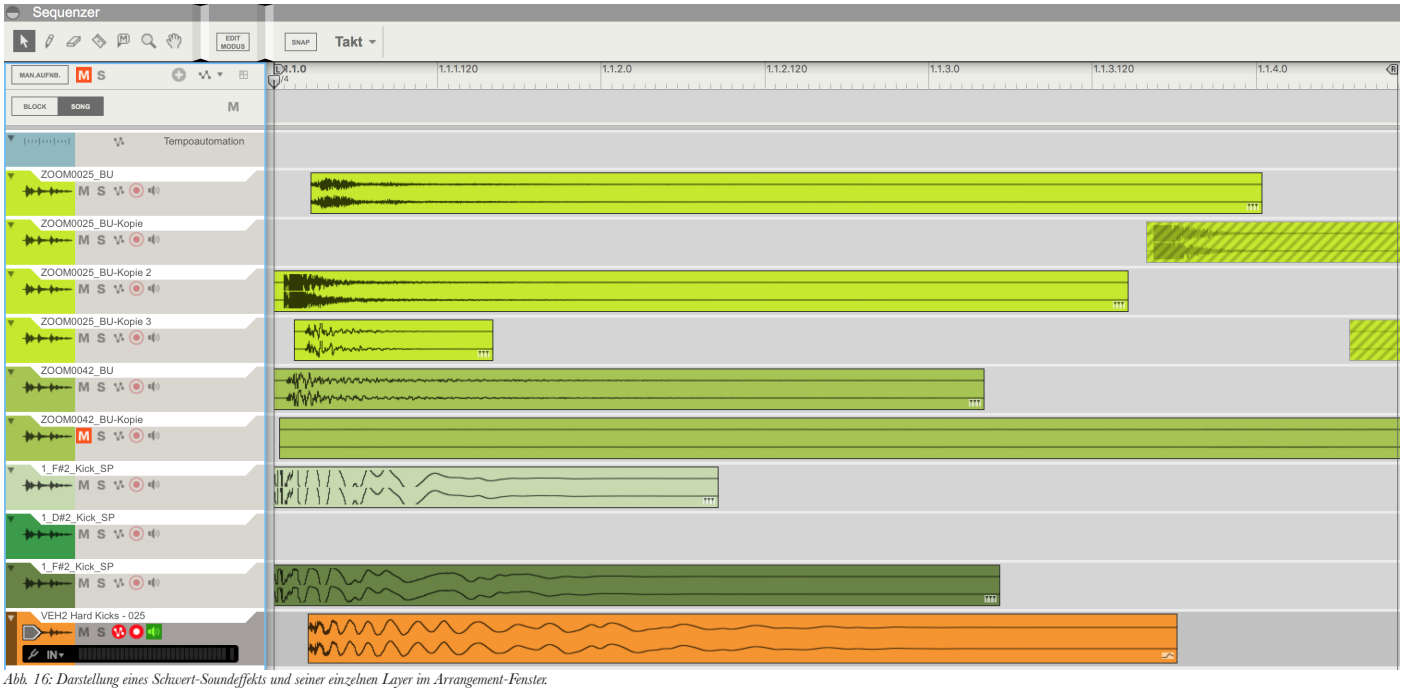

Abb. 16: Darstellung eines Schwert-Soundeffekts und seiner einzelnen Layer im Arrangement-Fenster.

10.1 WORKFLOW

der Items

Sämtliches Material wurde mit einem ZOOM-H6 in Verbindung mit einem Sennheiser ME66 Kondensatormikrofon aufgenommen. Platz der Aufzeichnungen war mein persönliches Foley-Studio. Die Voreinstellungen der Aufnahme beliefen sich auf eine Abtastrate von 48kHz und einer Bittiefe von 24Bit. Die Aufnahmen werden in die Audio-Software Reason gespeist und dort für die Weiterverarbeitung aufbereitet. Selektion erfolgt durch Vergleichen mit Referenzanimationen und Soundeffekten aus einschlägigen Fantasy-Videospielen wie „The Elder Scrolls: Online“ von Bethesda. Nach diesem Arbeitsschritt werden die einzeln ausgewählten Soundeffekte noch sorgsamer beschnitten, gelayert und in ihrer Position verändert um die gewünschte Tonalität zu erhalten. Ebenso wird bei manchen einzelnen Spuren die Tonhöhe angepasst um einen Schwertschlag mächtiger klingen zu lassen oder Fußstapfen dem Körpergewicht der Person anzupassen. Das anpassen von Fade-In's und Fade-Out's ist ebenso wichtig wie die Anpassung der Lautstärken der einzelnen Clips um Klicken oder anderen unerwünschten Störgeräuschen oder Artfakten vorzubeugen. Durch Equalizing werden zuletzt noch die essentiellen Frequenzen so moduliert, dass sie dem gewünschten Ergebnis zu entsprechen. Um eine Maximallautstärke von 0dB zu erreichen, wird eine Normalisation durchgeführt, da die einzelnen Effektverkettungen, in den einzelnen finalen Soundeffekten, stark variieren und diese alle eine Maximallautstärke von 0dB erreichen.

Weiters werden die Soundeffekte wie schon bei sämtlichen anderen Audiodateien zuvor, mit einer Abtastrate von 44.1kHz und einer Bittiefe von 16Bit exportiert, und sind somit bereit für die Einbindung in die Middleware Wwise.

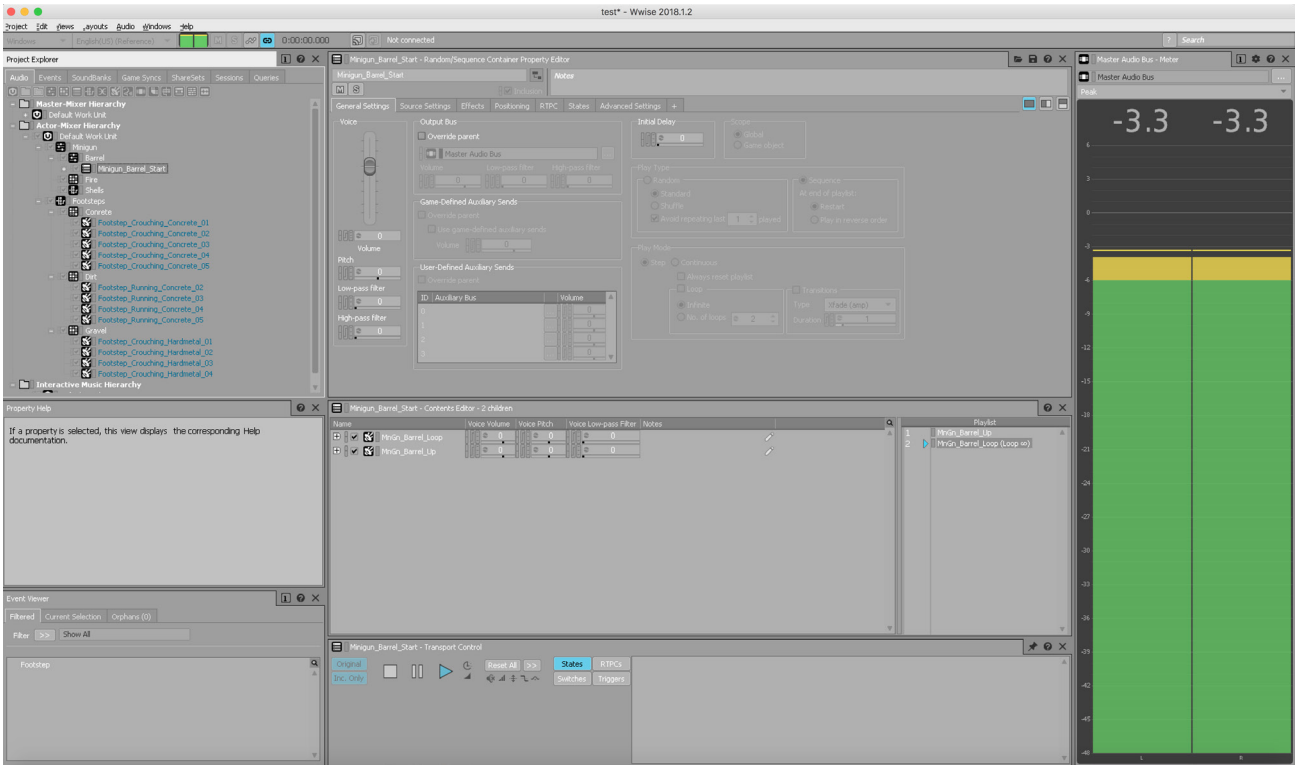

Abb. 17: Eine Wwise Session in der diverse Tests mit Soundeffekten durchgeführt wurden um den Arbeitsweise dementsprechend anzupassen.

11. SOUNDTRACK

Konzept

Dem Konzept liegt eine ausgiebige Recherche im Genre-Bereich Trip-Hop und dessen einzigartigem Sound zugrunde[36]. Zu den größten Inspirationen diente allerdings das bekannte Stück „Unfinished Sympathy“ von „Massive Attack“[37]. Zusätzlich zu Trip-Hop habe ich noch klassische House-Elemente einfließen lassen. Ersteres harmoniert wunderbar mit dem Gedanken eines tiefen, mysteriösen Verlieses welches vom Helden erkundet wird. House wird hier neben atmosphärischen Synthesizern als Akzent eingesetzt um die geheimnisvolle und ernstere Stimmung zu brechen. Als Tonart wurde die Phrygisch-dominante erwählt da diese, richtig eingesetzt, orientalisch anmutet, und hervorragend zu dem Schlagzeug im Trip-Hop Stil passt. Dieses zeichnet sich stark durch eine sehr gedämpfte, aber dennoch stark von Hall geprägte Weise aus. Diverse Synthesizer werden mit einer Bandmaschine großzügig bearbeitet um gezielt diese Tonalität zu erhalten. Die Tempi der besagten Musikstücke bewegen sich zwischen 80 bis zu 120 Schlägen pro Minute, obwohl sich Trip-Hop hauptsächlich zwischen 60 bis 90 Schlägen bewegt.

36) Johnson, Phil: Straight outa Bristol. Massive Attack, Portishead, Tricky and the Roots of Trip Hop. Hodder & Stoughton, London 1996
37) Massive Attack: Unfinished Sympathy. In: Massive Attack: Blue Lines. o.O. 1991, In: https://www.youtube.com/watch?v=ZWmrfgj0MZI (zuletzt aufgerufen am 26.8.2019)

Abb. 17: Ein Synthesizer welcher mit seinem Klangspektrum prägend für die Tonalität des Soundtracks war. Der „Thor - Polyphonic Synthesizer".

11.1 WORKFLOW

des Soundtracks

Als Basis
für die Komposition dient eine Bibliothek mit vielen einschlägigen Loops im angestrebten Tempo. Diese wurden gröber und großzügiger in das Arrangement eingebunden als ich es sonst in meinen Arbeitsprozessen halte. Neben der Kreation von Melodien mit Wiedererkennungswert wurden diese noch mit unheimlich anmutenden Atmosphären untermalt. Als Synthesizer habe ich hier hauptsächlich den „Thor - Polyphonic Synthesizer“ verwendet. Dieser zeichnet sich durch seine Klarheit und Brillanz im Klang aus welcher exzellent mit mehreren MIDI-Noten gleichzeitig harmoniert. Um den gewünschten unheimlichen Effekt zu erzielen wird dieser mit dem „Scream 4“ und seinem Bandmaschinen-Effekt belegt. Danach wird noch Hall und eine Dämpfung in den Höhen hinzugefügt, dieser wird mit dem „RV7000“ generiert. Die Besonderheit dieser Effekteinheit ist die Feineinstellungen auf Raumgrößen, Impulsantwort und diverser Entfernungsreglern um das gewünschte Ergebnis zu erhalten. Die Bandmaschine und der Hall wurden mit den selben Geräten auf sämtliche Perkussionsinstrumente angewandt. Im Falle der Synthesizer wurden diese durch Automation nur zeitweise aktiviert.

Für das Mastering
des Stückes „Space“ wurde hier zu allererst ein Compressor mit einer Grenze von -30dB geschalten welches das Signal mit einem Ratio von 1,79:1 bearbeitet. Mit einer Angriffszeit von 100ms und einer Releasezeit von 233ms wirkt dies noch ein wenig weicher. Der MClass-Compressor sorgt für ein dichteres Verhältnis der einzelnen Komponenten und leitet das Signal in einen Equalizer des gleichen Herstellers MClass. Mit einem Lo-Cut werden

sämtliche Frequenzen unter 30Hz beschnitten da diese nicht benötigt werden. Eine kleine Anhebung im Bassbereich und eine kleine steilere Absenkung der Mitten/Höhe wurde ebenso vorgenommen. Mit einem MClass-Maximizer wurde das Signal zuletzt noch verstärkt und in den richtigen Pegelbereich gebracht.

Ident mit den Sprachaufnahmen aus Punkt 9.2.1 wurden die einzelnen Musikstücke in Stereo exportiert um den Spieler mehr den Eindruck der Immersion zu verstärken. Die Abtastrate sowohl die Bitrate haben ebenso den Standard von 44.1kHz und 16 Bit, ebenso zu Gunsten der Performance des Videospiels.

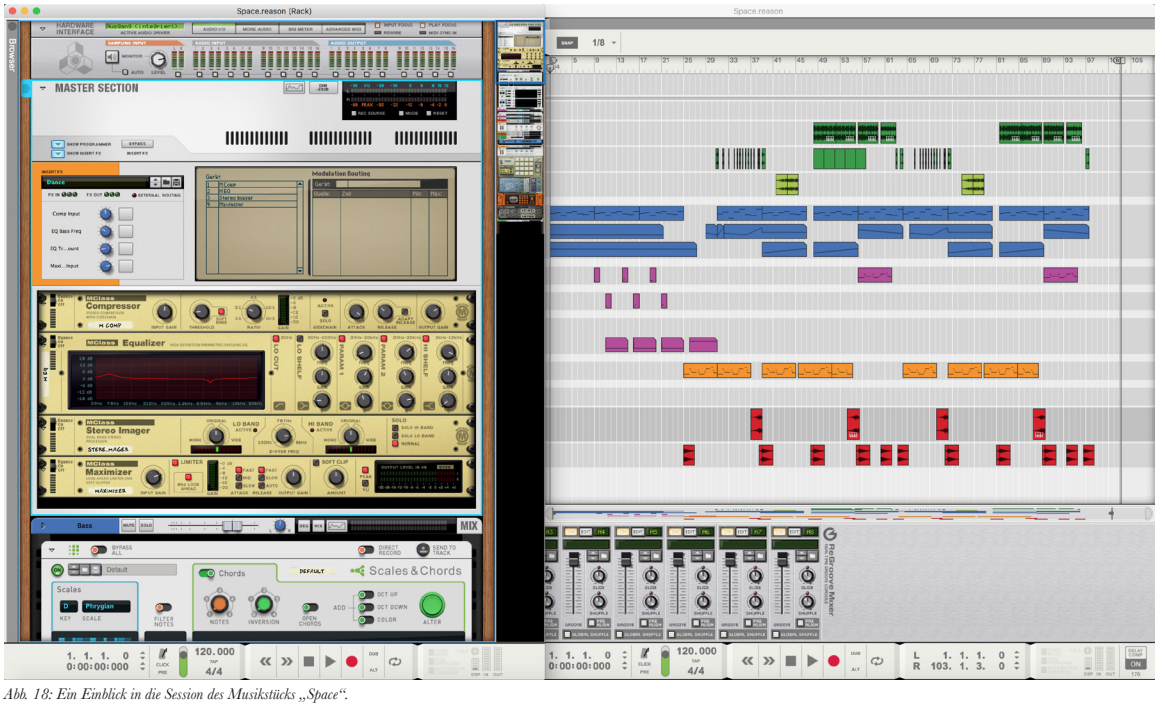

Abb. 18: Ein Einblick in die Session des Musikstücks „Space“.

LITERATUR VERZEICHNIS

1) Vgl. Dreyfus, Hubert L.: Die Grenzen künstlicher Intelligenz. Was Computer nicht können. Athenäum, Königstein 1985

2) Statista, und Norstat. „In immer mehr Bereichen des Lebens spielen digitale Sprachassistenten eine Rolle. Welche dieser Sprachassistenten kennen Sie?." Chart. 29. März, 2017 https://de.statista.com/statistik/daten/studie/739040/umfrage/umfrage-zur-bekanntheit-ausgewaehlter-sprachassistenten-in-deutschland/ (zuletzt aufgerufen am 09.7.2019)

3) Google Assistant: What it can do. In: https://assistant.google.com/learn/ (zuletzt aufgerufen am 09.7.2019)

4) Vgl. Wikipedia. Die freie Enzyklopädie (14.11.2012), s.v. Bibliothek, https://de.wikipedia.org/wiki/Künstliche_Intelligenz (zuletzt aufgerufen am 09.7.2019)

5) Vgl. Bostrom, Nick: Superintelligenz. Szenarien einer kommenden Revolution. Suhrkamp, Frankfurt am Main. 2016

6) Vgl. Lewandowski, Theodor: Linguistisches Wörterbuch. 4., neu bearbeitete Aufl. Quelle & Meyer, Heidelberg 1985

7) Vgl. Sapir, Edward: Language. An Introduction to the Study of Speech. Harcourt Brace, New York, 1921

8) Vgl. C. Lalueza-Fox u.a: The derived FOXP2 variant of modern humans was shared with Neandertals. In: Curr Biol. 17(21), 2007

9) Vgl. Grüner, Sebastian: Tacotron 2: Googles Sprachsynthese erreicht fast menschliche Qualität - Golem.de. In: golem.de. 21. Dezember 2017 (zuletzt aufgerufen am 09.7.2019)

10) Reichel, Uwe: Unit-Selection-Synthese. Datengetriebenes Vorgehen vs. Signalmanipulation. Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München, München 2017

11) Vgl. Örnek, Evin Pinar: medium. 31.1.2018, https://medium.com/@evinpinar/wavenet-implementation-and-experiments-2d2ee57105d5 (zuletzt aufgerufen am 17.7.2019)

12) Vgl. Soni, Devin: Towards: Data Science. 22.3.2018, https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d (zuletzt aufgerufen am 19.8.2019, Übersetzung des Verfassers)

13) Vgl. Audiokinetic: Wwise, The Enige Powering Interactive Audio. In: https://www.audiokinetic.com/products/wwise/ (zuletzt aufgerufen am 27.7.2019)

14) Vgl. Audiokinetic: Audiokinetic Video Channel, The interactive audio channel. In: https://www.audiokinetic.com/products/wwise/ (zuletzt aufgerufen am 27.7.2019)

15) Vgl. Audiokinetic: Audiokinetic Video Channel, The interactive audio channel. In: https://www.audiokinetic.com/learn/videos/-XZO9PgHmFA/ (zuletzt aufgerufen am 27.7.2019)

16) Handbuch der künstlichen Intelligenz. Görz, Günther; Schneeberger, Josef. Oldenburg Wissenschaftsverlag GmbH, München 2014
17) Human + Machine: Künstliche Intelligenz und die Zukunft der Arbeit. Daugherty, Paul R.; Wilson H., James. dtv Verlagsgesellschaft mbH & Co. KG, 2018
18) Intonation, Kurze Einführungen in die germanistische Linguistik. Peters, Jörg; Wilson H., James. Universitätsverlag Winter GmbH Heidelberg, 2014
19) https://www.ted.com/playlists/310/talks_on_artificial_intelligen (abgerufen am 8.1.2019)
20) Walter Murch. In: transom. A Showcase and Workshop for New Public Radio. Stand: 4.1.2005, https://transom.org/2005/walter-murch/ (zuletzt aufgerufen am 28. Juli 2019).
21) Bridgett, Rob: Dynamics of Narrative, Gamasutra.com, 24.9.2009, http://www.gamasutra.com/view/feature/132531/dynamics_of_narrativephp (zuletzt aufgerufen am 26. Juli 2019).
22) Fritsch, Melanie: History of Video Game Music, in: Music and Game. Perspectives on a Popular Alliance, hg. von Peter Moormann, Wiesbaden 2014, S. 11-40.
23) Paul Watzlawick: Man kann nicht nicht kommunizieren. Das Lesebuch. Hersg. Von Trude Trunk. Huber Verlag Bern, 2011
24) Markus Bandur: Melodia / Melodie [1998, 38 Seiten], in: Handwörterbuch der musikalischen Terminologie, hg. von H. H. Eggebrecht,
25) Franz Steiner, Wiesbaden, später Stuttgart, 1971–2006; CD-ROM, Stuttgart 2012
26) Overwatch - Game Audio Using Wwise. In: audiokinetic. Stand: 13.9.2016, https://blog.audiokinetic.com/Overwatch-game-audio-using-wwise-1/ (zuletzt aufgerufen am 28. Juli 2019).
27) Kramer, Gregory: Sound and Communication in Virtual Reality, in: Communication in the Age of Virtual Reality, hg. von Frank Biocca und Mark R. Levy, New Jersey 2005, S. 259-276.
28) Herzfeld, Gregor: Atmospheres at Play. Aesthetical Considedrations of Game Music, in: Music and Game. Perspectives on a Popular Alliance, hg. von Peter Moormann, Wiesbaden 2014, S. 147-158.
29) Demers, Joanna: Dancing Machines. Dance Dance Revolution, Cybernetic Dance and Musical Taste, in: Popular Music and Multimedia, hg. von Julie McQuinn, Lawrence 2011, S. 443-456.
30) Geräuschemacher. In: Wikipedia. Die freie Enzyklopädie. Stand: 30.7.2016, https://de.wikipedia.org/wiki/Geräuschemacher (zuletzt aufgerufen am 26. Juli 2019).
31) Jörg U. Lensing: Sound-Design - Sound-Montage - Soundtrack-Komposition: Über die Gestaltung von Filmton, 2. Auflage, Schiele & Schön, Berlin 2009
32) Audio Special: Foley for games. Stand: 27.4.2013, http://www.develop-online.net/analysis/audio-special-foley-for-games/0117620 (zuletzt aufgerufen am 27. Juli 2019).
33) Klangsynthese. In: Wikipedia. Die freie Enzyklopädie. Stand: 15.7.2017, https://de.wikipedia.org/wiki/Klangsynthese (zuletzt aufgerufen am 28. Juli 2019).
34) Maas, Andrew: Spoken Language Processing. Frühling 2017, https://web.stanford.edu/class/cs224s/lectures/224s.17.lec16.pdf (zuletzt aufgerufen am 10. August 2019)
35) Vgl. Aunkofer, Benjamin: Data Science Blog: Machine Learning vs. Deep Learning, https://data-science-blog.com/blog/2018/05/14/machine-learning-vs-deep-learning-wo-liegt-der-unterschied/ (zuletzt aufgerufen am 18.8.2019)
36) Johnson, Phil: Straight outa Bristol. Massive Attack, Portishead, Tricky and the Roots of Trip Hop. Hodder & Stoughton, London 1996
37) Massive Attack: Unfinished Sympathy. In: Massive Attack: Blue Lines. o.O. 1991. In: https://www.youtube.com/watch?v=ZWmrfgj0MZI (zuletzt aufgerufen am 26.8.2019)

ABBILDUNGS VERZEICHNIS

Abb. 9: Sourcecode des eines WaveNet-Projektes zur Auswertung in Python.
(Urheber Lukas Matthias Robausch)

Abb. 10: Erste Skizzen des Hauptcharakters.
(Urheber David Angelo Tschmuck)

Abb. 11: Weiterentwickelte Version des Willow mit sichtbaren Extremitäten.
(Urheber David Angelo Tschmuck)

Abb. 12: Finale Konzeptzeichnung des Videospielcharakter in verschiedenen Farbvarianten.
(Urheber David Angelo Tschmuck)

Abb. 13: Ein Referenzbild welches die Stimmung des fertigen Spiels wiederspiegeln soll.
Vgl.: https://image.winudf.com/v2/image/Y29tLm1hc3VyeS5ncmF2aXR5X3NjcmVlbl80XzE1MjI5NDg5NTlfMDgw/screen-4.jpg?fakeurl=1&type=.jpg

Abb. 14: Finale Farbauswahl des Protagonisten.
(Urheber David Angelo Tschmuck)

Abb. 15: Ansicht einer Voiceline in der dazugehörigen Reason-Arbeitsdatei.
(Urheber Lukas Matthias Robausch)
Abb. 16: Darstellung eines Schwert-Soundeffekts und seiner einzelnen Layer im Arrangement-Fenster.
(Urheber Lukas Matthias Robausch)

Abb. 17: Eine Wwise Session in der diverse Tests mit Soundeffekten durchgeführt wurden um den Arbeitsweise dementsprechend anzupassen.
(Urheber Lukas Matthias Robausch)

Abb. 17: Ein Synthesizer welcher mit seinem Klangspektrum prägend für die Tonalität des Soundtracks war. Der „Thor - Polyphonic Synthesizer“.
(Urheber Lukas Matthias Robausch)

Abb. 18: Ein Einblick in die Session des Musikstücks „Space“.
(Urheber Lukas Matthias Robausch)

Abb. 19: Beispiel von „freiem“ Lernen in welchem das System ohne Vorgaben selbst kategorisiert.
Vgl.: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d (zuletzt aufgerufen am 19.8.2019)

Abb. 20: Beispiel von „überwachtem Lernen in welchem durch Vorgaben Kreise und Kreuze getrennt werden oder ein sich der effektivste Pfad abbilden soll.
Vgl.: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d (zuletzt aufgerufen am 19.8.2019)

GGWP