Bachelorarbeit im Studiengang Audiovisuelle Medien (AMB7)

Entwicklung einer virtuellen Konferenz unter besonderer Berücksichtigung der Reproduktion von zuvor aufgenommener Sprache

Vorgelegt von Robin Hirt
Matrikelnummer: 27360
Vorgelegt am 03. April 2017
an der Hochschule der Medien Stuttgart
zur Erlangung des akademischen Grades eines „Bachelor of Engineering“
Erstprüfer: Prof. Oliver Curdt
Zweitprüfer: Prof. Jens-Helge Hergesell

Eidesstattliche Versicherung

„Hiermit versichere ich, Robin Hirt, an Eides Statt, dass ich die vorliegende Bachelorarbeit mit dem Titel: „Entwicklung einer virtuellen Konferenz unter besonderer Berücksichtigung der Reproduktion von zuvor aufgenommener Sprache“selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentliche oder in anderer Form als Prüfungsleistung vorgelegt worden.
Ich habe die Bedeutung der eidesstattlichen Versicherung und die prüfungsrechtlichen Folgen (§26 Abs. 2 Bachelor-SPO (6 Semester), § 23 Abs. 2 Bachelor-SPO (7 Semester) bzw. § 19 Abs. 2 Master-SPO der HdM) sowie die strafrechtlichen Folgen (gem. § 156 StGB) einer unrichtigen oder unvollständigen eidesstattlichen Versicherung zur Kenntnis genommen.“

Stuttgart, 1. April 2017 ———————————

(Name)

Kurzfassung

Diese Arbeit befasst sich mit der Erstellung einer virtuellen Konferenz, mit der anschließend additive Konferenzmikrofon-Systeme getestet werden können. Der Schwerpunkt liegt dabei auf der menschlichen Stimme und wie diese über Lautsprecher möglichst natürlich wiedergegeben werden kann. Der Leser erhält zunächst einen Einblick in die Stimmerzeugung und welche akustischen Eigenschaften die menschliche Stimme hat. Nach Erläuterung der Konzeption einer virtuellen Konferenz werden Aufnahmen von Sprache, Geräuschen und Atmos beschrieben. Die Signale werden anhand ermittelter Daten der Aufnahme- Wiedergabekette entzerrt und in einer Session organisiert. Für die Wiedergabe wird ein mehrkanaliges Lautsprecher-Setup mit einem drehbaren Lautsprecher entworfen. Über das Lautsprecher-Setup wird die erstellte virtuelle Konferenz wiedergegeben und mit einer realen Konferenz verglichen.

Abstract

This thesis is focusing on how to create a virtual conference for testing conference microphone systems. The main focus is put on the human voice and how this can be reproduced as natural as possible via loudspeakers. The reader is initially given an insight into the sound generation and the acoustic characteristics of the human voice. After explaining the conception of a virtual conference, recordings of speech, noise and atmos are described. The signals are equalized based on collated data of the recording and reproduction chain and organized in a session. A multi-channel speaker setup with one rotatable speaker is designed for playback. The speaker setup will play the created virtual conference and it will be compared to a real conference.

Inhaltsverzeichnis

1 Einführung **1**
1.1 Zielsetzung und Vorgehensweise 1

2 Grundlagen **3**
2.1 Akustik 3
2.1.1 Schallfeld 3
2.1.2 Direkt-, Diffusfeld und Hallradius 3
2.1.3 Schalldruck und Schalldruckpegel 4
2.1.4 Richtcharakteristik 5
2.1.5 Phantomschallquelle 6
2.2 Elektronik 6
2.2.1 Arduino 6
2.2.2 Operationsverstärker 7

3 Die menschliche Stimme **9**
3.1 Stimmerzeugung 9
3.2 Schalldruckpegel der gesprochenen Sprache 10
3.3 Frequenzspektrum der gesprochenen Sprache 13
3.4 Richtcharakteristik des Sprachapparats 15
3.5 Elemente der Stimme und Sprechweise 17
3.6 Sprechgeschwindigkeit und Pausen 19
3.7 Verdeckung von Sprache durch Geräusche 19

4 Konzeption einer virtuellen Konferenz **21**
4.1 Konferenz? 21
4.2 Eingrenzung 22
4.3 Aufnahme 23
4.3.1 Aus welchen Geräuschen besteht eine Konferenz 23
4.3.2 Auswahl der Sprecher und Sprachen 23
4.3.3 Erstellung eines Sprechertextes 24
4.3.4 Auswahl der Störgeräusch 26
4.3.5 Auswahl der Atmo 27
4.4 Entzerrung 27
4.5 Session 28
4.6 Wiedergabe 28
4.6.1 Lautsprecher-Setup 28
4.6.2 Beweglicher Lautsprecher 29

5 Die Aufnahme **31**
5.1 Der Aufnahmeraum 31
5.2 Auswahl des Sprechermikrofons 31

5.3 Abstand des Sprechers zum Mikrofon . . . 33
5.4 Die Aufnahmekette . . . 34
5.5 Aufnahme . . . 34
5.5.1 Aufnahme der Sprecher . . . 34
5.5.2 Aufnahme der Störgeräusche . . . 36
5.5.3 Aufnahme der Atmos . . . 36
5.5.4 Aufnahme der Impulsantwort . . . 37

6 Entzerrung der Signale 39
6.1 Werkzeuge zur Entzerrung . . . 40
6.1.1 Grafischer EQ . . . 40
6.1.2 Parametrische EQ . . . 40
6.2 Voxengo Deconvolver . . . 41
6.2.1 Entfaltung des Raums . . . 41
6.3 Entzerrung der aufgenommenen Signale . . . 42
6.3.1 Ermittlung der Frequenzgänge von Lautsprecher und Mikrofon . . . 42
6.3.2 Konfigurierung des EQs anhand der ermittelten Frequenzgänge . . . 45
6.3.3 Fazit . . . 49

7 Die Session 51
7.1 Einstellen der Schallpegel . . . 52

8 Das Wiedergabesystem 53
8.1 Verwendete Lautsprecher . . . 53
8.2 Anordnung der Lautsprecher . . . 53
8.3 Der drehbare Lautsprecher . . . 55
8.3.1 Aufbau des Drehtellers . . . 55
8.3.2 Signal zur Ansteuerung . . . 56
8.3.3 Aufbau der elektronischen Umgebung . . . 57
8.3.4 Aufbau der Programmierung . . . 59

9 Vergleich von realer und virtueller Konferenz 61

10 Fazit und Ausblick 63

11 Anhang 65
11.1 Sprechertext . . . 65
11.2 Technische Zeichnungen . . . 70
11.3 E-Plan . . . 74
11.4 Arduino Programm . . . 75
11.5 Danksagung . . . 78
11.6 Digitaler Datenträger . . . 78

Literaturverzeichnis 79

Abbildungsverzeichnis

2.1 Standard-Lautsprecheranordnung in einem gleichschenkligen Dreieck mit dem Hörer für Zweikanal-Stereowiedergabe. 6
2.2 Vergleich idealer und realer Operationsverstärker 7

3.1 Frequenz- Bewertungskurven für Schallpegelmessung 11
3.2 Terzpegel, unbewerteter und A-bewerteter Pegel von bandbegrenztem weißem Rauschen 11
3.3 Schallpegel (L) der Sprache als Funktion der Zeit (t) eines monoton gesprochenen Satzes, aufgezeichnet mit einem Pegelschreiber (Schreibgeschwindigkeit 3 cm/s) mit der Zeitbewertung Fast und Slow 12
3.4 Durchschnittliche Spektren deutscher Sprache von männlichen Sprechern, für leise, mittellaut und laut gesprochener Sprache, Schalldruckpegel je Terzbereich 14
3.5 Durchschnittliche Spektren deutscher Sprache von Sprecherinnen, für leise, mittellaut und laut gesprochener Sprache, Schalldruckpegel je Terzbereich . . 14
3.6 Charakteristische Frequenzgebiete der 1./2./3./ Formanten von Vokalen von Männern (m), Frauen (w) und Kindern (k) 15
3.7 Polardiagramm der menschlichen Stimme A-Bewertet von Mann und Frau. . 16
3.8 Polardiagramm der menschlichen Stimme von 160 Hz bis 8 kHz 17
3.9 Sprachverständlichkeit gegenüber S/N 20

4.1 Open Space Konferenz Kassel 21
4.2 Bsp: Telefonkonferenz 22

5.1 Aufbau für Sprachaufnahmen mit Schallpegelmesser 35
5.2 Beispiel einer Schallpegelmessung eines Sprechers 35

6.1 Frequenzgang Schoeps CCM 4V 42
6.2 Frequenzgang Genelec 1029A in 1m Entfernung bei 0°, 15°, 30° und 45° . . . 43
6.3 Messung von Genelec 1029A (lila) und Schoeps CCM 4V(grün) 44
6.4 Messung des Frequenzgangs des Genelec *1029A* und Schoeps *CCM 4V* im Aufnahmeraum 45
6.5 Der reine Sinus-Sweep und aufgenommene Sprache 46
6.6 Signale entzerrt nach Frequenzgang des im Aufnahmeraum aufgenommenem Sinus-Sweeps 47
6.7 Signale entzerrt nach Frequenzgang des *1029A* 47
6.8 Signale entzerrt nach Frequenzgang des *1029A* und *CCM 4V* 48
6.9 EQs konfiguriert anhand ermittelter Frequenzgänge 48

7.1 Zusammenstellung einer kleinen Konferenz 51
7.2 Zusammenstellung der großen Konferenz 52

8.1 Anordnung der Lautsprecher 54

8.2 Testaufbau des Lautsprecher-Setups 54
8.3 Drehteller mit Schrittmotor 56
8.4 Recht-Wechselspannung 1 Volt, 100 Hz 57
8.5 Einweg-Präzisionsgleichrichter 58
8.6 Gleichgerichtetes Rechtecksignal 5 Volt, 100 Hz 59

9.1 Reales Nachsprechen der virtuellen Konferenz 61

11.1 Sprechertext 1 von 5 65
11.2 Sprechertext 2 von 5 66
11.3 Sprechertext 3 von 5 67
11.4 Sprechertext 4 von 5 68
11.5 Sprechertext 5 von 5 69
11.6 Technische Zeichnung Drehkranz 70
11.7 Technische Zeichnung der Bodenplatte 1 von 2 71
11.8 Technische Zeichnung der Bodenplatte 2 von 2 72
11.9 Technische Zeichnung der Kunststoffscheibe 73
11.10 E-Plan 74
11.11 Arduino Programm Teil 1 von 3 75
11.12 Arduino Programm Teil 2 von 3 76
11.13 Arduino Programm Teil 3 von 3 77

Abkürzungsverzeichnis

cm Zentimeter

DAW Digital Audio Workstation

dB Dezibel

$dB(A)$ Dezibel A-Bewertet

$dB(SPL)$ Dezibel Sound Pressure Level

Hz Hertz

kHz Kilohertz

HdM Hochschule der Medien Stuttgart

$kOhm$ Kilo-Ohm

L Schalldruckpegel

m Meter

m^2 Quadratmeter

m^3 Kubikmeter

ms Millisekunden

Pa Pascal

S/N Signal to Noise Ratio

s Sekunden

1 Einführung

Diese Arbeit befasst sich mit der Erstellung einer virtuellen Konferenz. Es wird versucht eine Konferenz mit 4 anwesenden Personen so realistisch wie möglich zu simulieren. Besonders wichtig ist dabei eine möglichst authentische Wiedergabe der Sprache.

1.1 Zielsetzung und Vorgehensweise

Ziel dieser Arbeit ist es ein Konzept zu entwickeln und auszuarbeiten mit dem anschließend additive Mikrofonsysteme getestet werden können. Besonders interessant ist das für Systeme, mit denen bestimmte Schallquellen fokussiert werden sollen. Die Einsatzgebiete, in denen ein solches System Anwendung finden könnte, sind sehr groß und die zu fokussierenden Schallquellen können wie auch die auszublendenden Schallquellen sehr unterschiedlich sein. Ein Anwendungsgebiet bilden Konferenzen, bei denen eine oder mehrere Personen anwesend sind. Genau mit dieser Situation befasst sich diese Arbeit. Die menschliche Stimme wird dabei die Schallquelle sein die es zu fokussieren gilt und alle anderen Schallereignisse sollen die auszublendenden Störquellen sein. Zu Beginn dieser Arbeit wird die Hauptschallquelle, die menschliche Stimme, genauer untersucht, wie sie entsteht, wie sich gesprochene Sprache zusammensetzt und welche akustischen Eigenschaften sie hat. Anschließend wird ein Konzept erstellt wie eine Konferenz ablaufen könnte und wie sie möglichst realistisch simuliert werden kann. Dazu wird zunächst die Größe der Konferenz, bezogen auf anwesende Personen und Umfang auftretender Geräusche eingegrenzt und auf dieser Basis eine Liste möglicher vorkommender Störgeräusche erstellt. Auch die Dauer der Konferenz wird stark begrenzt. Dafür werden aber mehrere kleine Konferenzen mit unterschiedlichen Sprechern entworfen, die später in der Session zu einer langen Konferenz verbunden werden können. Die Texte der Sprecher werden auf der Basis der am häufigsten in der deutschen Sprache vorkommenden Wörter gebildet, zu zusammenhängenden Sätzen geformt und anschließend in mehrere Sprachen übersetzt. Vor den Sprach- und Geräuschaufnahmen werden die Eigenschaften des Aufnahmeraums ermittelt, sowie Position und Abstände von Mikrofon und Sprecher definiert. Die nach der Aufnahme vorliegenden Audiodateien werden analysiert und so entzerrt, dass sie bei der Wiedergabe möglichst nahe an dem aufgenommenen Schallereignis liegen. Mit diesem Vorgang soll ein möglichst realistisches Szenario geschaffen werden. Angepasst werden die Audiodateien an einen bestimmten Lautsprechertypen, der später auch bei den Tests des hier vorgesehenen additiven Mikrofonsystems verwendet wird. Die fertigen Audiodateien werden in einer Session so organisiert, dass der Nutzer später noch die Möglichkeit

hat, ohne großen Aufwand die Reihenfolge der gesprochenen Sätze, der Geräusch und der Atmos modifizieren zu können, um so das Testszenario zu verändern. Aus den vorhandenen Spuren wird ein Wiedergabesystem bzw. eine Lautsprecheranordnung konzipiert, die aus drei Schichten besteht. Die erste Schicht bilden die Lautsprecher für Störgeräusche, die sich nahe dem Mikrofonsystems ereignen. Die zweite bilden die Lautsprecher für die einzelnen Sprecherstimmen und die dritte bilden Lautsprecher, mit denen bei Bedarf unterschiedliche Atmos der Konferenz zugemischt werden können. Alle Lautsprecher werden nach dem Mikrofonsystem ausgerichtet, was besonders bei den Sprechern den Optimalfall darstellt. Um aber auch eine nicht optimale Situation simulieren zu können, wird für einen Sprecher-Lautsprecher eine motorisierte Drehvorrichtung entworfen, die über die Session angesteuert werden kann und den Lautsprecher von seiner Ausrichtungsposition wegdreht.

2 Grundlagen

In diesem Kapitel werden einige Grundlagen beschrieben, die für den weiteren Verlauf dieser Arbeit benötigt werden.

2.1 Akustik

2.1.1 Schallfeld

Schall ist eine erzwungene Schwingung einer Materie, der Schallquelle, bei der benachbarte Moleküle eines Mediums aus ihrer Ruhelage bewegt werden [Goe15] [1]. „Wenn eine Schallquelle das sie umgebende Medium, z. B. Luft, zum Mitschwingen anregt, so entsteht um die Schallquelle eine sich ausbreitende Schallwelle, ein Schallfeld.“ [Dic14] [2] In z.B. dem uns als Luft bekannten Medium findet eine Folge von Luftverdichtungen und -verdünnungen statt, die sich periodisch durch den Raum fortpflanzen [Goe02] [3]. Ist kein Medium vorhanden, wie in einem Vakuum, kann es auch zu keiner Verdichtung oder Verdünnung kommen und somit auch kein Schallfeld erzeugt werden [Dic14] [2].

2.1.2 Direkt-, Diffusfeld und Hallradius

Innerhalb eines Raumes, wie beispielsweise einer Halle, lässt sich das Schallfeld als Summe des direkt abgestrahlten Schalls und des diffusen Schalls, der näherungsweise im gesamten Raum gleich stark ist, betrachten. „Im sog. freien Schallfeld kann sich der Schall völlig ungehindert in alle Richtungen ausbreiten, er trifft nicht auf Hindernisse; er erreicht den Hörer oder das Mikrofon nur auf direktem Wege, daher auch Direktschall genannt [Dic14] [2]. “Innerhalb eines Raums gibt es einen begrenzten Bereich um eine Schallquelle, in dem der Pegel des Direktschalls überwiegt, das Direktfeld. Der Bereich in dem der Direkt- und der Diffusschall den gleichen Pegel aufweisen nennt sich Hallradius oder Hallabstand. Er berechnet sich wie folgt:

[1] Vgl. Görne, 2015
[2] Vgl. Dickreiter, 2014, S.9
[3] Vgl. Görne, 2002

$$r_H = 0,057 * \sqrt{\frac{V}{T}} \qquad [\text{Goe15}]^4 \tag{2.1}$$

Der Hallradius wird in m angegeben. T beschreibt die Nachhallzeit in s und V das Volumen des Raums in m^3. Der Hallradius r_H ist allerdings für gleichmäßig in alle Richtungen abstrahlende Schallquellen und ebenso gleichen Schallempfang definiert. Für die Berechnung gerichteter Schallquellen und Empfänger muss der Bündelungsgrad γ berücksichtigt werden:

$$r_H = 0,057 * \sqrt{\gamma} * \sqrt{\frac{V}{T}} \qquad [\text{Dic14}]^5 \tag{2.2}$$

Außerhalb dieses Hallradius liegt das Diffusfeld. In diesem Bereich überwiegt der diffuse Schall, der durch mehrere Reflexionen des Schalls von den Wänden, Boden, Decke und anderen Gegenständen entsteht.

2.1.3 Schalldruck und Schalldruckpegel

In z.B. einem gasförmigen Medium, wie dem Gasgemisch Luft, herrscht im Ruhezustand ein bestimmter Druck, der atmosphärische Gleichdruck. Wird eine Schallquelle zum Schwingen angeregt, entsteht ein Wechseldruck, dessen Schwankungen sich mit dem atmosphärischen Gleichdruck überlagern [Dic14] [5]. Da das menschliche Ohr auf diese Druckschwankungen reagiert, ist der Schalldruck die für uns wohl wichtigste Schallfeldgröße [Goe15] [1]. Seine SI-Einheit ist Pascal, mit dem Einheitszeichen Pa.

Da der menschliche Gehörsinn keine absoluten Sinnesreize, sondern näherungsweise Reizverhältnisse als konstant empfindet, erweist sich eine lineare Darstellung als nicht sehr vorteilhaft [Goe15] [1]. Zur besseren Erfassung des Schalldrucks wird deshalb generell der absolute Schalldruckpegel L genannt. Dieser setzt sich aus dem 20-fach logarithmierten Verhältnis von Schalldruck und einem vereinbarten Bezugsschalldruck p_0 zusammen. Seine Maßeinheit ist dB.

$$L = 20 * \log \frac{p}{p_0} \qquad [\text{Dic14}]^2 \tag{2.3}$$

[4] Vgl. Goerne, 2015, S.86
[5] Vgl. Dickreiter, 2014, S.41

2.1.4 Richtcharakteristik

Eine Schallquelle, die klein gegenüber der von ihr abgestrahlten Wellenlänge des Schalls ist, strahlt den Schall zum größten Teil gleichmäßig in alle Richtungen ab. Ist eine Schallquelle größer als die Wellenlänge des von ihr erzeugten Schalls ist das nicht mehr der Fall. Liegen die Dimensionen der Schallquelle im Bereich der für uns hörbaren Frequenzen, erfolgt ein für uns hörbarer und messbarer Unterschied in der Abstrahlung des Schalls [Dic14] [6]. Die Wellenlänge der für uns hörbaren Frequenzen liegen ungefähr zwischen 15 m und 2 cm. Da die Abstrahlung von der Beschaffenheit der Schallquelle und der abgestrahlten Frequenz abhängt kann die Bündelung sehr komplexe Formen annehmen.

Neben den Schallquellen gibt es auch Schallempfänger die, bedingt durch ihre Konstruktion, eine Richtwirkung aufweisen. Je stärker die Richtwirkung ausgeprägt ist desto kleiner wird der Bereich in dem das Mikrofon den Schall aufnimmt.

Beschrieben wird dieses Verhalten mit:

Richtungsfaktor bei Mikrofonen:
Verhältnis des Schalldrucks, der unter einem bestimmten Winkel zur Bezugsrichtung einfällt, zum Schalldruck der Bezugsrichtung. [Dic14] [7]

Richtungsfaktor bei Lautsprechern:
Verhältnis des Schalldrucks, der unter einem bestimmten Winkel zur Bezugsrichtung abgestrahlt wird, zum Schalldruck der Bezugsrichtung. [Dic14] [8]

Richtungsmaß:
20-facher Zehnerlogarithmus des Richtungsfaktors. [Dic14] [8]

Bündelungsgrad bei Mikrofonen:
Verhältnis zwischen freiem und diffusem Schall im Vergleich zu einem Mikrofon mit gleichem Freifeld-Übertragungsmaß und kugelförmiger Charakteristik. [Dic14] [8]

Bündelungsgrad bei Lautsprechern:
Verhältnis der akustischen Leistung eines ungerichteten zu der eines gerichteten Lautsprechers, deren Übertragungsmaß auf der Bezugsachse, in gleicher Entfernung, gleich ist. [Dic14] [8]

Bündelungsmaß:
10-facher Logarithmus des Bündelungsgrads. [Dic14] [8]

[6] Vgl. Dickreiter, 2014, S.97
[7] Vgl. Dickreiter, 2014, S.637
[8] Vgl. Dickreiter, 2014, S.641

2.1.5 Phantomschallquelle

Als Phantomschallquelle wird ein psychoakustischer Effekt bezeichnet, bei dem ein Schallereignis aus einer Richtung wahrgenommen wird, in der sich keine Schallquelle befindet. Eine Grundbedingung zur Erzeugung dieses Effektes ist, dass mindestens zwei Schallquellen vorhanden sind [Fri14] [9]. Wenn beide Lautsprecher in einem bestimmten Abstand zueinander aufgestellt werden und sie gleichzeitig das exakt selbe Signal abstrahlen, lokalisiert der Hörer nicht 2 getrennte, sondern eine virtuelle Schallquelle in der Mitte der beiden Schallquellen. Durch das Verändern von Pegel- und Laufzeitdifferenzen zwischen den Signalen kann die Phantomschallquelle entlang der Basis der Lautsprecher bewegt werden [Dic14] [10].

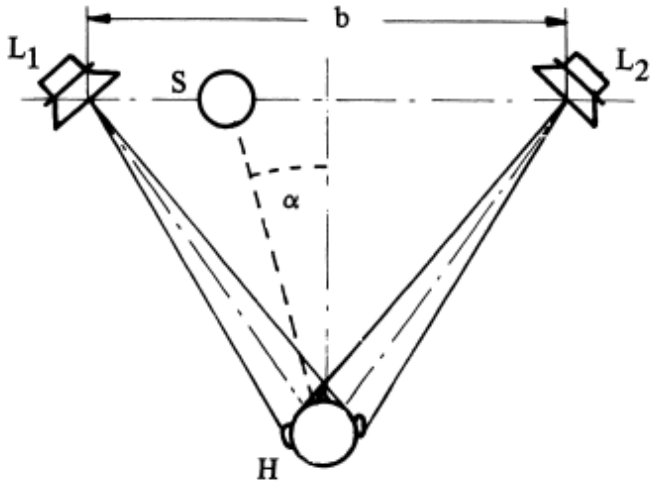

Abbildung 2.1: Standard-Lautsprecheranordnung in einem gleichschenkligen Dreieck mit dem Hörer für Zweikanal-Stereowiedergabe. [10]

2.2 Elektronik

2.2.1 Arduino

Die Arduino Reihe ist eine sogenannte „Physical-Computing-Plattform“ bestehend aus Soft- und Hardware [Wik17a] [11]. Eines der kleinsten Boards der Reihe, der „Arduino Uno“, ist ein auf dem ATmega328P basierendes Microcontroller-Board mit 14 digitalen Ein-Ausgangs Pins.

[9] Vgl. Friesecke, 2014, S.145
[10] Vgl. Dickreiter, 2014, S.221
[11] `https://de.wikipedia.org/wiki/Arduino_(Plattform)`

Mit einer Taktzeit von 16MHz und einem 32KB Flash-Speicher lassen sich allerdings schon erstaunlich umfangreiche Projekte realisieren. Programmiert wird er über eine eigene integrierte Entwicklungsumgebung (IDE), bei der es sich wiederum um eine JAVA-Anwendung handelt. Zusätzlich lassen sich in C und C++ programmierte avr-gcc Libraries in die Programme einbinden, auf deren Funktionen vom Code-Editor aus zugegriffen werden kann [Ard17] [12].

2.2.2 Operationsverstärker

Bei einem Operationsverstärker handelt es sich um ein elektronisches Bauteil, das aus einer sehr komplexen Schaltung von Transistoren, Kondensatoren und Widerständen besteht. Ein Operationsverstärker ist ein mehrstufiger, hochverstärkender Differenzverstärker, mit dem sowohl Gleich- als auch Wechselspannungen enorm verstärkt werden können [EKum] [13]. Im Idealfall hat ein Operationsverstärker eine unendlich große Verstärkung, einen unendlich großen Eingangswiderstand, einen Ausgangswiderstand gleich Null und einen Frequenzbereich von Null bis unendlich. Technisch ist das leider nicht in diesem Maße zu realisieren. Abbildung 2.2 zeigt einen Vergleich zwischen idealem und realem Operationsverstärker.

Kenngröße	Idealer Operationsverstärker	Realer Operationsverstärker
Verstärkungsfaktor V	unendlich	ca. 1.000.000
Eingangswiderstand R_e	unendlich Ω	1 MΩ bis 1000 MΩ
Untere Grenzfrequenz f_{min}	0 Hz	0 Hz
Unitity-Gain-Frequenz-Bandbreite	unendlich Hz	> 100 MHz
Gleichtaktverstärkung V_{Gl}	0	ca. 0,2
Gleichtaktunterdrückung G	unendlich	ca. 5.000.000
Rausch-Ausgangsspannung U_{rausch}	0 V	ca. 3 µV

Abbildung 2.2: Vergleich idealer und realer Operationsverstärker [14]

Operationsverstärker werden meist über eine negative und positive Versorgungsspannung betrieben und können, abhängig von ihrer Beschaltung, zum Eingangssignal gleich- und gegenpolige Spannungen am Ausgang erzeugen.

[12] `https://www.arduino.cc/en/Main/ArduinoBoardUno`
[13] https://www.elektronik-kompendium.de/sites/bau/0209092.htm
[14] `https://www.elektronik-kompendium.de/sites/bau/0209092.htm`

3 Die menschliche Stimme

3.1 Stimmerzeugung

Im Laufe der Evolution hat der Mensch gelernt das schwache Geräusch, das strömende Luft beim Ein- und Ausatmen erzeugt, durch gezielte Artikulation dazu zu verwenden, Lautsignale zu erzeugen, die zur Kommunikation benutzt werden können. Die dabei verwendeten Mechanismen sind sehr komplex. Zur Erzeugung von Lauten kontrollieren wir Formungen der Mundöffnung, Verlagerung der Zunge in der Mundhöhle, Verschiebung des Gaumensegels und Formung der beiden Stimmlippen im Stimmlippenapparat, auch Glottis genannt [The98] [1].

Die Sprachsignale, also die vom Sprechenden erzeugten Schalldrucksignale, lassen sich grob in folgende Gruppen aufteilen:

Vokale:

Auch Selbstlaute genannt, bestehen aus einem Grundton und Obertönen. Der Grundton wird hierbei durch die Glottisschwingung erzeugt, während im Vokaltrakt (u.a. Stimmritze, Mund und Nasenraum) der Obertonbereich erzeugt wird [Wik17b] [2]. Zu den Vokalen zählen unter anderem die Laute a, e, i, o, u und Verwandte.

Nasale:

Sie sind wie die Vokale stimmhaft. Weil aber im Gegensatz zu den Vokalen der Mund weitgehend geschlossen ist, erfolgt die Schallabstrahlung hauptsächlich über die Nase. Eine wichtige Rolle bei der Erzeugung unterschiedlicher Nasale spielen Mund- und Zungenposition [The98] [1]. Zu den Nasalen zählen Laute wie n, m, ng und Verwandte.

Frikative:

Bei der Erzeugung von Frikativen formt sich die Glottis zu einem schmalen Spalt. Auch Rachenwand, Gaumen und Zähne sind maßgeblich an der Erzeugung beteiligt [The98] [1]. Im Gegensatz zu Vokalen und Nasalen lassen sich Frikative mit Zisch und Rauschlauten umschreiben. Zu ihnen zählen die Laute f, s, sch und Verwandte.

Plosivlaute:

Ihre Erzeugung beruhen überwiegend auf einem zeitlichen Ablauf. Zuerst wird der Atemluftstrom durch die Lippen oder Zunge blockiert und anschließend wieder freigesetzt, wobei ein explosionsartiger Laut entsteht [Wik16a] [3]. Zu ihnen zählen Laute wie

[1] Vgl. Therhardt, 1998
[2] `https://de.wikipedia.org/wiki/Vokal`
[3] `https://de.wikipedia.org/wiki/Plosiv`

b, g, d, p, k, t und Verwandte.

3.2 Schalldruckpegel der gesprochenen Sprache

Akustisch und informationstechnisch betrachtet ist die Sprache ein Schallereignis, bei dem durch Schalldruckänderungen über Zeit in bestimmten Mustern Informationen übertragen werden, die vom Zuhörer wahrgenommen und entschlüsselt werden. Diese Schalldruckänderungen lassen sich messtechnisch erfassen, mit unterschiedlichen Methoden auswerten und veranschaulichen. Hierzu wird zum einen ein Empfänger benötigt, der auf Änderungen des Schalldrucks reagiert, wie z.B. ein Mikrofon. Zum anderen wird entsprechende Elektronik benötigt, die Schalldruckänderungen auswerten und veranschaulichen kann. Um den Schalldruckpegel erfassen zu können gibt es sogenannte *Schallpegelmesser*. Diese erfassen den Schalldruckpegel über einen definierten Frequenzbereich und ermitteln daraus einen Gesamtpegel. Für die Auswertung des Schalldruckpegels wurden unterschiedliche Bewertungen definiert, die sich auf unterschiedliche Gewichtung von Frequenzen beziehen, die Frequenzbewertungen und auf die Mittelung der Messwerte zeitlich schwankender Schallpegel, die Zeitbewertungen. Der Hintergrund unterschiedlicher Frequenzbewertungen ist der, dass das menschliche Gehör nicht alle Frequenzen mit gleichem Schalldruckpegel als gleich laut empfindet. Aus diesem Grund ist es sinnvoll den Schalldruckpegel nicht nur linear sondern auch im Verhältnis zum empfundenen Lautstärkepegel auswerten zu können. Besonders berücksichtigt wird hierbei die geringe Empfindlichkeit des menschlichen Ohrs für tiefe Frequenzen [Sen02] [4]. Dennoch ist der bewertete Schallpegel keine physiologische oder physikalische Messgröße. Die *DIN EN 61672-1:2014-07* beschreibt die Eigenschaften und Berechnung der Bewertungsfilter. Abbildung 3.1 zeigt Unterschiedliche Bewertungskurven [Moe15] [5]. Auffallend ist hier besonders der Unterschied der A- und C- Bewertung. Während die A-Bewertung tiefe Frequenzen wesentlich schwächer bewertet, ist sie bei der C-Bewertung, über einen großen Teil des hörbaren Frequenzspektrums, weitgehend linear.

[4] http://www.sengpielaudio.com/BerechnungDerBewertungsfilter.pdf
[5] Vgl. Möser, 2015, S.12

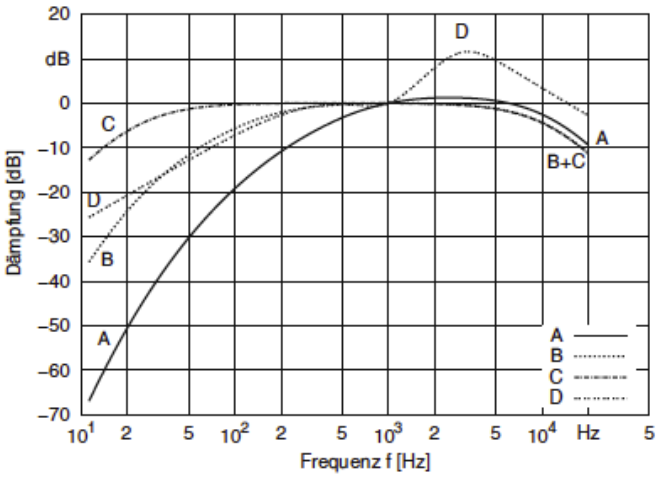

Abbildung 3.1: Frequenz- Bewertungskurven für Schallpegelmessung [6]

In Abbildung 3.2 wird ein weißes Rauschen als Signal verwendet. Von diesem Signal werden Terzpegel, der unbewertete Gesamtpegel (L_{in}) und der A-bewertete Gesamtpegel (A) bestimmt [Moe15] [6].

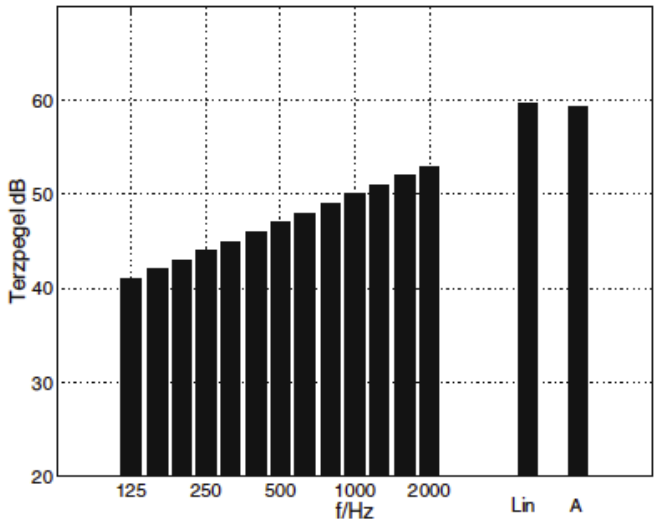

Abbildung 3.2: Terzpegel, unbewerteter und A-bewerteter Pegel von bandbegrenztem weißem Rauschen [6]

Um den Schalldruckpegel auch subjektiv nachempfinden und beurteilen zu können, wird im weitere Verlauf dieser Arbeit, sowohl in der Theorie als auch folgend in der praktischen

[6] Vgl. Möser, 2015, S.12

Umsetzung, wenn nicht explizit auf einen anderen Bewertungsfilter hingewiesen wird, der Bewertungsfilter A verwendet, der sich am meisten mit dem empfundenen Lautstärkepegel deckt.

Auch die Zeitbewertung folgt genormten Vorgaben. Sie beschreibt die Mittelung des Messwerts über einen bestimmten Zeitraum bei zeitlich schwankendem Schalldruckpegel. Häufig verwendet sind die Zeitbewertungen F (fast = schnell), deren Bewertungszeit 125 *ms* beträgt und S (slow = langsam), deren Bewertungszeit 1 *s* beträgt. Bei der Schalldruckpegelmessung von Sprache ist zu beachten, dass Sprache kein durchgehendes Schallereignis, sondern eine Aneinanderreihung von Schallereignissen und Pausen ist. Bei einem Schallpegelmesser, der in zeitlich sehr kurzen Abständen misst, wird ein visualisiertes Signal sehr große Unterschiede der gemessenen Schalldruckpegel aufweisen. Um dem entgegen zu wirken, kann auf die langsame Zeitbewertung zurückgegriffen werden. Abbildung 3.3 zeigt zwei unterschiedliche Zeitbewertungen, die mit einem Pegelschreiber aufgezeichnet wurden.

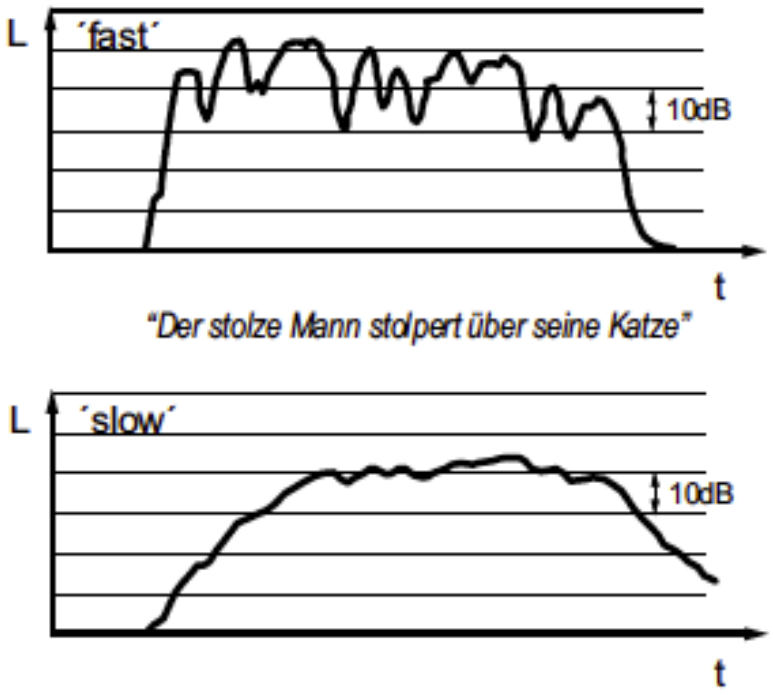

Abbildung 3.3: Schallpegel (L) der Sprache als Funktion der Zeit (t) eines monoton gesprochenen Satzes, aufgezeichnet mit einem Pegelschreiber (Schreibgeschwindigkeit 3 cm/s) mit der Zeitbewertung Fast und Slow [7]

[7] Vgl. Möser, 2010, S.58

Abbildung 3.3 veranschaulicht die starken Schallpegelschwankungen bei der schnellen Zeitbewertung F. In dieser Arbeit ist die Betrachtung gemittelter Spitzenwerte über einen größeren Zeitraum interessant, wofür sich die langsame Zeitbewertung S anbietet.

In einer Entfernung von 60 cm von Schallpegelmesser zum Mund des Sprechers liegt der Schallpegel der gemessenen und gemittelten Spitzenwerte, bei normal laut gesprochener Sprache, bei etwa 60 $dB(A)$ bei Männern und 58 $dB(A)$ bei Frauen. Bei lautem Sprechen erhöht er sich um etwa 6 dB. Bei sehr lautem Sprechen erreicht er Werte von ungefähr 76 $dB(A)$ bei Männern, bzw. 68 $dB(A)$ bei Frauen.Verringert sich der Abstand um die Hälfte auf 30 cm zeigt der Schalldruckpegel eine Erhöhung von ungefähr 4 dB. Die Dynamik kann, durch z.B. leises Murmeln oder lautes Schreien, um etwa 5 dB nach oben oder unten erweitert werden. [Dic14] [8] Somit liegt die Dynamik der Sprache, mit ihren Extremformen, bei ungefähr 26 dB bei Männern und 21 dB bei Frauen. Ohne die Extremformen sind es bei Männer etwa 16 dB und bei Frauen 10 dB.

3.3 Frequenzspektrum der gesprochenen Sprache

Der Frequenzbereich der menschliche Stimme umfasst ein sehr breites Spektrum. Da es je nach Alter, Geschlecht und körperlicher Beschaffenheit stark variieren kann, lässt es sich nur schwer in einen genauen Bereich einordnen. Deutlich macht das bereits die Stimmlippen-Oszillationsfrequenz, die bei Frauen-, Männer- und Kinderstimmen in einem Bereich von 70 - 600 Hz liegen kann [The98] [9]. Der Mittelwert liegt bei Frauen bei 240 Hz und bei Männern bei 120 Hz [9]. Durch die in Kapitel 3.1 beschriebenen Variationen des Stimmapparats lassen sich Frequenzen bis zu ungefähr 10 kHz messen [LSS$^+$07] [10]. Über 10 kHz liegen lediglich Spektralkomponenten von Frikativen (siehe 3.1). Wegen der Komplexität der Spektren bei unterschiedlichen Sprachlauten ist es durchaus sinnvoll sich zur besseren Veranschaulichung auf statistisch ermittelte Durchschnittswerte zu beziehen. Abbildung 3.4 und 3.5 zeigen die durchschnittlichen Spektren von Sprechern und Sprecherinnen bei unterschiedlich lauter Aussprache. Es sei darauf hingewiesen, dass es sich in der Abbildung 3.4 und 3.5 um B-Bewertete Spektren handelt.

[8] Vgl. Dickreiter, 2014, S.70
[9] Vgl. Therhardt, 1998, S.32
[10] Vgl.Lazarus, 2007, S.51

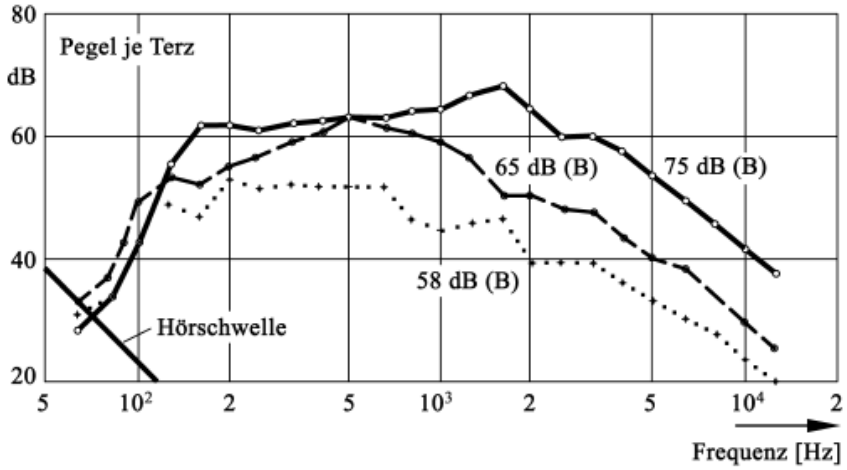

Abbildung 3.4: Durchschnittliche Spektren deutscher Sprache von männlichen Sprechern, für leise, mittellaut und laut gesprochener Sprache, Schalldruckpegel je Terzbereich [11]

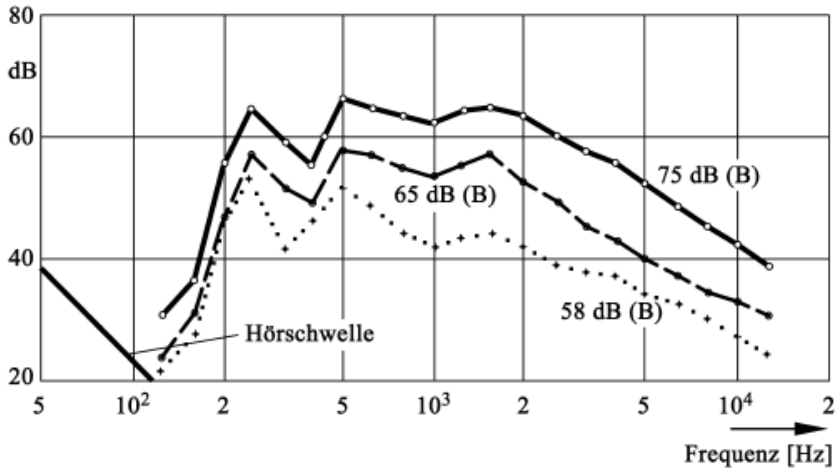

Abbildung 3.5: Durchschnittliche Spektren deutscher Sprache von Sprecherinnen, für leise, mittellaut und laut gesprochener Sprache, Schalldruckpegel je Terzbereich [11]

Wie beschrieben finden sich die ersten Maxima bei der Sprechtonhöhe. Bei Männern liegt sie zwischen 120 und 150 *Hz*, bei Frauen bei ungefähr 240 *Hz*. Weitere Maxima liegen bei beiden Geschlechtern bei 1500 *Hz*, wobei sich nur bei Frauen noch ein weiteres bei 500 *Hz* zeigt. Zu erwähnen sei hier auch, dass wenn sich die Lautstärke der Stimme noch weiter erhebt,

[11] Vgl. Dickreiter, 2014, S.69

wie im Extremfall bei lautem Schreien und Brüllen, eine Verlagerung der Maxima bei beiden Geschlechtern in den Bereich um die 1200 bis 1500 *Hz* stattfindet. [Bri16] [12]

Mitbestimmt werden die Charakteristika vor allem durch unterschiedliche Sprachlaute. In einem direkten Vergleich von Vokalen und Konsonanten zeigen sich enorme Unterschiede im Frequenzverlauf. Vokale sind durch mehrere Frequenzgebiete gekennzeichnet, die Formanten. Ihre Frequenzbereiche bleiben, für einzelne Vokale, unabhängig von der Grundfrequenz erhalten. [LSS+07] [13] Abbildung 3.6 Zeigt Formantfrequenzen einzelner Vokale.

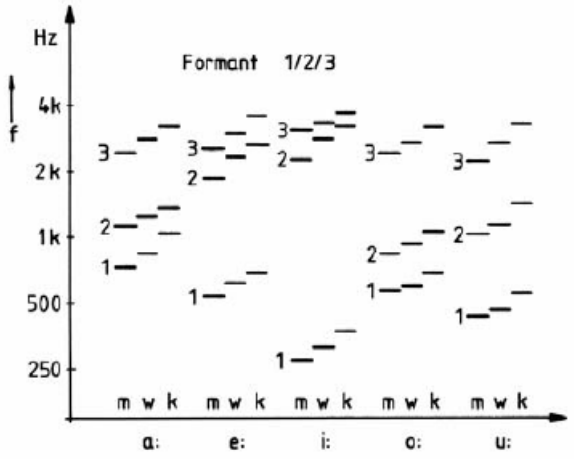

Abbildung 3.6: Charakteristische Frequenzgebiete der 1./2./3./ Formanten von Vokalen von Männern (m), Frauen (w) und Kindern (k) [13]

Das Frequenzspektrum von Konsonanten dagegen umfasst einen wesentlich breiteren Frequenzbereich und fällt unregelmäßig aus.

3.4 Richtcharakteristik des Sprachapparats

Wie in Kapitel 2.1.4 beschrieben, kommt es zu einer Bündelung des abgestrahlten Schalls, wenn die Wellenlänge klein gegenüber der Schallquelle ist. Das betrifft natürlich auch den Sprachapparat. Beeinflusst wird die Richtcharakteristik der Sprachen zum einen durch die Schallabschattung durch den Kopf und zum anderen durch die Trichterwirkung des Mundes. Generell ist der Schallpegel vor dem Mund des Sprechers höher als hinter oder seitlich von

[12] `http://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility`
[13] Vgl.Lazarus, 2007, S.54

ihm. Das gilt sowohl für die horizontale als auch für die vertikale Richtung. Die Schallbündelung nimmt dabei ab 1 kHz enorm zu. Abbildung 3.7 zeigt das horizontale und das vertikale, A-Bewertete, Polardiagramm der menschlichen Stimme, bei einem Meter Entfernung zum Mikrofon. Dabei wurden die Werte für Männer und Frauen gemittelt. Alle Sprecher saßen bei den Aufnahmen. Eine deutliche Abnahme des Schallpegels zur Seite hin und weiter bis unmittelbar hinter den Sprecher ist deutlich zu erkennen. Außerdem lässt sich bei ungefähr 330° vertikaler Darstellung eine Erhöhung des Schallpegels feststellen, wofür die Reflexionen an der Brust als der Verursacher gelten. [Bri16] [14]

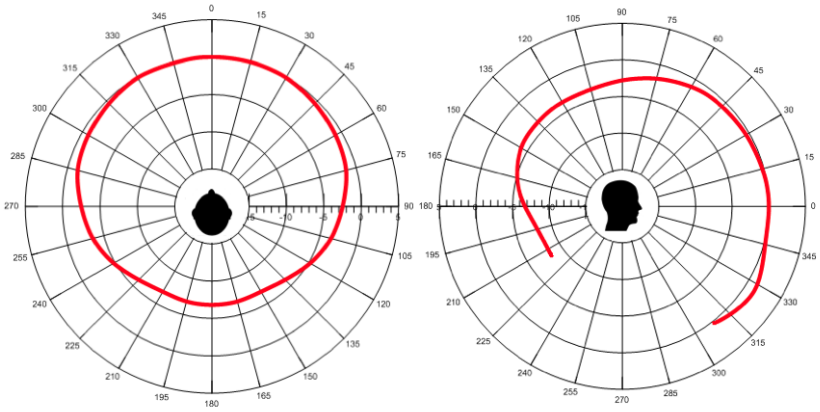

Abbildung 3.7: Polardiagramm der Menschlichen Stimme A-Bewertet von Mann und Frau. [15]

Abbildung 3.8 zeigt die Polardiagramme, die unter den selben Bedingungen erstellt wurden wie die in Abbildung 3.7, mit dem Unterschied, dass nicht die A-Bewerteten, sondern die Schallpegel unterschiedlicher Frequenzbänder untersucht wurden. Die rot gefärbte Linie kennzeichnet das tiefste Frequenzband. Die blaue Linie, die höchste Frequenz. In horizontaler Ebene lässt sich, mit Zunahme der Höhe der Frequenz, eine deutliche Abnahme des Schallpegels zu den Seiten und nach hinten beobachten. In vertikaler Ebene erfährt der Schallpegel, bei hohen Frequenzen, eine sehr starke Abschwächung ab ungefähr 105°. Auffallend ist auch hier wieder die teils starke, durch Reflexionen an der Brust verursachte Anhebung des Schallpegels bei etwa 330°, die aber nicht bei allen Frequenzen gleich stark ausgeprägt ist.

[14] `http://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility`
[15] `http://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility`

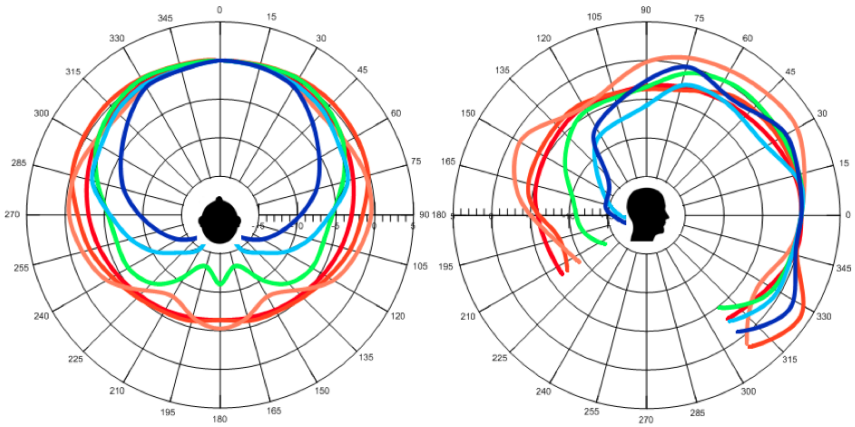

Abbildung 3.8: Polardiagramm der menschlichen Stimme von 160 Hz bis 8 kHz. [16]

3.5 Elemente der Stimme und Sprechweise

Akustisch betrachtet bestehen Sprachäußerungen aus einer Folge von Schallereignissen und Stilleperioden. Aneinander gereihte Kombinationen von Vokalen, Frikativen, Nasalen und Plosivlaute ergeben einzelne Wörter, die durch kleine Pausen getrennt, zu ganzen Sätzen verbunden werden. Betrachten wir die Stimme und die Sprechweise genauer, lässt sie sich in unterschiedliche Klassen einteilen. [Sch82] [17]

- Habituelle Stimmqualitäten
- Individuelle Verlaufsqualitäten
- Formale oder akzessorische Qualitäten

Unter habituellen Stimmqualitäten verstehen sich Eigenschaften bezogen auf den Klang einer Stimme, die mitunter die Grundstruktur der Persönlichkeit einer Stimme aufzeigen. Darin enthalten sind [Sch82] [17]:

Tonhöhe (Stimmlage):

Unter anderem abhängig von der Dimensionierung der Stimmlippen und Größe des Kehlkopfs.

[16] `http://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility`
[17] Vgl. Scherer, 1982, S.138

Lautstärke:
Bedingt durch die Amplitude der Luftschwingungen. Je höher diese ist, desto lauter das erzeugte Schallereignis.

Stimmfülle (Volumen):
Als eine große Stimmfülle empfinden wir ein, meist kombiniert mit einer hohen Amplitude, dichtes Gewebe von sich überlagernden Frequenzen, bestehend aus Grundton und Obertönen. Im Gegensatz dazu empfinden wir eine Stimmfülle als geringer wenn sie eine kleinere Amplitude und ein weniger stark ausgeprägtes Obertonspektrum aufweist.

Klangfarbe:
Sie ist eine Mischung aus Grundton und Partialtönen. Obertonarme Stimmen nehme wir als dunkel war. Stimmen bei denen die Obertöne dominieren dagegen als hell.

Individuelle Verlaufsqualitäten sind Sprachmerkmale, die stark von der individuellen Entwicklung und Auseinandersetzung mit der Umwelt geformt werden. Sie beinhalten:

Sprechtempo:
Gemeint sind damit sowohl die Geschwindigkeit der gesprochenen Worte als auch die Dauer der Pausen zwischen ihnen. Unter anderem lässt sich das Sprechtempo in Silben pro Minute einteilen und kann grob in Bereiche von 100 (sehr langsam) bis zu 400 (sehr schnell) Silben pro Minute eingeteilt werden.

Rhythmischer Ablauf:
Er ist sehr schwer zu erfassen und kann sich sowohl in Mustern der Betonung als auch im Sprechtempo wiederfinden.

Akzentuierung (Hervorhebung, Schwerpunkt):
Akzentuierungen können z.B. Tonhöhenbewegungen sein, wie sie beispielsweise am Ende einer Frage vorkommen. Genauso können es auch Veränderungen der Lautstärke, des Tempos, der Klangfarbe oder des rhythmischen Ablaufs sein.

Artikulation:
Lässt sich grob als Grad der Sprechdeutlichkeit verstehen.

Die formalen Qualitäten beinhalten Kriterien, wie Unterschiede in Sprechstilen, wie sie sich z.B. in Monologen und Dialogen finden. Berücksichtigt werden auch Satzbau und die Wortwahl der sprechenden Person.

3.6 Sprechgeschwindigkeit und Pausen

Weitere charakteristische Merkmale der Sprache sind die Sprechgeschwindigkeit, sowie die Pausen zwischen den Schallereignissen. Messungen von Lazarus et al. [18] ergaben, dass in englischer Sprache deutlich und isoliert gesprochene Einsilber bei T = 0.5 - 0.7 *s* liegen. Vokale und Konsonanten zu Beginn und zum Schluss eines Einsilbers haben eine Sprechdauer von T = 0.2 - 0.4 *s*, T = 0.1 *s* und T = 0.15 - 0.25 *s*. Es stellte sich auch heraus, dass die selben Einsilber im Satz fast doppelt so schnell gesprochen wurden wie isoliert gesprochene Einsilber. [LSS+07] [18] Folgende Werte wurden hierfür ermittelt:

- Kurze Vokale: 62-73 *ms*
- Lange Vokale: 115-140 *ms*
- Konsonanten: 50-120 *ms*
- Plosive: stimmhaft 50-57 *ms* und stimmlos 48-57 *ms*
- Affrikate (Enge Verbindung von Plosiv und Frikativ): stimmhaft 45-65 *ms* und stimmlos 42-39 *ms*

Wichtig zu erwähnen ist hierbei, dass mit ansteigendem Sprechpegel die Sprechgeschwindigkeit oft verlangsamt wird. Damit verändert sich auch die Dauer der Laute. Des weiteren ist die Sprechgeschwindigkeit sehr stark von Sprechweise, Situation, emotionaler Lage und natürlich auch vom Sprecher selbst abhängig, wie sich im Verlauf dieser Arbeit herausstellen wird. Pause bezeichnet den Zeitraum, indem der Spechpegel praktisch verschwindet. Sie weisen eine Größenordnung von 0.1-3 *s* auf.

3.7 Verdeckung von Sprache durch Geräusche

Finden außer der Sprache noch weiter Schallereignisse andere Schallquellen statt, überlagern sich diese Signale. Die Schallereignisse, die von anderen Schallquellen verursacht werden, können dabei das Sprachsignal verdecken. Die verdeckende Wirkung der Geräusche richtet sich dabei nach Intensität, Frequenzbereich und Zeitstruktur. [LSS+07] [19] Je nach dem wie diese ausfallen, können sie sich negativ auf die Sprachverständlichkeit auswirken. Abbildung 3.9 zeigt wie sich die Sprachverständlichkeit gegenüber der S/N verhält. Dabei wird ersichtlich, dass bis zu einem gewissen Grad, auch bei negativer S/N, bei der das Geräusch 10 dB lauter ist als der Sprechpegel, Sprache vom Hörer verstanden werden kann.

[18] Vgl. Lazarus, 2007, S.58
[19] Vgl. Lazarus, 2007, S.154

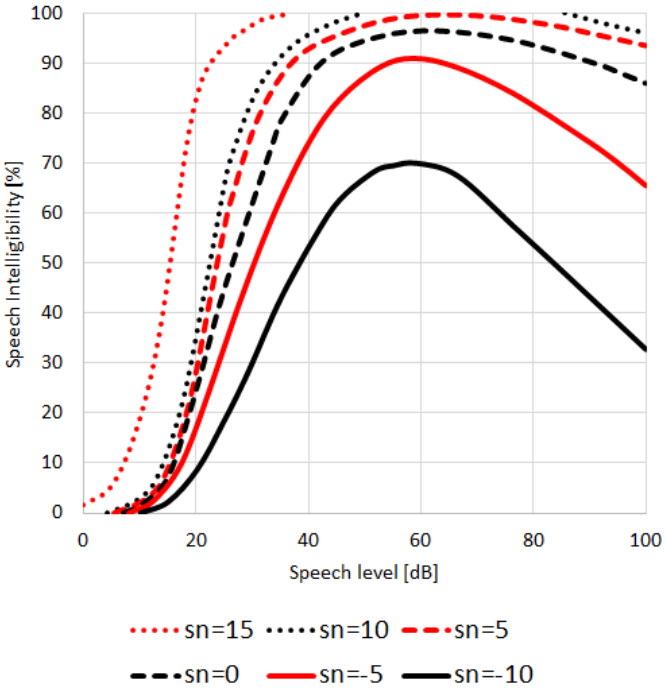

Abbildung 3.9: Sprachverständlichkeit gegenüber S/N [20]

[20] http://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility

4 Konzeption einer virtuellen Konferenz

4.1 Konferenz?

Bevor mit der Erstellung eines Konzepts für eine virtuelle Konferenz begonnen werden kann, muss an dieser Stelle zuerst eine Eingrenzung des Begriffs stattfinden. Der Begriff Konferenz, dessen Ursprung sich im Lateinischen findet (conferre), bedeutet wörtlich übersetzt „zusammentragen, vergleichen"[Wik16b] [1]. Heute wird er für die Zusammenkunft von Personen zur Beratung oder Austausch von Informationen verwendet [Dud17] [2]. Angaben im Bezug auf Anzahl der Personen, räumliche Umgebung oder verwendete Technologien, beinhaltet dieser Begriff nicht. Eine Konferenz kann also ein persönliches Gespräch zweier Personen in einem kleinen Raum, aber auch eine große Veranstaltung mit sehr viel mehr Personen, in einem sehr viel größeren Raum sein. Um bei großen Konferenzen gewährleisten zu können, dass alle anwesenden Zuhörer den Vortragenden gut verstehen können, ist es notwendig, das Sprachsignal mittels eines Mikrofon-Lautsprecher-Systems zu verstärken.

Abbildung 4.1: Open Space Konferenz Kassel [3]

Eine Konferenz kann aber auch ein Geschehen sein, bei dem sich die Teilnehmer an räumlich getrennten Orten befinden, die über Kommunikationstechniken miteinander kommunizieren, wie es z.B. bei Telefonkonferenzen der Fall ist.

[1] https://de.wikipedia.org/wiki/Konferenz
[2] http://www.duden.de/rechtschreibung/Konferenz
[3] http://www.kassel-1100.de/aktuelles/bildergalerien/bilder/106418/index.html

Abbildung 4.2: Bsp: Telefonkonferenz [4]

4.2 Eingrenzung

Das Mikrofonsystem, das mit dieser virtuellen Konferenz getestet wird, zielt mehr auf kleinere Konferenzen, wie z.B. der in Abbildung 4.2 abgebildeten Telefonkonferenz, ab. Speziell zu Beginn der Entwicklung eines solchen Systems ist es sinnvoll, die Anzahl der sprechenden Personen und der auftretenden Störgeräusche einzugrenzen. In dieser Arbeit wird die Anzahl der anwesenden, sprechenden Personen auf 4 begrenzt. Diese nehmen unterschiedliche Rollen in der Konferenz ein. Für die Wiedergabe über Lautsprecher wird später jede Person einem Lautsprecher zugeordnet:

- Person 1: Sie übernimmt die Rolle der aktuell vortragenden Person
- Person 2: Sie wird von Person 1 angesprochen und führt mit ihr einen kurzen Dialog
- Person 3 und 4: Sie führen im Hintergrund eine leise Unterhaltung

Die Reihenfolge der gesprochenen Sätze wird so festgelegt, dass sich Person 1 und 2 nicht ins Wort fallen und somit keine Überlagerung der beiden Sprachsignale stattfindet. Der Dialog der Personen 3 und 4 im Hintergrund dagegen soll sich mit der Sprache der Personen 1 und 2 überlagern. Eine räumliche Eingrenzung wird nicht explizit definiert, jedoch zielt auch die Erstellung des Sounddesigns und der Atmos auf eine kleine Konferenz ab, z.B. eine Telefonkonferenz.

[4] `http://zeitgewinn-hamburg.de/telefonkonferenz-richtigen-anbieter-finden/`

4.3 Aufnahme

4.3.1 Aus welchen Geräuschen besteht eine Konferenz

Ausgehend davon, dass eine Konferenz an den unterschiedlichsten Orte in unterschiedlichen Dimensionen stattfinden kann, wird sehr schnell klar, dass es einen gigantischen Umfang an möglichen auftretenden Geräuschen geben kann. Wichtig ist jedoch zuerst einmal zu definieren, was denn von einem Konferenzmikrofon in erster Linie übertragen werden soll, nämlich die Stimme der sprechenden Person oder Personen. Wie in Kapitel 3 beschrieben, ist die Sprache ein sehr komplexes, dynamisches Instrument, das möglichst in voller Dynamik und dem gesamten Frequenzbereich vom Mikrofon erfasst werden soll. Im besten Fall finden außer dem Sprachsignal keine anderen Schallereignisse statt, die das Sprachsignal überlagern. Davon ist aber nicht auszugehen. Aus diesem Grund werden in Kapitel 4.3.4 einzeln auftretende Störgeräusche definiert, die sich in diesem Szenario störend mit dem Sprachsignal überlagern sollen.

Ein weiteres Geräusch, beziehungsweise Gemisch aus Geräuschen, ist das Grundrauschen des Raumes. Es setzt sich aus unterschiedlichen Geräuschen, die von Schallquellen im Raum, wie z.B. Lüfter eines PCs und von Schallquellen außerhalb des Raumes erzeugt werden, wie z.B. Wind oder vorbeifahrende Autos, zusammen. Außer in speziell konstruierten, akustisch entkoppelten Räumen, ist immer ein deutlich wahrnehmbares und auch messbares Grundrauschen vorhanden. Wird die virtuelle Konferenz nicht in solch einem Raum installiert ist somit schon eine Atmo vorhanden. Dennoch werden für Testzwecke unterschiedliche Atmos aufgenommen, die bei Bedarf zugemischt werden können.

4.3.2 Auswahl der Sprecher und Sprachen

Die menschliche Stimme ist, in dieser Arbeit, die wichtigste Schallquelle. Da, wie in Kapitel 3 beschrieben, die akustischen Eigenschaften der Stimme stark variieren können, ist es wichtig möglichst viele unterschiedliche Stimmen aufzunehmen. Für diese Arbeit werden deshalb Sprecher ausgesucht, die sich möglichst in habitueller Stimmqualität und individueller Verlaufsqualität (Kapitel 3.5) unterscheiden. Große Unterschiede habitueller Stimmqualität zeigen sich bei Mann und Frau. Besonders betrifft das die Tonhöhe sowie die Klangfarbe. Betreffend der individuellen Verlaufsqualität werden solche Sprecher ausgesucht, die sich in Ihrem Sprechtempo deutlich unterscheiden.

Da der Verfasser dieser Arbeit sich im deutschsprachigen Raum befindet, werden die meisten Sprecher auch auf Deutsch einsprechen. Die Dialekte variieren dabei nicht sehr stark und haben ihren Ursprung meist in der süddeutschen Region. In einem Ranking der am meisten

als Muttersprache gesprochenen Sprachen befindet sich Deutschland allerdings auf Rang 11 [noa15] [5]. Da in dieser Arbeit nicht davon ausgegangen wird, dass ein Konferenzmikrofon speziell nur für den deutschsprachigen Raum entwickelt wird, ist es wichtig auch andere Sprachen zu berücksichtigen. Sehr interessant sind dabei die Sprachen, die sich in der Phonetik stark unterscheiden, wie es z.B. in der chinesischen und arabischen Sprache der Fall ist.

Chinesisch ist die am meisten als Muttersprache gesprochene Sprache. [noa15] [5]Die Hochsprache basiert auf dem nordchinesischen Dialekt Mandarin. Chinesisch ist eine sehr tonale Sprache. Inklusive Umlaute beinhaltet die deutsche Sprache 8 Vokale. Im chinesischen sind es dagegen 21 Vokale. Hinzu kommen die 4 unterschiedlichen Betonungen von Silben. Mit ihnen entstehen sehr tonale Klangabläufe, die die chinesische Sprache sehr melodisch klingen lassen.

Ein starker Kontrast zur chinesischen Sprache ist die arabische Sprache, die Platz 5 der am meisten als Muttersprache gesprochenen Sprachen belegt. Im Hocharabischen existieren lediglich die 3 Vokale a, i und u. Diese werden jedoch von den sie umgebenden Konsonanten beeinflusst und variiert. Hocharabisch verfügt dafür über 28 Konsonantenphoneme, von denen sich ganze 13 zu den Frikativen zählen. Die arabische Sprache ist stark durch Laute geprägt, die durch Zähne, Gaumen und Kehle gebildet werden, wodurch ein sehr rauer, kehliger Klangeindruck entsteht.

Gerade diese genannten Unterschiede sind für das Testen von Konferenzsystemen wertvoll. Zusätzlich werden noch Sprecher in spanischer Sprache aufgenommen. Spanisch belegt Platz 2 im Ranking der am meisten als Muttersprache gesprochenen Sprachen [noa15] [5]. Zwar ähnelt sie, in Anzahl an Vokalen und Konsonanten, sehr der deutschen Sprache, weist aber teils starke Unterschiede in der Aussprache auf. Im Spanischen werden viele Reibe- und Dentallaute verwendet.

Der in Kapitel 4.3.3 erstellte Text wird zuerst in deutscher Sprache entworfen und anschließend, so weit es möglich ist, in die eben genannten Sprachen übersetzt.

4.3.3 Erstellung eines Sprechertextes

Vor dem Entwurf des eigentlichen Sprechertextes wurde ein Versuch durchgeführt, um herauszufinden, wie sich ungeübte Sprecher vor dem Mikrofon bewegen. Dazu wurde ein Notenständer zwischen Sprecher und Mikrofon platziert, auf den ein Ausschnitt eines Dialogs gelegt wurde. Aufgabe der Sprecher war es nun den Text in das Mikrofon zu sprechen. Ein

[5] Vgl. Wunderlich, 2015, S.36

sehr großes Problem bei der Sprachaufnahme von ungeübten Sprechern ist, dass sie dazu neigen sich zum Text hin zu richten, was sich teils sehr stark auf das über das Mikrofon aufgenommene Signal auswirkt. Das war auch der Fall, wenn sie darum gebeten wurden, sich einen Satz zu merken und diesen anschließend in das Mikrofon zu sprechen. Anschließend wurde der Notenständer seitlich, außerhalb des Blickwinkels platziert. Die Sprecher wurden darum gebeten sich einen Satz zu merken, ihren Blick auf eine Markierung an der Wand hinter dem Mikrofon zu richten und den Satz dann zu sprechen. Die Ergebnisse verbesserten sich durch diese Maßnahme enorm. Je nach Satzlänge waren allerdings mehrere Anläufe nötig bis der ganze Satz vollständig gesprochen werden konnte.

Auffallend ist zudem, dass bei ungeübten Sprechern sehr deutlich hörbar wird, ob sie gerade frei sprechen oder ablesen. Ein typisches Beispiel ist eine Tonhöhenbewegung nach unten am Ende eines Satzes. Das verbesserte sich nachdem der Text einige Male wiederholt gesprochen wurde. Wie viel Wiederholungen dazu nötig sind ist sehr stark abhängig von der Person und eine große Verbesserung auch nicht immer garantiert. Trotz der Verbesserung lässt sich, bei den meisten ungeübten Sprechern, ein Unterschied zwischen frei gesprochener und abgelesener Sprache feststellen, auch dann, wenn der Text sehr zusammenhängend geschrieben ist.

Für das Erstellen des Textes war der Zusammenhang des Inhalts deshalb nicht das ausschlaggebende Kriterium. Der Fokus wird darum auf die verwendeten Wörter an sich gelegt. Der Wortumfang, der in einer Konferenz vorkommen kann, ist sehr stark davon abhängig wer worüber spricht. Wichtig für diese Arbeit ist es Wörter zu verwenden, die, unabhängig von z.B. Fachbereichen, wahrscheinlich vorkommen können. Deshalb werden für die Erstellung des Textes Wörter verwendet, die laut Duden [Dudum] [6] und gut1.de [eaum] [7], zu den 100 am häufigsten vorkommenden Wörter in deutschsprachigen Texten zählen. Um aus diesen zusammenhängende Sätze zu konstruieren gibt es vereinzelt Abweichungen, wie auch Konjugationen einzelner Wörter.

Der finale Text beinhaltet alle Sätze, die von den in Kapitel 4.2 beschriebenen Personen gesprochen werden, inklusive Dialog und der leisen Unterhaltung im Hintergrund. Bei den Aufnahmen wird der gesamte Text von jedem Sprecher eingesprochen. Aus diesen Aufnahmen werden, in der Session, anschließend der Dialog, wie auch die leise Unterhaltung im Hintergrund, aus unterschiedlichen Sprechern zusammengeschnitten (Sprechertext in Anhang 11.1).

[6] http://www.duden.de/sprachwissen/sprachratgeber/die-haeufigsten-woerter-in-deutschsprachigen-texten

[7] http://www.gut1.de/grundwortschatz/grundwortschatz_100.html

4.3.4 Auswahl der Störgeräusch

Die erste Idee Störgeräusche in die virtuelle Konferenz einzubauen bestand darin, das Sprachsignal möglichst zu stören. Dieser Ansatz stößt allerdings auf ein Problem, das Lombard [Lom11] [8] erforschte. Er bemerkte, dass Personen mit gesunden Ohren, die über Kopfhörer mit Geräuschen beschallt wurden, anfingen lauter zu sprechen oder sogar zu schreien. Das ist ein Nachweis dafür, dass sprechende Personen mit gesunden Ohren die Lautstärke ihrer Sprache automatisch an den Geräuschpegel ihrer Umgebung anpassen. Die Aufnahmen der Sprecher für die virtuelle Konferenz finden allerdings in ruhiger Umgebung statt, um möglichst störungsfreie Signale zu erhalten. Werden nun Geräusche in die virtuelle Konferenz integriert, auf Grund derer die Sprecher ihren Sprachpegel erhöhen, würde dadurch der Realitätsgrad verfälscht.

Deshalb werden für diese Konferenz bestimmte Geräusche definiert, die nach Erfahrung des Autors oft in Konferenzen mit wenigen Personen vorkommen. Diese haben keine wissenschaftliche Festigkeit. Folgende Geräusche werden für die Konferenz aufgenommen und anschließend in die Session integriert:

- Bewegungen sitzend auf einem Schreibtischstuhl mit Rollen
- Schreiben auf Papier
- Blättern auf die nächste Seite eines Blocks
- Abreißen eines Blatts aus dem Block
- Zerknüllen eines Papiers
- Abstellen einer Tasse auf einen Tisch
- Abstellen einer Tasse auf eine Untertasse
- Rühren in der vollen Kaffeetasse
- Ablegen eines Löffels
- Klingeln und Vibrieren eines Mobiltelefons
- Kleidungsgeräusche
- Räuspern und husten

Den Nutzern der später erstellten Session steht es natürlich frei den Umfang der Geräusche nach Belieben zu erweitern.

[8] Vgl. Lombard, 1911

4.3.5 Auswahl der Atmo

Für diese Arbeit werden zwei Räume gesucht, die sich in einer lauten Umgebung befinden. Wichtig ist dabei, dass sie sowohl mit Fenstern als auch Türen ausgestattet sind, die sich während der Aufnahme öffnen und schließen lassen. Der erste Raum befindet sich nahe einer Baustelle. Der zweit befindet sich in der *HdM* Stuttgart nahe der Kantine.

4.4 Entzerrung

Ziel der Erstellung der virtuellen Konferenz ist es, eine reale Situation möglichst authentisch zu simulieren. In dieser Arbeit ist es besonders wichtig, dass die aufgenommene Sprache realistisch wiedergegeben wird. Um das erreichen zu können ist es notwendig, die Eigenschaften der Aufnahmekette und des Wiedergabesystems zu ermitteln. Im besten Fall weisen Aufnahmekette und Wiedergabesystem einen exakt linearen Frequenzgang und ein extrem schnelles Impulsverhalten auf. Der Aufnahmeraum sollte keine reflektierenden Wände haben. Das wären die besten Voraussetzungen für optimale Aufnahme und Wiedergabe. Technisch ist dies aber kaum umzusetzen. Für diese Arbeit stehen weder optimale Aufnahme- und Wiedergabegeräte, noch ein reflexionsarmer Raum zur Verfügung. Deshalb entstehen, bei der Aufnahme wie auch bei der Wiedergabe, Verfälschungen gegenüber der ursprünglichen Schallquelle. Besonders gravierend sind dabei die Veränderungen des Frequenzganges.

Um dennoch möglichst nahe an die originalen Sprachsignale zu kommen sind einige Eingriffe notwendig. Zum einen wird versucht die frühe Reflexionen und den Diffusschallanteil, die durch Reflexionen an den Wänden entstehen, durch Faltung mit einer aufgenommenen Impulsantwort von den aufgenommenen Sprachsignalen zu entfernen. Anschließend werden die Signale, anhand der ermittelten Frequenzgänge von Aufnahme- und Wiedergabesystem, mit digitalen Filtern entzerrt. Um den Einfluss dieser Eingriffe zu überprüfen, werden die bearbeiteten Signale wieder aufgenommen und analysiert. Die Aufnahmen finden unter den selben Bedingungen statt, die auch bei den Aufnahmen der Sprecher herrschten. Anstelle der Sprecher wird ein Lautsprecher die Signale wiedergeben.

Da die Aufnahmen der Störgeräusche unter unterschiedlichen Bedingungen stattfanden, wird auf das Entfernen des Raumanteils verzichtet. Die Bearbeitung mittels Filter wird dagegen wie oben durchgeführt.

4.5 Session

Als *DAW* wird die Software *Pro Tools* von Avid verwendet [9]. Die bearbeiteten Audiodateien werden anschließend in einer Session organisiert. Zunächst wird der Ablauf für eine Kombination von Person 1, 2, 3 und 4 (4.2) festgelegt. Da alle Sprecher bei der Aufnahme den selben Text einsprechen, können Dialog und leise Unterhaltung im Hintergrund von unterschiedlichen Sprechern zusammengeschnitten werden. In dem Zeitraum, in dem diese Personen sprechen, werden auch alle Störgeräusche einmal abgespielt. Dieser Ablauf stellt eine Konferenz dar, deren Dauer ungefähr 50 Sekunden betragen wird. Auf dieser Basis werden anschließend unterschiedliche Kombinationen von Sprechern erstellt. Die Störgeräusche werden in jeder Konferenz an die zeitlich gleiche Stelle gesetzt. Für die finale Session werden die so entstandenen Konferenzen hintereinander zu einer großen Konferenz zusammengefügt.

Da der Ablauf der einzelnen Konferenzen immer gleich bleibt, kann so geprüft werden, wie das zu testende Mikrofonsystem sich bei unterschiedlichen Stimmen verhält.

4.6 Wiedergabe

Eine reale Konferenz besteht, alleine schon auf die Sprecher bezogen, aus mehreren beweglichen Schallquellen. Ein Beispiel dafür wurde schon in Kapitel 4.3.3 beschrieben. Hier neigt sich der Sprecher beim Ablesen zum Dokument hin und richtet anschließend den Blick wieder Richtung Mikrofon. Ebenso können Bewegungen zur Seite, zu anderen Teilnehmern einer Konferenz, vorkommen. Für eine realistische Simulation müssten also, alleine für jeden hier vorgesehenen Sprecher, bewegliche Lautsprecher gefertigt und Bewegungsabläufe definiert werden. Die Umsetzung ist allerdings sehr kostspielig, weshalb in dieser Arbeit überwiegend statisch platzierte Lautsprecher verwendet werden.

In dieser Arbeit werden keine Bewegungsabläufe durch Panning simuliert. Bei einem Menschen würde das zwar wegen des in Kapitel 2.1.5 beschriebenen Phänomens funktionieren, bei Mikrofonsystemen aber nicht.

4.6.1 Lautsprecher-Setup

Die bearbeiteten Audiodateien werde in einer Session organisiert und definierten Ausgängen zugewiesen. Jede Spur wird dabei einem Ausgang und somit einem Lautsprecher zugewie-

[9] `https://www.avid.com/de/pro-tools`

sen. Das Setup wird sich also aus vielen Mono-Schallquellen zusammensetzen, was auch der Realität entspricht. Wichtig ist an dieser Stelle auch zu berücksichtigen, wo im Raum sich das zu testende Mikrofonsystem befindet. Konferenzsysteme, speziell Mikrofone für Telefonkonferenzen, befinden sich meist auf einem Tisch. Für die hier vorgesehenen Tests befindet sich das Mikrofon in der Mitte eines Tisches. Von dieser Situation ausgehend wird das Lautsprecher-Setup in drei Schichten eingeteilt, die sich an der räumlichen Anordnung der wiederzugebenden Klänge orientiert.

Schicht 1:

Die Schicht, die sich am nächsten zum Mikrofon befindet. Viele der aufgenommenen Geräusche sind solche, die auf einem Tisch stattfinden werden. Dazu zählen z.B. das Schreiben auf Papier oder Geräusche, die mit einer Tasse erzeugt werden (4.3.4). Für diese Schicht wird lediglich ein Lautsprecher verwendet. Dieser wird auf dem Tisch, nahe der Tischkante positioniert.

Schicht 2:

Die Schicht mit den Lautsprechern, über die die Sprache wiedergegeben wird. Die Lautsprecher werden ungefähr auf Kopfhöhe einer sitzenden Person nahe der Tischkante positioniert. Neben der Sprache werden von ihnen auch Geräusche von Kleidungsstücken und Schlürfen wiedergegeben. Für jeden simulierten Sprecher ist ein Lautsprecher vorgesehen.

Schicht 3:

Die Schicht über die Atmos und Geräusche wie Tür/Fenster öffnen und schließen wiedergegeben werden. Für diese Schicht sind vier Lautsprecher vorgesehen, die nach dem Prinzip der Vierkanalstereofonie wiedergegeben werden.

Alle Lautsprecher werden in Richtung Mikrofon ausgerichtet. Die hier definierten drei Schichten bilden den Grundaufbau, den der Autor vorgesehen hat. Es steht dem Nutzer natürlich frei die Anordnung nach eigenen Vorstellungen zu verändern.

4.6.2 Beweglicher Lautsprecher

Wie schon in Kapitel 4.6 beschrieben besteht das Problem, dass sich Personen beim Sprechen bewegen können. Die bisher vorgesehenen Lautsprecher befinden sich aber fest montiert auf ihren Positionen. Dadurch entsteht eine Abweichung zu einer realen Konferenz. Um dem entgegen zu wirken wird ein Drehteller entworfen, mit dem Kopfbewegungen zur Seite simuliert werden können. Damit kann geprüft werden wie sich das zu testenden Mikrofonsystem verhält, wenn sich ein simulierter Sprecher von der optimalen, zum Mikrofon

hin gerichteten Position, herausdreht. Dabei muss natürlich berücksichtigt werden, dass sich die Richtwirkung von Mensch und Lautsprecher unterscheiden.

Wichtig für diese Arbeit ist, dass das ganze Projekt in einer Session organisiert wird. Deshalb wird nach einer Methode gesucht, wie sich der bewegliche Lautsprecher über die Session ansteuern lässt (8.3).

5 Die Aufnahme

5.1 Der Aufnahmeraum

Die Aufnahmen der Sprecher und der meisten Störgeräusche fanden im Studio E der Hochschule der Medien Stuttgart statt. Die Grundfläche dieses Raumes beträgt 12 m^2 und sein Raumvolumen 32,4 m^3. Die Wände bestehen aus einer Kombination von unterschiedlich stark absorbierenden Oberflächen. Außerdem wurden beim Bau des Studios rechte Winkel vermieden. Die Nachhallzeit beträgt 0,2 s.[Flo16] [1] Anhand der Formel 5.2 lässt sich mit diesen Werten der Hallradius ermitteln. Ohne Berücksichtigung des Bündelungsgrades von Schallquelle und Schallempfänger, beträgt der Hallradius 0,73 m.

5.2 Auswahl des Sprechermikrofons

Ziel der Sprachaufnahme ist es die Sprache möglichst natürlich aufzunehmen. Wichtig ist auch, dass so wenig Reflexionen wie möglich aufgenommen werden. Hierfür wurden vor den eigentlichen Sprachaufnahmen Mikrofone mit unterschiedlichen Empfängerprinzipien und Richtcharakteristika getestet. Bei allen hier getesteten Mikrofone handelt es sich um elektrostatische Wandler oder auch Kondensatormikrofone genannt. Die Auswertung des Tests erfolgte rein subjektiv.

Schoeps MK2:

Das Schoeps MK 2 [2] ist ein Druckempfänger. Bei Druckempfängern bildet eine Mikrofonmembran eine Seite einer akustisch dichten Mikrofonkapsel [Dic03] [3]. Die Membran wird durch den Schalldruck bewegt. Dabei ist sie für alle Schalleinfallsrichtungen gleich empfindlich, solange das Mikrofon wesentlich kleiner als die Wellenlänge ist. Daraus resultiert eine Kugelcharakteristik über einen sehr großen Frequenzbereich. Das Mikrofon zeichnet eine sehr ausgewogener Frequenzgang und ein sehr niedriger Rauschwert aus. Sprache wird mit diesem Mikrofon sehr genau abgebildet.

Schoeps CCM 4V:

Das Schoeps CCM 4V [4] ist ein Druckgradientenempfänger. Bei Druckgradientenemp-

[1] Vgl. Kontny, 2016, S.30
[2] `http://www.schoeps.de/de/products/mk2`
[3] Vgl. Dickreiter, 2003, S.90
[4] `http://www.schoeps.de/de/products/ccm4v`

fänger wird die Membran durch Druckdifferenzen zwischen Membranvorderseite und Membranrückseite bewegt. Das Mikrofon ist so konstruiert, dass daraus eine Nierencharakteristik resultiert. Eine leichte Anhebung bei 10 *kHz* trägt, laut Hersteller, zur besseren Sprachverständlichkeit bei, weshalb es sich besonders für Sprachaufnahmen eignet.

Sennheiser MKH416

Das Sennheiser MKH416 [5] ist ein Mikrofon mit einer stark ausgeprägten Richtcharakteristik, die sich zwischen Superniere und Keule bewegt. Das Mikrofon zeichnet aus, dass es besonders von vorne empfindlich ist, aber nicht von den Seiten und nur wenig von hinten. Auch bei diesem Mikrofon gibt es eine leichte Anhebung des Frequenzgangs bei ungefähr 10 kHz.

Mit allen Mikrofonen wurden Tests in 15 *cm* und 60 *cm* Abstand zum Sprecher durchgeführt. Das hat den Grund, dass bei Druckgradientenempfängern, bei Entfernungen ungefähr kleiner als 60 *cm*, der Nahbesprechungseffekt auftritt. Dieser wirkt sich in einer Anhebung tiefer Frequenzen aus. Beim Druckempfänger tritt dieser nicht auf. Zur Vermeidung von Popgeräuschen wurde ein Popschutz eingesetzt, der auch als Abstandhalter fungierte.

Für einen direkten Vergleich wurde der in Kapitel 4.3.3 erstellte Text mit unterschiedlichen Mikrofonen und unterschiedlichem Abstand von Mikrofon zum Sprecher aufgenommen. Der Nahbesprechungseffekt, der bei 15 *cm* Abstand bei den Mikrofonen Schoeps CCM 4V und Sennheiser MKH416 auftritt, wurde mit einem digitalen Filter ausgeglichen. Als Referenz dienten hierbei die Aufnahmen mit den gleichen Mikrofonen bei 60 *cm* Abstand. Auch Pegelunterschiede wurden ausgeglichen. Anschließend wurde das Mikrofon aus dem Aufnahmeraum entfernt. Der Verfasser dieser Arbeit stellte sich an die Stelle, an der vorher das Mikrofon stand. Der Kopf war dabei nach vorne zum Sprecher hin gerichtet. Der Abstand zwischen dem Verfasser dieser Arbeit und dem Sprecher betrug 60 *cm*. Nun wurde abwechselnd die aufgenommene Sprache über Kopfhörer und der „real“vor dem Autor stehende Sprecher angehört und verglichen. Zu beachten ist hierbei, dass die Unterschiede oder Ähnlichkeiten nur näherungsweise erfasst werden können, was einer Reihe an Gründen zu verschulden ist. Zum einen wurde bei den Aufnahmen jeweils nur ein Mikrofon verwendet. Der Mensch hat allerdings mit seinen zwei Ohren sozusagen zwei Mikrofone. Am Kopf treten außerdem unterschiedliche Beugungs- und Reflexionseffekte auf, wie es bei einem Mikrofon der Fall ist. Auch die Reflexionen des Raumes werden im Gegensatz zum Mikrofon aus zwei Richtungen empfangen. Ein weiterer Grund ist, dass der verwendete Kopfhörer selbst auch akustische Eigenschaften hat, die das wiedergegebene Signal beeinflussen. Dennoch kann so

[5] http://de-de.sennheiser.com/richtrohrmikrofon-reportage-spielfilm-funk-fernsehen-mkh-416-p48u3

ein Eindruck gewonnen werden, welches der getesteten Mikrofone dem Original am nächsten kommt.

Das Schoeps MK2 bildet die Sprache sehr authentisch ab. Leider bildet es die im Raum auftretenden Reflexionen, aus fast allen Richtungen, ebenso genau ab. Bei kurzer Entfernung nimmt dieser Effekt zwar wegen des größeren Schallpegelabstands von Direkt- zu Diffusschall ab, dennoch sind die Reflexionen deutlich hörbar.

Im Gegensatz dazu steht das Sennheiser MKH416. Durch seine starke Richtwirkung werden weniger Reflexionen aufgenommen. Im Gegensatz zum MK2 bildet es die Sprache allerdings nicht so authentisch ab. Besonders bei Frequenzen ab 4 kHz treten Klangverfälschungen auf.

Das Schoeps CCM 4V klingt ähnlich dem MK2 sehr authentisch. Lediglich die Anhebung bei 10 kHz führt zu Verfälschungen. Dafür nimmt es die Reflexionen des Raums weniger stark auf, wie es bei dem MK2 der Fall ist. Aus diesem Grund wird im weiteren Verlauf dieser Arbeit das Schoeps CCM 4V verwendet. Die Anhebung bei 10 kHz wird in der Nachbearbeitung ausgeglichen.

5.3 Abstand des Sprechers zum Mikrofon

Anhand der in Kapitel 5.2 durchgeführten Probeaufnahmen wurde untersucht, welcher Abstand, von Mikrofon zu Sprecher, sich am besten dazu eignet, die Sprache möglichst authentisch aufzunehmen. Beim Schoeps CCM 4V ist, bei einem Abstand von 15 cm, eine deutliche Anhebung tiefer Frequenzen um ungefähr 200Hz zu hören. Für eine möglichst authentische Aufnahme ist dies zunächst ein Nachteil. Der Vorteil bei nahen Aufnahmen ist, dass der Schallpegel des Direktschalls höher ist als bei Aufnahmen in 60cm Abstand. Dadurch vergrößert sich der Abstand des Schallpegels von direktem zu diffusem Schall. Um der Anhebung tiefer Frequenzen entgegen zu wirken wurde ein digitales Filter verwendet. Damit kann der Effekt zwar abgeschwächt werden, trotzdem klingt das Signal nicht so natürlich, wie es beim Abstand von 60cm der Fall ist. Ein weiteres Problem bei kurzen Entfernungen zum Mikrofon ist, dass Schmatz und Klickgeräusche des Munds sehr präzise erfasst werden.

Bei einem Abstand von 60cm klingt die Sprache wesentlich natürlicher. Der Nachteil ist allerdings, dass der Abstand des Schallpegels von Direkt- und Diffusschall kleiner ist als es bei Aufnahmen mit 15cm Abstand der Fall ist. Dennoch werden bei 60cm Abstand die authentischeren Ergebnisse erzielt, weshalb alle Sprecher, die im Rahmen dieser Arbeit aufgenommen wurden, aus dieser Entfernung aufgenommen wurden.

5.4 Die Aufnahmekette

Von der Schallquelle ausgehend bestand die gesamte Aufnahmekette aus folgenden Komponenten:

- Popschutz von König & Meyer
- Mikrofon Schoeps CCM 4V
- Audiointerface RME Fireface UFX
- Digital Audio Workstation ProTools 12

5.5 Aufnahme

5.5.1 Aufnahme der Sprecher

Alle Sprachaufnahmen wurden in dem in Kapitel 5.1 beschriebenen Aufnahmeraum durchgeführt. Da nicht alle Aufnahmen an einem Tag durchgeführt werden konnten, wurden die vorgesehene Position des Mikrofons und des Sprechers ausgemessen. Für die in Kapitel 6 beschriebene Nachbearbeitung ist es wichtig, dass Mikrofon und Sprecher immer an der selben Position stehen. Lediglich die Höhe, in der das Mikrofon aufgestellt wurde variierte je nach Größe des Sprechers. Mund des Sprechers und Mikrofon waren stets auf gleicher Höhe. Für jeden Sprecher wurde eine individuelle Markierung an der Wand platziert, auf die sie sich bei der Aufnahme hin ausrichten sollten.

Bei der ersten Aufnahme erhielt der Sprecher den Text im Studio kurz vor der Aufnahme. Wie sich herausstellte, verursachte die ungewöhnliche Satzkonstellation des in Kapitel 4.3.3 erstellten Textes anfänglich Probleme. Einzelne Sätze des Textes hängen nicht unbedingt zusammen. Das resultierte anfangs in einer sehr unnatürlichen Aussprache des Sprechers. Nachdem der Text mehrere Male laut gelesen wurde trat dieses Problem nicht mehr auf. Deswegen erhielten alle folgenden Sprecher den Text mindestens einen Tag vor den Aufnahmen und wurde darum gebeten, diesen mehrere Male laut zu lesen. Die Sprecher gewöhnen sich so schon einmal an den etwas ungewöhnlichen Text, was die benötigte Aufnahmezeit enorm verkürzt.

Vor der eigentlichen Aufnahme der Sprecher wurde für jeden Sprecher eine Schallpegelmessung durchgeführt. Für eine realistische Wiedergabe ist es sehr wichtig, den Schallpegel bei der Aufnahme zu überprüfen, um ihn später mit dem selben Schallpegel wiederzugeben.

Verwendet wurde der Schallpegelmesser 8922 RS 232 von AZ Instruments [6]. Über die Software des Herstellers konnte der Schallpegel aufgezeichnet werden. Diese gibt jede Sekunde einen gemittelten Wert aus. Der Schallpegelmesser wurde direkt über dem Sprechermikrofon positioniert. Die Messungen wurden für jeden Sprecher zwei mal durchgeführt. Das Gute an diesem Vorgehen ist, dass sie die Sprecher schon einmal an das Studio gewöhnen. Sie werden dadurch etwas lockerer und sprechen dann auch etwas natürlicher.

Abbildung 5.1: Aufbau für Sprachaufnahmen mit Schallpegelmesser

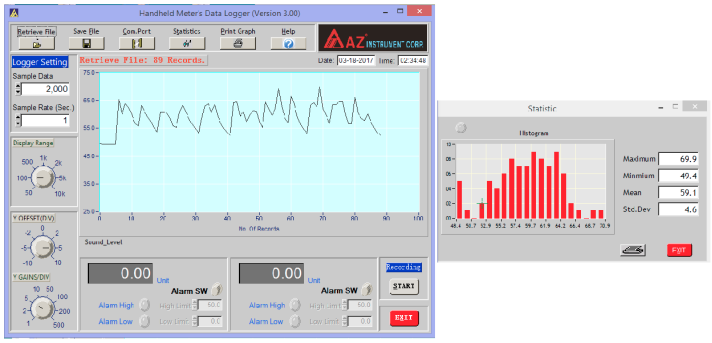

Abbildung 5.2: Beispiel einer Schallpegelmessung eines Sprechers

Um sicher zu gehen, dass die Sprecher den Abstand zum Mikrofon einhalten, wurde ein Mikrofonstativ, an dem auch der Popschutz installiert wurde, als Abstandhalter eingesetzt (siehe Abb. 5.1). Um Verfälschungen des Signals durch Reflexionen am Schallpegelmesser zu vermeiden, wurde dieser für die Aufnahmen wieder entfernt. Außerdem wurde ohne

[6] http://www.az-instrument.com.tw/az-instrument/en/productsinfo/148.html

Monitoring gearbeitet. Die Sprecher sollten sich selbst so hören, wie sie es von Natur aus gewohnt sind. Im Gegensatz zu Abb. 5.1 befand sich der Notenständer nicht unterhalb des Mikrofons sondern neben dem Sprecher. Damit konnte vermieden werden, dass sich die Sprecher ungewollt nach unten zum Text hin neigten (siehe Kapitel 4.3.3). Zur Sicherheit wurde bei den Aufnahmen beobachtet, ob sich die Sprecher dennoch, versehentlich zum Text hin wegdrehen.

Alle Sprachaufnahmen wurden mit einer Samplerate von 192 kHz und einer Auflösung von 32 Bit float durchgeführt. Das war notwendig, um bei der in Kapitel 6.2.1 beschriebenen Entfaltung optimale Ergebnisse zu erzielen.

5.5.2 Aufnahme der Störgeräusche

Alle Störgeräusche wurden ebenfalls mit dem Schoeps CCM 4V und soweit möglich, auch im Studio E der *HdM* Stuttgart aufgenommen. Auch hier wurde wieder darauf geachtet, dass der Abstand von der Schallquelle zum Mikrofon 60cm betrug. Die Aufnahme wurde in zwei Schritten durchgeführt. Zuerst wurde der Schallpegel mit dem AZ 8922 RS 232 aufgezeichnet. Anschließend wurde die eigentliche Aufnahme durchgeführt.

Aufnahmen von Geräuschen wie Fenster öffnen und schließen konnten nicht im Studio E durchgeführt werden. Diese wurden zusammen mit der Atmo in der Wohnung des Autors und in Vorlesungsräumen der *HdM* aufgenommen.

Alle Störgeräusche wurden mit einer Samplelrate von 48 kHz und einer Auflösung von 24 Bit durchgeführt.

5.5.3 Aufnahme der Atmos

Die Wiedergabe der Atmos erfolgt über ein Vierkanal-Stereosystem. Für die Aufnahmen der Atmos wurde deshalb ein Doppel-MS System der Firma Schoeps verwendet. Das Doppel-MS System ist eine Erweiterung der, aus der Stereofonie bekannten, MS-Aufnahmetechnik. Der Abkürzung MS steht hierbei für Mid/Side, zu Deutsch Mitte/ Seite. Für die MS-Aufnahme werden zwei Mikrofone benötigt. Meist werden dafür ein Mikrofon mit der Richtcharakteristik einer Acht und ein Mikrofon mit der Richtcharakteristik einer Kugel, Niere oder Hyperniere dicht übereinander platziert. Das Mikrofon mit der Kugel oder Hypernieren-Charakteristik wird zur Schallquelle gerichtet. Es bildet das Mittensignal. Das Mikrofon mit der Acht-Charakteristik wird so ausgerichtet, dass die positive Seite nach links zeigt. [Goe02]

[7] Es bildet das Seitensignal. Um die Wiedergabe von MS-Aufnahmen auf einer herkömmlichen stereofonen Lautsprecheranordnung wiedergeben zu können, müssen aus den Signalen die Informationen für den rechten (R) und linken (L) Lautsprecher gewonnen werden. Das geschieht durch Summierung und Differenzierung der beiden Signale.

$$L = M + S \tag{5.1}$$

$$R = M - S \tag{5.2}$$

Beim Doppel-MS System kommt nun noch ein weiteres, nach hinten gerichtetes Mikrofon dazu. Das nach vorne und auch das nach hinten gerichtete Mikrofon weisen in diesem Fall jeweils die Richtcharakteristik einer Niere auf. Aus den drei Mikrofonsignalen lassen sich nun die Informationen für das Vierkanal-Stereosystem gewinnen. Das Signal für die vorderen Lautsprecher wird aus dem nach vorne gerichteten Mikrofon und der Acht gewonnen. Die Signale für die hinteren zwei Lautsprecher werden aus dem nach hinten gerichteten Mikrofon und der Acht gewonnen.

Die Aufnahmen fanden in zwei unterschiedlichen Räumen statt. Der erste Raum befand sich in unmittelbarer Umgebung einer Baustelle, der zweite nahe einer Kantine. Beide Räume waren mit Fenstern und Türen ausgestattet, die während der Aufnahme geöffnet und kurz darauf wieder geschlossen wurden. Zum einen konnten so die Geräusche, die Türen und Fenster beim Öffnen und Schließen erzeugen, aufgenommen werden und zum anderen konnten die Schallereignisse von außen bei geöffneten Fenstern und Türen besser in den Raum eindringen, wodurch sich der Schallpegel im Raum deutlich erhöhte.

Alle Atmos wurden mit einer Samplelrate von 48 kHz und einer Auflösung von 24 Bit durchgeführt.

5.5.4 Aufnahme der Impulsantwort

Um in der Nachbearbeitung die frühen Reflexionen und den Diffusschall von den Sprachaufnahmen (siehe Kapitel 6.2.1) entfalten zu können, ist es notwendig die Impulsantwort des Raums zu ermitteln. Dazu muss zuerst der Raum durch einen kurzen Impuls angeregt werden. Für eine optimale Impulsantwort müsste dieser Impuls unendlich kurz sein und einen unendlich hohen Schallpegel haben [noa03] [8]. In der Realität ist dies nicht umzusetzen.

[7] Vgl. Goerne, 2003, S.106
[8] Vgl. Sandmann, 2003, S.65

Um sich der optimalen Impulsantwort zu nähern gibt es unterschiedliche Methoden. Beispielsweise kann das Platzen eines Luftballons als kurzer Impuls verwendet werden [noa03] [9]. Eine weitere Methode ist das Anregen des Raums durch einen Lautsprecher der einen Sinus-Sweep, also einen Sinuston durch alle Frequenze über eine definierte Zeit, wiedergibt. In dieser Arbeit wird hierzu die Applikation *Deconvolver* von Voxengo verwendet [10]. Dieser stellt unterschiedliche Testtöne zur Verfügung, mit denen, laut Hersteller, die ihm bestmöglichen Ergebnisse erzielt werden. Ein solcher Testton wird über einen Lautsprecher wiedergegeben und mit einem Mikrofon aufgenommen.

Um den Raum möglichst so anzuregen wie es die Sprecher bei den Aufnahmen taten, wurde hierfür der Lautsprecher an exakt die selbe Position gestellt, an der sich auch die Sprecher befanden. Auch die Aufnahme sollte unter den gleichen Bedingungen wie bei den Sprachaufnahme stattfinden. Deshalb wurde auch hier das Schoeps CCM 4V verwendet, das sich an exakt gleicher Position wie bei den Sprachaufnahmen befand. Der Popschutz wurde nicht verwendet. Wiedergegeben wurde ein vom *Deconvolver* generierter, dreisekündiger Sinus-Sweep mit einer Abtastrate von $192kHz$ und eine Auflösung von 32 Bit Float (Gleitkomma). Zu Versuchszwecken wurde dieser sowohl mit dem durchschnittlichen Sprechpegel ($60dB(A)$) als auch mit dem maximal möglichen Schallpegel des Lautsprechers wiedergegeben.

[9] Vgl. Sandmann, 2003, S.65
[10] http://www.voxengo.com/product/deconvolver/

6 Entzerrung der Signale

Als Filter wurden ursprünglich Schaltungen bezeichnet, deren Betrags-Frequenzgang über oder unter einer bestimmten Grenzfrequenz stetig frequenzabhängig sinkt. Die Stärke dieser Dämpfung, die sogenannte Flankensteilheit, wird dabei in dB pro Frequenzverdopplung, pro Oktave oder pro Dekade angegeben. Bestimmt wird die Stärke der Dämpfung durch die Ordnungszahl n. Bei einem Filter n-ter Ordnung ergibt sich die Flankensteilheit durch n*6dB pro Oktave beziehungsweise n*20dB pro Dekade. In der Regel ist bei klassischen Filtern die Grenzfrequenz einstellbar. Sie ist die Frequenz, bei der die Dämpfung 3 dB beträgt. [Dic14] [1]

Klassische Filter bestanden zumeist aus RC-Schaltungen und Operationsverstärken und sind deshalb auch in der umsetzbaren Flankensteilheit eingegrenzt. Die heutige Digitaltechnik erlaubt dagegen die Entwicklung von Filtern mit wesentlich steileren Flanken und komplexeren Eigenschaften.

Entzerrer, auch Equalizer (kurz EQ) genannt, haben im Gegensatz zu Filtern die Eigenschaft, dass neben einer Absenkung auch eine Anhebung des Pegels innerhalb definierter Frequenzbereiche möglich ist. Sie weisen keine feste Flankensteilheit auf, sondern erlauben eine gezielte Verstärkung von Frequenzbändern. EQs sind meist als aktive RC-Netzwerke mit Operationsverstärkern aufgebaut oder digital realisiert [Dic14] [2]. EQs werden in Anlehnung an die Form des Betragsfrequenzgangs für L $\neq$ 0 dB unterschieden in Glockencharakteristik und Kuhschwanz- bzw. Schelfcharakteristik.

Anwendung finden EQs in unterschiedlichen Bereichen. Für die Aufnahme von Instrumenten oder Gesang werden sie häufig eingesetzt um Unebenheiten im Frequenzgang der Aufnahmegeräte auszugleichen. In der Veranstaltungstechnik werden sie, neben der Einzelsignalbearbeitung, auch dazu verwendet, den Gesamtklang einer Beschallungsanlage an die Örtlichkeiten anzupassen. Möglich ist es auch gezielt Frequenzbereiche für das Monitoring bei Livekonzerten abzusenken um so Rückkopplungen zu vermeiden. Anwendung finden EQs auch in der Klanggestaltung von z.B. Musikmischungen um einzelne Signale harmonisch in die Gesamtmischung einzufügen.

[1] Vgl. Dickreiter, 2014, S.388
[2] Vgl. Dickreiter, 2014, S.393

6.1 Werkzeuge zur Entzerrung

Heute ist der Markt voll von digitalen und analogen Filtern, die sich in Qualität und Funktionsumfang unterscheiden. Grob lassen sie sich in zwei Gruppen unterteilen.

6.1.1 Grafischer EQ

Bei grafischen EQs wird jedem beeinflussbaren Frequenzband ein Regler zugewiesen. Typisch ist hierbei eine $1/3$ Oktave breite Aufteilung über das hörbare Frequenzspektrum. Grafische EQs werden eher zur Entzerrung von Beschallungsanlagen als für Einzelsignale verwendet. Im Gegensatz zu parametrischen EQs sind die Frequenzbänder fest zugeordnet. Deshalb ist es mit ihnen nicht immer möglich das gewünschte Frequenzband genau zu treffen. Bei der Bearbeitung von Einzelsignalen kann das ein großer Nachteil sein. Ein weiteres Problem bei der Bearbeitung von Einzelsignalen ist, besonders bei analogen grafischen EQs, dass das Signal durch alle Filternetzwerke durchgeschleift wird. Diese lassen sich nicht einzeln abschalten. Dadurch addieren sich Rauschanteil, Verzerrungen und Phasenfehler.

6.1.2 Parametrische EQ

Parametrische EQs bestehen aus einer Gruppe von Filtern, die Einstellungen mehrerer Parameter erlauben. Unterschieden wird hier zwischen teilparametrischen und vollparametrischen EQs. Teilparametrische EQs bieten Einstellungsoptionen für Verstärkungspegel und Mittenfrequenz. Bei vollparametrischen EQs kommt zu den zwei genannten Einstellungsoptionen die Güte und gelegentlich eine Umschaltung von Glocken- auf Schelfcharakteristik hinzu. Die Güte beschreibt wie stark die Frequenzbänder um die Mittenfrequenz beeinflusst werden. Der große Vorteil bei parametrischen EQs ist, dass sich gezielt Frequenzbänder verstärken lassen.

6.2 Voxengo Deconvolver

Mit dem *Deconvolver* von Voxengo lassen sich Impulsantworten von beliebigen Räumen ermitteln. In der Klanggestaltung wird diese Methode gerne zur künstlichen Hallerzeugung verwendet. Durch Faltung der Impulsantwort mit einer trocken aufgenommenen Schallquelle lässt sich diese sozusagen in den Raum stellen, aus dem die Impulsantwort gewonnen wurde. Die Berechnung folgt dabei dem mathematischen Verfahren der Faltung [noa03] [3].

Um eine Impulsantwort mit dem *Deconvolver* zu generieren muss, wie in Kapitel 5.5.4 beschrieben, der Raum durch ein Schallereignis angeregt werden. Laut Hersteller erzielt der *Deconvolver* die besten Ergebnisse wenn dafür ein von ihm generierter Sinus Sweep verwendet wird [Dec17] [4]. Mittels FFT (Fast-Fourier-Transfom, zu Deutsch schnelle Fourier-Transformation) berechnet er aus original Sweep und aufgenommenem Sweep die Impulsantwort. Für die Berechnung stehen unterschiedliche Methoden zur Verfügung. In dieser Arbeit wurden die besten Ergebnisse mit der sogenannten *Reversed technique* erzielt, deren Funktion leider vom Hersteller nicht weiter erläutert wird [Dec17] [5].

6.2.1 Entfaltung des Raums

Im Gegensatz zu der in Kapitel 6.2 beschriebenen Möglichkeit mit der Impulsantwort einen künstlichen Raum zu erzeugen wird in dieser Arbeit versucht, den Raumanteil der Sprachaufnahmen durch Entfaltung zu entfernen. Theoretisch ist das möglich. Umgesetzt wurde die Entfaltung in dieser Arbeit mit der Applikation *MATLAB* von MathWorks [6]. Diese Methode befindet sich noch in der Testphase und wurde nicht vom Autor dieser Arbeit entwickelt, weshalb hier nur kurz das Prinzip erläutert wird.

Generell bietet Matlab die Möglichkeit einer Deconvolution mit dem Befehl „deconv“an. Leider stößt „deconv“bei großen Datenmengen an seine Grenzen und kann aktuell nur für einfache Funktionen genutzt werden. Da die in „deconv“interne Datenverarbeitung nicht klar ist bzw. geändert werden kann, wird die Deconvolution „manuell“mit Hilfe einer FF-Transformation durchgeführt werden. Dazu wird in Matlab eine Audiodatei und die generierte Impulsantwort in den FFT-Raum überführt. Hier wird nun die Audiodatei durch die Impulsantwort geteilt und das Ergebnis wieder in die Zeitdomäne übertragen. Die Berechnung basiert auf dem fundamentalen Theorem des Filterns: Der Faltungsvorgang im

[3] Vgl. Sandmann, 2003, S.64
[4] http://www.voxengo.com/product/deconvolver/
[5] https://www.voxengo.com/doc/deconvolver/
[6] `https://de.mathworks.com/products/matlab.html?s_tid=hp_products_matlab`

Frequenzraum ist äquivalent zu einer Multiplikation im Fourierraum oder bei Entfaltung einer Divison [Pol07] [7].

6.3 Entzerrung der aufgenommenen Signale

6.3.1 Ermittlung der Frequenzgänge von Lautsprecher und Mikrofon

In dieser Arbeit wird mit dem Einsatz von EQs versucht die Unebenheiten im Frequenzgang auszugleichen, die von dem verwendeten Mikrofon und den verwendeten Lautsprechern verursacht werden. Zur Ermittlung des Frequenzgangs wurden in dieser Arbeit drei unterschiedliche Methoden verwendet:

Methode 1:
Ermittlung anhand der Datenblätter des Hersteller.

Methode 2:
Ermittlung durch Messungen von Lautsprecher und Mikrofon in einem reflexionsarmen Raum.

Methode 3:
Ermittlung durch Messungen im Aufnahmeraum.

Bei dem vom Hersteller angegebenen Frequenzgang besteht oft das Problem, dass sie sehr stark geglättet werden. Ein gutes Beispiel dafür ist der von Hersteller angegebene Frequenzgang des Schoeps *CCM 4V* (siehe Abbildung 6.1). Wichtig ist allerdings, dass trotz dieser Glättung erkenntlich wird, welcher Frequenzbereich stärker beeinflusst wird. Im Fall des Schoeps *CCM 4V* betrifft dies eine Anhebung der Frequenzen um 10 kHz und die Absenkung ab Frequenzen tiefer 150 Hz und höher 15 kHz.

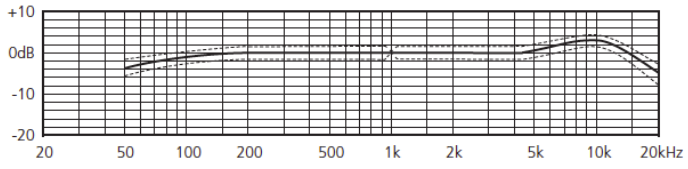

Abbildung 6.1: Frequenzgang Schoeps CCM 4V [8]

[7] http://www.astro.rug.nl/~vdhulst/SignalProcessing/Hoorcolleges/college03.pdf
[8] http://www.schoeps.de/de/products/ccm4v

Der vom Genelec angegebene Frequenzgang des Lautsprechers vom Typ *1029A* ist wesentlich schwächer geglättet (siehe Abbildung 6.2). Gemessen aus 0° fallen hier besonders eine schmalbandige Anhebung bei ungefähr 75 *Hz*, zwei etwas breitbandigere Anhebungen bei 1,5 *kHz* und 2,5 *kHz* und eine sehr breitbandige Anhebung bei 9 *kHz* auf. Ab 70 *Hz* fällt der Pegel stark ab.

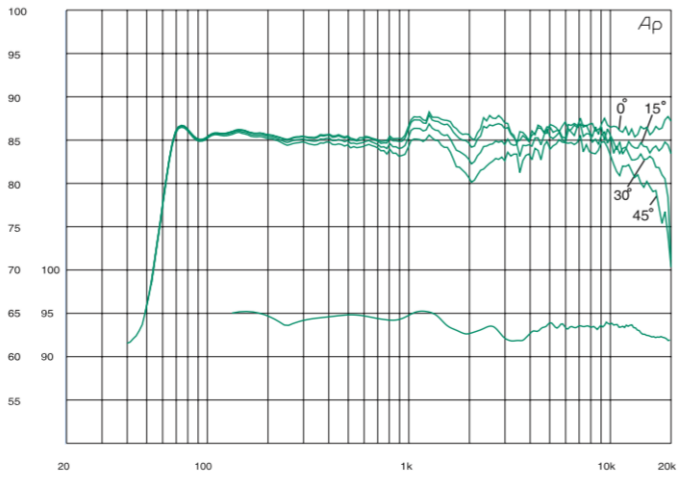

Abbildung 6.2: Frequenzgang Genelec 1029A in 1m Entfernung bei 0°, 15°, 30° und 45° [9]

Zur Überprüfung der vom Hersteller angegebenen Daten wurden das Mikrofon, das in dieser Arbeit für die Aufnahmen verwendet wurde und einer der Lautsprecher, aus denen das Lautsprecher-Setup bestand, in einem reflexionsarmen Raum gemessen. Zur Messung des Lautsprechers wurde das *M 371* Messmikrofon von Microtech Gefell verwendet. Anschließend wurde mit dem *1029A* das *CCM 4V* gemessen. Der Abstand von Lautsprecher zum Mikrofon betrug in beiden Fällen 4 *m*. Verwendet wurde für beide Messungen ein dreisekündiger Sinus Sweep. Abbildung 6.3 zeigt den gemessenen Frequenzgang des *1029A* (lila) und den aus mehreren Richtungen gemittelten Frequenzgang des *CCM 4V* (grün). Die gemessenen Frequenzgänge kommen sehr nahe an die des Herstellers heran. Lediglich die breitbandige Anhebung bei 9 *kHz* und die schmalbandige Anhebung bei ungefähr 75 *Hz* des *1029A* konnten hier nicht nachvollzogen werden.

[9] `http://www.genelec.com/sites/default/files/media/Previous%20models/2-way%20loudspeakers/1029A/ds1029a.pdf`

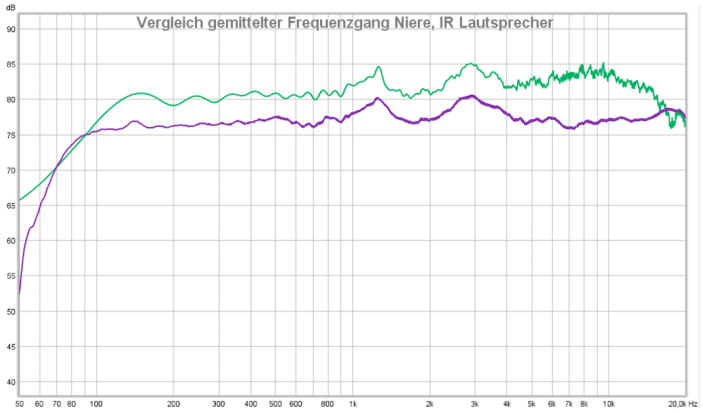

Abbildung 6.3: Messung von Genelec 1029A (lila) und Schoeps CCM 4V(grün)

Zuletzt wurde noch eine Messung von Lautsprecher und Mikrofon zusammen unter den Umstände durchgeführt, die auch bei der Ermittlung der Impulsantwort des Aufnahmeraumes herrschten (siehe Kapitel 5.5.4). Im Gegensatz zu den Messungen im reflexionsarmen Raum stand hier nicht exakt der selbe Lautsprecher, sonder ein anderer der selben Baureihe, zur Verfügung. Als Testton wurde ein zehnsekündiger Sinus-Sweep verwendet. Wiedergabe und Aufnahme wurden mit einer Abtastrate von 192kHz und einer Auflösung von 32 Bit Float (Gleitkomma) durchgeführt. Im Gegensatz zu den Messungen im reflexionsarmen Raum beeinflussen hier die durch den Raum verursachten Reflexionen das Signal. Abbildung 6.4 zeigt den gemessenen Frequenzgang. Deutlich zu erkennen sind auch hier eine Anhebung des Pegels bei 1,5 kHz und 2,5 kHz, die vermutlich vom Lautsprecher verursacht werden. Auch eine sehr ausgeprägte Anhebung bei ungefähr 9 kHz ist sichtbar, die vermutlich vom *CCM 4V* und dem *1029A* verursacht wird. Im Gegensatz zu den Herstellerdaten und den Messungen im reflexionsarmen Raum kommt es im Bereich zwischen 80 Hz und 200 Hz zu starken Pegelschwankungen, die vermutlich auf die akustischen Eigenschaften des Aufnahmeraums zurückzuführen sind.

Abbildung 6.4: Messung des Frequenzgangs des Genelec *1029A* und Schoeps *CCM 4V* im Aufnahmeraum

6.3.2 Konfigurierung des EQs anhand der ermittelten Frequenzgänge

Ziel der Entzerrung in dieser Arbeit ist es, zum einen den Frequenzgang des Mikrofons der aufgenommenen Signale und zum anderen den Frequenzgang des Lautsprechers bei der Wiedergabe auszugleichen. Um das zu erreichen wurde für die in Kapitel 6.3.1 ermittelten Frequenzgänge EQs konfiguriert. Der Frequenzgang des verwendeten Audiointerfaces ist sehr linear und wurde deshalb vernachlässigt. Für einen objektiven Vergleich wurde der in Kapitel 5.5.4 erstellte Sweep mit den erstellten EQs bearbeitet, über den Lautsprecher *1029A* wiedergegeben und mit dem Mikrofon *CCM 4V* aufgenommen. Lautsprecher und Mikrofon befanden sich an den gleichen Positionen wie bei der Aufnahme zur Ermittlung der Impulsantwort des Aufnahmeraums (siehe Kapitel 5.5.4). Durch das Ausgleichen der Unebenheiten der Frequenzgänge durch den EQ, sollte es somit möglich sein, wieder einen näherungsweise linearen Sweep zu erhalten. Näherungsweise deshalb, weil diese Tests wieder im in Kapitel 5.1 beschriebenen Aufnahmeraum und nicht in einem reflexionsarmen Raum durchgeführt wurden. Reflexionen im Raum beeinflussen die Genauigkeit des Tests. Um den Raum nur so stark anzuregen wie es bei den Sprachaufnahmen der Fall war, wurde der wiedergegebene Sinus-Sweep nach dem durchschnittlichen Sprachpegel von 60 $dB(A)$ eingestellt. Zusätzlich zu dem Sinus-Sweep wurden auch Aufnahmen eines Sprechers über die Lautsprecher wiedergegeben und wieder aufgenommen.

Neben dem objektiven Vergleich wurde auch stets subjektiv verglichen. Hierzu stellte sich der Autor an die Stelle, an der sich das Mikrofon befand und verglich das mit den EQs bearbeitete Signal mit der Originalaufnahme. Da ein Sinus-Sweep die Ohren doch sehr belastet, wurde hier nur die Sprachaufnahme verglichen.

Zur Entzerrung wurde der digitale *EQ III* von Avid verwendet [10]. Der *EQ III* ist ein vollparametrischer 5-Band-EQ mit zusätzlichem Hoch- und Tiefpass und gehört zur Standardausstattung von Pro Tools. Nutzer der Session müssen also keine Erweiterungen dazukaufen.

Um die Ergebnisse der EQ-Konfigurationen besser vergleichen zu können zeigt Abbildung 6.5 den Frequenzgang des reinen Sinus-Sweeps und den Frequenzgang eines gesprochenen Satzes der Originalaufnahme.

Abbildung 6.5: Der reine Sinus-Sweep und aufgenommene Sprache

Zuerst wurde das Signal anhand des im Aufnahmeraum (siehe Kapitel 5.1) aufgenommenen Sweeps (siehr Abbildung 6.4) entzerrt. Dieser Ansatz liefert zwar sehr gute Ergebnisse, bringt aber auch gewisse Probleme mit sich. Dieser Filter würde nur bei genau der hier verwendeten Konstellation von Lautsprecher, Mikrofon und Raum funktionieren. Verwendet der Nutzer der virtuellen Konferenz nun z.B. andere Lautsprecher, passen die Einstellungen des Filters nicht mehr zu dessen Frequenzgang. Abbildung 6.6 zeigt das Resultat des hier eingestellten EQs.

[10] https://www.avid.com/de/plugins/eq-iii

Abbildung 6.6: Signale entzerrt nach Frequenzgang des im Aufnahmeraum aufgenommenem Sinus-Sweeps

Um den Nutzer der virtuellen Konferenz nicht an genau die hier verwendete Testumgebung zu binden, wurde der EQ im zweiten Versuch zuerst nach den vom Hersteller angegebenen Frequenzgang des *1029A* und anschließend nach den in Kapitel 6.3.1 gemessen Frequenzgang des *1029A* konfiguriert. Zuletzt wurde noch eine Kombination der beiden Filter erstellt. Wie sich zeigte lieferte diese auch die besten Ergebnisse.

Abbildung 6.7: Signale entzerrt nach Frequenzgang des *1029A*

Im dritten Versuch wurde zu dem eben erstellten EQ noch ein weiterer EQ konfiguriert, der den gemessen Frequenzgang des *CCM 4V* ausgleicht.

Abbildung 6.8: Signale entzerrt nach Frequenzgang des *1029A* und *CCM 4V*

Objektiv und subjektiv betrachtet liegen die Ergebnisse des nach dem im Aufnahmeraum gemessenen Frequenzgangs und des nach den ermittelten Frequenzgängen des *CCM 4V* und *1029A* konfigurierten EQs sehr nahe beieinander. Beide klingen beinahe natürlich. Lediglich der Bereich um die 400 *Hz* ist etwas zu dominant und wird für den finalen EQ um 2 *dB* reduziert.

Für die virtuelle Konferenz werden letztendlich die Filter aus dem dritten Versuch verwendet. Es wird also einen EQ für die Entzerrung des Lautsprechers und einen EQ für die Entzerrung des Mikrofons mit der Absenkung bei 400 *Hz* geben. Sollten andere wie die hier verwendeten Lautsprecher benutzt werden, müssen die EQs für den Lautsprecher modifiziert werden.

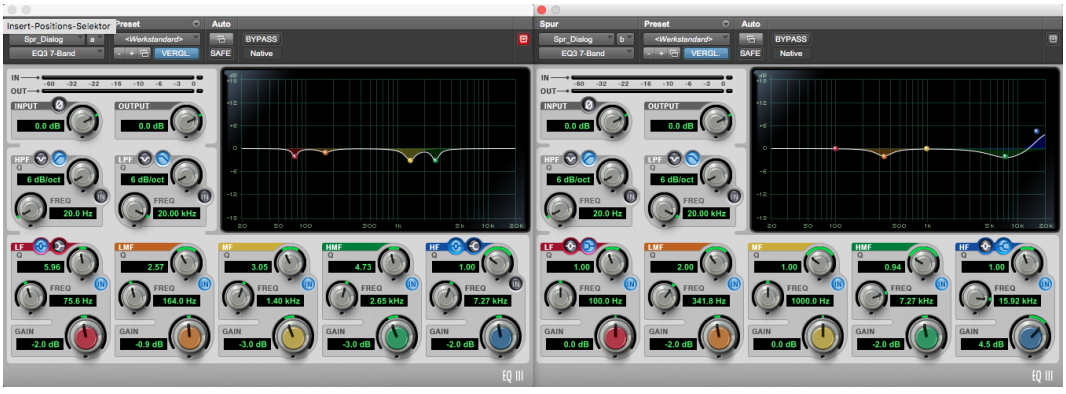

Abbildung 6.9: EQs konfiguriert anhand ermittelter Frequenzgänge

Wegen der Pegelreduzierung der betreffenden Frequenzbänder ist es wichtig nochmals Schallpegelmessungen durchzuführen, um so wieder die Originallautstärke reproduzieren zu können.

6.3.3 Fazit

Die Entfaltung mit *MATLAB* bewirkt eine deutliche Abnahme des auf den Aufnahmen hörbaren Diffusschalls. Die besten Ergebnisse lieferte dabei die Faltung mit der in durchschnittlicher Sprachlautstärke aufgenommenen Impulsantwort. Wurde die gesamte Impulsantwort vom Audiosignal entfaltet, wirkte sich dies in deutlich hörbaren Artefakten, vor allem in den Bereichen ab 4 kHz aus. Das verbesserte sich extrem, nachdem bei der Impulsantwort das letzte Viertel abgeschnitten und entfernt wurde. Die Abnahme des hörbaren Diffusschalls blieb dabei erhalten.

Bei einem direkten Vergleich zwischen den originalen Aufnahmen und den über den Lautsprecher wiedergegebenen Aufnahmen sind trotz EQs Unterschiede zu hören. Die originalen Aufnahmen klingen transparenter als die wiederaufgenommenen. Das verbesserte sich auch bei mehreren und härteren EQ-Eingriffen nicht, sondern wirkte dadurch eher künstlich.

7 Die Session

Die Session wurde in 48 *kHz* und 24 Bit erstellt. Die Sprachaufnahmen, die in 192 *kHz* und 32 Bit float aufgenommen und entfaltet wurden, mussten ein Downsampling durchlaufen. Die Session wurde so organisiert, dass jeder Sprecher (Person 1 bis 4 nach Kapitel 4.2) und jedes Geräusch bzw. jede Geräuschgruppe eine eigene Spur erhält. Dabei kommt es vor, dass mehrere Spuren später auf den selben Ausgang geroutet wurden. Der Ablauf der einzelnen Audiodateien wurde für eine *kleine* Konferenz festgelegt und anschließend auf alle anderen übertragen. Die einzelnen *kleinen* Konferenzen wiederum wurden in einer Session zu einer *großen* Konferenz zusammengefasst. Abbildung 7.2 zeigt den Ablauf einer *kleinen* Konferenz.

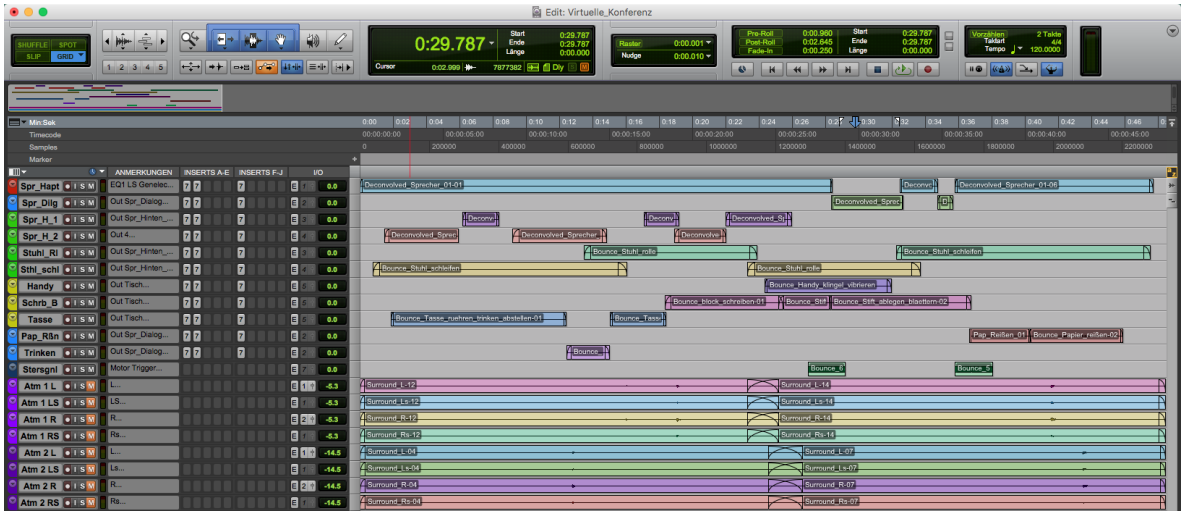

Abbildung 7.1: Zusammenstellung einer kleinen Konferenz

Die Reihenfolge der „Person 1 und Person 2 Sprecher“folgt der Reihenfolge des erstellten Sprechertextes (siehe Anhang 11.1). Wichtig war hierbei, dass Unterschiede der Sprechgeschwindigkeit unterschiedlicher Person 1 Sprecher durch das Verlängern oder Verkürzen der Pausen zwischen den Sprechern ausgeglichen wurden, so dass die Sätze immer an gleicher Stelle beginnen. Die gesamte Dauer einer *kleinen* Konferenz bleibt immer gleich. Dadurch, dass jedes Geräusch bzw. jede Geräuschgruppe einer eigenen Spur zugewiesen wurde, hat der Nutzer der Session die Möglichkeit die Reihenfolge der Geräusche in der *großen* Konferenz zu verändern. In den *kleinen* Konferenzen bleibt der Ablauf zeitlich identisch.

Zu erwähnen ist an dieser Stelle noch, dass der Dialog, den die Person 3 und 4 führen, zweimal während eines Durchlaufs abgespielt wird, genauso wie das Rollen und Schleifen von

Stühlen. Die zwei Quadrofonen Atmos wurden gruppiert und können bei Bedarf zugeschaltet werden.

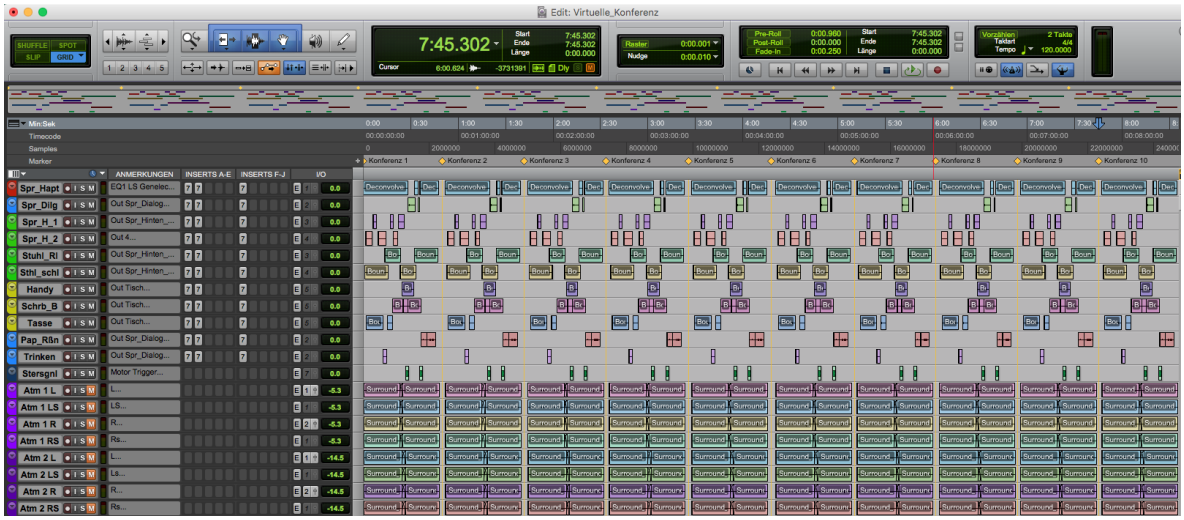

Abbildung 7.2: Zusammenstellung der großen Konferenz

7.1 Einstellen der Schallpegel

Wichtig für eine realistische Wiedergabe ist, dass der Schalldruckpegel mit dem des aufgenommenen Signals übereinstimmt. Bei den Aufnahmen wurden deshalb stets Schallpegelmessungen durchgeführt. Um diese wieder reproduzieren zu können wurde wie folgt vorgegangen. Zuerst wurden ein aktiver Studiomonitor und der Schallpegelmesser wie in Kapitel 5.5.4 aufgestellt. Der aktive Studiomonitor wurde auf ungefähr mittlere Verstärkung eingestellt. Der verwendete Kanal in der Session und im Mixer des Audiointerfaces wurden auf 0 dB gestellt. Die Audiodateien wurden nun über den Lautsprecher wiedergegeben und mit der Clipgain-Funktion der DAW auf den Schallpegel eingestellt, der bei der Aufnahme gemessen wurde. Diese Prozedur wurde mit allen aufgenommenen Dateien durchgeführt.

Wegen des stark schwankenden Schallpegels der unterschiedlichen Geräusche ist dies sehr mühselig. Damit das nicht bei jedem Aufbau der virtuellen Konferenz wiederholt werden muss, wurde eine Spur in die Session integriert, die ein weißes Rauschen enthält. Der Clipgain wurde für diese Spur so eingestellt, dass unter den oben genannten Bedingungen exakt 60 $dB(A)$ gemessen wurde. Gemessen wurde aus 60 cm Entfernung. Mit dieser Spur lassen sich nun die Lautstärkeeinstellungen an den für die virtuellen Konferenz verwendeten Lautsprechern vornehmen.

8 Das Wiedergabesystem

8.1 Verwendete Lautsprecher

Für das gesamte Lautsprecher-Setup wurden Lautsprecher des Typs *1029A* von Genelec verwendet [1]. Die Verwendung des gleichen Lautsprechertyps ist in dieser Arbeit von sehr hoher Bedeutung. Wie in Kapitel 6.3 beschrieben wurde die Entzerrung auch anhand des ermittelten Frequenzgangs dieses Lautsprechertyps durchgeführt. Der *1029A* ist ein kompakter 2-Wege aktiv Nahfeldmonitor mit je einer Verstärkerstufe pro Chassis. Mit einem maximalen Schalldruckpegel von 98 $dB(SPL)$ ist er absolut ausreichend dimensioniert. Dennoch bringt ein 2-Wegesystem auch gewisse Nachteile mit sich. Wegen der räumlichen Anordnung der Chassis können Verzerrungen durch Zeit- und Phasenfehler auftreten. Um dies auszuschließen würden sich Breitbandlautsprecher anbieten, die für diesen Versuch allerdings nicht in der benötigten Anzahl zur Verfügung standen.

8.2 Anordnung der Lautsprecher

Die Anordnung wurde, wie in Kapitel 4.6.1 beschrieben, in drei Schichten eingeteilt. Für die hier durchgeführten Tests wurden die Lautsprecher nach Abbildung 8.1 angeordnet. Die hier definierten Abstände und Winkel müssen nicht unbedingt eingehalten werden, sie dienen lediglich der Orientierung.

[1] `http://www.genelec.com/sites/default/files/media/Previous%20models/2-way%20loudspeakers/1029A/ds1029a.pdf`

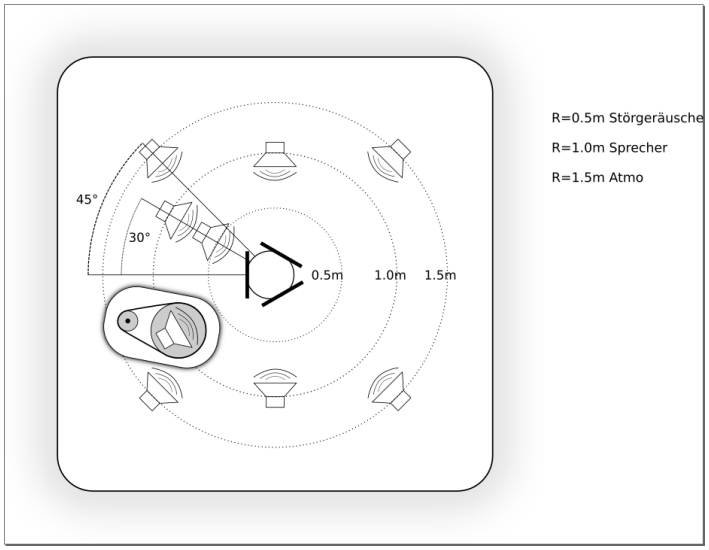

Abbildung 8.1: Anordnung der Lautsprecher

Das Mikrofon befand sich in der Mitte zweier zusammengeschobener Tische. Die zwei Tische zusammen hatten eine Breite von 1,4 *m* und eine Länge von 1,7 *m*. Die Abstände zu den Lautsprechern wurde von der Mitte der Tische gemessen. Zur Ermittlung der Winkel diente die Gerade, welche die beiden zusammengeschobenen Tische in der Mitte bilden. Die Höhe der Lautsprecher betrug, vom Boden bis zur Mitte des Lautsprechers gemessen, 1,15 *m* für die Lautsprecher der Sprecher und 1,35 *m* für die Lautsprecher der Atmo. Alle Lautsprecher wurden zum Mikrofon hin ausgerichtet.

Abbildung 8.2: Testaufbau des Lautsprecher-Setups

Abbildung 8.2 zeigt den realen Testaufbau der virtuellen Konferenz. Leider stand bei dem

Test keine Monitorständer mit einer ebenen Fläche zur Verfügung, auf den der Drehteller für den *Person 1* Sprecher (siehe Kapitel 4.2) platziert werden konnte. Zu Demonstrationszwecken wurde dieser deshalb für den Lautsprecher der Schicht 1 verwendet. Dieser sollte eigentlich direkt auf den Tisch gestellt werden.

8.3 Der drehbare Lautsprecher

Um Bewegungen eines Sprechers simulieren zu können wurde ein Drehteller entworfen, auf den der Lautsprecher des *Person 1* Sprechers (siehe Kapitel 4.2) gestellt wird. Mit diesem kann eine Drehung des Kopfes um 90° zur Seite simuliert werden. Zu beachten ist an dieser Stelle, dass sich Lautsprecher und Sprecher in ihrer Richtcharakteristik unterscheiden, wodurch eine Verfälschung zu einer realen Situation entsteht (siehe Abbildung 6.2 und Abbildung 3.8).

Um den Drehteller bewegen zu können wurde in dieser Arbeit ein Schrittmotor [2] verwendet. Die Ansteuerung des Schrittmotors erfolgt über ein *Arduino Uno* [3], das mit dem *Motor Shield V2* [4] von Adafruit erweitert wurde. Die Programmierung für die Ansteuerung des Schrittmotors wurde so konstruiert, dass mit ihm zwei absolute Positionen angefahren werden können.

8.3.1 Aufbau des Drehtellers

Bei der Konstruktion des Drehtellers stand zu Beginn ein Schrittmotor zur Verfügung. Wegen des knappen Budgets wurde nach einer möglichst kostengünstigen Lösung gesucht, um diesen Schrittmotor für die Konstruktion des Drehtellers verwenden zu können. Der Ansatz der hierbei verfolgt wurde stammt von motorisierten Drehtellern, die häufig zur Präsentation von Produkten verwendet werden.

Auf die Welle des Schrittmotors wurde eine Gummirolle montiert. Die Gummirolle wird gegen eine runde Kunststoffscheibe gedrückt, die auf einen Drehkranz mit Kugellager montiert wurde. Die Kunststoffscheibe wurde passend zum *1029A* von Genelec dimensioniert. Schrittmotor und Drehkranz wurden auf einer MDF-Plattenkonstruktion befestigt. Der Schrittmotor wurde dabei nur mit einer Schraube befestigt. Damit der Druck auf den Drehteller annähernd

[2] `https://www.sparkfun.com/products/9238`
[3] `https://www.arduino.cc/en/Main/ArduinoBoardUno`
[4] `https://learn.adafruit.com/adafruit-motor-shield-v2-for-arduino/overview`

konstant bleibt wurde der Schrittmotor mit einem Gummi eingespannt. Um dem Lautsprecher genug Halt auf der Kunststoffscheibe zu geben wurde diese mit Antirutschmatten versehen.

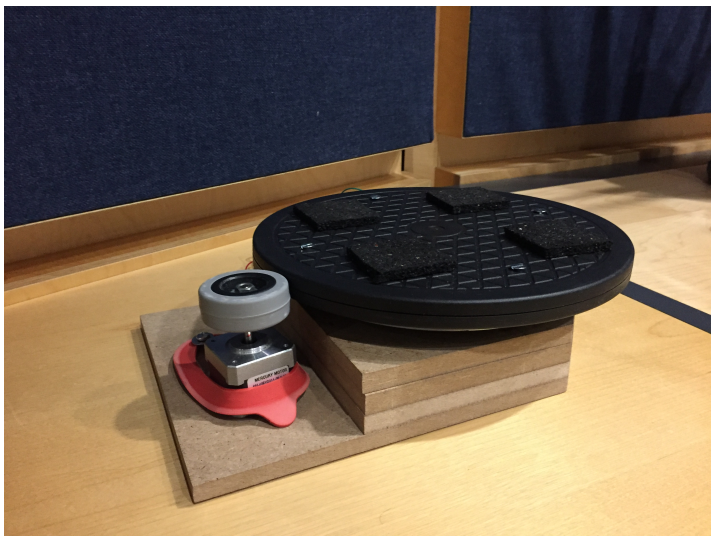

Abbildung 8.3: Drehteller mit Schrittmotor

Ein Dauertest von 100 Drehungen um 90° und zurück zeigte, dass die Konstruktion sehr zuverlässig funktioniert. Es konnte lediglich ein Versatz von 1-2° gemessen werden. Die Kosten des gesamten Drehtellers mit Schrittmotor belaufen sich auf 45 €(technische Zeichnungen im Anhang 11.2).

8.3.2 Signal zur Ansteuerung

Wichtig für diese Arbeit ist, dass neben der Wiedergabe der Audiosignale auch der Schrittmotor über die Session angesteuert werden kann. Dafür wurden mit einem Frequenzgenerator zwei Rechtecksignale unterschiedlicher Frequenz generiert, aufgenommen und in die Session integriert. Um am Audiointerface eine Ausgangsspannung von 1 Volt Spitze zu erhalten wurde im Mixer des Audiointerfaces und in der *DAW* der entsprechende Kanal auf 0 *dB* gestellt. Anschließend wurde der Clipgain der in der Session vorliegenden Audiodateien angepasst.

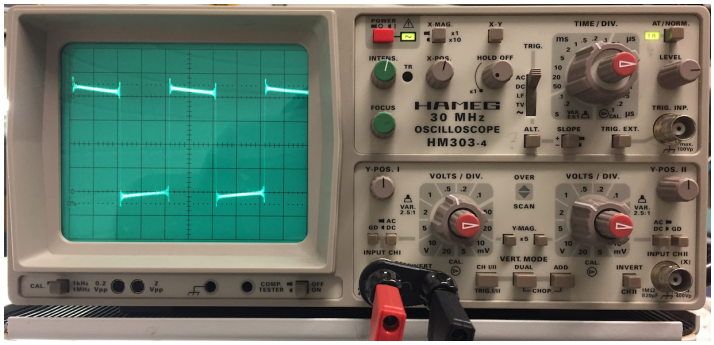

Abbildung 8.4: Recht-Wechselspannung 1 Volt, 100 Hz

8.3.3 Aufbau der elektronischen Umgebung

Bei der Entwicklung der elektronischen Umgebung mussten mehrere Punkte berücksichtigt werden. Zum einen ist es wichtig, dass der verwendete Line-Ausgang des Audiointerfaces nicht zu stark belastet und dadurch beschädigt wird. Zum anderen muss das Steuersignal auf die Spannung verstärkt werden, die vom *Arduino Uno* erkannt wird. Außerdem wird für das Ermitteln der Frequenz des Steuersignals ein *PWM-Input* des Arduinos genutzt. Damit dieser richtig arbeitet ist es wichtig, dass das Steuersignal zwischen 0 Volt und 5 Volt pendelt.

Alle genannten Punkte lassen sich durch die richtige Beschaltung eines Operationsverstärkers umsetzen. Da der Eingangswiederstand eines Operationsverstärkers sehr hoch ist, wird der Audioausgang des Audiointerfaces kaum belastet. Wird der Operationsverstärker um einige Widerstände und Dioden erweitert, kann so eine gezielte Verstärkung und Gleichrichtung umgesetzt werden. Diese Kombination von Bauteilen ist auch als Präzisionsgleichrichter bekannt. Hierbei ist aber darauf zu achten, dass der Eingangswiderstand nun durch den, für die Schaltung verwendeten Widerstand R1 (siehe Abbildung 8.5) bestimmt wird. Die Eingangsimpedanz von Line-Eingängen liegt typischerweise bei 10 *kOhm*, weshalb dieser Wert auch für R1 verwendet wird.

Die Abbildung 8.5 zeigt einen Einweg-Präzisionsgleichrichter. Wird diese Schaltung mit einer Wechselspannung gespeist, wird diese vom Operationsverstärker invertiert. Bei der positiven Sinushalbwelle leitet die Diode G1 und die Diode G2 sperrt. Die Ausgangsspannung bleibt somit auf 0 Volt. Bei der negative Halbwelle wird der Ausgang des Operationsverstärkers positiv, wodurch Diode G2 den Strom über den Widerstand R2 zurück an den invertierenden Eingang leitet. Die negative Halbwelle wird durch das Widerstandsverhältnis R2 / R1 verstärkt und erscheint als invertierte Ausgangsspannung Ua.

Um einen Operationsverstärker optimal betreiben zu können, wird eine positive und negative Versorgungsspannung benötigt. Die positive Versorgungsspannung wird aus einem externen 12 Volt Gleichspannungsnetzteil gewonnen, das auch als Versorgungsspannung für *Arduino Uno* und *Motor Shield V2* verwendet wird. Um die negative Versorgungsspannung erzeugen zu können wird der Spannungsinverter *ICL7660* [5] verwendet, der nach dem Prinzip einer Ladungspumpe konstruiert ist. Bei dem eingesetzten Operationsverstärker handelt es sich um den *TL081* [6].

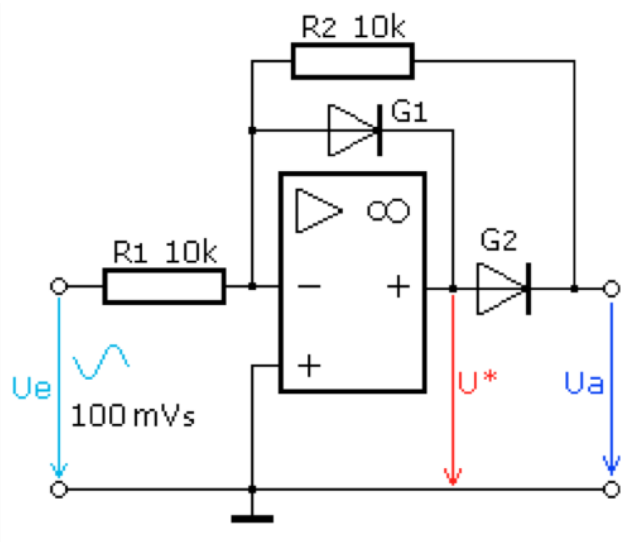

Abbildung 8.5: Einweg-Präzisionsgleichrichter [7]

Um die in unserem Fall generierte Wechselspannung von 1 Volt Spitze auf 5 Volt Spitze zu verstärken, muss das Verhältnis der Widerstände 5/1 betragen. Durch den für R1 bestimmten Widerstand von 10 *kOhm* ergibt sich für R2 ein Widerstand von 50 *kOhm* (E-Plan im Anhang 11.10).

[5] `http://www.intersil.com/content/dam/Intersil/documents/icl7/icl7660.pdf`
[6] `http://www.ti.com/lit/ds/symlink/tl082.pdf`
[7] `http://elektroniktutor.de/analogverstaerker/aktivglr.html`

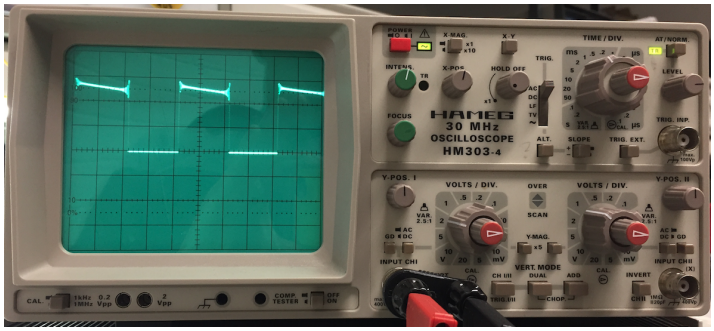

Abbildung 8.6: Gleichgerichtetes Rechtecksignal 1 Volt, 100 Hz

8.3.4 Aufbau der Programmierung

Für die Programmierung des Arduino wurden drei Bibliotheken integriert:

Adafruit MotorShield.h:

Bibliothek, die Funktionen zur Ansteuerung des *Motor Shield V2* beinhaltet [Ada17b] [8].

AccelStepper.h:

Bibliothek, die Funktionen für Beschleunigungs- und Bremsrampen beinhaltet [Ada17a] [9].

AccelStepper.h:

Bibliothek, die Funktionen für die Frequenzzählung beinhaltet [Naw12] [10].

Das in Kapitel 8.3.3 gleichgerichtete Rechtecksignal wird in den *PWM-Input* des Arduino Uno gespeist (Eingang 5). Der Frequenzzähler zählt nun die Pulse und gibt das Ergebnis als Integer-Wert aus. Die Auswertung des Frequenzzählers benötigt 100 *ms*, wodurch ein kleiner Versatz bei der Ansteuerung entsteht. Je nach Frequenz wird nun eine der zwei absoluten Positionen angefahren. Um ruckartige Bewegungen zu vermeiden werden bei Beschleunigung und Abbremsung Rampen gefahren (Programmcode im Anhang 11.4).

[8] https://github.com/adafruit/Adafruit_Motor_Shield_V2_Library/blob/master/Adafruit_MotorShield.h

[9] https://github.com/adafruit/AccelStepper/blob/master/AccelStepper.h

[10] http://interface.khm.de/index.php/lab/interfaces-advanced/arduino-frequency-counter-library/

9 Vergleich von realer und virtueller Konferenz

Da der Schwerpunkt dieser Arbeit auf das möglichst natürliche Wiedergeben der Sprache liegt, wurde hierfür ein direkter Vergleich zwischen realer und wiedergegebener Sprache durchgeführt. Zuerst wurde dafür das Lautsprecher-Setup wie in Kapitel 8.2 beschrieben aufgebaut. Über zwei Schicht 2 Lautsprecher (siehe Kapitel 4.6.1) wurde die Sprache der Person 1 und 2 Sprecher (siehe Kapitel 4.2) wiedergegeben und über das Mikrofon in der Mitte des Tisches aufgenommen.

Anschließend wurden die zwei Lautsprecher von den Stativen entfernt. Die Stative blieben auf der selben Position stehen. Die Personen, die auch die eben wiedergegebene Sprache eingesprochen hatten, setzten sich nun an die Stelle, an der sich die Lautsprecher befanden. Die Münder der Personen waren dabei so ausgerichtet wie es die Lautsprecher auch waren. Nun wurde die Konferenz von den zwei Personen „in echt “eingesprochen und über das Mikrofon aufgenommen.

Abbildung 9.1: Reales Nachsprechen der virtuellen Konferenz

Beim direkten Vergleich stellte sich schnell heraus, dass die Höhe der aufgestellten Lautsprecher das vom Mikrofon empfangene Signal stark beeinflussen. Gerade in dem hier verwendeten kleinen Abstand von 1 m von den Sprecher-Lautsprechern zum Mikrofon wird das Signal sehr basslastig, wenn die Hauptachse des Lautsprechers gegenüber dem Mikrofon höher liegt. Es sollte also bei der Aufstellung der Lautsprecher darauf geachtet werde, dass die Hauptachse aller Lautsprecher auf der Höhe des Mikrofons sind. Dadurch verbessert sich das Ergebnis erheblich. Dennoch sind Unterschiede zu einer echten Konferenz wahrnehmbar. Die echt aufgenommenen Sprecher klingen transparenter, wie die simulierten.

Der Versuch zeigt auch, dass ein direkter Vergleich sehr schwer umzusetzen ist. Sprechern fällt es extrem schwer immer exakt gleich zu sprechen, wie sie es bei der Aufnahme getan haben. Dennoch dient er zur Orientierung. Es stellte sich heraus, dass der Pegel bei 150 Hz noch um 2,5 dB schmalbandig abgesenkt werden muss. Damit konnte ein zufriedenstellendes Ergebnis erzielt werden.

10 Fazit und Ausblick

Ziel dieser Arbeit war es, eine möglichst reale Konferenz zu simulieren. Besonderes Augenmerk wurde dabei auf die Reproduzierbarkeit der Sprache gelegt. Dafür wurden unterschiedliche Sprecher und Geräusche in einer kontrollierten Umgebung aufgenommen. Um den Diffusschallanteil der Signale zu entfernen wurde ein Entfaltungsalgorithmus angewandt. Die entfalteten Signale wurden analysiert und anhand ermittelter Frequenzgänge von Mikrofon und Lautsprecher entzerrt. Zur Überprüfung der Signale wurden diese unter den selben Bedingungen, die auch bei den Sprachaufnahmen herrschten, über einen Lautsprecher wiedergegeben und mit dem Sprechermikrofon aufgenommen. Unterschiede, die bei der Überprüfung auffielen, wurden ausgeglichen. Nach der Bearbeitung wurde eine Session konstruiert, in der alle aufgenommenen Signale zu der finalen virtuellen Konferenz zusammengefügt wurden. Für die Wiedergabe wurde ein Lautsprecher-Setup mit einem automatisch drehbaren Lautsprecher konstruiert und mit diesem ein Vergleich zwischen echter und virtueller Konferenz durchgeführt.

Es zeigte sich, dass es nur näherungsweise möglich ist Sprache mit Lautsprechern wieder authentisch zu reproduzieren. Es stellte sich auch heraus, dass es absolut notwendig ist beim Aufbau des Lautsprecher-Setups die Hörachse der verwendeten Lautsprecher auf das Mikrofonsystem hin auszurichten. Wurde das nicht berücksichtigt entstand ein Ungleichgewicht im Frequenzgang. Eine weitere Abweichung zu einer realen Situation entsteht durch die fixe Position der Lautsprecher. In einer „real“nachgespielten Konferenz mussten sich die Teilnehmer sehr darauf konzentrieren den Kopf über die gesamte Konferenz zum Mikrofon hin auszurichten. In einer echten Situation ist dies nicht der Fall.

Die hier erstellte Konferenz bildet dennoch eine sehr gute Testumgebung für die Entwicklung eine Konfernzmikrofons. Dank des automatisch drehbaren Lautsprechers lassen sich Drehungen des Sprechers zur Seite simulieren. Für die Weiterentwicklung der Konferenz wäre es sinnvoll Sprachaufnahmen und die Entzerrung der Aufnahme- Wiedergabekette in einem reflexionsarmen Raum durchzuführen. Unter diesen Bedingungen lassen sich objektive Vergleiche mit stärkerer Aussagekraft durchführen. Sinnvoll wäre auch die Konstruktion von beweglichen, mehrachsigen Lautsprecherhalterungen, mit denen komplexe Bewegungsabläufe automatisiert werden können.

11 Anhang

11.1 Sprechertext

Guten Tag zusammen,

الكل الناس عالسلامة

大家好，

dà jiā hǎo

Buenos dias a todos

Wie sie sehen werden, haben wir heute sehr viel vor.

نعملو هم باش حاجات برشة عندنا ،تشوفوا كيما

如您所见，今天我们有许多事要做。

rú nín suǒ jiàn，wǒ men jīn tiān yǒu xǔ duō shì yào zuò.

Como verán, hoy tenemos mucho por delante

Wenn wir in diesem Jahr so weiter machen, werden wir große Probleme bekommen.

مشاكل برشا نلقاو بش ،هذايا العام هكا نكملو كان

如果我们今年继续这样做，我们将会面临很大的问题。

rú guǒ wǒ men jīn nián jì xù zhè yàng zuò，wǒ men jiāng hùi miàn lín hěn dà de wèn tí.

Si seguimos asi este año tendremos muchos Problemas

Es ist immer schön nichts zu tun und stehen zu bleiben, statt mit der Zeit zu gehen.

الوقت معا تمشي يمالل خير شي تعمل ما و واقف تقعد

无所事事站着不动比忙碌要舒服。

wú suǒ shì shì zhàn zhe bú dòng bǐ máng lù yào shū fu.

Es mas comodo no hacer nada y quedarse quieto, en vez de seguir adelante.

Abbildung 11.1: Sprechertext 1 von 5

Schauen sie, wie lange ich auf und ab laufen kann ohne zu fallen. Und dabei sehe ich gar nicht alt aus.

نطيح مغرما نوخر و نقدم نجم قداش أغزرو

您看我可以多久来回走动不停歇，与此同时我看上去并不老。

nín kàn wǒ kě yǐ duō jǐ lái húi zǒu dòng bù tíng xīe， yǔ cǐ tóng shí wǒ kàn shàng qù

bìng bù lǎo

Observen como sigo corriendo de un lado a otro, sin caer, y al mismo tiempo no parezco tan viejo.

Jetzt finden sie wir alle dürfen so viel nehmen wie wir brauchen.

نحبوا ما قد نهزوا انجموا الي يعرفو توه

现在您发现我们可以拿所有我们需要的。

xiàn zài nín fā xiàn wǒ men kě yǐ ná suǒ yǒu wǒ men xū yào de

Ahora, ustedes creen? Que podemos tomar tanto como necesitemos?

Aber darauf kann man keine Häuser bauen.

ديار عليه تبني تنجمش ما ما ا

但是人们不能在上面盖房子。

dàn shì rén men bù néng zài shàng miàn gài fáng zi.

Pero con eso no se pueden construir casas

Glauben sie, dass es schon wieder so sein wird?

فاتت الي المرة كيما تكون باش تتصور

您觉得，又会这样么？

nín jué de， yòu hùi zhè yàng me？

Cree usted, que volvera a ser así?

Abbildung 11.2: Sprechertext 2 von 5

Ja ich bin mir sicher.

متأكد

是的我确定。

shì de wǒ què dìng.

Si estoy seguro

Wir müssen noch etwas ganz Neues kaufen!

جديدة حاجة نشتري نزيدوا لازمنا

我们必须再买些新的！

wǒ men bì xū zài mǎi xiē xīn de.

Tenemos que comprar algo nuevo

Ja, das kann ich gerne für dich machen.

اي، نعملهولك إنجم

好的，我乐意为你效劳。

hǎo de， wǒ lè yì wèi nǐ xiào láo.

Claro con gusto lo hago por ti

Morgen?

غدوة

明天？

míng tián？

Mañana?

Abbildung 11.3: Sprechertext 3 von 5

Nein, heute.

لا، اليوم

不，今天。

bù，jīn tiān.

No, hoy!

Kommen wir zum nächsten Thema!

ميعدوا الي للموضوع نتعداو

让我们进入下一个话题！

ràng wǒ men jìn rù xià yí gè huà tí.

Llegamos al siguiente Tema

Ich denke, dass er mit dem Auto schnell zur Firma fahren wollte um dort weiter daran arbeiten zu können.

يخدم يزيد ينجم باش بالكر هية للخدمة فيسع يمشي يحب نتصور

我想，他想要开车快点到公司然后可以在那里继续工作。

wǒ xiǎng，tā xiǎng yǎo kā chē kuài diǎn dào gōng sī rán hòu zài nà lǐ jì xù gōng zuò.

Creo que el queria ir con el carro a la empresa y desde ahí seguir trabajando.

Vielen Dank und auf Wiedersehen.

بالسلامة و شكرا

非常感谢，再见.

fēi cháng gǎn xiè，zài jiàn

Muchas gracias y Hasta la proxima

Abbildung 11.4: Sprechertext 4 von 5

Verstehst du was er uns da sagt?

يقو للنا قاعد فاش تفهم

你明白他对我们说的么？

nǐ míng bái tā dùi wǒ men shuō de me？

Entiendes lo que el dice?

Nein ich habe keine Ahnung wie er das meint.

يقصد شنية فكرة حتى عندي ما نعرفش ما

不明白，我不知道他什么意思。

bù míng bai，wǒ bù zhī dào tā shén mé yì si.

No, ni idea que es lo que quiere

Vielleicht kann er uns das ja später nochmal in Ruhe erklären.

يريض كي مبعد يفسر لنا ينجم بالكشي

也许他会晚点再和我们详细的解释一遍。

yě xǔ tā hùi wǎn diǎn zài hé wǒ men xiáng xì de jiē shì yí biàn.

A lo mejor nos puede explicar despues con calma, que es lo que quiere.

Legende:

Schwarz:	**Sprecher 1**
Rot:	**Sprecher 2**
Blau:	**Sprecher 3**
Grün:	**Sprecher 4**

Abbildung 11.5: Sprechertext 5 von 5

11.2 Technische Zeichnungen

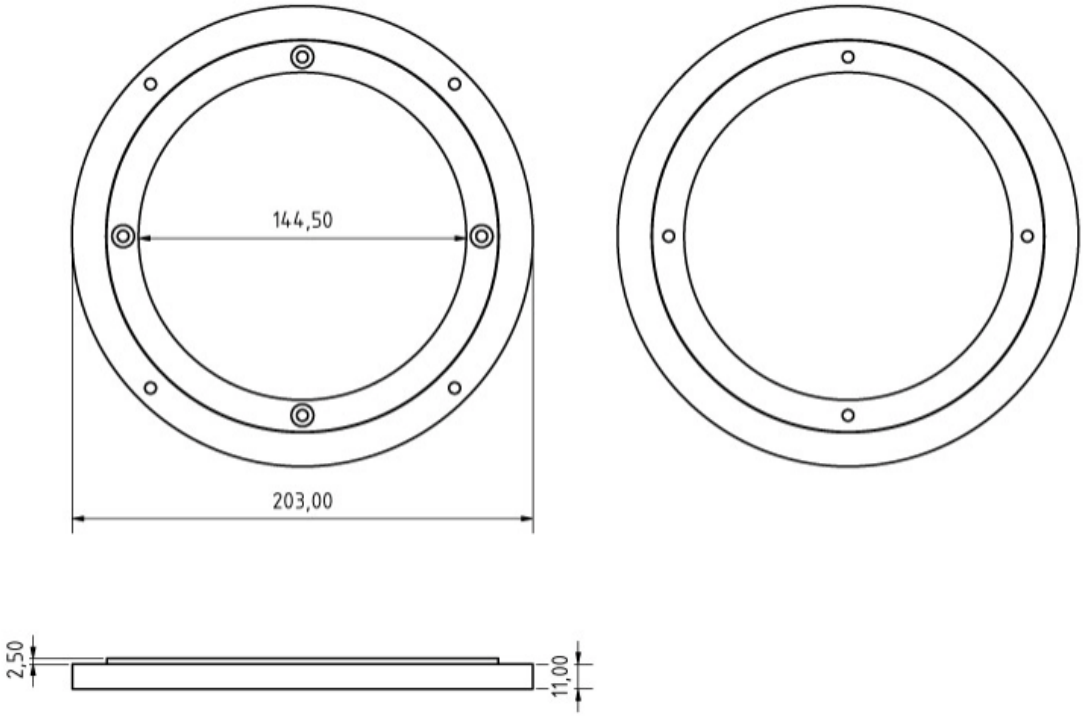

Abbildung 11.6: Technische Zeichnung Drehkranz

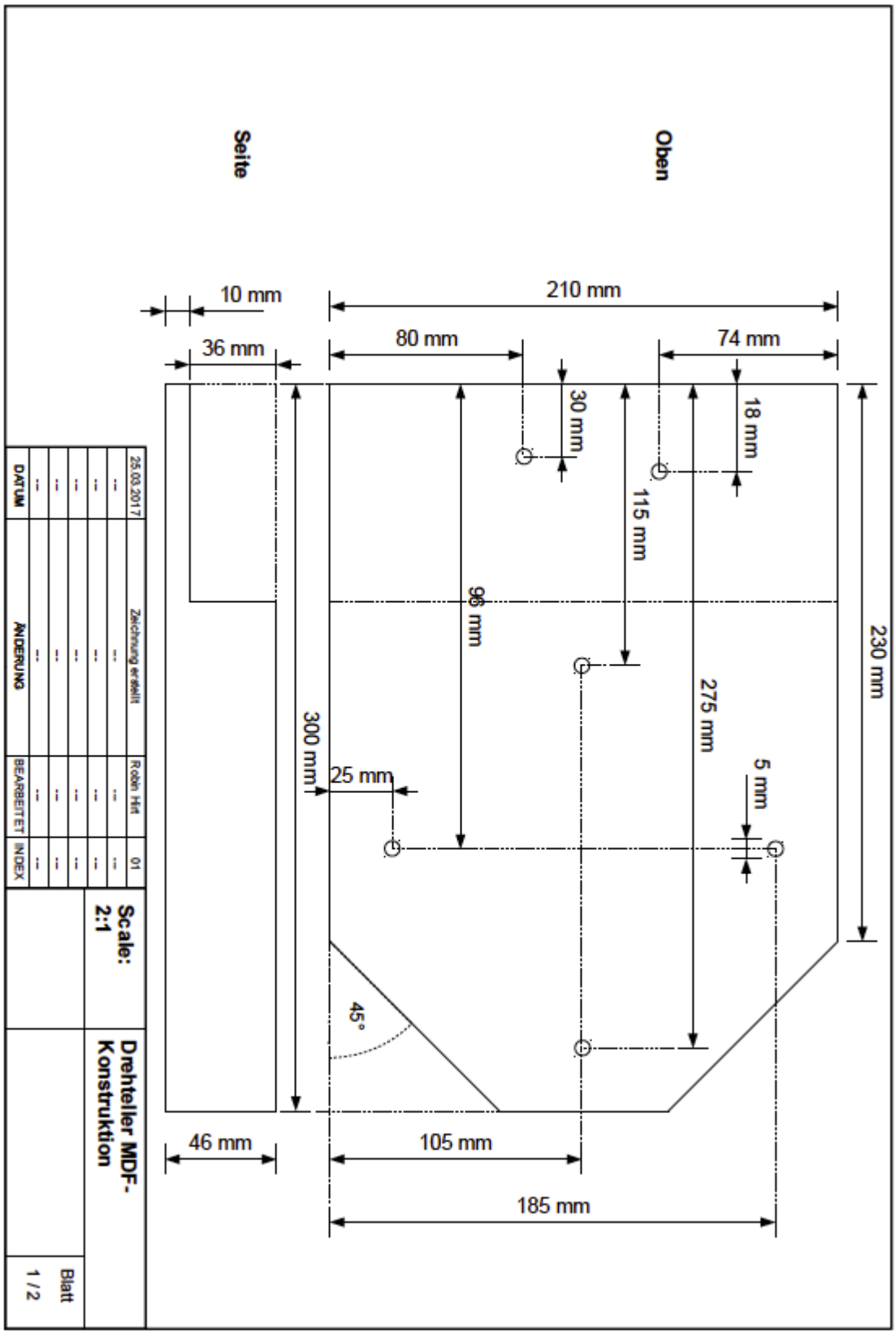

Abbildung 11.7: Technische Zeichnung der Bodenplatte 1 von 2

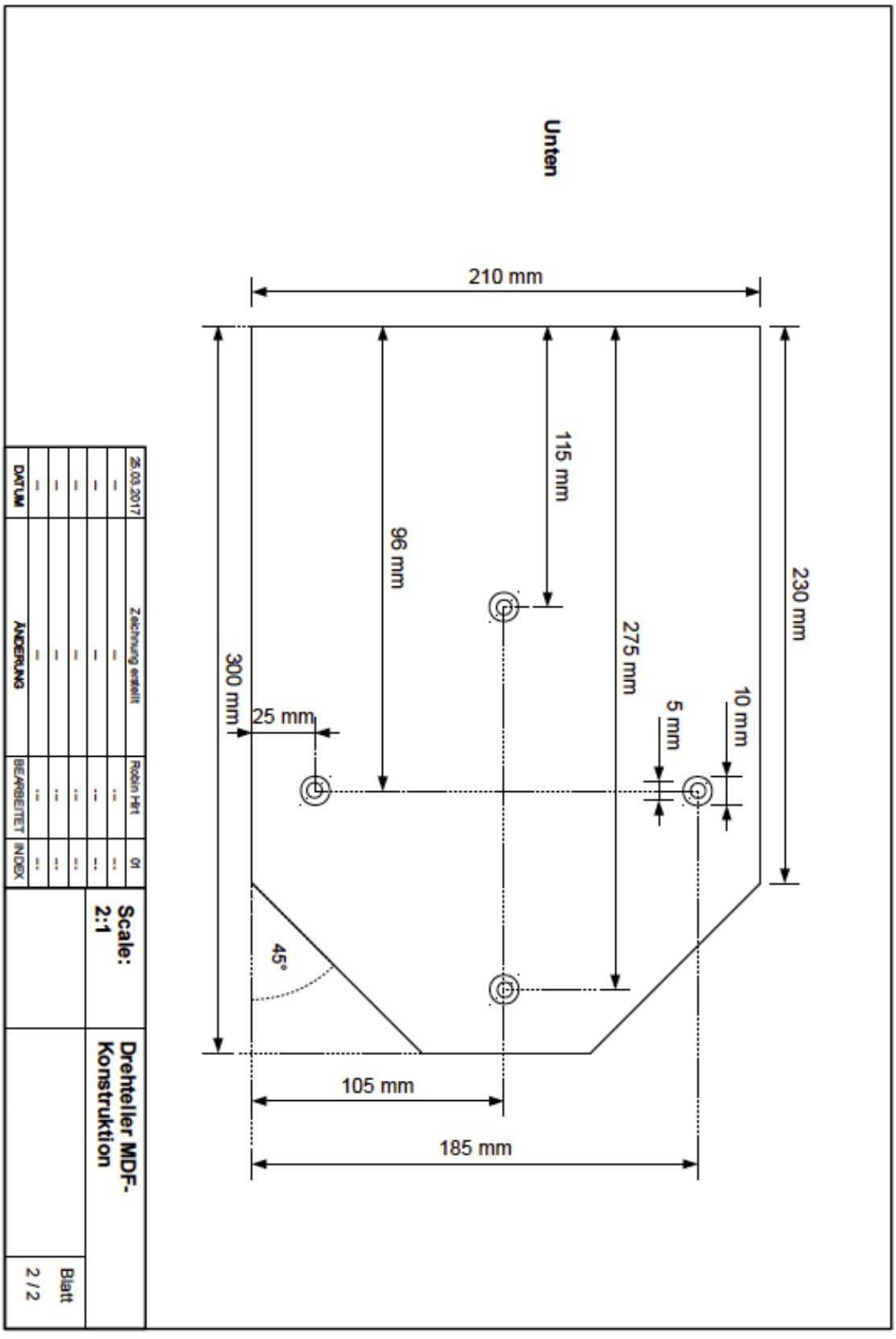

Abbildung 11.8: Technische Zeichnung der Bodenplatte 2 von 2

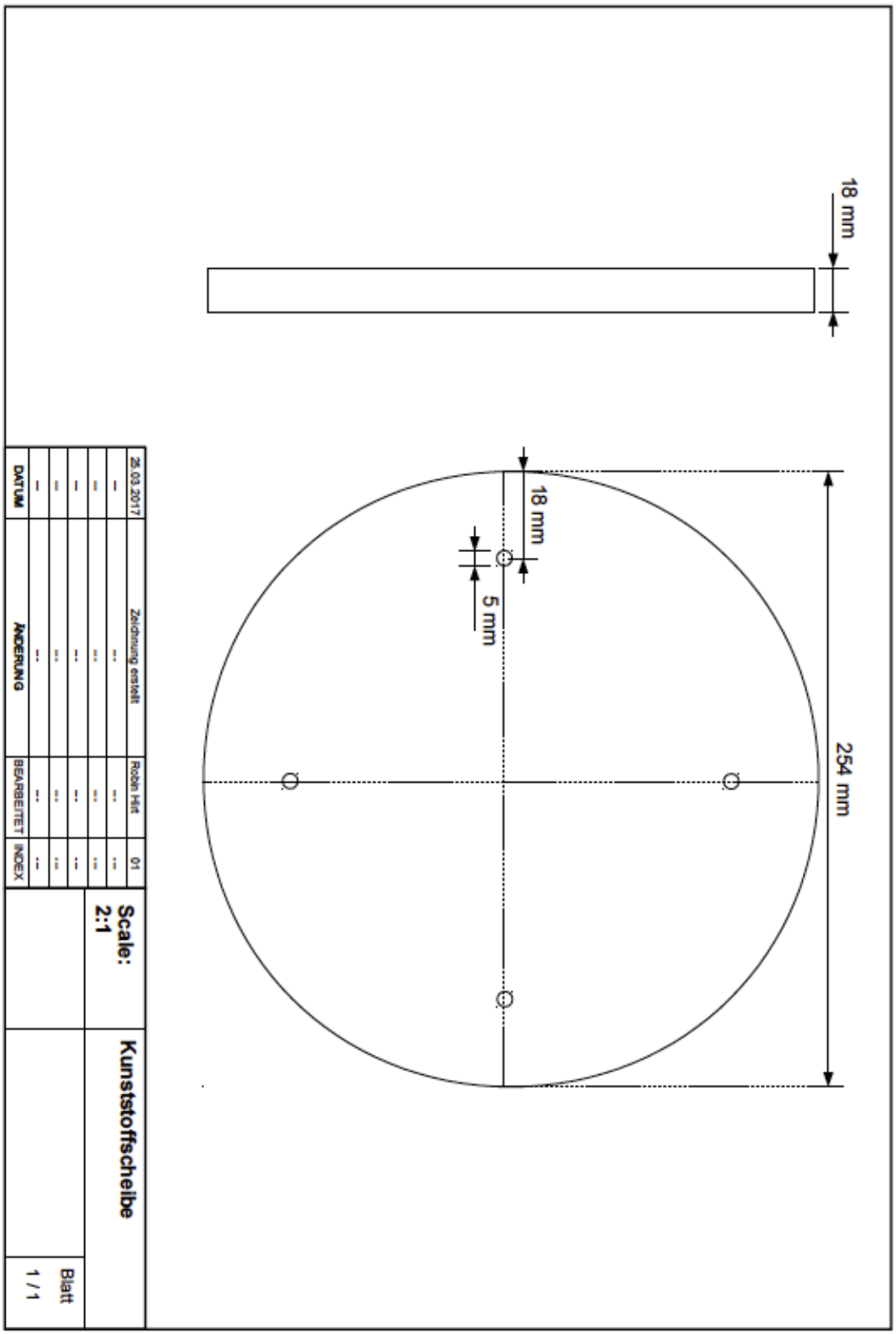

Abbildung 11.9: Technische Zeichnung der Kunststoffscheibe

11.3 E-Plan

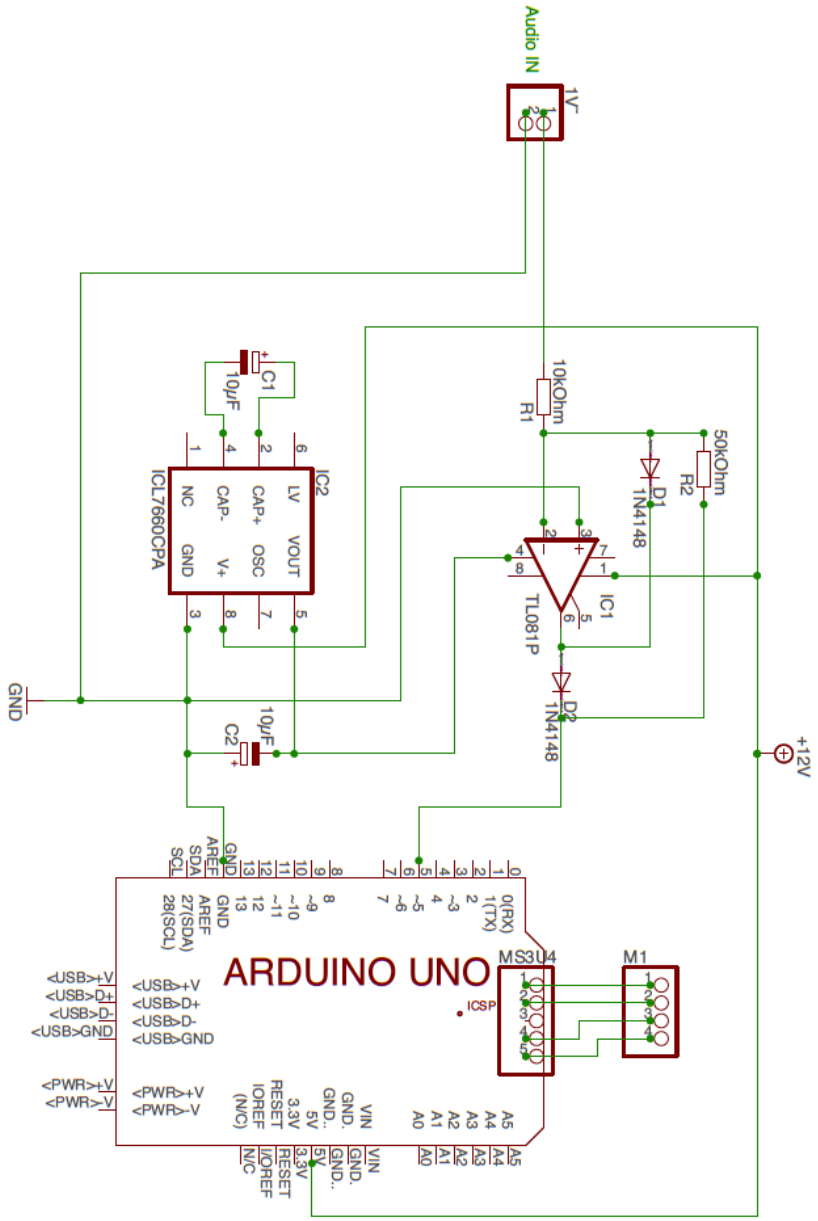

Abbildung 11.10: E-Plan

11.4 Arduino Programm

TheProject | Arduino 1.8.1

TheProject

```
#include <AccelStepper.h>
#include <Wire.h>
#include <Adafruit_MotorShield.h>
#include <FreqCounter.h>

Adafruit_MotorShield AFMSbot(0x61); // Rightmost jumper closed
Adafruit_MotorShield AFMStop(0x60); // Default address, no jumpers

// Connect two steppers with 200 steps per revolution (1.8 degree)
// to the top shield
Adafruit_StepperMotor *myStepper = AFMStop.getStepper(200, 2);

//****************************************   Variablen ****************************************
int maxdata = 0;
int motionactive = 0;
int actpos = 0;
int toggle = 0;
long int freq;
long int freqrough;

//****************************************   Methoden ****************************************

// you can change these to DOUBLE or INTERLEAVE or MICROSTEP!

void forwardstep() {
  myStepper->onestep(FORWARD, DOUBLE);
}
void backwardstep() {
  myStepper->onestep(BACKWARD, DOUBLE);
}

//****************************************   Wrapper  ****************************************

// Now we'll wrap the stepper in an AccelStepper object
AccelStepper stepper(forwardstep, backwardstep);
```

Abbildung 11.11: Arduino Programm Teil 1 von 3

```
//****************************************   Setup    ****************************************
void setup()
{
  Serial.begin(9600);
  AFMSbot.begin(); // Start the bottom shield
  AFMStop.begin(); // Start the top shield


  stepper.setMaxSpeed(4000.0);
  stepper.setSpeed(3000.0);
  stepper.setAcceleration(300.0);
//  stepper.moveTo(500);


}
```

Abbildung 11.12: Arduino Programm Teil 2 von 3

```
//****************************************   Main   ****************************************

void loop()
{

  if (Serial.available() > 0){
    maxdata = Serial.parseInt(); // lesen AN falls serielle Schnittstelle verfügbar
  //  Serial.println(maxdata);
  }

  if (motionactive == 0) {
    FreqCounter::f_comp= 8;             // Kompensation einstellen
    FreqCounter::start(100);           // Zähler starten für 100ms
    while (FreqCounter::f_ready == 0)   // warten bis der Zähler fertig ist

    freqrough=FreqCounter::f_freq;           // Ergebnis lesen
    freq = freqrough*10;                  // Frequenz mit 10 multiplizieren wegen Frequenz Library
    Serial.println(freq);                // und ausgeben
  }


  if (freq > 4000 && freq < 7000 && toggle == 0) {  //Position 1
    stepper.moveTo(300);
    maxdata = 0;
    motionactive = 1;
    toggle = 1;
    Serial.println (maxdata);
    Serial.println ("position xx");
  }
  if (freq > 300 && freq < 600 && toggle == 1) {      //Position 2
    stepper.moveTo(0);
    maxdata = 0;
    motionactive = 1;
    toggle = 0;
    Serial.println (maxdata);
    Serial.println ("position 0");
  }

  actpos = stepper.currentPosition();
  if ((actpos == 0 && toggle == 0) || (actpos >= 300 && toggle == 1)){
    motionactive = 0;
    Serial.println("reset");
  }
    stepper.run();

}
```

Abbildung 11.13: Arduino Programm Teil 3 von 3

11.5 Danksagung

Im Rahmen dieser Abschlussarbeit möchte ich mich ganz herzlich bei allen beteiligten Personen bedanken. Im Besonderen bei:

Prof. Oliver Curdt für die Betreuung.

Jörg Bauer und Heiko Schulz für die Bereitstellung von Studio und technischem Equipment.

Hendrik Paukert und Jonathan Ziegler für tatkräftige Unterstützung und Zuarbeit des Deconvolvers und der Messdaten.

Allen geduldigen Kommilitonen, die sich für die Sprachaufnahmen bereiterklärt haben.

11.6 Digitaler Datenträger

Auf dem beigelegten Datenträger befinden sich:

- Vorliegende Arbeit als PDF-Version
- Die Pro Tools Session „Virtuelle Konferenz“
- Eine Hörprobe des Vergleichs der virtuellen und realen Konferenz
- Arduino Programmcode

Literaturverzeichnis

[Ada17a] Adafruit. AccelStepper Library. https://github.com/adafruit/AccelStepper/blob/master/AccelStepper.h, (2017). Abgerufen am 25. März 2017.

[Ada17b] Adafruit. Motor Shield V2 Library. https://github.com/adafruit/Adafruit_Motor_Shield_V2_Library/blob/master/Adafruit_MotorShield.h, (2017). Abgerufen am 25. März 2017.

[Ard17] Arduino. Arduino Uno. https://www.arduino.cc/en/Main/ArduinoBoardUno, 2017. Abgerufen am 06. März 2017 von Arduino.cc:.

[Avi17] Avid. Pro Tools. https://www.avid.com/de/pro-tools, (2017). Abgerufen am 25. März 2017.

[Bri16] E. B. Brixen. Facts about speech intelligibility. http://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility, 2016. Abgerufen am 06. März 2017 von DPA Microphones.

[Dec17] V. Deconvolver. Voxengo Deconvolver. http://www.voxengo.com/product/deconvolver/, (2017). Abgerufen am 25. März 2017.

[Dic03] *Mikrofon-Aufnahmetechnik: Aufnahmeräume, Schallquellen, Mikrofone, Räumliches Hören, Aufnahmeverfahren, Aufnahme einzelner Instrumente und Stimmen*. Hirzel, Stuttgart; Leipzig, 3., neu bearb. und erw. aufl. edition, 2003.

[Dic14] *Handbuch der Tonstudiotechnik*. de Gruyter, Berlin [u.a.], 8., völlig überarb. und erw. aufl. edition, 2014.

[Dud17] Duden. Duden Konferenz. http://www.duden.de/rechtschreibung/Konferenz, 2017. Abgerufen am 06. März 2017 von Duden.

[Dudum] Duden. Die häufigsten Wörter in deutschsprachigen Texten. http://www.duden.de/sprachwissen/sprachratgeber/die-haeufigsten-woerter-in-deutschsprachigen-texten, (kein Datum). Abgerufen am 08. März 2017 von Duden.

[eaum] G. et al. Grundwortschatz 100. http://www.gut1.de/grundwortschatz/grundwortschatz_100.html, (kein Datum). Abgerufen am 08. März 2017 von Duden.

[EKum] Elektronik-Kompendium. Operationsverstärker. https://www.elektronik-kompendium.de/sites/bau/0209092.htm, (kein Datum). Abgerufen am 08. März 2017 von Elektronik Kompendium.

[Flo16] K. Florian. ECHTHALL UND HALLEMULATIONEN BEI MUSIKPRODUKTIONEN. https://www.hdm-stuttgart.de/~curdt/Kontny.pdf, (2016). Abgerufen am 20. März 2017.

[Fri14] *Die Audio-Enzyklopädie: ein Nachschlagewerk für Tontechniker*. de Gruyter Saur, Berlin, 2. aufl. edition, 2014.

[Gen03] Genelec. 1029A Data Sheet. `http://www.genelec.com/sites/default/files/media/Previous%20models/2-way%20loudspeakers/1029A/ds1029a.pdf`, (2003). Abgerufen am 25. März 2017.

[Goe02] *Mikrofone in Theorie und Praxis*. Elektor-Verl., Aachen, 6. aufl. edition, 2002.

[Goe15] *Tontechnik: Hören, Schallwandler, Impulsantwort und Faltung, digitale Signale, Mehrkanaltechnik, tontechnische Praxis; mit 33 Tabellen*. Fachbuchverl. Leipzig im Carl-Hanser-Verl., München, 4., aktualisierte aufl. edition, 2015.

[Lom11] *Le signe de l'élévation de la voix. Annales des Maladies de L'Oreille et du Larynx 37: 101–119*. 1911.

[LSS+07] H. Lazarus, C. Sust, R. Steckel, M. Kulka, and P. Kurtz. *Akustische Grundlagen sprachlicher Kommunikation*. Springer Berlin Heidelberg, 2007.

[Moe15] *Technische Akustik*. Springer-Verlag GmbH Berlin Heidelberg, Berlin, Heidelberg, 10. aufl. 2015 edition, 2015.

[Naw12] M. Nawrath. Arduino Frequency Counter Library. `http://interface.khm.de/index.php/lab/interfaces-advanced/arduino-frequency-counter-library/`, (2012). Abgerufen am 25. März 2017.

[noa03] *Effekte & dynamics: das Salz in der Suppe jeder Musik-Mischung: Technik und Praxis der Effekt- und Dynamikbearbeitung*. PPV, Presse-Project-Verl., Bergkirchen, 3., erw. aufl. edition, 2003.

[noa15] *Sprachen der Welt: warum sie so verschieden sind und sich doch alle gleichen*. WBG (Wiss. Buchges.), Darmstadt, 2015.

[Pol07] R. Polikar. Discrete Convolution. `http://www.astro.rug.nl/~vdhulst/SignalProcessing/Hoorcolleges/college03.pdf`, (2007). Abgerufen am 25. März 2017.

[Sch82] *Vokale Kommunikation: nonverbale Aspekte des Sprachverhaltens*. Beltz, Weinheim, 1982.

[Sch17] Schoeps. CCM 4V. `http://www.schoeps.de/de/products/ccm4v`, (2017). Abgerufen am 25. März 2017.

[Sen02] U. B. Sengpiel. Berechnung der Bewertungsfilter-Kurven. `http://www.sengpielaudio.com/BerechnungDerBewertungsfilter.pdf`, 2002. Abgerufen am 07. März 2017 von Sengpielaudio.

[The98] *Akustische Kommunikation : Grundlagen mit Hörbeispielen. - BUCH.* Springer Berlin Heidelberg, Berlin, 1998.

[Wik16a] Wikipedia. Plosiv. `https://de.wikipedia.org/wiki/Plosiv`, 2016. Abgerufen am 06. März 2017 von Wikipedia.

[Wik16b] Wikipedia. Wikipedia Konferenz. `https://de.wikipedia.org/wiki/Konferenz`, 2016. Abgerufen am 06. März 2017 von Wikipedia.

[Wik17a] Wikipedia. Arduino (Plattform). `https://de.wikipedia.org/wiki/Arduino_(Plattform)`, 2017. Abgerufen am 06. März 2017 von Wikipedia.

[Wik17b] Wikipedia. Vokal. `https://de.wikipedia.org/wiki/Vokal`, 2017. Abgerufen am 06. März 2017 von Wikipedia.

All links were last followed on 1. April 2017