Universität Rostock

Traditio et Innovatio

Masterarbeit

Ultraschall Datenschutz

Vorgelegt von:
Jan Heisenberg
Matrikel-Nr.: 217203561

Eingereicht am:
18.10.2023

Betreuer:
Dr. Jonas Flint

Erstgutachter:
Prof. Dr. rer. nat. Clemens H. Cap

Zweitgutachter:
Prof. Dr.-Ing. habil. Gero Mühl

UNIVERSITÄT ROSTOCK | FAKULTÄT FÜR INFORMATIK UND ELEKTROTECHNIK
Lehrstuhl für Informations- und Kommunikationsdienste
IEF / Informatik
Albert-Einstein-Straße 22, 18059 Rostock | Tel: +49 381 498-7501

Vorwort

Ich möchte mich bei Prof. Dr. Clemens H. Cap dafür bedanken, dass ich dieses Thema bearbeiten durfte. Es ist ein Thema, mit dem ich während meines Studiums nicht direkt in Berührung gekommen bin. Ebenso möchte ich mich für die Weiterleitung an Dr. Jonas Flint von der DEJ Technology GmbH bedanken. Dr. Flint hat mich bei der Bearbeitung des Themas betreut und mir während der gesamten Arbeit seinen Rat gegeben. Ebenso waren Dr. Flint und Prof. Cap jederzeit bereit, Feedback zum aktuellen Stand zu geben. Ein weiteres Dankeschön an Bisrat aus der DEJ Technology GmbH. Bisrat hat mir einen Schnellkurs über neuronale Netze gegeben und war immer für meine Fragen zu dessen bereit.

Betreuer: Dr. Jonas Flint

Tag der Ausgabe: 31.05.2023
Tag der Abgabe: 18.10.2023

Inhaltsverzeichnis

Abkürzungsverzeichnis

CER	Character Error Rate
CNN	Convolutional Neural Network
VCTK	CSTR VCTK Corpus
WER	Word Error Rate

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Der Begriff Datenschutz hat in den letzten Jahren stetig an wachsender Bedeutung gewonnen und diese Entwicklung beschränkt sich keineswegs ausschließlich auf die IT-Branche. Allerdings wird der Datenschutz eher seltener in Zusammenhang mit Ultraschall gebracht. Der Begriff Ultraschall wird umgangssprachlich eher für die medizinische Sonografie verwendet. Eine andere Bedeutung aber ist der Frequenzbereich des Schalles, welcher nicht mehr von Menschen wahrgenommen werden kann. Diese Form des Ultraschalles wird von vielen elektronischen Geräten unseres Alltages teilweise unterstützt. Die meisten generischen Lautsprecher und Mikrofone sind in der Lage, einen kleinen Teil des Frequenzbereichs von Ultraschall zu senden bzw. aufzunehmen. Dementsprechend fallen auch Smartphones und Sprachassistenten in diese Kategorie. Amazon und Google verwenden bereits Ultraschall in ihren Sprachassistenten, um zu erkennen, ob sich Personen innerhalb eines Raumes befinden und wie weit entfernt diese vom Gerät sind [Tuo21]. Aus der weit verbreiteten Unterstützung dieser Technologie ergeben sich einige Probleme im Bereich des Datenschutzes.

1.1 Hinweise und Annahmen

Diese Arbeit baut auf einer existierenden Bachelorarbeit über die Erstellung eines Entwurfes für eine Ultraschall-Web-API auf. Die Idee hinter der API ist das Trennen von Ultraschall und Hörschall in Webanwendungen. Der Grund für diese Trennung ist ebenfalls der Datenschutz. Da keine der üblichen Betriebssysteme sowie keine der üblichen Browser zwischen Hörschall und Ultraschall unterscheiden können, kommt es zu dem Problem, dass Ultraschallanwendungen vollständige Mikrofonrechte einfordern müssen, um ihre Funktionen auszuführen. Der Endnutzer weiß dabei natürlich nicht, dass solche Anwendungen die vollständigen Rechte auch für böswillige Taten ausnutzen können. Die direkte Trennung dieser beiden Schallbereiche für das Einfordern der Mikrofonrechte könnte somit die Nutzung der Daten des jeweils anderen Schallbereiches verhindern. Da diese Arbeit auf der Konzeptebene darauf aufbauen soll, wird die Annahme gemacht, dass die hier verwendeten Geräte in der Lage sind, Ultraschall von Hörschall zu trennen. Somit können die Geräte mit Hilfe einer Audioaufnahme ausschließlich den Ultraschallbereich aufnehmen. Die Bachelorarbeit macht außerdem die Annahme, dass die Grenze zwischen den Schallbereichen bei 16 kHz liegt. Diese Annahme stammt aus der DIN 1320 [Deu09] und wird ebenfalls in diese Arbeit übernommen.

Ein weiterer wichtiger Teil dieser Arbeit ist die Indoor Positioning Lösung **Koopango**

[Kooa]. Koopango erlaubt das einfache Navigieren von Personen innerhalb von Gebäuden über eine Smartphone App. Dies könnten beispielsweise unter anderem der Einzelhandel oder Krankenhäuser sein. Koopango verwendet für die Positionsbestimmung Ultraschallsignale. Die Signale werden über Lautsprecher innerhalb der Gebäude übertragen und von der App des Smartphones aufgenommen. Mit Hilfe eines Algorithmus ist Koopango in der Lage, anhand der Ultraschallsignale die genaue Position zu berechnen. Somit kann eine Person in Echtzeit das ausgestattete Gebäude navigieren.

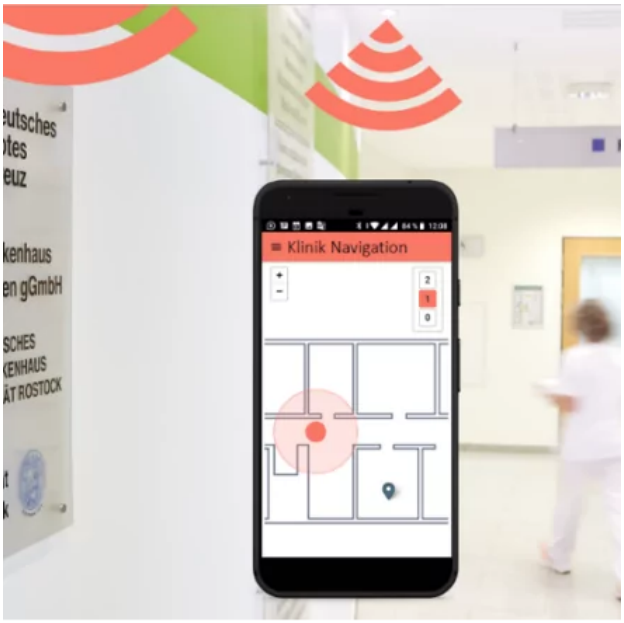

Abbildung 1.1: Koopango Beispiel [Koob]

1.2 Problemstellung

Folgendes Szenario wird für den Verlauf dieser Arbeit eine wichtige Rolle spielen: Person A wurde nach einem Unfall in ein Krankenhaus eingeliefert und kann mittels Koopango das Krankenhaus problemlos navigieren. Nun hat Person A einen Termin zur ärztlichen Untersuchung in Raum 207. Koopango navigiert Person A in Richtung des Raumes. Auf dem Weg dorthin unterhalten sich Person B und Person C auf dem Flur und sprechen dabei über vertrauliche Informationen. Koopango benötigt für die Navigation einen permanenten Zugriff auf das Mikrofon des Smartphones von Person A. Unter der Annahme aus Abschnitt 1.1, dass das Smartphone ausschließlich Ultraschall aufnehmen kann, stellt sich nun die folgende Frage. Bricht Person A den Datenschutz, indem die Ultraschallsignale von Person B und Person C aufgenommen werden? Hierfür muss ein Blick in das Gesetz geworfen werden, denn laut diesem verstößt die Aufnahme gegen die **Vertraulichkeit des Wortes**:

§ 201 Verletzung der Vertraulichkeit des Wortes
(1) Mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe wird bestraft,

1. wer unbefugt das nichtöffentlich gesprochene Wort eines anderen auf einen Tonträger aufnimmt oder

2. eine so hergestellte Aufnahme gebraucht oder einem Dritten zugänglich macht.

siehe § 201 Absatz 1 StGB.

1.3 Fragestellungen

An diesem Punkt ist nun immer noch unklar, ob solch eine Ultraschallaufnahme zu der Aufnahme des Wortes zählt. Dies geht allerdings über diese Arbeit hinaus. Nichtsdestotrotz ist dies eine Anregung für die erste Fragestellung dieser Arbeit. Diese lautet wie folgt: **Können aus Ultraschallsignalen eines Gespräches die Sprachsignale rekonstruiert werden?** Das Ziel dieser Frage ist es herauszufinden, ob die Sprache überhaupt rekonstruiert werden kann. Wenn sich herausstellt, dass dies möglich ist, muss untersucht werden, wie akkurat und konsistent die Sprache wiederhergestellt werden kann. Für diese Frage wird die Literatur nach existierenden Ultraschallsystemen durchsucht. Zusätzlich werden bestehende Spracherkennungssysteme mit Ultraschallaufnahmen getestet. Nach Bearbeitung dieser Fragestellung wird das Augenmerk wieder mehr auf den Einsatz von Koopango gerichtet. Wenn sich die Rekonstruktion als möglich herausstellt, wäre dies ein Problem für Koopango. Daraus ergibt sich die zweite Fragestellung, welche wie folgt lautet: **Wie kann die Aufnahme limitiert werden, sodass die Gespräche nicht mehr rekonstruiert werden können ohne dabei die Genauigkeit der Positionsbestimmung zu stark einzuschränken?** Ziel hinter dieser Frage ist natürlich der Versuch, den Datenschutz wiederherzustellen. Dafür wird untersucht, ob beispielsweise über das Hinzufügen von Rauschen die Rekonstruktion verhindert werden kann. Dabei muss allerdings beachtet werden, dass die Positionsbestimmung von Koopango weiterhin möglich ist. Denn auch der Algorithmus der Positionsbestimmung könnte unter dem Rauschen leiden.

2 Stand der Technik

2.1 Sprachrekonstruktion aus Ultraschallsignalen

Die Verarbeitung von Audiodaten für die Rekonstruktion von Sprache ist ein wichtiger Bestandteil unseres Alltags. Dazu gehört das Nutzen von Sprachassistenten, welche die Stimmen der Nutzer in Kommandos für den Sprachassistenten umwandeln und somit die Steuerung des Gerätes mittels Sprache erlauben. Smartphones verwenden Transkription, um die eingesprochenen Stimmen in Text umzuwandeln. Ebenso ist die Umwandlung in die entgegengesetzte Richtung möglich mit Text-to-Speech Anwendungen. Text-to-Speech ist eine Methode, um geschriebenen Text in hörbaren Ton umzuwandeln. All diese Techniken verwenden in der Verarbeitung der Audiodaten Frequenzen unterhalb des Ultraschallbereiches und fallen somit nicht in den zu untersuchenden Teil dieser Arbeit. Allerdings gibt es auch wissenschaftliche Forschungen im Bereich der Sprachrekonstruktion aus Ultraschall.

2.1.1 Lippenlesen

Das Lesen von Lippen mit Hilfe von Ultraschall ist eine der verbreitetsten Techniken, um Sprache aus Ultraschall Audio zu rekonstruieren. Es gibt bereits einige Forschungen, welche erfolgreich Sprache mit sehr hohen Genauigkeiten rekonstruieren konnten. Dazu gehören Systeme wie SVoice [Fu+22], EchoWhisper [Gao+20] und SoundLip [Zha+21]. Die Grundidee hinter dieser Technik ist immer dieselbe. Die Testperson hält ein Smartphone sehr nahe an den Mund und spricht die vorgegebenen Sätze ein. Das Smartphone sendet währenddessen durchgehend Ultraschallsignale ab. Die gesendeten Signale prallen dabei von dem Gesicht des Sprechers ab und werden wieder vom Smartphone empfangen. Die Unterschiede in den empfangenen Signalen spiegeln die Eigenschaften des Sprechens wider. Zu den Eigenschaften gehören zum Beispiel die Position des Kinns, der Lippen und der Zunge. Die Motivation hinter diesen Forschungen ist der Datenschutz. Wenn zwei Menschen persönliche Informationen untereinander in der Öffentlichkeit austauschen, können Personen in der Nähe diese Informationen mitschneiden. Die Idee, welche von den Forschungen verfolgt wird, ist das lautlose Einsprechen der Informationen in ein Smartphone, indem nur die Lippen bewegt werden. Somit kann verhindert werden, dass andere Personen die Informationen hören können. UltraSpeech ist eine weitere Forschung, welche dieselbe Technik verwendet, um die Rekonstruktion von Sprache an Orten mit viel Rauschen zu verbessern [Din+22]. UltraSpeech ist in der Hinsicht anders, dass nicht ausschließlich Ultraschall für die Rekonstruktion verwendet wird.

Die genannten Forschungen erreichen mit Hilfe dieser Systeme sehr hohe Genauigkeiten in der Rekonstruktion von Wörtern. Für die Darstellung der Ergebnisse werden 2 unterschiedliche Parameter verwendet. Einige Forschungen verwenden die Character Error Rate (CER), während andere die Word Error Rate (WER) benutzen. Für die Berechnungen der Werte wird die Levenshtein-Distanz genutzt. Die CER besagt, wie viel Prozent der gesamten Buchstaben falsch klassifiziert wurden, und die WER besagt, wie viel Prozent der gesamten Wörter falsch klassifiziert wurden. Die WER ist somit immer größer als die CER aber niemals kleiner, da die Anzahl an Buchstaben innerhalb eines Satzes immer größer ist als die Anzahl der Wörter. Dies sollte bedacht werden, wenn Vergleiche zwischen unterschiedlichen Ergebnissen evaluiert werden.

In Tabelle 2.1 sind die erreichten Error Rates der Forschungen aufgelistet. UltraSpeech ist in dieser Tabelle nicht eingetragen, da es nicht ausschließlich mit Hilfe von Ultraschall Sprache rekonstruiert und somit nicht wirklich vergleichbar ist. Zusätzlich geben die Autoren keinen definitiven Wert für die WER oder CER an und konzentrieren sich mehr auf die Reduzierung des Rauschens. Nichtsdestotrotz lässt sich aus den dargestellten Graphen eine WER ablesen, die schätzungsweise zwischen 45% und 65% liegt. Das ausschlaggebende Ergebnis bei UltraSpeech ist die Verbesserung der Audiorekonstruktion um 40% an Orten mit viel Rauschen.

Forschung	**WER/CER**
SVoice	7.62% CER
EchoWhisper	8.33% WER
SoundLip	12.56% CER

Tabelle 2.1: WER und CER der einzelnen Forschungen

Die Werte für SVoice und SoundLip wurden direkt aus [Fu+22] übernommen, während der Wert von EchoWhisper aus dem entsprechenden Paper übernommen wurde [Gao+20]. Diese Systeme verfolgen alle dasselbe Ziel in der Erstellung einer lautlosen Schnittstelle. Der Unterschied liegt in der Methodik und der Aktualität der Forschungen. SVoice ist das aktuellste dieser Paper und erreicht mit 7.62% CER die beste Performance.

2.1.1.1 Limitierungen des Lippenlesens

Die Technik des Lippenlesens mittels Ultraschall erzielt erstaunliche Ergebnisse. Allerdings hat auch diese Technik ihre Limitierungen, denn die Ergebnisse wurden in sehr kontrollierten Umgebungen erreicht. Einige der Limitierungen sind dabei folgende:

1. **Reichweite**
   Die Reichweite ist die größte Einschränkung, welche das Lippenlesen limitiert. Die einzelnen Forschungen verwenden nicht immer die exakt gleichen Distanzen. Trotzdem kann aus den Forschungen entnommen werden, dass die Reichweite auf 2cm bis maximal 10cm beschränkt ist. Wenn das Smartphone zu nahe am Mund ist, kann der Ultraschall die Mundbewegungen nicht mehr vollständig erfassen [Fu+22]. Diese Limitierung ist für die Anwendungsgebiete der Forschungen akzeptabel. Für den in dieser Arbeit untersuchten Anwendungsfall ist auf Grund dieser Limitierung das Lippenlesen ungeeignet.

2. **Neigungsgrad**
   Nicht nur die Reichweite ist sehr begrenzt, sondern auch der Winkel zwischen Smartphone und Mund spielt eine wichtige Rolle. SVoice ist unter optimalen Bedingungen in der Lage, mit einer maximalen Spanne von -60° bis 60° Wörter zu reproduzieren. Allerdings sind die Ergebnisse bei solch einem Neigungswinkel nicht mehr konsistent. Somit liegt die wahre Spanne, um stabile Ergebnisse zu erhalten, über die unterschiedlichen Forschungen bei -15° bis 15°.

3. **Bewegungen**
   Die sprechenden Personen sitzen bei den Tests still. Die Bewegungen der Testpersonen führen zu ungewollten Artefakten im Ultraschallbereich und verringern somit die Genauigkeit der Rekonstruktion.

2.2 Sprachrekonstruktion aus Hörschall

Die Rekonstruktion von Sprache aus Ultraschall Audio limitiert sich im Stand der Technik völlig auf das Lippenlesen. Auch wenn diese Technik unter optimalen Bedingungen gute Ergebnisse liefert, ist es keine Lösung für das Problem, welches bei der Nutzung von Koopango entsteht. Spracherkennungsmethoden im menschlich hörbaren Bereich sind bereits sehr ausgeprägt. Zu den verbreitetsten und effektivsten Methoden für die Spracherkennung gehören neuronale Netze im Bereich der künstlichen Intelligenz. Es gibt bereits sehr viele Forschungen, welche unterschiedliche Konfigurationen von Netzwerkmodellen untersuchen und evaluieren. Darunter unter anderem Quartznet [Kri+20], ContextNet [Han+20], BigSSL [Zha+22], Conformer [Gul+20] und Whisper [Rad+22]. Keines der genannten Systeme verwendet dabei Ultraschalldaten für die Rekonstruktion. Ebenso lassen sich keine Forschungen finden, welche auf diese Art und Weise versuchen Stimmen mit Ultraschall zu rekonstruieren.

2.3 Menschliche Sprache im Ultraschallbereich

An der Stelle stellt sich die Frage, ob die menschliche Sprache im Ultraschallbereich überhaupt repräsentiert wird. Dieselbe Frage haben sich die Autoren von SUPERVOICE [Guo+22] gestellt. SUPERVOICE verfolgt das Ziel, unterschiedliche Sprecher zu identifizieren. Hierfür gibt es bereits funktionierende Systeme, allerdings verwenden diese ausschließlich den hörbaren Schall. SUPERVOICE verwendet ebenso hörbaren Schall, um Personen zu identifizieren mit dem Unterschied, dass zusätzlich zum hörbaren Schall Ultraschall verwendet wird, um die Performance noch ein wenig zu verbessern. Somit ist SUPERVOICE eine Kombination aus Ultraschall und hörbarem Schall.

Der für diese Arbeit interessante Teil dieser Forschung ist die Untersuchung, ob Menschen Ultraschall produzieren können. Die Forschung geht dabei ins Detail über die verschiedenen Kategorien an Tönen, die Menschen produzieren können und welche davon Ultraschall produzieren. Die folgenden Informationen aus SUPERVOICE spielen dabei eine wichtige Rolle für diese Arbeit. Menschliche Sprache wird in Vokale und Konsonanten unterteilt. Der wichtige Unterschied ist, dass Vokale fast ausschließlich über die Stimmbänder produziert werden, während Konsonanten hauptsächlich mit Hilfe von Luftströmen gebildet werden. Diese Luftströme werden beim Sprechen durch Engpässe geführt, die Menschen innerhalb des Mundes bilden. Dadurch entsteht Reibung und die dabei ausgesprochenen Konsonanten werden **Frikative** genannt. Weiterhin können Konsonanten gebildet werden, indem die Luftströme kurzzeitig gestoppt und dann plötzlich freigelassen werden. Die dabei entstehenden Konsonanten werden **Plosive** und **Affrikate** genannt. Das Paper stellt als wichtiges Ergebnis dar, dass genau diese Frikative, Affrikate und Plosive Ultraschall produzieren können.

2.3.1 Untersuchung des menschlichen Ultraschalles

Der Stand der Technik für die Erkennung von Sprache im Ultraschallbereich limitiert sich zu diesem Zeitpunkt auf das Lippenlesen. Diese Technik bringt auch noch einige Limitierungen mit sich, sodass weitere Untersuchungen in diesem Themenbereich nötig sind. Im Folgenden wird die Feststellung von SUPERVOICE, dass Menschen mit Hilfe von Frikativen, Affrikativen und Plosiven Ultraschall produzieren, untersucht. Für die Untersuchungen werden mehrere Spektrogramme von selbst eingesprochenen Audioaufnahmen erstellt.

2.3.1.1 Spektrogramme

Eine Art, um Informationen über Audio darzustellen, ist die Konvertierung der Daten in ein Spektrogramm. Spektrogramme sind wie eine Heatmap und stellen die Intensität der Töne dar. Die Skala für die Intensität ist in vielen Fällen Dezibel. Das heißt, die heller dargestellten Frequenzen sind lauter als die dunkleren. Die Achsen sind in den meisten Fällen die Frequenz und die Zeit. Spektrogramme visualisieren Informationen,

die in den reinen Daten nicht erkennbar sind. Somit können akustische Eigenschaften hervorgehoben werden. Darunter unter anderem die Tonhöhe oder besondere Muster. Viele der folgenden Abbildungen werden die Audioeigenschaften mit Spektrogrammen darstellen.

2.3.1.2 Vokale

Die Vokale der menschlichen Sprache beinhalten die Töne für die Buchstaben **A, E, I, O** und **U**. Diese Töne werden fast ausschließlich mit den Stimmbändern produziert und tragen laut SUPERVOICE wenig bis keine Energie im Ultraschallbereich. In Abbildung 2.1 werden die Vokale in einem Spektrogramm dargestellt. Dieses Spektrogramm stellt den Frequenzbereich von 0 Hz bis 24.000 Hz dar. Das Mikrofon nimmt in dem Aufnahmeraum immer ein Standardrauschen wahr. Das Standardrauschen existiert von 0 Hz bis ungefähr 21 kHz. Ein wichtiger Begriff aus der Signaltheorie ist die sogenannte Nyquist-Frequenz. Die Nyquist-Frequenz entspricht der Hälfte der Abtastrate und solange das Audiosignal eine Frequenz kleiner als die Nyquist-Frequenz hat, entstehen keine Verzerrungen bei der Abtastung [Sha49]. Im Umkehrschluss heißt dies, dass die Abtastrate mindestens doppelt so hoch sein sollte wie die Frequenz des aufzunehmenden Audiosignals. Daraus kann beispielsweise abgeleitet werden, dass die Aufnahme mit einer Abtastrate von 44.100 Hz erstellt wurde. Innerhalb des Spektrogrammes ist bei 16 kHz eine Grenze eingezeichnet. Dies ist die Grenze zwischen hörbarem Schall und Ultraschall, welche für diese Arbeit zu Beginn als Annahme festgelegt wurde. Zusätzlich werden die einzelnen Vokale mit weißen Vierecken markiert. Dies soll dabei helfen zu erkennen, wann ein Ton in den Ultraschallbereich einschreitet. Der Großteil der Informationen liegt anscheinend unterhalb von 5 kHz. Mit steigender Frequenz wird die Menge an Informationen aber immer geringer. Ab einer Frequenz von 10 kHz ist die Informationsmenge schon verschwindend gering. Keines der Vokale hat einen Ausschlag, welcher größer als 16 kHz ist.

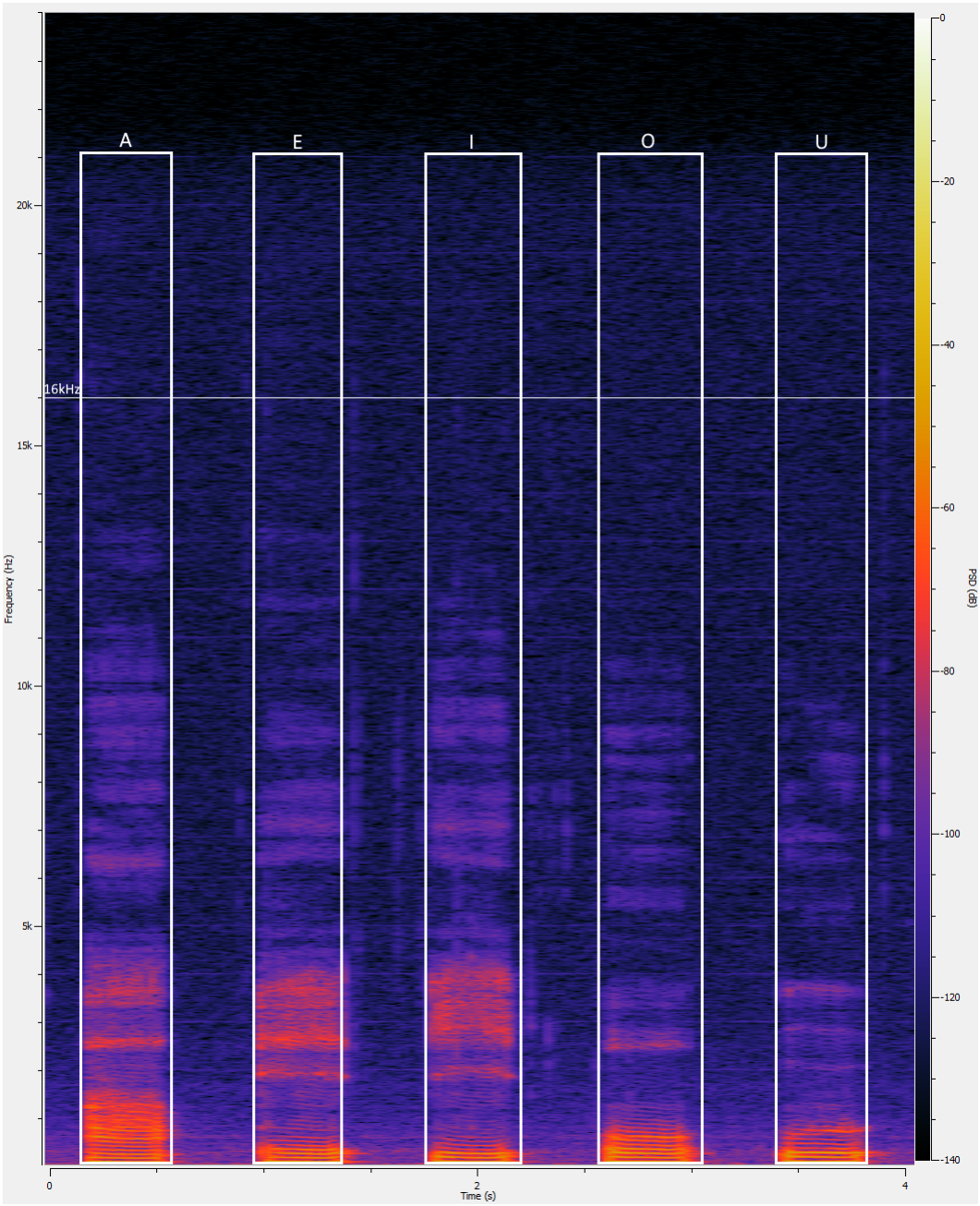

Abbildung 2.1: *A E I O U* Spektrogramm(Vokale)

2.3.1.3 Konsonanten

Für die Konsonanten werden beim Sprechen zusätzlich zu den Stimmbändern die Luftströme innerhalb des Mundes verwendet. Durch die dabei verursachte Reibung wird Energie im Ultraschallbereich generiert. Eine Form der Konsonanten sind die Plosive. Dabei wird der Luftstrom im Mund für eine kurze Weile blockiert und dann plötzlich freigelassen. Damit können unter anderem Töne wie **T, D, K, P, Q** produziert werden. Abbildung 2.2 zeigt dieselbe Art von Spektrogramm wie bei den Vokalen, nur diesmal mit einigen Plosiven. Hier ist gut zu erkennen, dass die Plosive auf jeden Fall weiter in den Ultraschallbereich dringen als die Vokale. Nichtsdestotrotz ist die Menge an Informationen in diesem Bereich immer noch verschwindend gering.

Das erstaunlichste Spektrogramm entsteht allerdings bei dem Test der Frikative und Affrikate. Frikative sind Töne wie **S, F, V**, welche durch Reibung mit einem konstanten Luftstrom entstehen. Affrikate nehmen sich nicht viel von Frikativen, da sie eine einfache Kombination aus einem Plosiv gefolgt von einem Frikativ sind. Bei einem Blick auf Abbildung 2.3 ist sofort zu erkennen wie viel Energie von Frikativen und Affrikaten im Ultraschallbereich vorhanden ist, da der Unterschied zwischen dieser Abbildung und der restlichen sehr stark ist. Dennoch sollte beachtet werden, dass die Aufnahmen der drei ersten Spektrogramme laut und direkt in das Mikrofon hinein gesprochen wurden. Bei einem komplett normal gesprochenen Satz sind die Ausschläge geringer.

Aufgrund dessen wurde ein weiteres Spektrogramm erstellt, um die Darstellung eines vollständigen Satzes abzubilden. Der hier ausgesprochene Satz lautet: **Ich spreche für die Testaufnahme.** Innerhalb der Abbildung 2.4 sind dieses Mal zusätzlich die wichtigsten Konsonanten lila markiert. Diese Abbildung zeigt sehr eindeutig, dass innerhalb des Ultraschallspektrums definitiv Informationen über die Sprache eines Menschen vorhanden sind. Zu beachten ist hier der Unterschied oberhalb und unterhalb der 16 kHz Grenze. Die reine Informationsmenge ist oberhalb der Grenze immer noch extrem gering im Vergleich zum Bereich unterhalb der Grenze.

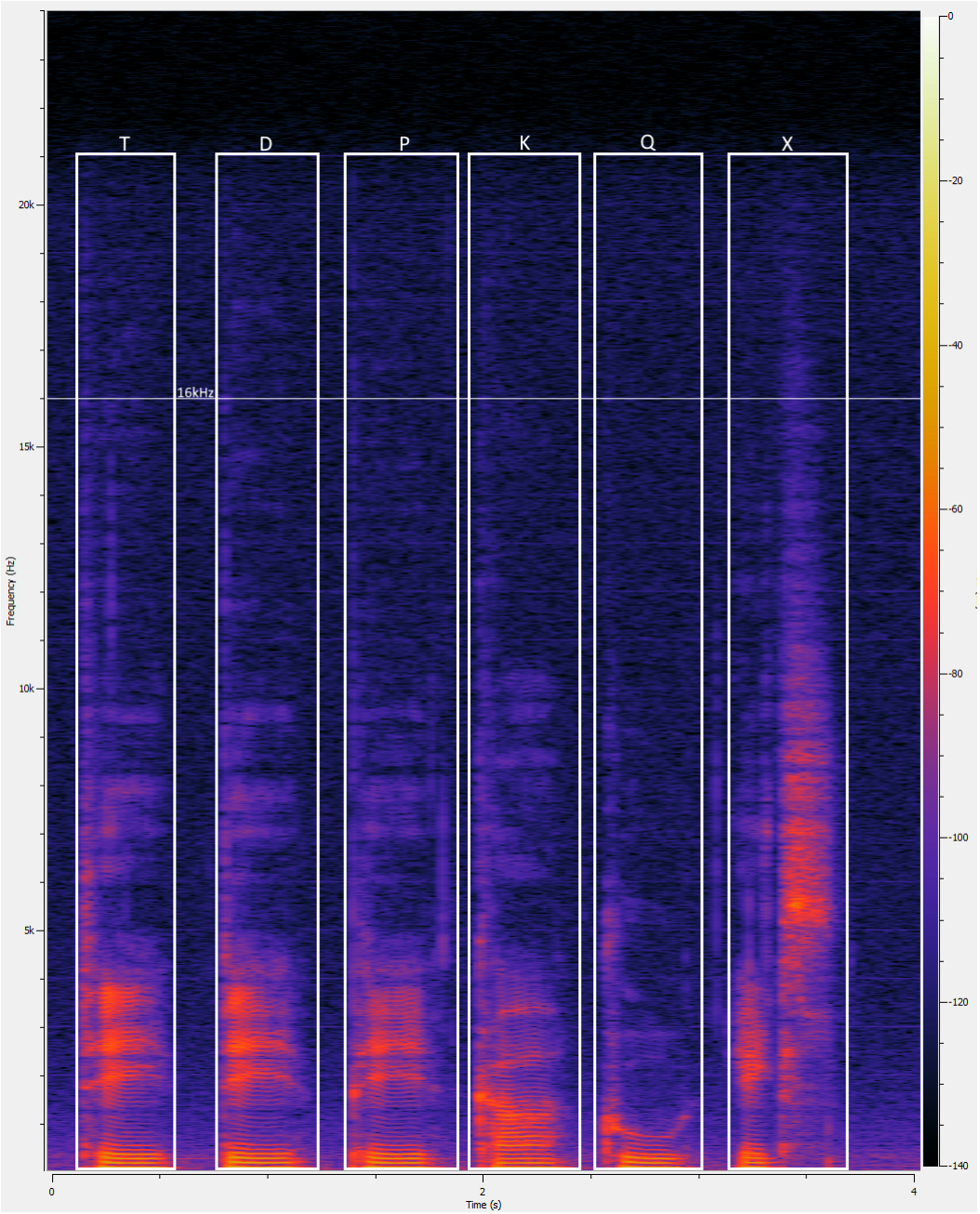

Abbildung 2.2: *T D P K Q X* Spektrogramm(Plosive)

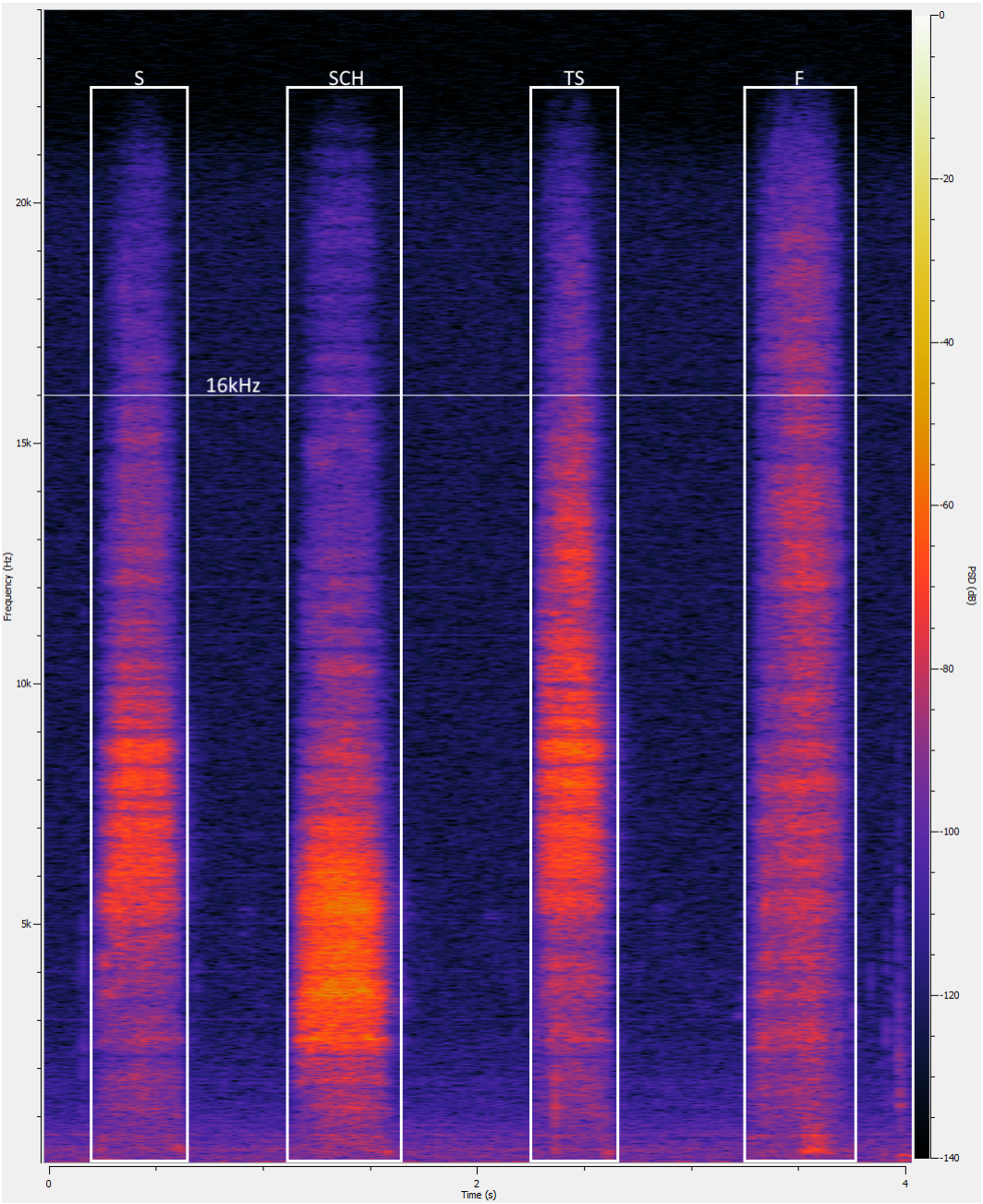

Abbildung 2.3: *S SCH TS F* Spektrogramm(Affrikate und Frikative)

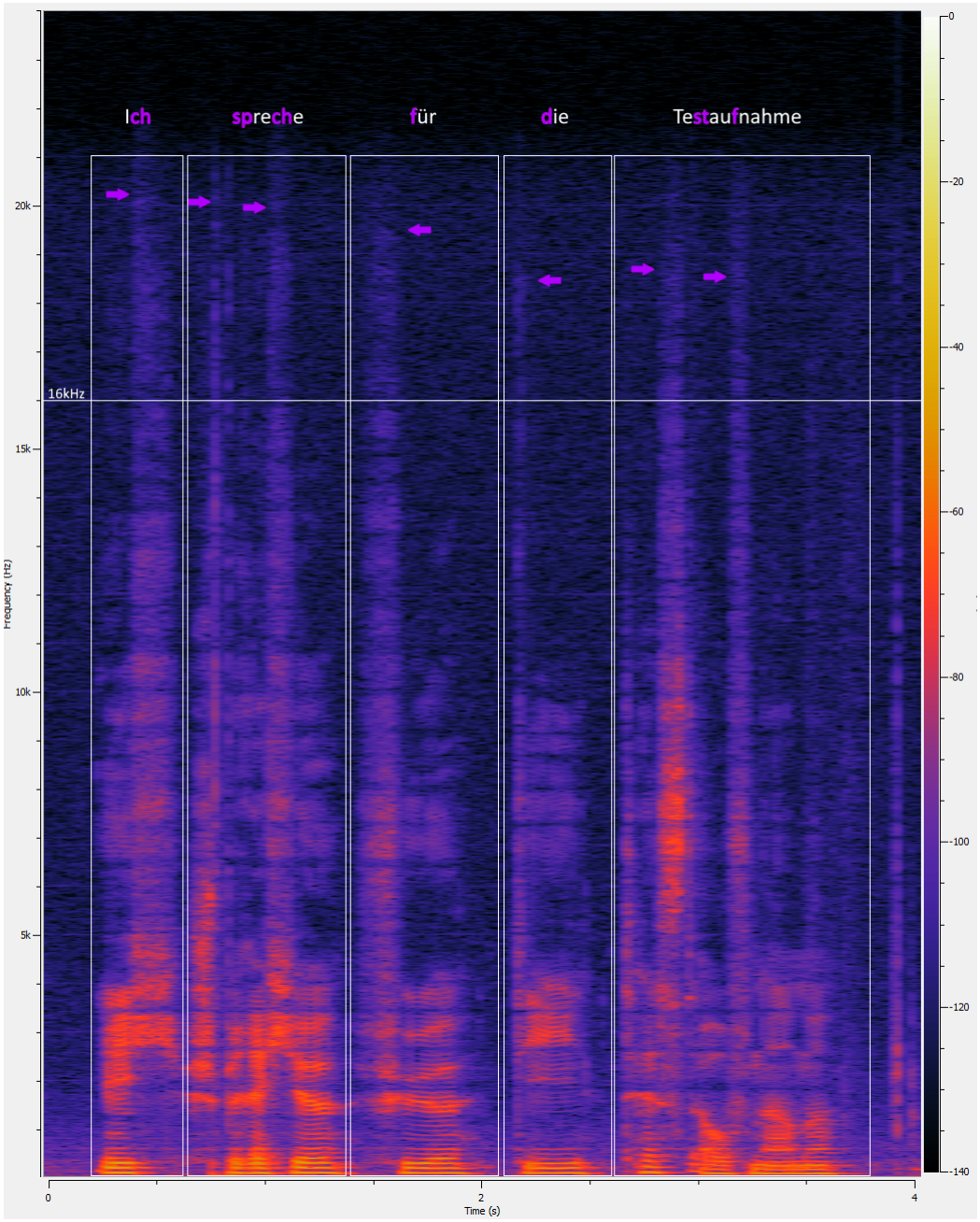

Abbildung 2.4: *Ich spreche für die Testaufnahme* Spektrogramm

Die Tests haben gezeigt, dass menschliche Sprache definitiv im Ultraschallspektrum repräsentiert wird. Die wichtige Frage an dieser Stelle ist: **Warum wurde Ultraschall bis jetzt noch nicht für die Spracherkennung verwendet?** Es gibt einige wenige Forschungen, welche Ultraschall verwenden, um existierende Systeme zu erweitern und minimal zu verbessern. Aber keines davon verwendet ausschließlich Ultraschall.

SUPERVOICE versucht, mit Hilfe von Ultraschall unterschiedliche Sprecher zu identifizieren. Dabei wurden Tests an existierenden Systemen durchgeführt, um zu untersuchen, wie diese auf Ultraschalldaten reagieren. Darunter befinden sich VGGVox [Nag+20], SincNet [RB18], GMM-UBM [RQD00] und GE2E [Wan+18]. Bei jedem dieser Systeme wurden Daten mit Frequenzbereichen von 0 - 8 kHz und 0 - 24 kHz getestet. Diese Systeme sind nicht auf Ultraschall ausgelegt und die Performance sinkt in jedem Test beim Sprung von 8 kHz auf 24 kHz. Es scheint als würden die zusätzlichen Informationen aus dem Ultraschallbereich sich ausschließlich negativ auf die Performance auswirken. Diese Systeme verfolgen zwar als Ziel die Identifikation der Sprecher und nicht die Rekonstruktion der Sprache, aber trotzdem könnte dies bereits ein Hinweis auf die Antwort für die soeben gestellte Frage sein.

3 Testen existierender Spracherkennungsmethoden

Aufgrund der Feststellungen der vorherigen Kapitel werden nun einige existierende Spracherkennungssysteme mit Ultraschallaufnahmen getestet. Was passiert, wenn reiner Ultraschall als Eingabe in Spracherkennungssysteme gegeben wird, die dem Stand der Technik entsprechen? Wenn die Systeme in der Lage sein sollten, aus reinem Ultraschall Sprache zu rekonstruieren, wäre dies ein großes Problem für den Datenschutz. Zu Beginn der Arbeit wurde in Abschnitt 1.1 besprochen, dass die Trennung zwischen Hörschall und Ultraschall das Ausnutzen der Daten des jeweils anderen Spektrums verhindern soll. Was aber, wenn eine Anwendung mit Positionierung mittels Ultraschall wirbt und dieselben Ultraschalldaten zur Sprachrekonstruktion verwendet werden können? Dies ist die direkte Motivation für die erste Forschungsfrage. Die Ergebnisse der einzelnen Systeme werden im Folgenden evaluiert und dabei werden einige Probleme beschrieben, die dabei entstehen.

3.1 Erstellen von Ultraschallaufnahmen

Bevor mit dem Testen losgelegt werden kann, wird nun erst mal beschrieben, wie überhaupt Ultraschallaufnahmen generiert werden. Der Grund dafür ist die Tatsache, dass die Trennung zwischen Ultraschall und Hörschall immer noch eine Annahme ist. Die hier verwendeten Smartphones sind nativ noch nicht in der Lage, spezifische Frequenzbereiche aufzunehmen. Das heißt, eine Audioaufnahme eines Smartphones enthält immer den kompletten unterstützten Frequenzbereich. Da der erwartete Aufwand für das Erstellen einer Applikation, welche ausschließlich Ultraschall aufnimmt, zu hoch ist, wird stattdessen untersucht, wie eine normale Aufnahme mit Nachbearbeitung einer Ultraschallaufnahme so nahe wie möglich kommt.

3.1.1 Elektrische Filter

Die einfachste Methode, eine digitale Aufnahme auf bestimme Frequenzen zu limitieren, ist die Anwendung von elektrischen Filtern. [Sil18] hat einige Filter und deren Eigenschaften beschrieben. Es gibt viele unterschiedliche Arten, darunter Bandpass-, Tiefpass-, Hochpass- und Bandspeer-Filter. Jeder dieser Filter limitiert auf eine unterschiedliche Weise das Audiosignal. Bandpass-Filter lassen nur Frequenzen auf einem

festgelegten Frequenzband zu und unterdrücken die restlichen. Bandspeer-Filter machen genau das Gegenteil. Tiefpass-Filter lassen Frequenzen unterhalb einer Frequenz passieren und Hochpass-Filter lassen Frequenzen oberhalb einer Frequenz passieren, während die restlichen in beiden Fällen blockiert werden. Filter bringen allerdings auch einige Schwierigkeiten mit sich.

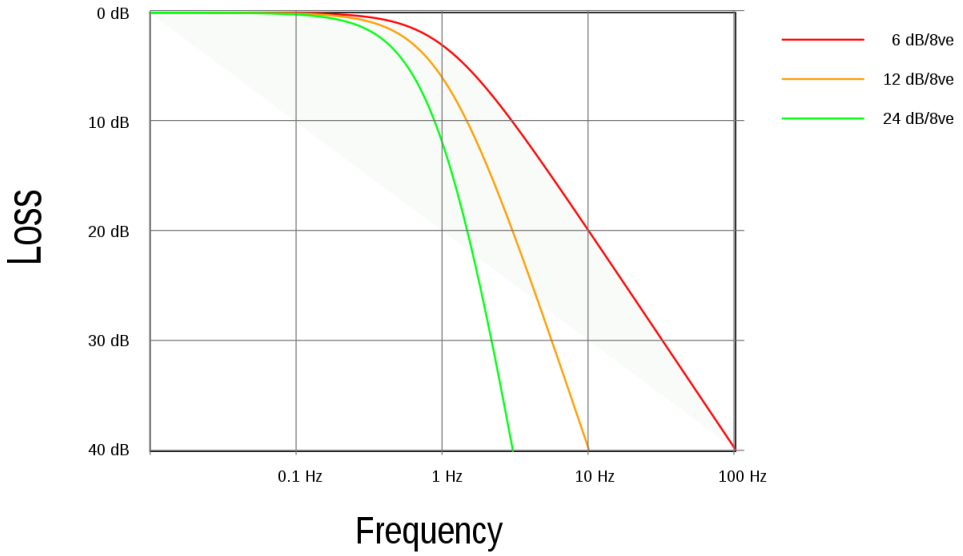

Abbildung 3.1: Filter unterschiedlicher Ordnungen

Quelle: `https://en.wikipedia.org/w/index.php?curid=23488139` (besucht am 06. 09. 2023)

Abbildung 3.1 zeigt einen Filter mit unterschiedlichen Ordnungen. In diesem Fall ist es 3 mal der gleiche Filter mit steigender Ordnung von 1 bis 3. Dabei ist zu erkennen, dass mit steigender Ordnung die Kurven immer steiler werden. Die Steilheit wird auch Roll-Off genannt. Der wichtige Punkt an dieser Stelle ist, dass elektrische Filter die Frequenzen nicht abrupt abschneiden, sondern diese schrittweise abdämpfen. Durch das abrupte Abschneiden einer Frequenz entstehen unerwünschte Artefakte in dem daraus resultierenden Audiosignal. Allerdings sind auch die schrittweisen Filter nicht von unerwünschten Artefakten befreit. In diesem Fall werden diese allerdings auf ein Minimum reduziert. Mit steigender Ordnung werden die Artefakte allerdings immer stärker.

3.1.2 Anwenden der Filter

Elektrische Filter können auf unterschiedliche Weise auf Audiosignale angewendet werden. Dazu gehören analoge und digitale Filter. Für diese Arbeit wird sich auf die Anwendung von digitalen Filtern konzentriert, da analoge zu einem höheren Aufwand und

höheren Kosten führen würden. Aber auch digitale Filter unterscheiden sich je nach Anwendungsart voneinander. Eine Möglichkeit für die Anwendung dieser wären bereits existierende Programme, welche auf Audioverarbeitung spezialisiert sind. Ein Beispiel für ein solches Programm ist z.B. die kostenlose Software Audacity [Aud]. Audacity bietet eine Vielfalt von Möglichkeiten zur Bearbeitung von Audiodateien. Darunter auch die Anwendung von Hochpass- und Tiefpass-Filtern.

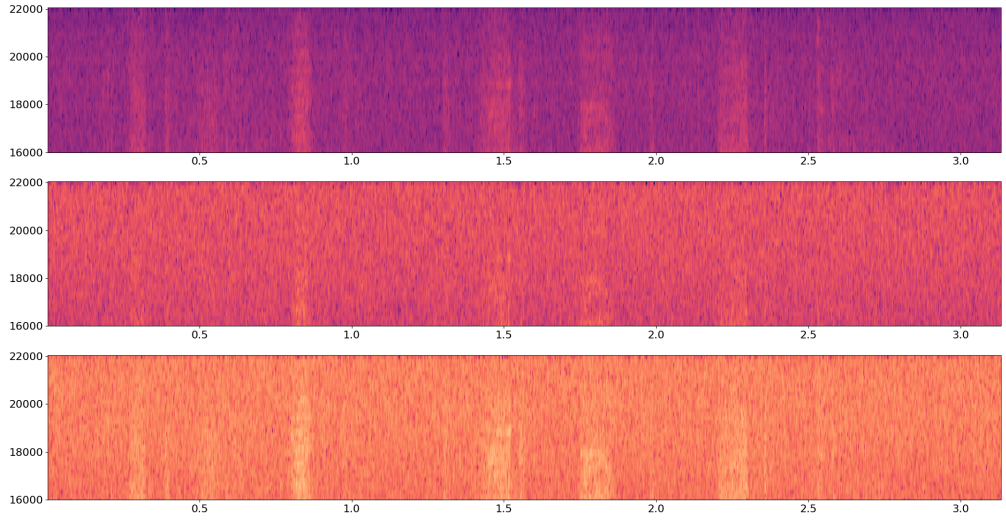

Abbildung 3.2: Unterschiedliche Filter Roll-Offs 0 dB, 6 dB, 12 dB - Audacity

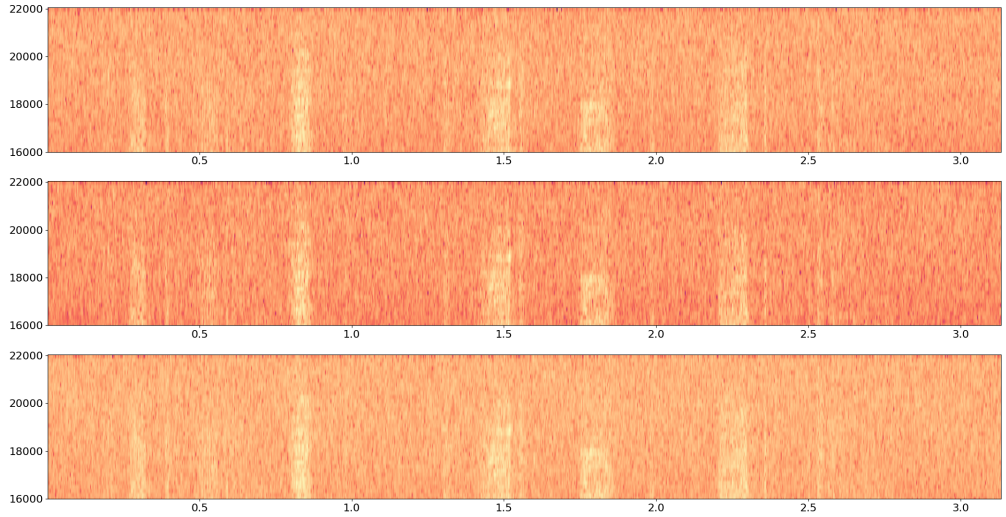

Abbildung 3.3: Unterschiedliche Filter Roll-Offs 24 dB, 36 dB, 48 dB - Audacity

Abbildung 3.2 und Abbildung 3.3 zeigen die Anwendung des Audacity Hochpass-Filters. Die Filter in den Spektrogrammen werden von oben nach unten immer stärker. Das oberste Spektrogramm stellt die unbearbeitete Audiodatei dar. Die Audiodatei wurde mit den Roll-Off Stärken 6 dB, 12 dB, 24 dB, 36 dB und 48 dB bei 16 kHz gefiltert. Diese Roll-Off Stärken entsprechen den zuvor erwähnten Ordnungen. Somit entspricht 6 dB einem Filter erster Ordnung, 12 dB einem zweiter Ordnung usw. Das Problem bei der Anwendung ist, dass die Implementation innerhalb von Audacity nicht einsehbar ist. Es wurde zuvor bereits angesprochen, dass steilere Filter zu unerwünschten Artefakten in Audiosignalen führen können. Auch in dieser Abbildung ist bereits zu erkennen, wie das Rauschen mit steigender Ordnung immer präsenter wird. Die Implementation der Filter kann ebenfalls zu unterschiedlichen Arten von Rauschen führen. Dies liegt an der Verwendung von Fourier Transformationen und der dabei gewählten Fenstergrößen. Deshalb wird eine weitere Implementation getestet, um Vergleiche heranzuziehen. Die zweite Implementierung erfolgt in Form einer Python-Bibliothek namens SciPy [Vir+20]. SciPy ist Open-Source und erlaubt Einblick in die Implementation. SciPy verwendet in ihrer Implementation die Form eines Butterworth Filters. Dies ist eine von vielen unterschiedlichen Filterfamilien und wird meist als maximal flacher Filter beschrieben [Sil18]. In Abbildung 3.4 und Abbildung 3.5 werden dieselben Darstellungen wie zuvor verwendet, jedoch diesmal unter Verwendung der SciPy Bibliothek anstelle von Audacity. Dabei ist zu erkennen, dass die Sprachmerkmale stärker vom Rauschen abgehoben werden. Während die Merkmale in Audacity gerade an der oberen Grenze manchmal im Rauschen untergehen, sind dieselben Merkmale nach Anwendung des SciPy Filters immer noch gut zu erkennen.

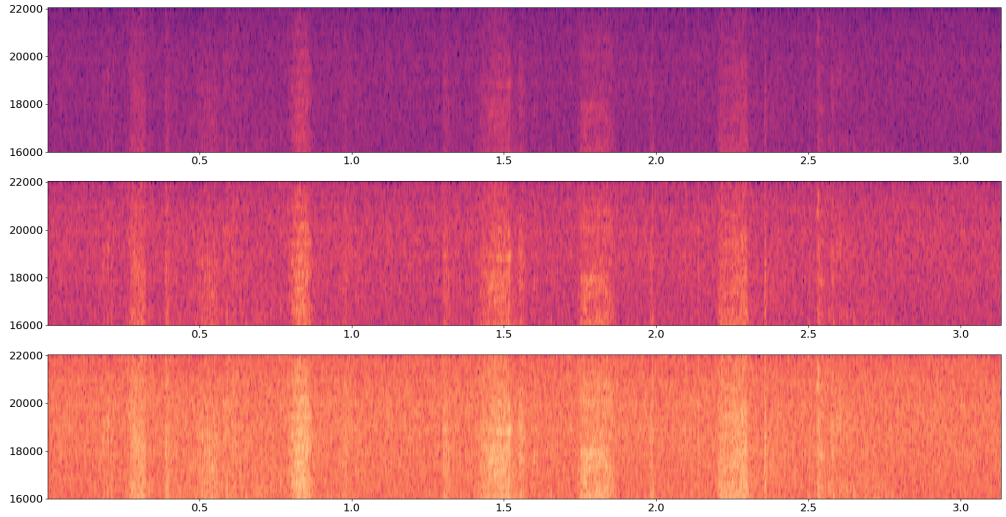

Abbildung 3.4: Unterschiedliche Filter Roll-Offs 0 dB, 6 dB, 12 dB - SciPy

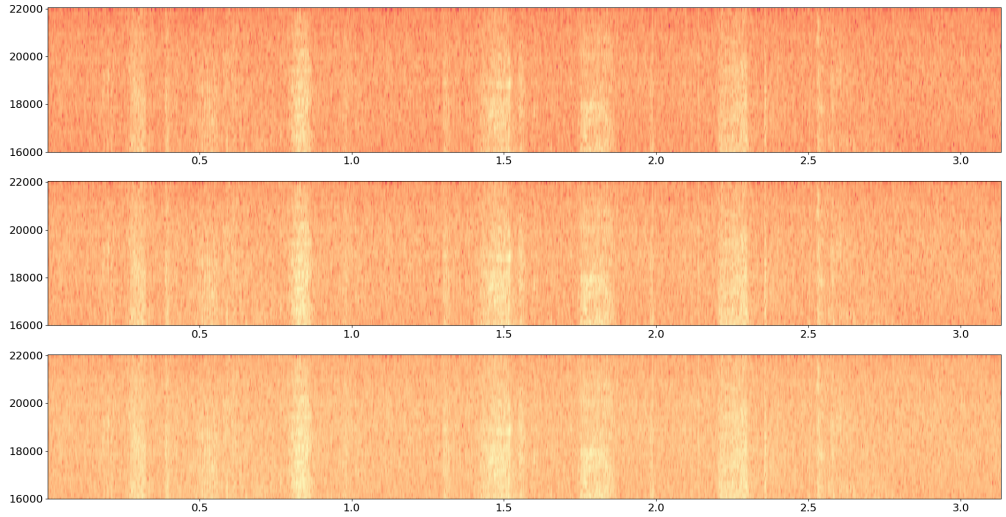

Abbildung 3.5: Unterschiedliche Filter Roll-Offs 24 dB, 36 dB, 48 dB - SciPy

Nach Anwendung der Filter und Abspeicherung der daraus resultierenden Audiosignale entstehen nun Audiodateien, mit denen existierende Spracherkennungsmethoden getestet werden können.

3.1.3 Frequency Shifting

Eine weitere Idee, um Ultraschallaufnahmen zu erhalten, ist das Nutzen von Frequency Shifting. Dabei wird das Ziel verfolgt, ein Audiosignal in einen anderen Frequenzbereich zu verschieben. Für die Anwendung des Frequency Shiftings wurde erneut Audacity verwendet. Diesmal allerdings wird die Funktion über ein Plugin durchgeführt [Zac].

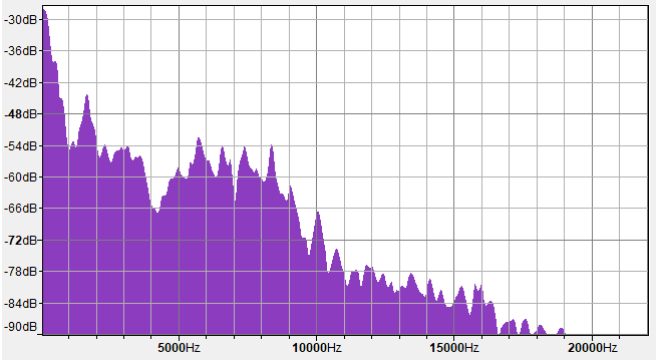

Abbildung 3.6: Audacity Spektrum - unbearbeitet

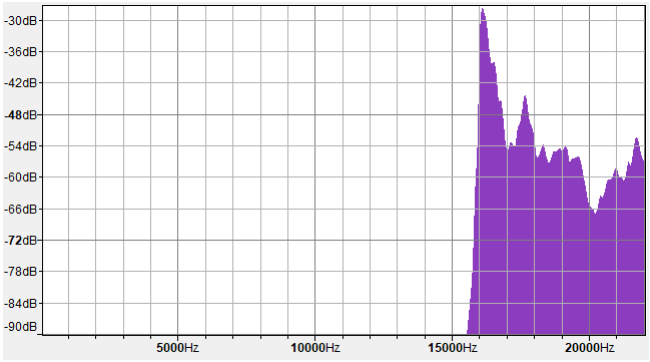

Abbildung 3.7: Audacity Spektrum - Frequency Shifter

Audacity berechnet ein Spektrum, indem die schnelle Fourier Transformation auf das Audiosignal angewendet wird. Damit kann ein Diagramm erstellt werden, welches das Vorhandensein der unterschiedlichen Frequenzen darstellt. Im Vergleich von Abbildung 3.6 und Abbildung 3.7 ist gut zu sehen, dass das Audiosignal um 16 kHz nach vorne verschoben wurde. Somit entsteht eine weitere Audiodatei, um die existierenden Spracherkennungsmethoden auf Ultraschall zu testen.

3.2 Testen der Ultraschallaufnahmen

Keines der folgenden getesteten Systeme erwähnt die Anwendung von Ultraschall. Aufgrund dessen werden diese Systeme nun mit den zuvor erstellten Audiodateien getestet

und die Ergebnisse evaluiert. Alle Systeme wurden sowohl auf deutsch als auch auf englisch getestet. Da die Modelle, auf denen die Systeme trainiert wurden, im Falle der englischen Sprache meistens deutlich größer waren, wird somit gleichzeitig untersucht, ob dies einen Unterschied in der Qualität der Ergebnisse macht.

Damit die Ergebnisse besser untereinander verglichen werden können, wurden für jedes System dieselben Audiodateien verwendet. Die eingesprochenen Sätze für die Tests lauten wie folgt:

- Englisch: **I am testing the Whisper AI with this recording.**
- Deutsch: **Mit dieser Aufnahme teste ich die Whisper KI.**

3.2.1 Whisper

Whisper ist eine von OpenAI entwickelte KI zur Spracherkennung. Das Besondere an Whisper ist die Skalierung der Trainingsdatensätze auf ungefähr 680.000 Stunden Rohmaterial. Whisper erreicht damit eine Performance, die dem Stand der Technik entspricht [Rad+22]. Die KI wird über eine Kommandozeile bedient. Dabei wird angegeben, welche Modellgröße, Sprache und Audiodatei verwendet werden sollen. Daraufhin arbeitet die KI ein paar Sekunden und gibt dann das Ergebnis innerhalb der Kommandozeile aus. Whisper bietet unterschiedliche Modelle mit unterschiedlichen Größen an. Für diese Arbeit wurde das *medium* Modell für alle Tests verwendet. OpenAI bietet zusätzlich zu *medium* noch das *large* Modell an. Dieses Modell wurde auf noch mehr Daten trainiert und liefert allgemein bessere Ergebnisse. Allerdings wurde auf Grund der hohen Hardwareanforderungen das *medium* Modell gewählt. In der späteren Evaluation wird erläutert, warum das *large* Modell nicht unbedingt bessere Ergebnisse geliefert hätte. Im Folgenden wird eine Reihe von Tabellen dargestellt, welche die Ergebnisse der einzelnen Filterstärken darstellt. Die Tabellen zeigen, welchen Text die jeweiligen Spracherkennungsmethoden für die Eingaben der unterschiedlichen Roll-Off Stärken erkennen. Dies entspricht der Ausgabe innerhalb der Kommandozeile. Durch die Filterung der deutschen und englischen Audiodateien ergeben sich 12 Aufnahmen, davon 6 mal der deutsche Satz und 6 mal der englische Satz mit jeweils 6 unterschiedlichen Roll-Off Stärken. Jede Tabelle zeigt auch den Roll-Off von 0 dB. Dies ist die unbearbeitete Audioaufnahme ohne jeglichen Filter.

Roll-Off Stärke	Erkannter Text
0dB	Mit dieser Aufnahme teste ich die Respec AI.
6dB	Mit dieser Aufnahme teste ich die Respec AI.
12dB	Mit dieser Aufnahme teste ich die WistLKI.
24dB	Vielen Dank für's Zuschauen.
36dB	Das war's für heute. Bis zum nächsten Mal. Tschüss.
48dB	Das war's für heute. Bis zum nächsten Mal. Tschüss.

Tabelle 3.1: Whisper Spracherkennung - deutsche Ergebnisse

Roll-Off Stärke	Erkannter Text
0dB	I am testing the wristband AI with this recording.
6dB	I am testing the wristband AI with this recording.
12dB	I am testing the risk that I have for this recording.
24dB	you
36dB	you
48dB	you

Tabelle 3.2: Whisper Spracherkennung - englische Ergebnisse

In den ersten Ergebnissen von Whisper kann bereits ein Muster erkannt werden. Die Ausgaben der Modelle von den ersten 3 Roll-Off Stärken liegen dem wahren Wert sehr nahe. Der Roll-Off von 6 dB ist in beiden Fällen identisch zu der Audioaufnahme ohne Filter. Die Ausgabe vom 12 dB Roll-Off Filter ist der Punkt, an dem die Wörter beginnen, sich zu verändern. Sowohl im deutschen Modell als auch im englischen Modell ist ab einer Roll-Off Stärke von 24 dB nichts mehr von der eigentlichen Eingabe zu erkennen. Besonderes Augenmerk sollte auch auf die Ausgabe gelegt werden. Die deutsche Ausgabe fokussiert sich auf eine Art Abschiedsspruch für alle Roll-Off Stärken über 24 dB. Währenddessen gibt das englische Modell konsistent *you* aus. Diese Ausgaben sind komisch, da sie einerseits komplett zufällig erscheinen und andererseits konsistent sind. Dies scheint allerdings ein Whisper spezifisches Problem zu sein, da dies in den späteren Tests nicht der Fall ist. Die Autoren von Whisper haben in ihrem Paper erwähnt, dass die KI zusammenhanglose Ergebnisse liefern kann, falls keine spezifischen Wörter ausgemacht werden können [Rad+22]. Dies ist sehr wahrscheinlich das, was hier passiert ist.

3.2.2 Smartphone Tastatur

Eine weitere Spracherkennungsmethode sind die eingebauten Speech-to-Text Funktionen innerhalb von Smartphone Tastaturen. Diese erlauben das sofortig Umwandeln der Stimme in geschriebenen Text. Auch diese Methode wird mit Audiodateien getestet,

auf welche Hochpass-Filter angewendet wurden. Die zwei getesteten Systeme bestehen aus der nativen Tastatur des Apple Betriebssystemes iOS und der Tastatur der Google GBoard Anwendung [Gooa]. Es wurden erneut beide Systeme auf Deutsch und Englisch getestet. Für die Tests wurden die Audiosignale über Kopfhörer abgespielt und wiederum von einem Smartphone aufgenommen. Die verwendeten Kopfhörer sind die Sennheiser HD 650 und unterstützen laut Spezifikation einen Frequenzbereich bis 41.000 Hz [Sen].

Roll-Off Stärke	**Erkannter Text**
0dB	mit dieser aufnahme teste ich dieses RKI
6dB	mit dieser aufnahme test ich
12dB	bitte
24dB	-
36dB	-
48dB	-

Tabelle 3.3: iOS native Tastatur - deutsche Ergebnisse

Roll-Off Stärke	**Erkannter Text**
0dB	I am testing this recording
6dB	I am testing this recording
12dB	-
24dB	-
36dB	-
48dB	-

Tabelle 3.4: iOS native Tastatur - englische Ergebnisse

Roll-Off Stärke	**Erkannter Text**
0dB	mit dieser aufnahme teste ich die Whisper KI
6dB	mit dieser aufnahme teste ich die beste KI
12dB	-
24dB	-
36dB	-
48dB	-

Tabelle 3.5: Google GBoard Tastatur - deutsche Ergebnisse

Roll-Off Stärke	Erkannter Text
0dB	I am testing the respirator eye with this recording
6dB	this recording
12dB	-
24dB	-
36dB	-
48dB	-

Tabelle 3.6: Google GBoard Tastatur - englische Ergebnisse

Die Tabellen zeigen, dass die eingebauten Speech-to-Text Methoden bereits bei einer Roll-Off Stärke von 6 dB an Korrektheit verlieren. Spätestens bei einem Roll-Off von 12 dB ist von der eigentlichen Eingabe nichts mehr zu erkennen. Die schlechtere Qualität dieser Ergebnisse im Vergleich zu der Whisper KI könnte mehrere Gründe haben. Einerseits kann es natürlich sein, dass die verwendeten Modelle einfach schlechter sind und nicht auf die Qualität von Whisper kommen. Andererseits könnte die Art der Eingabe ein großes Problem sein. Die Tatsache, dass die Audiodateien erst über Kopfhörer abgespielt werden und dann von dem Smartphone aufgenommen werden, führt zu einem zusätzlichen Übertragungsmedium, in dem Informationsverlust stattfinden kann. Im Vergleich dazu können bei Systemen wie Whisper die Audiosignale direkt als Eingabe in das System gegeben werden.

3.2.3 NVIDIA NeMo

Die letzten hier getesteten Systeme sind eine Vielfalt von Modellen, welche über NVIDIA NeMo bereitgestellt werden [NVI]. NVIDIA NeMo ist ein Framework für das Erstellen von KI Modellen für viele unterschiedliche Anwendungsgebiete. Darunter Speech Processing, Text-to-Speech und Natural Language Processing. Für diese Arbeit wird der Fokus auf den *Automatic Speech Recognition* Teil aus der Speech Processing Branche gelegt. Viele Wissenschaftler haben bereits eine Vielzahl an unterschiedlichen vortrainierten Modellen erstellt, welche innerhalb von NVIDIA NeMo frei zur Verfügung stehen. Dazu gehören Conformer [Guo+22], ContextNet [Han+20], Quartznet [Kri+20] und Jasper [Li+19]. Diese Modelle funktionieren ähnlich wie Whisper. In diesen Fällen können die Audiosignale direkt als Eingabe in die Systeme eingegeben werden. Im Folgenden werden in einer Reihe von Tabellen die Ergebnisse der einzelnen Modelle dargestellt. Unter den Modellen befinden sich 10 englische und 5 deutsche Modelle.

Roll-Off Stärke	Erkannter Text
0dB	i am testing the wisper eye with this recording
6dB	i am testing the rispery with this recording
12dB	i am testing the risk that iy with this recording
24dB	-
36dB	-
48dB	-

Tabelle 3.7: stt_en_conformer_ctc_large

Roll-Off Stärke	Erkannter Text
0dB	i am testing the wspare i worth this recording
6dB	i am testing in the wispter i wth this recording
12dB	i detesting that youester hom with this morning
24dB	ess
36dB	-
48dB	-

Tabelle 3.8: stt_en_conformer_ctc_large_ls

Roll-Off Stärke	Erkannter Text
0dB	i am testing the whisper eye with this recording
6dB	i am testing the wrist with this recording
12dB	i am testing the wrist with this recording
24dB	-
36dB	-
48dB	-

Tabelle 3.9: stt_en_contextnet_1024

Roll-Off Stärke	Erkannter Text
0dB	i am testing the wispy eye with this recording
6dB	i am testing the wisp eye with this recording
12dB	i am testing the risk with this recording
24dB	-
36dB	-
48dB	-

Tabelle 3.10: stt_en_conformer_transducer_large

Roll-Off Stärke	Erkannter Text
0dB	i am testing the whisper eye with this recording
6dB	i am testing the wrist with this recording
12dB	i am testing the wrist with this recording
24dB	-
36dB	-
48dB	-

Tabelle 3.11: stt_en_contextnet_1024

Roll-Off Stärke	Erkannter Text
0dB	i am testing the wrisbai with this recording
6dB	i am testing the whisper i with this recording
12dB	i am testing the misper is recording
24dB	-
36dB	-
48dB	-

Tabelle 3.12: stt_en_contextnet_1024_mls

Roll-Off Stärke	Erkannter Text
0dB	i am testing theispa i with this recording
6dB	i am testing theisp i with this recording
12dB	i am testing the wister i with thisjoing
24dB	-
36dB	-
48dB	-

Tabelle 3.13: stt_en_citrinet_1024_gamma_0_25

Roll-Off Stärke	Erkannter Text
0dB	i am testing the wspare i with this recording
6dB	i am testing at the risk there i with this recording
12dB	i am testing an exid i with for this retorting.
24dB	-
36dB	-
48dB	-

Tabelle 3.14: stt_en_citrinet_1024

Roll-Off Stärke	Erkannter Text
0dB	i am testing as the whisper i with this recording
6dB	i am testing the whisper i with this recording
12dB	i'm testing an wisten alay with this recording
24dB	-
36dB	-
48dB	-

Tabelle 3.15: stt_en_jasper10x5dr

Roll-Off Stärke	Erkannter Text
0dB	mit dieser aufnahme teste ich die wisberk i
6dB	mit dieser aufnahme teste ich die wizbälkeie
12dB	mit dieser aufnahme teste ich die bisder kanale
24dB	is a
36dB	-
48dB	-

Tabelle 3.16: stt_de_quartznet15x5

Roll-Off Stärke	Erkannter Text
0dB	mit dieser aufnahme teste ich die whsperci
6dB	mit dieser aufnahme teste ich die whisperci
12dB	mit dieser aufnahme teste ich die whiststerkie
24dB	muss die lichis
36dB	-
48dB	-

Tabelle 3.17: stt_de_citrinet_1024

Roll-Off Stärke	Erkannter Text
0dB	mit dieser aufnahme teste ich die whisberkai
6dB	mit dieser aufnahme teste ich die whisber ki
12dB	mit dieser aufnahme teste ich die whister ki
24dB	u diese interessantisen
36dB	-
48dB	-

Tabelle 3.18: stt_de_conformer_ctc_large

Roll-Off Stärke	Erkannter Text
0dB	mit dieser aufnahme teste ich die whisbecai
6dB	mit dieser aufnahme teste ich die whisper ki
12dB	mit dieser aufnahme teste ich die whist der charlodie
24dB	ist
36dB	-
48dB	-

Tabelle 3.19: stt_de_contextnet_1024

Roll-Off Stärke	Erkannter Text
0dB	mit dieser aufnahme teste ich die whisbacker
6dB	mit dieser aufnahme teste ich die whiskerki
12dB	mit dieser aufnahme teste ich die whistelkali
24dB	ist die sendung ist die single
36dB	eine
48dB	-

Tabelle 3.20: stt_de_conformer_transducer_large

Die Spracherkennungsmodelle ergeben unterschiedlich gute Ergebnisse für die Audiosignale. Auch in diesen Modellen sind sehr eindeutig Muster erkennbar. Mit steigender Stärke des Roll-Offs wird die Qualität des ausgegebenen Textes immer schlechter. Ab einem Roll-Off von 24 dB werden entweder gar keine Wörter oder nur noch zusammenhanglose Wörter ausgegeben.

3.3 Frequency Shifting Test

Zuvor wurde das Frequency Shifting als eine weitere Methode für die Erstellung von Ultraschallaufnahmen erwähnt. Mit Hilfe dieser Methode wurden die beiden getesteten Sätze um 16 kHz nach oben verschoben. Keines der getesteten Systeme hat ein positives Ergebnis dieser Dateien ausgegeben. Im Falle der Smartphone Tastaturen und NVIDIA NeMo war ausnahmslos jede Ausgabe leer, während Whisper erneut Abschiedssprüche ausgegeben hat.

3.4 Fehleranalyse der Tests

Über die verschiedenen Tests hat sich herausgestellt, dass die bereits existierenden Spracherkennungsmethoden nicht in der Lage sind, Ultraschall zu verarbeiten. Daraus ergibt sich die Frage: **Warum sind die Spracherkennungsmethoden nicht in der**

Lage, Wörter aus Ultraschall zu reproduzieren? In diesem Kapitel werden einige Gründe dafür genannt. Der erste Punkt, der untersucht wird, ist der Unterschied zwischen den Audiosignalen mit einem 12 dB Roll-Off und einem 24 dB Roll-Off. Mit Ausnahme der Smartphone Tastaturen und einigen wenigen NVIDIA NeMo Modellen ist der Sprung von 12 dB auf 24 dB der entscheidende Punkt, ab dem keine Ausgabe mehr entsteht. Die folgenden Abbildungen zeigen Spektren von den unterschiedlichen Audiosignalen der verschiedenen Roll-Off Stärken.

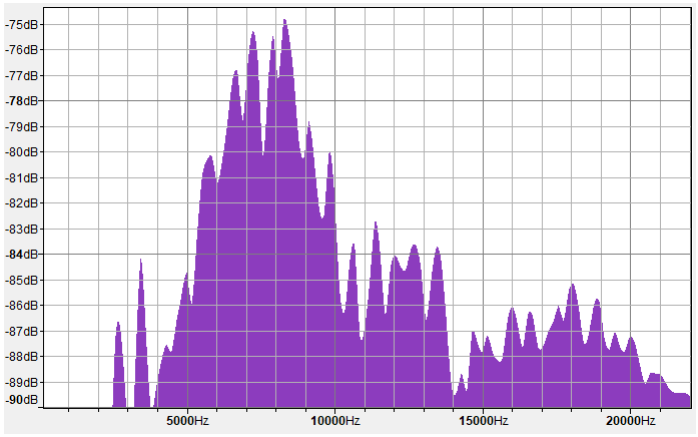

Abbildung 3.8: 12dB Hochpass-Filter Frequenzspektrum - Audacity

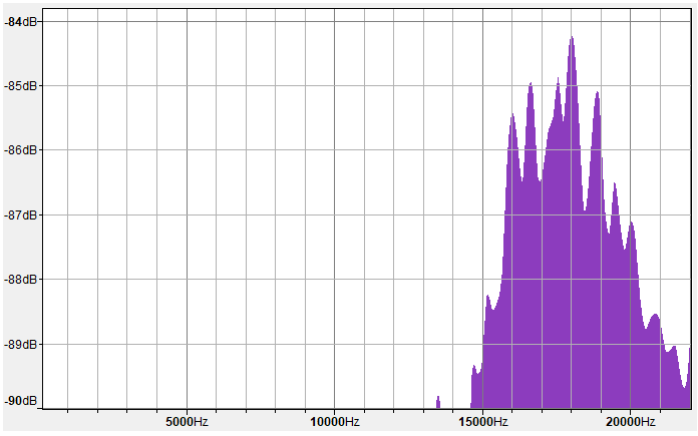

Abbildung 3.9: 24dB Hochpass-Filter Frequenzspektrum - Audacity

Abbildung 3.8 zeigt das Frequenzspektrum des eingesprochenen deutschen Satzes mit Anwendung eines 12 dB Roll-Off Hochpass-Filters. Dabei ist sehr gut zu erkennen, dass trotz des Filters immer noch eine große Menge an Informationen unterhalb der 16 kHz Grenze vorhanden sind. Der Grund dafür ist die zuvor genannte Steilheit der Filter. Da die Filter keinen glatten Schnitt machen, bleiben einige Überreste der Frequenzen unterhalb der Grenze vorhanden. Im Falle des 12 dB Roll-Offs sind dies immer noch genug

Informationen für viele der Rekonstruktionsmethoden, um die Sprache wiederherzustellen. Wenn nun ein Blick auf Abbildung 3.9 geworfen wird, ist zu sehen, dass dies der Punkt ist, wo nur noch minimale Artefakte aus dem Frequenzbereich unterhalb der 16 kHz Grenze vorhanden sind. Das Audiosignal besteht fast nur noch aus Ultraschallinformationen. Diese sind allerdings nicht ausreichend für die Spracherkennungsmethoden, um die Wörter zu rekonstruieren.

Eine weitere Visualisierung für dieses Problem wird in Abbildung 3.10 dargestellt. Hier sind die unterschiedlichen Roll-Off Stärken 0 dB, 12 dB, 24 dB und 36 dB in Form von Spektrogrammen zu sehen. Die Roll-Off Stärken werden von oben nach unten größer. Auch in dieser Visualisierung ist gut zu erkennen, dass selbst ein Roll-Off von 12 dB immer noch eine große Menge an Informationen aus dem hörbaren Spektrum hinterlässt. Erst bei einem Roll-Off von 24 dB entsteht eine große Lücke an Informationen innerhalb des Spektrogrammes.

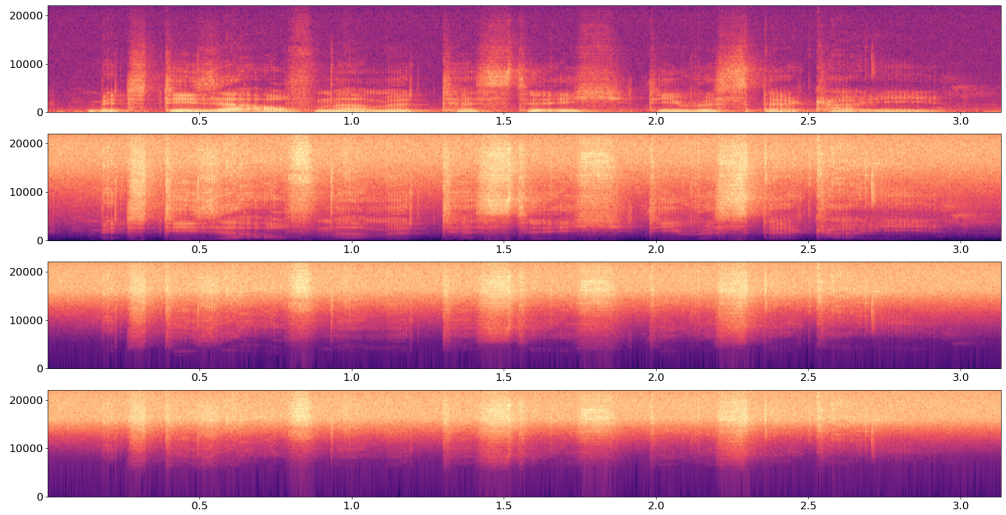

Abbildung 3.10: Unterschiedliche Roll-Off Stärken 0, 12, 24, 36 dB - SciPy Hochpass-Filter

Die übrig gebliebene Menge an Informationen ist für keines der getesteten Systeme ausreichend, um korrekte Sprachfragmente zu erkennen. In den meisten Fällen sind die Ausgaben komplett leer. In Unterabschnitt 2.3.1 wurde beschrieben, dass Menschen Informationen über ihre Sprache innerhalb des Ultraschallspektrums ausgeben. **Warum also können diese Systeme damit nicht umgehen?** Google schreibt in den Leitfäden ihres Speech-to-Text Systems folgendes:

"Wenn Sie bei der Codierung des Quellmaterials eine Wahl haben, erfassen

> Sie Audio mit einer Abtastrate von 16.000 Hz. Niedrigere Werte können die Spracherkennungsgenauigkeit beeinträchtigen, höhere Werte haben keine nennenswerte Auswirkung auf die Spracherkennungsqualität [Goob]."

Dieser Leitfaden vermittelt den Eindruck, dass Audiosignale mit einer Abtastrate von mehr als 16 kHz gar nicht verarbeitet werden. Ein weiterer Blick in die Trainingsvorgänge der zuvor getesteten Systeme enthüllt die niedrigen Abtastraten der Datensätze. Die Autoren der Whisper KI erwähnen, dass die Datensätze vor dem Training auf 16 kHz herunter getaktet werden. Ein weiterer Datensatz, der dem Stand der Technik entspricht und von vielen der NVIDIA NeMo Modelle verwendet wird, ist der LibriSpeech Datensatz [Pan+15]. Auch diese Daten sind bereits auf 16 kHz herunter getaktet. Wenn die Modelle der Spracherkennungsmethoden auf Daten mit einer Abtastrate von 16 kHz trainiert wurden, heißt das, dass auch die Eingaben in die fertigen Modelle auf 16 kHz herunter getaktet werden. Somit ist es kein Wunder, dass die Modelle die Ultraschallinformationen nicht verarbeiten können.

Ein weiteres Argument gegen die erfolgreiche Rekonstruktion ist die von SUPERVOICE [Guo+22] untersuchte Thematik der Konsistenz von Ultraschall. Die Autoren dieses Papers haben Tests durchgeführt, bei denen die Energie der menschlichen Sprache auf den Frequenzbereichen von 20 Hz bis 96 kHz untersucht wurde. Die Tests haben ergeben, dass es einen großen Unterschied zwischen Ultraschallenergie und Hörschallenergie bei Menschen gibt. Die Energie von **unterschiedlichen** Sätzen einer **einzelnen** Person ist unterhalb von 16 kHz stark zerstreut, während die Energie oberhalb von 16 kHz konsistent ist. Auf der anderen Seite ist die Energie von **mehreren** Personen, die alle **denselben** Satz sprechen, unterhalb von 16 kHz konsistent und oberhalb von 16 kHz zerstreut. Abbildung 3.11 zeigt eine Darstellung aus SUPERVOICE. Die Autoren nutzen diese Eigenschaft aus, um die Sprecher zu identifizieren. Sie erreichen eine erfolgreiche Verbesserung der Identifizierung von Personen in existierenden Systemen, indem die Ultraschallinformationen zusätzlich verarbeitet werden.

Wenn die Energie einer einzelnen Person oberhalb von 16 kHz auf Grund der Konsistenz für die Identifikationserkennung verwendet wird, lässt dies darauf schließen, dass dieselbe Energie für die spezifische Spracherkennung ungeeignet sein könnte.

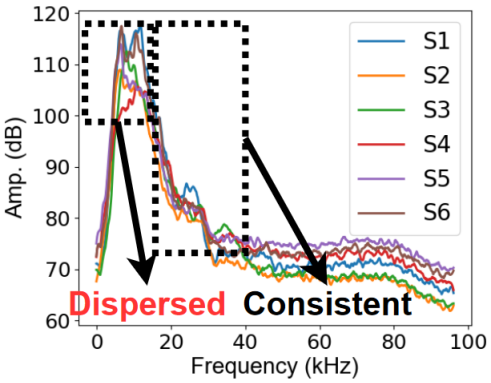

Abbildung angepasst von Abbildung 4 in [Guo+22]. Übersetzung der Bildunterschrift.

(a) Energie von unterschiedlichen Sätzen derselben Person

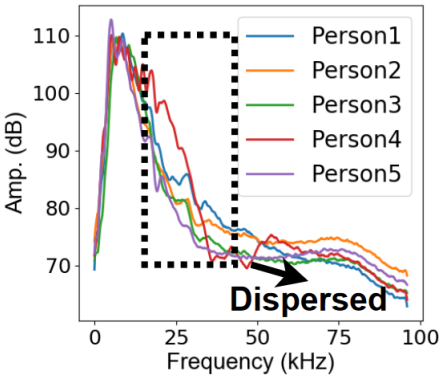

Abbildung angepasst von Abbildung 4 in [Guo+22]. Übersetzung der Bildunterschrift.

(b) Energie der gleichen Sätze von unterschiedlichen Personen

Abbildung 3.11: Vergleich der Energien der menschlichen Sprache

4 Konzepte

Es wurde festgestellt, dass die getesteten Spracherkennungssysteme die Ultraschalldaten nicht verarbeiten können. Dies reicht leider nicht aus, um die erste Forschungsfrage zu beantworten. Die Systeme sind nicht darauf ausgelegt, Ultraschall zu verarbeiten. Was passiert, wenn solch ein System von vornherein darauf ausgelegt wird? Hierfür werden einige Konzepte und Ideen vorgestellt, um zu beschreiben, auf welche Art und Weise die Ultraschalldaten verarbeitet werden könnten.

4.1 Trainieren von Speech-to-Text Modellen mit Ultraschall

Es hat sich in den Untersuchungen herausgestellt, dass die Abtastraten der Datensätze, die für das Training der Spracherkennungssysteme genutzt werden, nicht groß genug sind, um Signale aus dem Ultraschallbereich zu erfassen. Dementsprechend können die Systeme nicht mit Ultraschallsignalen als Eingabe umgehen. Da stellt sich nun die Frage: **Können diese Art von Systemen mit Datensätzen trainiert werden, die eine genügend große Abtastrate besitzen, um Ultraschallinformationen zu verarbeiten?** Dabei würde mit Hilfe eines Filters der Großteil der Informationen unterhalb von 16 kHz herausgefiltert werden, sodass ausschließlich mit Ultraschall trainiert wird. Mit der Umsetzung dieses Konzeptes könnte untersucht werden, ob die Ultraschallinformationen alleine bereits ausreichen, um Wörter zu reproduzieren.

Aufwand: Zu hoch.
Das Trainieren eigener Modelle für die Erkennung von Sprache im Ultraschallbereich ist für den Umfang dieser Arbeit einfach zu hoch. Die reine Menge an Sprachinformationen innerhalb des Ultraschallbereiches ist im Vergleich zu den Informationen des Hörschalles extrem gering. Zusätzlich wurde bereits in Abschnitt 3.4 die Eigenschaft erwähnt, dass die produzierten Ultraschallsignale von Menschen stark unterschiedlich zueinander sind und es sich somit eher für die Identifizierung von Personen eignet. Es ist zu erwarten, dass die selbst trainierten Modelle nicht in der Lage wären, spezifische Wörter zu reproduzieren.

4.2 Identifizieren von Personen mit Ultraschall

Eine der bereits zuvor erwähnten wissenschaftlichen Forschung SUPERVOICE untersucht, wie Ultraschall bei der Identifizierung von Personen helfen kann. Dabei wurde festgestellt, dass die Ultraschallinformationen tatsächlich bei der Identifizierung von Nutzen sind. Allerdings nutzt SUPERVOICE nicht ausschließlich Ultraschall. Es wird eine bereits vorhandene Identifizierungsmethode mit Hilfe der Ultraschalldaten ein wenig verbessert. Somit ergibt sich ein weiteres Konzept zur Identifizierung ausschließlich durch die Verarbeitung von Ultraschallsignalen.

Aufwand: Zu hoch.
SUPERVOICE verwendet mehrere neuronale Netze, um die Identifikation unterschiedlicher Personen zu erreichen. Zusätzlich werden nicht ausschließlich Ultraschallinformationen verwendet. Mit einer geringeren Menge an Daten wird es noch schwieriger, korrekte Ergebnisse zu erzielen. Außerdem werden für die Aufgabe Datensätze benötigt, welche einen großen Anteil an unterschiedlichen Personen enthalten.

4.3 Identifizierung der Existenz von Sprache

Da die anderen Konzepte in ihrem Aufwand den Umfang dieser Arbeit übersteigen, wird nun das Konzept untersucht, ob anhand von Ultraschallsignalen entschieden werden kann, ob in der Umgebung des Aufnahmegerätes gesprochen wird. Dies ist eine Abstrahierung der anderen Konzepte, um dem Umfang dieser Arbeit gerecht zu werden. Im gleichen Atemzug wird hiermit auch der Fokus dieser Arbeit etwas angepasst. Anstatt zu untersuchen, wie spezifische Sprachsignale extrahiert werden können, wird nun untersucht, ob überhaupt zwischen dem Vorhandensein von Sprachsignalen und dem Nichtvorhandensein unterschieden werden kann. Es soll ein neuronales Netz erstellt und trainiert werden, um bei der Eingabe eines Ultraschallsignals zu entscheiden, ob dieses Signal menschliche Stimmen enthält oder nicht. Somit wird die Problemstellung von der Erkennung von Sprache auf ein binäres Klassifizierungsproblem reduziert. Für diese Aufgabe werden Datensätze benötigt, welche eine genügend große Abtastrate besitzen. Die Abtastrate sollte mindestens 44.100 Hz betragen, damit eine ausreichende Menge an Informationen aus dem Ultraschallbereich aufgenommen werden kann. Die Daten müssen normalisiert werden, damit jeder Datensatz ein konsistentes Format aufweist.

5 Umsetzung und Ergebnisse

Dieses Kapitel stellt die Umsetzung des dritten Konzeptes über die Entscheidung, ob Sprache in einem Ultraschallsignal vorhanden ist, vor.

5.1 Datensätze

Für das Trainieren eines neuronalen Netzes werden sehr viele Daten benötigt. Viele der vorhandenen Spracherkennungsmethoden, die dem Stand der Technik entsprechen, verwenden Datensätze wie LibriSpeech [Pan+15], um ihre Modelle zu trainieren. Das Problem dabei ist die niedrige Abtastrate von 16 kHz. Aufgrund dessen wurden die folgenden Datensätze für dieses Problem herangezogen:

- CSTR VCTK Corpus(VCTK) [YVM19], welches menschliche Sprache von 110 unterschiedlichen Personen enthält. Jede Person spricht dabei ungefähr 400 Sätze aus und jeder Satz wird mit zwei Mikrofonen aufgenommen. Der Datensatz hat eine Abtastrate von 48 kHz und enthält ungefähr 88.000 einzelne Datenproben.
- UrbanSound8K [SJB14], welches städtische Hintergrundgeräusche enthält. Die Geräusche sind in 10 Klassen unterteilt und der Datensatz enthält keine Sprache. Die Abtastrate liegt bei 44.1 kHz und der Datensatz enthält ungefähr 8.000 Datenproben.
- FSD50K [Fon+20], welches Hintergrundgeräusche aus 200 Klassen enthält. Der Datensatz enthält ungefähr 50.000 Datenproben. Die Abtastrate beträgt 44.1 kHz.
- Einige selbst aufgenommenen Audiodaten. Mehr dazu später in Unterabschnitt 5.3.3

5.2 Spektrogramme von Ultraschallsignalen

Hier werden nun einige Spektrogramme für die Visualisierung von Ultraschallsignalen herangezogen. Die Spektrogramme wurden alle in der Programmiersprache Python mit der *matplotlib* [Hun07] Bibliothek erstellt. Die Spektrogramme stellen den Frequenzbereich oberhalb von 16 kHz dar. Die obere Grenze beträgt je nach Beispiel entweder 22.050 Hz oder 24.000 Hz. In diesen Fällen wurden die Audiosignale **nicht** zuvor gefiltert. Es wird lediglich der gewünschte Frequenzbereich aus dem Spektrogramm genommen und neu auf die volle Größe skaliert. Ein Beispiel für diesen Vorgang wird in Abbildung 5.1

dargestellt. Das obere Spektrogramm stellt den gesamten Frequenzbereich von 0 Hz bis 24 kHz dar, während das untere Spektrogramm den Frequenzbereich von 16 kHz bis 24 kHz darstellt.

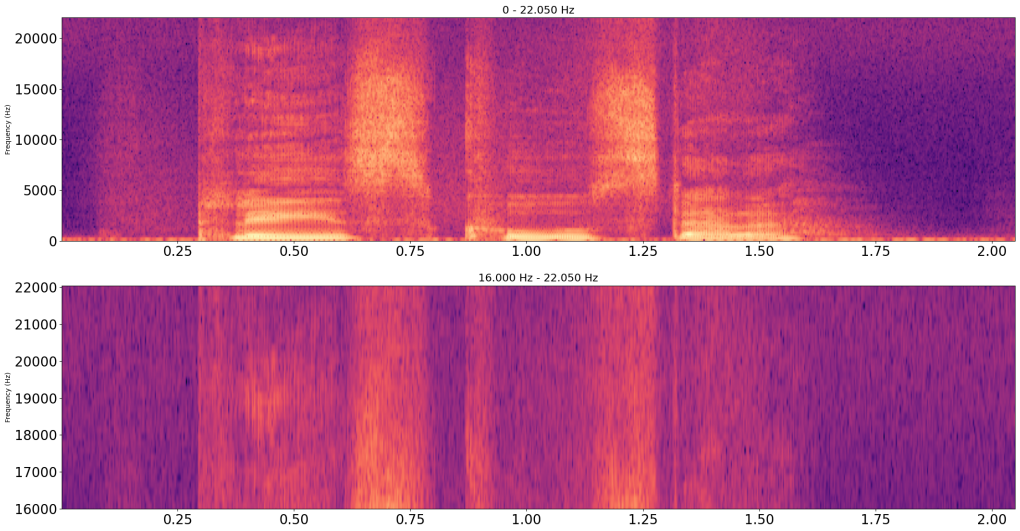

Abbildung 5.1: Unterschiedliche Skalierungen - dieselbe Audiodatei

Mit diesem Hintergrundwissen werden nun einige Proben aus den in Abschnitt 5.1 erwähnten Datensätzen verglichen. Begonnen wird mit dem VCTK Datensatz, welcher aus sauberen Aufnahmen von menschlichen Stimmen besteht. In Abbildung 5.2 werden 4 Spektrogramme untereinander dargestellt. Jedes Spektrogramm ist einer unterschiedlichen Person zugewiesen. Das Besondere dabei ist, dass jeder dieser Personen denselben Satz ausspricht. Die Aufnahmen enthalten keine Hintergrundgeräusche und stellen ausschließlich die Stimmen der Personen dar. Obwohl jedes Spektrogramm denselben Satz darstellt, sind die Aufnahmen nicht alle von identischer Länge. Die Personen sprechen in unterschiedlichen Tempos, wodurch sich die Aufnahmen um bis zu einer Sekunde voneinander unterscheiden können. Nichtsdestotrotz ist in dieser Darstellung ein deutlicher Zusammenhang zwischen den verschiedenen Personen erkennbar. Jedes Spektrogramm weist dieselbe Reihenfolge von Konstrukten auf.

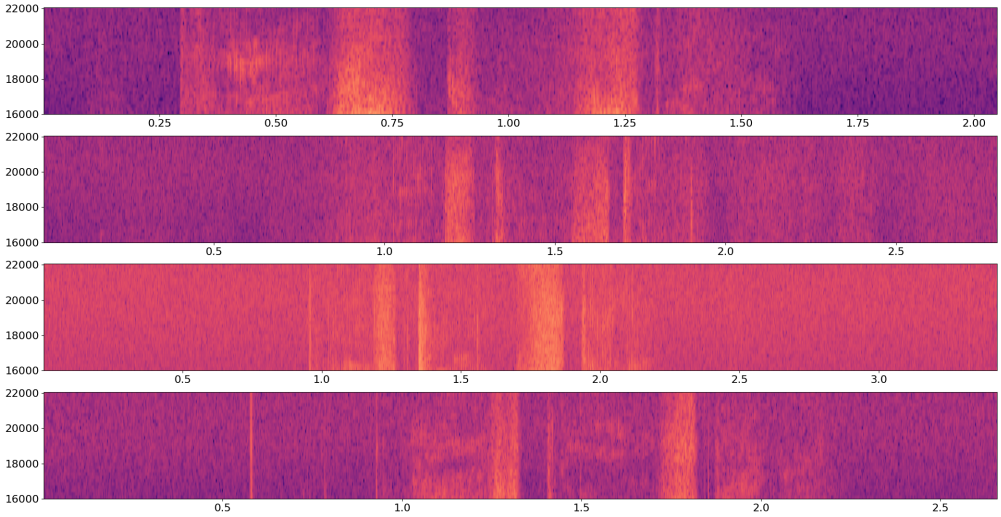

Abbildung 5.2: 4 unterschiedliche Personen sprechen denselben Satz aus

Das nächste Beispiel in Abbildung 5.3 zeigt erneut 4 Spektrogramme. Diesmal zeigt jedes Spektrogramm dieselbe Person, wobei diese Person jedes Mal einen anderen Satz ausspricht. Hier sind die Unterschiede zwischen den einzelnen Darstellungen deutlich erkennbar. Trotzdem muss erwähnt werden, dass die einzelnen Ausschläge innerhalb der Spektrogramme sich nicht allzu stark voneinander unterscheiden. Eher die Anordnung der Ausschläge und die dazwischen liegenden Abstände machen die Unterschiede deutlich.

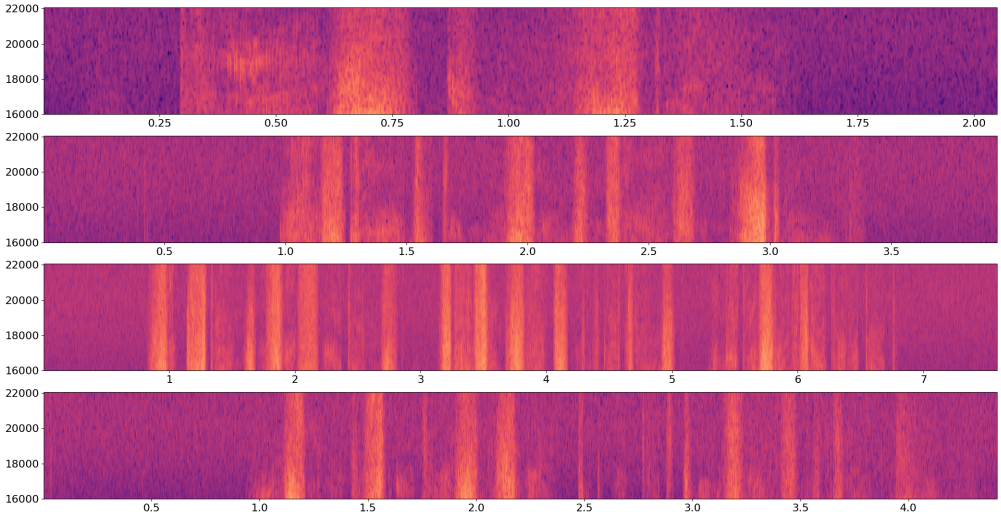

Abbildung 5.3: Eine Person spricht 4 unterschiedliche Sätze aus

Die Visualisierung kann allerdings auch einen falschen Eindruck hinterlassen. Die Audiosignale sind von sehr unterschiedlicher Länge, wodurch die Abstände und Menge der Konstrukte innerhalb der Spektrogramme täuschend sein können. Zusätzlich benötigt das spätere neuronale Netz Eingaben von derselben Länge. Deswegen wird in Abbildung 5.4 erneut ein Vergleich zwischen 4 unterschiedlichen Sätzen derselben Person dargestellt. Diesmal wurden per Hand 4 Proben herausgesucht mit einer ungefähr identischen Länge. In diesem Beispiel ist glücklicherweise immer noch ein deutlicher Unterschied zwischen den 4 Sätzen zu erkennen. Allerdings ist dieser Unterschied nicht mehr so stark wie in der vorherigen Abbildung. Die einzelnen Ausschläge innerhalb der Spektrogramme sehen sich isoliert voneinander sehr ähnlich. Dies kann auf die in Unterabschnitt 2.3.1 besprochene Thematik, dass der Ultraschall der menschlichen Sprache sehr limitiert ist, zurückgeführt werden. Die Menge an Informationen über die menschliche Sprache ist im Ultraschallbereich sehr gering. Nur eine kleine Anzahl an Konsonanten wird überhaupt erkannt, wodurch die Variation der Ausschläge nicht sehr groß ist.

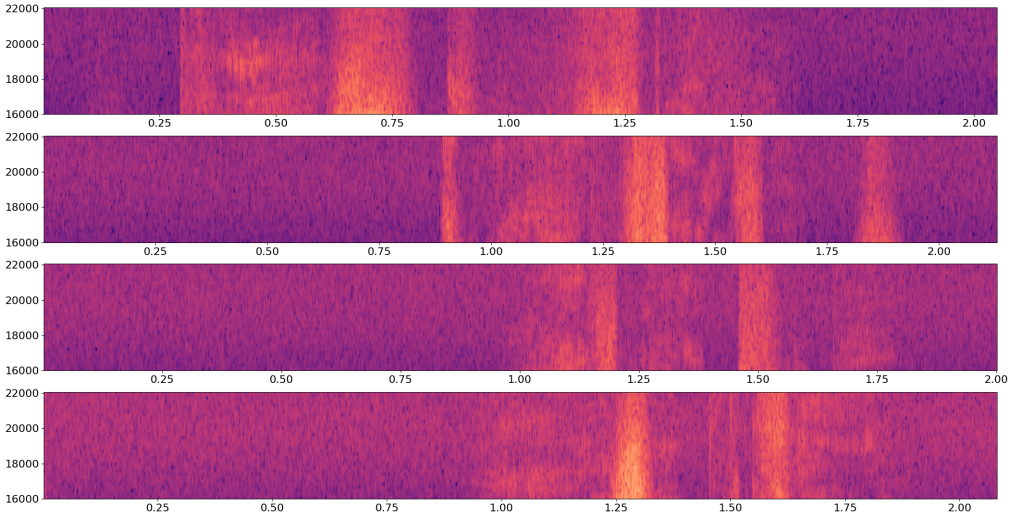

Abbildung 5.4: Eine Person spricht 4 unterschiedliche Sätze aus - ungefähr selbe Länge

Das letzte Beispiel zeigt erneut einen Vergleich zwischen 4 Spektrogrammen. Dieses Mal wurden die Proben allerdings aus einem anderen Datensatz entnommen. Die 4 Audiosignale entstammen aus dem FSD50K Datensatz, welcher aus Hintergrundgeräuschen besteht. In Abbildung 5.5 werden von oben nach unten folgende Hintergrundgeräusche dargestellt:

1. Eine knisternde Plastiktüte
2. Schütteln einer Sprühdose mit Mischkugel
3. Eine Autohupe
4. Ein arbeitender Warenautomat

Diese Töne stehen im Gegensatz zu den vorherigen Beispielen, da hier keine menschliche Sprache enthalten ist. Die Töne besitzen außerdem ebenfalls eine ungefähre Länge von 2 Sekunden. Der Unterschied zwischen den Hintergrundgeräuschen und der menschlichen Sprache ist im Vergleich zu den vorherigen Abbildungen eindeutig. Dabei muss allerdings bedacht werden, dass dies nur eine kleine Auswahl an Hintergrundgeräuschen ist. Was die Abbildung ebenfalls gut darstellt, ist die Repräsentation der Töne im Ultraschallbereich. Während die Plastiktüte und die Sprühdose gut erkennbar sind, sind die Autohupe und der Warenautomat kaum auszumachen. In beiden Fällen sind zwar einige Artefakte zu sehen, wenn genau hin geschaut wird allerdings ist es keinesfalls so offensichtlich wie bei den ersten beiden.

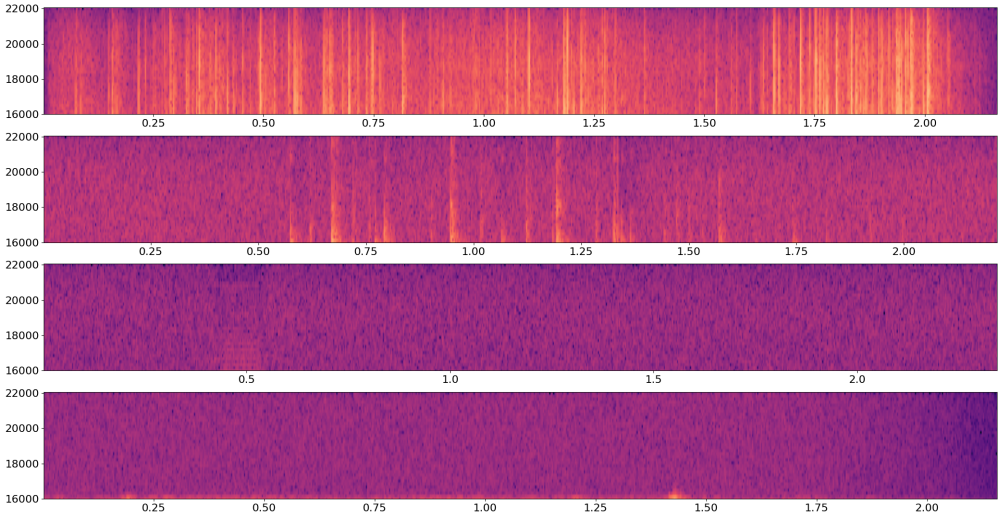

Abbildung 5.5: 4 Hintergrundgeräusche ohne Sprache

5.3 Neuronales Netz zur Klassifizierung

Neuronale Netze eignen sich sehr gut, um Klassifizierungsprobleme zu lösen. Es gibt viele unterschiedliche Formen von neuronalen Netzen. Für die Klassifizierung von Bildern und Audio eignet sich eine Form von neuronalen Netzen besonders. Dies sind die Convolutional Neural Networks (CNN). Sie eignen sich besonders gut, da Bild und Audio in zweidimensionaler Form in die Netzwerke als Eingabe gegeben werden. Es gibt bereits Forschungen, welche CNN nutzen, um Audio zu klassifizieren. Sowohl in [Pic15] als auch in [SB17] wird der bereits zuvor erwähnte UrbanSound8K Datensatz erfolgreich auf die 10 unterschiedlichen Klassen klassifiziert. So wie es auch in den sonstig genannten Forschungen der Fall war, haben auch diese Forschungen dabei keinen Ultraschall verwendet. Nichtsdestotrotz sind die neuronalen Netze und die dahinter liegende Architektur für die hier bearbeitete Problemstellung anwendbar.

5.3.1 Daten Pre-Processing

Das Pre-Processing von den Daten ist ein sehr wichtiger Schritt im Aufbau eines künstlichen neuronalen Netzes. Wenn die Daten zuvor nicht korrekt verarbeitet wurden, kann auch ein noch so komplexes neuronales Netz nicht in der Lage sein, die Merkmale der Daten zu lernen.

5.3.1.1 Data Augmentation Methoden

Der erste wichtige Schritt ist Data Augmentation. Data Augmentation wird in den meisten Fällen verwendet, um die Menge der vorhandenen Daten zu erhöhen oder diese robuster gegen bestimmtes Rauschen zu machen. Wenn beispielsweise ein Audiosignal etwas langsamer gemacht wird, ist das daraus resultierende Signal eine neue Datenprobe. Somit kann die Menge an vorhandenen Daten immens gesteigert werden. Die Autoren in [SB17] haben für die Klassifizierung des UrbanSound8K Datensatzes einige Data Augmentation Methoden genannt und verwendet. Darunter befinden sich die folgenden Methoden:

- **Time-Stretching:** Das Audiosignal wird schneller oder langsamer gemacht. Dadurch verlängert bzw. verkürzt sich auch die Länge des Signales. Wichtig dabei ist, dass die Tonlage des Signals unverändert bleibt.
- **Pitch-Shifting:** Die Tonlage des Signals wird erhöht oder verringert. Dabei wird die Länge des Signals nicht verändert.
- **Dynamic Range Compression:** Die Lautstärke des Audiosignals wird normalisiert. Die lauten Töne werden leiser gemacht, während die leisen Töne lauter gemacht werden.
- **Background Noise:** Es werden Hintergrundgeräusche in das Audiosignal eingefügt.

Dies sind keinesfalls alle Data Augmentation Methoden. Weitere Methoden, die nicht in der spezifischen Forschung genannt wurden, sind unter anderem das Time-Shifting [Dos21] und SpecAugment [Par+19]. Das Time-Shifting verschiebt das Audiosignal auf der Zeitachse um einen zufälligen Wert nach hinten oder vorne. Während die bis jetzt genannten Data Augmentation Methoden direkt auf die Audiosignale angewendet wurden, wird SpecAugment erst auf das später erstellte Spektrogramm angewendet. Dabei werden Frequenz- und Zeitmasken über das Spektrogramm gelegt. Dies sind Balken, die über dem Spektrogramm liegen und die jeweiligen Frequenzen oder Zeitstempel verdecken. Die Frequenzmasken werden horizontal aufgelegt und die Zeitmasken werden vertikal aufgelegt. Die Autoren haben mit Hilfe dieser Methoden Ergebnisse erreicht, die dem Stand der Technik entsprechen. SpecAugment verbessert die Performance von neuronalen Netzwerken und macht diese robuster, wenn die Audiosignale Fehler enthalten und Informationsverlust aufweisen [PC19].

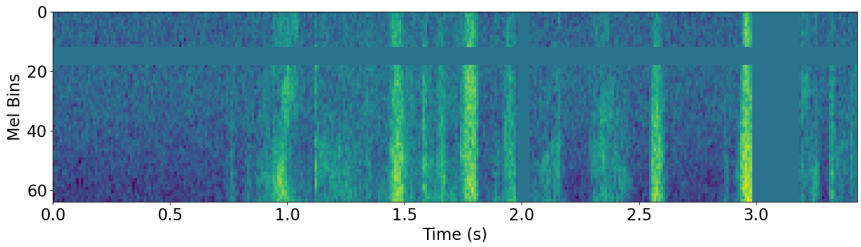

Abbildung 5.6: SpecAugment Frequenz- und Zeitmasken

Ein Beispiel für die Anwendung von SpecAugment wird in Abbildung 5.6 dargestellt.

5.3.1.2 Mel Spektrogramm

Der zweite wichtige Schritt des Pre-Processings ist die Konvertierung des Signals in ein Spektrogramm. Das Besondere hier ist nun, dass das Signal in ein Mel Spektrogramm umgewandelt wird. Mel Spektrogramme sind Spektrogramme, welche die Frequenzskala auf der y-Achse mit der Mel-Skala ersetzen. Während die Frequenzskala von den vorherigen Spektrogrammen eine lineare Skalierung hat, verwendet die Mel-Skala eine logarithmische Skalierung. Der Hintergrund für die logarithmische Eigenschaft der Skala ist die menschliche Wahrnehmung von Tönen. Die Abstände zwischen Tönen auf der Mel-Skala soll den gleichwertigen Abstand darstellen, den Menschen beim Abspielen von Tönen wahrnehmen [Ped65]. Die Skala wird in Form von Mel-Bins angegeben. In der Beispielabbildung besitzt das Mel Spektrogramm 64 Mel-Bins. Das heißt, die vorherige Frequenzspanne wurde in 64 gleichgroße Abschnitte aufgeteilt. Gleichgroß heißt in diesem Kontext, dass jeder Abschnitt den gleichen Unterschied in der Wahrnehmung der Menschen widerspiegelt. Der Unterschied zwischen 200 Mels und 300 Mels sollte sich identisch zu dem Unterschied zwischen 2600 und 2700 Mels anfühlen [Ped65].

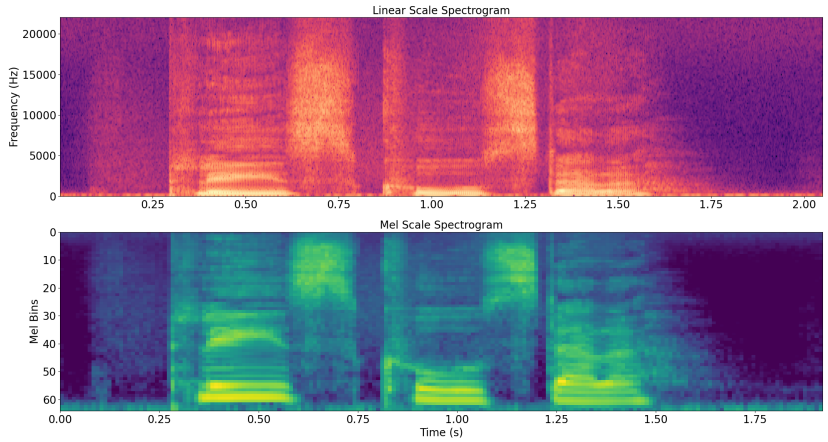

Abbildung 5.7: Vergleich: Mel Spektrogramm und "normales" matplotlib Spektrogramm

Abbildung 5.7 vergleicht eines der vorherigen Spektrogramme mit einem Mel Spektrogramm. In diesem Fall stellen die Spektrogramme dieselben Audiosignale dar. Die unterschiedlichen Skalierungen sind sehr gut zu erkennen. Zusätzlich haben Mel-Spektrogramme eine deutliche niedrigere Auflösung. Dies ist später für die Eingabe in das künstliche neuronale Netz von Bedeutung. Einerseits wird die Geschwindigkeit des Trainings erhöht. Andererseits wird der benötigte Speicher für die Datensätze verringert.

5.3.1.3 Pre-Processing der verwendeten Daten

Das Pre-Processing für die in dieser Arbeit verwendeten Daten wurde in Python durchgeführt. Dafür wurden die Bibliotheken *torchaudio* [Yan+21] und *librosa* [McF+23] verwendet. Der Ablauf der Datenverarbeitung sah dabei wie folgt aus:

1. Die Abtastraten der Audiosignale wurden normalisiert. Alle Audiosignale sind auf 44.1 kHz herunter getaktet.
2. Die Anzahl der Audiokanäle wurde normalisiert. Alle Audiosignale wurden von Stereo auf Mono reduziert. Einerseits ist die Anzahl der Kanäle von den Rohdaten unterschiedlich, andererseits verringert dies den benötigten Speicherplatz und somit auch die benötigte Verarbeitungszeit des neuronalen Netzes.
3. Alle Audiosignale wurden auf eine konsistente Länge von 4 Sekunden gekürzt bzw. verlängert. Bei der Verlängerung der Audiosignale wurden diese mittels *zero padding* gestreckt. Dabei werden die Vektoren mit Nullen gefüllt, bis die gewünschte Länge erreicht wird. Die Nullen wurden dabei zufällig nach vorne und hinten aufgeteilt.
4. Aus jedem Audiosignal wurde ein Mel Spektrogramm erstellt.
5. Die für diese Arbeit selbst aufgenommenen Daten wurden mit Hilfe der genannten Data Augmentation Methoden ergänzt, um die Menge der Daten zu erhöhen.

5.3.2 Erstellen eines Convolutional Neural Networks

Das Erstellen von Convolutional Neural Networks für die Rekonstruktion von Sprache ist ein bereits gelöstes Problem für das menschlich hörbare Spektrum. Für den Aufbau des hier folgenden neuronalen Netzes wurden zum Großteil Referenzen aus [SB17] und [Dos21] gezogen. Das neuronale Netz ist durch sehr viele Iterationen gegangen und dabei sind eine Menge an unterschiedlich guten Ergebnissen entstanden. In dieser Arbeit wird lediglich das “beste” Ergebnis dargestellt.

5.3.2.1 Abstraktion des Problems

Die zuvor genannten existierenden Forschungen klassifizieren erfolgreich Mehrklassenprobleme. Im Falle des UrbanSound8K Datensatzes sind dies 10 Klassen. Das hier beschriebene Problem limitiert sich allerdings auf nur 2 Klassen. Damit wird aus dem Mehrklassenproblem ein binäres Klassenproblem. In diesem Fall bestehen die beiden Klassen aus:

- Das Audiosignal enthält menschliche Sprache
- Das Audiosignal enthält **keine** menschliche Sprache

Im Kontext eines neuronalen Netzes ist ein wichtiger Unterschied dieser Probleme die Wahl der Verlustfunktion. Die Verlustfunktion berechnet für jede Entscheidung, die das neuronale Netz fällt, einen Fehler. Dieser Fehler gibt an, wie weit die Entscheidung von dem wahren Wert entfernt ist. Das Ziel ist also, den Fehler der Verlustfunktion zu minimieren, damit das neuronale Netz die besten Entscheidungen treffen kann. Im Falle eines binären Klassenproblems fällt die Wahl der Verlustfunktion auf die *BinaryCrossEntropyLoss* Funktion.

5.3.2.2 Trainingsdaten

Die Trainingsdaten für das neuronale Netz wurden so gewählt, dass die Klassen ungefähr balanciert sind. Das bedeutet, dass die Datensätze, in denen menschliche Sprache enthalten ist, ungefähr der Anzahl der Datensätze entsprechen, in denen keine menschliche Sprache enthalten ist. Datenproben werden für das Training von neuronalen Netzen mit sogenannten Labels versehen. Labels sagen über die Datenproben aus, welcher Klasse diese Probe zuzuordnen ist. Damit weiß das neuronale Netz nach einer Entscheidung, ob es damit richtig lag oder nicht bzw. wie weit die Entscheidung von der Wahrheit entfernt war. Im Falle von binären Klassifizierungsproblemen werden die Labels meist als 0 und 1 gesetzt. Für dieses neuronale Netz erhalten die Datenproben mit der menschlichen Sprache das Label 1 und die Datenproben ohne menschliche Sprache das Label 0.

Für die Klasse 1 wurde der der VCTK Datensatz herangezogen. Dieser Datensatz enthält insgesamt 88.328 Datenproben. Davon wurden für dieses neuronale Netz 15.910 Datenproben für das Training ausgewählt. Innerhalb dieser Teilmenge sind 20 unterschiedliche Sprecher vertreten mit Aufnahmen von jeweils 2 unterschiedlichen Mikrofonen. Die restlichen 72.418 Datenproben werden später für das Testen des fertigen Modelles genutzt. Für die Klasse 0 wurde der FSD50K Datensatz verwendet. Dieser besteht von vornherein bereits aus einem Trainings- und Testdatensatz. Der Trainingsteil enthält 40.966 Datenproben, wovon 16.350 für das Training verwendet wurden. Der Testdatensatz sowie der Rest des Trainingsdatensatzes werden ebenfalls später für das Testen des fertigen Modelles verwendet.

5.3.2.3 Modellaufbau

Für den Aufbau des Modells wurden die Modellgrößen von [SB17] und [Dos21] evaluiert und für dieses Problem adaptiert. Die spezifische Größe des Modelles lautet wie folgt:

1. **Convolutional Layer:** Eingabegröße - 1, Ausgabegröße - 12, Kernel Size 5x5, Strides (5,5)
2. **MaxPooling:** Kernel Size 2, Strides 1
3. **Aktivierungsfunktion:** ReLU
4. **Convolutional Layer:** Eingabegröße - 12, Ausgabegröße - 24, Kernel Size 3x3, Strides (3,3)
5. **Flatten des Inputs**
6. **Lineares Layer:** Eingabegröße 1584, Ausgabegröße 48
7. **Aktivierungsfunktion:** ReLU
8. **Lineares Layer:** Eingabegröße 48, Ausgabegröße 1
9. **Aktivierungsfunktion:** Sigmoid

Da dieses Netzwerk ein binäres Klassifizierungsproblem lernen soll, ist die Ausgabegröße der letzten Ebene 1. Das heißt, dass das Netzwerk einen einzigen Wert als Ergebnis für die Entscheidung der Klassen ausgibt. Die Ausgabe der sigmoiden Aktivierungsfunktion hat einen reellen Wert zwischen 0 und 1. Ist der Wert größer oder gleich 0,5 wird dieser als Klasse 1 klassifiziert und ist der Wert kleiner als 0,5 wird dieser als Klasse 0 klassifiziert. Zusätzlich kann anhand dieses Wertes festgestellt werden, mit welcher Wahrscheinlichkeit das neuronale Netz die Klasse vorhersagt. Das heißt, ein Wert, der sehr nahe bei 0 oder 1 liegt, besagt, dass das Netzwerk sich in der Entscheidung der Klasse sehr sicher ist. Wenn sich der Wert in der Nähe von 0,5 befindet, kann die Aussage zwar immer noch korrekt sein, aber das Netzwerk ist sich dabei sehr unsicher.

5.3.2.4 Training und Test

Das Modell wird auf 40 Epochen trainiert. Der gewählte Optimizer ist *Stochastik Gradient Descent mit Momentum* mit einer Learning Rate von 0,001. In Abbildung 5.8 ist die Entwicklung des Trainings- und Validierungsfehlers über die 40 Epochen dargestellt. Der Graph zeigt eine gute Entwicklung des Fehlers und deutet auf ein wenig underfitting hin, da der Validierungsfehler etwas über dem Trainingsfehler liegt.

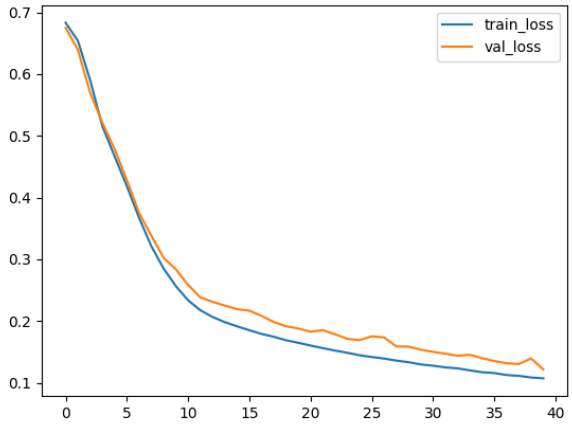

Abbildung 5.8: Vergleich Trainings- und Validierungsfehler über 40 Epochen

Nach dem abgeschlossenen Trainingsdurchlauf wird das Modell abgespeichert, um es auf anderen Daten zu testen. Die Testdaten bestehen aus den restlichen Daten der genutzten Trainingsdatensätze sowie weiteren komplett ungenutzten Datensätzen. Die genaue Zusammensetzung der Testdaten sieht wie folgt aus:

1. Ein Testdatensatz, der während des Trainingsprozesses von den Trainingsdaten abgespalten wird. Es werden dafür zufällig 10% der Trainingsdaten gewählt. Somit ist dieser Datensatz nach jedem Trainingsdurchlauf unterschiedlich. Dies entspricht ungefähr ~2800 Datenproben.

2. Der komplette UrbanSound8K Datensatz. Dieser enthält 8732 Geräusche aus städtischem Umfeld.

3. FSD50K_testing, dies ist ein im Vorfeld abgespaltener Teil aus dem FSD50K Datensatz. Dieser enthält 10.231 Proben.

4. Die restlichen ungenutzten Datenproben aus dem VCTK Datensatz, dieser enthält die restlichen 89 Sprecher und somit 71.622 Datenproben.

5. Küchenaufnahme, ein selbst aufgenommener Datensatz mit 900 Datenproben. Enthält drei Personen, die in einer Küche sprechen.

6. Essen_Rechner, ein selbst aufgenommener Datensatz mit 1208 Datenproben. Enthält Hintergrundgeräusche einer Person, welche isst und Geräusche am Schreibtisch macht. Dabei ist keine Sprache enthalten.

5.3.3 Ergebnisse

Die Ergebnisse des Modells werden in diesem Kapitel vorgestellt. Zuerst werden in Abbildung 5.9 die Genauigkeiten der Anwendung des Modells auf die ersten 4 Datensätze dargestellt. Diese Datensätze hat das Modell noch nie gesehen und haben somit nicht zu dem Training beigetragen. Die Genauigkeiten auf den Testdatensätzen sind erstaunlich hoch. Das Modell erkennt in 87,33% der Fälle die Stimmen im VCTK Datensatz und weist diesen Datenproben die Klasse 1 zu. Bei dem internen Testdatensatz sowie dem FSD50K_testing Datensatz erreicht das Modell sogar eine Genauigkeit von rund 96%. Das Erstaunlichste ist allerdings der UrbanSound8K Datensatz. Hier erreicht das Modell eine Genauigkeit von über 98%. Das Besondere daran ist, dass der UrbanSound8K Datensatz der einzige Datensatz ist, welcher komplett unabhängig von den Trainingsdaten ist. Auch wenn das Modell die Testteilmenge aus VCTK noch nie gesehen hat, sind diese trotzdem aus demselben Datensatz wie die Trainingsteilmenge und weisen somit gewisse Zusammenhänge zu den Trainingsdaten auf, wie zum Beispiel die Verwendung desselben Mikrofons.

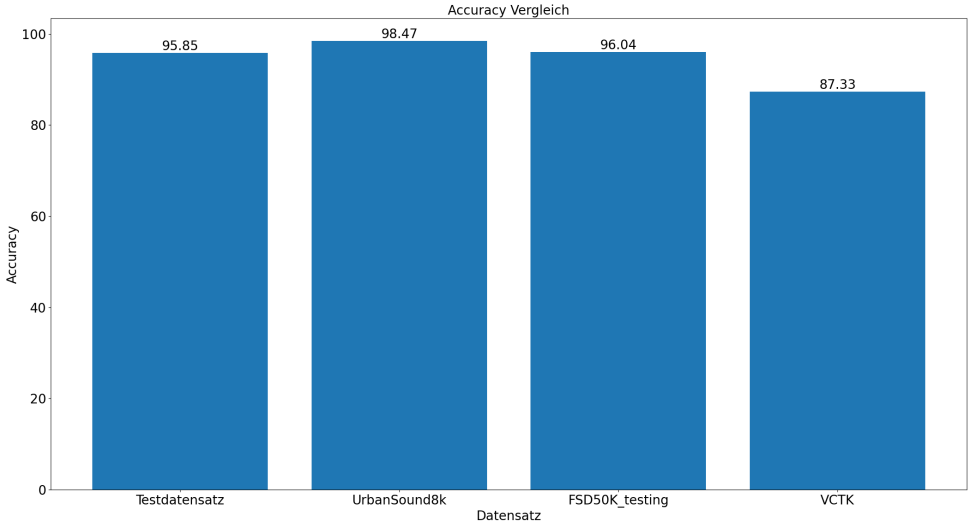

Abbildung 5.9: Vergleich Genauigkeiten ohne eigene Aufnahmen

Das Modell scheint also in der Lage zu sein, mit ziemlich hohen Genauigkeiten zu bestimmen, ob innerhalb einer Ultraschallaufnahme Stimmen enthalten sind oder nicht. Auf den ersten Blick sieht dieses Ergebnis wie ein Erfolg aus. Die Tests bestanden allerdings aus weiteren Testdatensätzen, die in der oberen Abbildung ausgelassen wurden. Wenn nun ein Blick auf Abbildung 5.10 geworfen wird, sieht die Situation schon anders aus. Die Genauigkeiten für die selbst aufgenommenen Datensätze sind 0% für *Kuechenaufnahme* und 99,83% für *Essen_Rechner*. Das Modell entscheidet also für diese beiden

Datensätze zu fast 100%, dass sie der Klasse 0 angehören. Im Falle der *Kuechenaufnahme* führt dies zu einer Genauigkeit von 0%, während die Genauigkeit des anderen Datensatzes somit fast 100% ist. Die Genauigkeit hängt somit nicht davon ab, dass das Modell so gut ist, sondern davon, dass der Datensatz zufälligerweise vollständig eine Klasse vertritt. An diesem Punkt stellt sich die Frage, ob die Genauigkeit des UrbanSound8K Datensatzes vielleicht auf Grund desselben Fehlers entstanden ist. Auch in diesem Fall ist der wahre Wert die Klasse 0. Es könnte also sein, dass auch hier das Modell zu fast 100% die 0 vorhersagt und durch einige Ausnahmen eine Genauigkeit von fast 99% erreicht wurde.

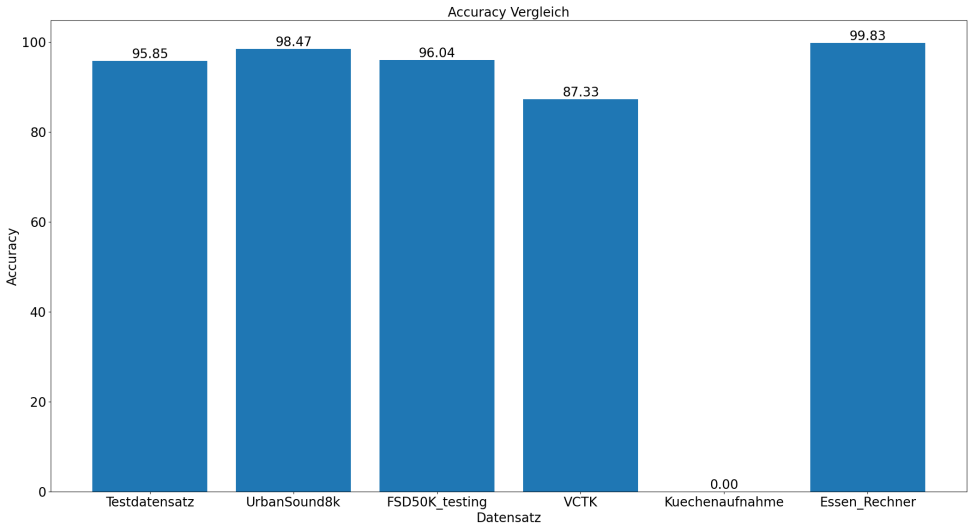

Abbildung 5.10: Vergleich Accuracies mit eigenen Aufnahmen

5.3.4 Probleme

Das Modell ist nicht in der Lage, die Existenz von Stimmen in ungesehenen Datensätzen zu erkennen. Dies ist offensichtlich ein Problem. Dieses Problem wurde in dieser Arbeit konsistent über alle Konfigurationen des neuronalen Netzes beobachtet. Es scheint, als würde das Netzwerk nicht die korrekten Merkmale über die Audiosignale lernen. Das Netzwerk lernt einen Unterschied zwischen den Datensätzen, welcher nicht auf Anhieb erkennbar ist. Eine Möglichkeit für das Lernen der inkorrekten Merkmale könnte durch den Moiré-Effekt entstehen. Der Moiré-Effekt kann bei der Überlagerung von zwei Mustern entstehen und die Muster müssen dabei nicht identisch sein [Boo01].

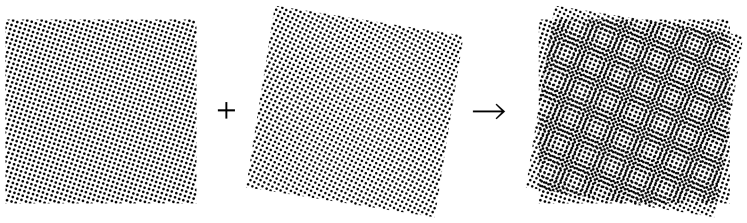

Abbildung 5.11: Beispiel Moiré-Effekt

Quelle: *https://upload.wikimedia.org/wikipedia/commons/5/52/Moir%C3%A92.png* (besucht am 13. 08. 2023).

Abbildung 5.11 zeigt ein Beispiel für das Entstehen eines Moiré-Effekts durch die Überlagerung zweier Muster. Dieser Effekt kann auch innerhalb der Datenverarbeitung für das neuronale Netz entstehen. Einige der möglichen Orte für den Moiré-Effekt sind unter anderem:

1. Abtastraten
2. Short Time Fourier Transform
3. Mel Skala

Die Abtastraten werden zu Beginn des Prozesses auf 44.1 kHz normalisiert. Ebenso wird dieser Wert in keinem der Schritte verändert. Die Short Time Fourier Transform wird für die Erstellung der Spektrogramme verwendet. Bei der Verwendung dieser Funktion spielt die gewählte Fenstergröße eine wichtige Rolle. Um zu überprüfen, ob die Fenstergröße möglicherweise ungewollte Artefakte in das Spektrogramm einbringt, werden 2 Abbildungen dargestellt. Die erste Abbildung zeigt vier verschiedene Fenstergrößen. Dabei ist zu beachten, dass die Fenstergrößen spezifische Unterschiede haben. In Abbildung 5.12 bestehen die Fenstergrößen aus 2er Potenzen. Dies liegt daran, dass der Algorithmus diese deutlich effizienter berechnen kann [lib]. In Abbildung 5.13 wurden die Fenstergrößen extra auf Werte gelegt, die keine 2er Potenz widerspiegeln.

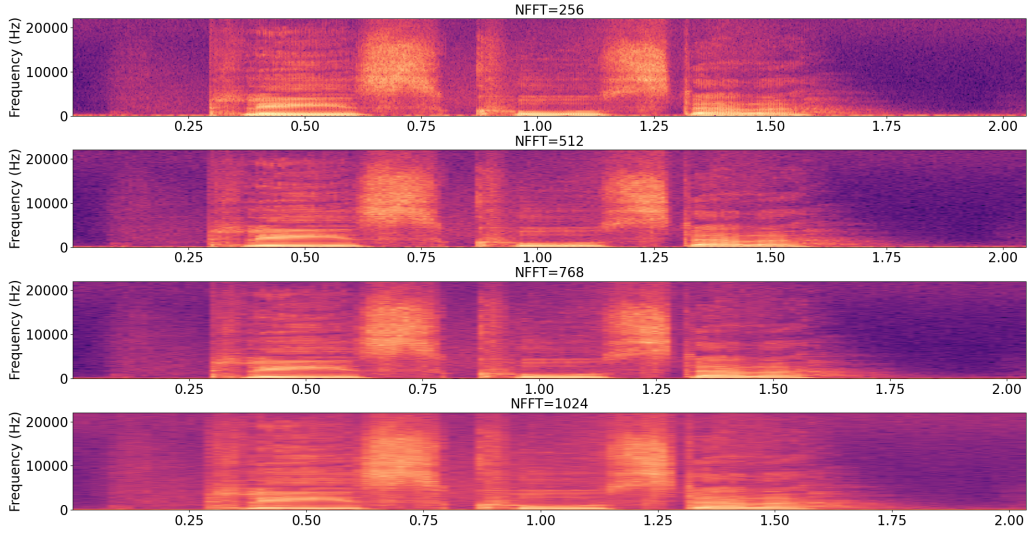

Abbildung 5.12: Spektrogramme mit unterschiedlichen STFT Fenstergrößen - 2er Potenzen

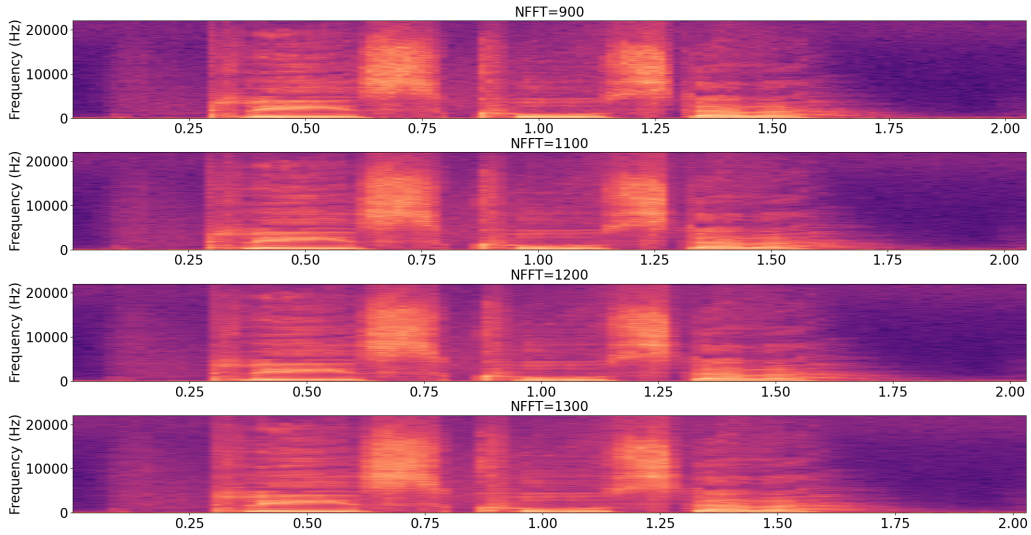

Abbildung 5.13: Spektrogramme mit unterschiedlichen STFT Fenstergrößen - keine 2er Potenzen

Die unterschiedlichen Fenstergrößen sind innerhalb der Spektrogramme zu erkennen.

Allerdings sind in keinem der Spektrogramme besondere Muster auf Grund eines Moiré-Effektes zu sehen.

Der letzte Punkt in der Untersuchung wäre die Nutzung der Mel Skala. Wie bereits in vorherigen Kapiteln erwähnt, wird die Mel Skala verwendet, um die Daten mit Hilfe des Logarithmus näher an die menschliche Wahrnehmung von Tönen zu bringen. Der Grund für das Verwenden von Mel Spektrogrammen in dieser Arbeit war die Tatsache, dass dies der Stand der Technik für Audiorekonstruktion von Sprache ist. Im Nachhinein ergibt sich die Frage, ob der Logarithmus nicht sogar von Nachteil ist, da die zu erkennenden Töne im Ultraschallbereich liegen. Innerhalb dieses Spektrums tritt sowieso keine menschliche Wahrnehmung mehr auf. Dementsprechend müssen die Daten nicht mehr mittels Logarithmus skaliert werden. Die Skalierung soll eigentlich dabei helfen, die Elemente aus dem niedrigen Frequenzbereich hervorzuheben. Dabei werden allerdings auch die Elemente aus den höheren Frequenzbereichen komprimiert. Dies könnte die Menge an wichtigen Merkmalen innerhalb der Spektrogramme, die aus dem Ultraschallbereich kommen, verringern.

5.4 Verschleierung der Sprachinformationen

Ein weiterer zu untersuchender Teil dieser Arbeit ist die Verschleierung der Sprachinformationen. Ist es irgendwie möglich, die Erkennung von Sprache zu limitieren? Da die erzielten Ergebnisse der Sprachrekonstruktion leider nicht vollständig erfolgreich waren, wird dieses Problem ebenfalls ein wenig abstrahiert. Es konnten leider keine spezifischen Wörter oder Sprachsignale rekonstruiert werden und somit können diese auch nicht mittels Rauschen verhindert werden. Dennoch ist das neuronale Netz immerhin in der Lage, ausschließlich mit Ultraschall zwischen 2 Datensätzen zu unterscheiden. Daraus stellt sich nun folgende Fragestellung:

Wie verhält sich die Genauigkeit des neuronalen Netzes, wenn die von Koopango verwendeten Ultraschallsignale für die Ortung mit den Testdatensätzen überlappt werden?

Kann die Genauigkeit des neuronalen Netzes mit Hilfe der Ortungssignale verringert werden? Dafür wurden die von Koopango genutzten Ortungssignale für diese Arbeit bereitgestellt. In Abbildung 5.14 werden 3 Spektrogramme dargestellt. Das erste Spektrogramm ist eine einfache Datenprobe aus dem VCTK Datensatz. Das zweite Spektrogramm zeigt eine der Ortungsproben von Koopango. Dabei sind die von dem Ortungsalgorithmus verwendeten künstlichen Töne gut zu erkennen. Das letzte Spektrogramm ist die Kombination der beiden Datenproben. Für die Überlappung der beiden Audiosignale wurde die Python Bibliothek *AudioSegment* [Str] verwendet. Auf diese Art und Weise wurden nun die kompletten Testdatensätze VCTK und FSD50K_testing bearbeitet, sodass jede Datenprobe zufällig mit einer Ortungsprobe von Koopango überlappt wurde.

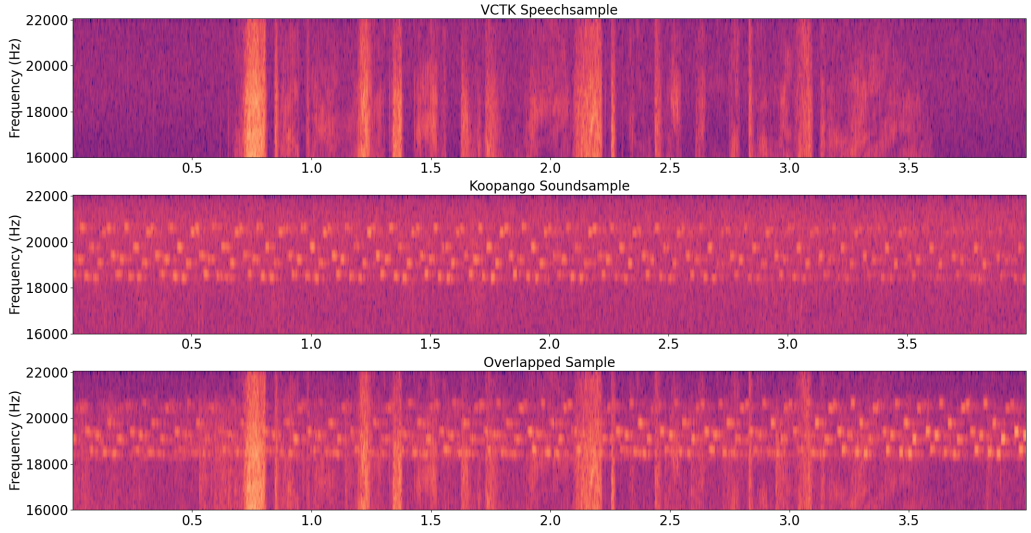

Abbildung 5.14: VCTK Sprachprobe - Koopango Ortungsprobe - Kombinierte Datenprobe

Die daraus entstehenden Testdatensätze wurden als Eingabe in das bereits trainierte Modell des neuronalen Netzes gegeben. Die daraus resultierenden Genauigkeiten sowie ein Vergleich zu den vorherigen Testdatensätzen werden in Abbildung 5.15 dargestellt.

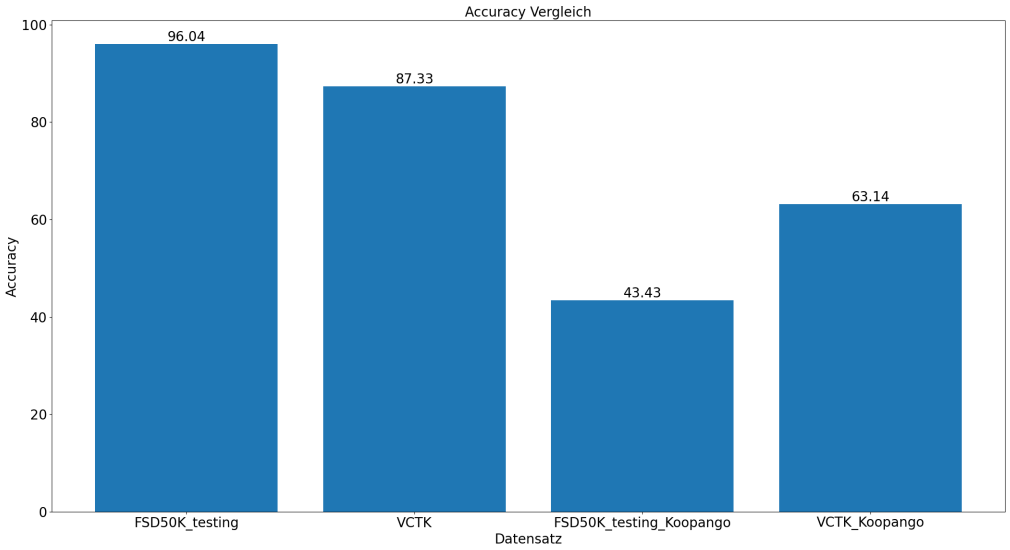

Abbildung 5.15: Vergleich Testdatensätze mit und ohne Koopango Ortungsproben

In der Abbildung ist zu erkennen, dass die Genauigkeiten für die beiden Testdatensätze stark gesunken sind. Der FSD50K_testing Datensatz, welcher nur aus Hintergrundgeräuschen besteht, sinkt von 96,04% auf 43,43%. Währenddessen sinkt die Genauigkeit des VCTK Datensatzes, welcher nur Stimmen enthält, von 87,33% auf 63,14%. Das Schöne an diesem Ergebnis ist, dass das Modell trotz der Veränderung der Daten immer noch eine gewisse Glaubwürdigkeit aufweist, indem es nicht an 0% oder 100% grenzt.

Das reine Hinzufügen der Koopango Ortungssignale führt bereits zu der Verringerung der Genauigkeit des neuronalen Netzes. Dies ist sehr wichtig für Koopango, da somit das Problem des Datenschutzes bereits teilweise gelöst wird. Es muss natürlich trotzdem weiterhin in Richtung Datenschutz geforscht werden, da dies nur der erste von vielen Schritten in die korrekte Richtung ist.

6 Fazit

6.1 Zusammenfassung

In dieser Arbeit wurde untersucht, wie die aktuellen Methoden für die Rekonstruktion von Sprache innerhalb des Ultraschallspektrums aussehen und wie diese Methoden für einen spezifischen Problemfall mit dem Ortungssystem Koopango ausgenutzt werden können. Die Rekonstruktion von Sprache aus dem Ultraschallbereich ist in der wissenschaftlichen Forschung bislang nur wenig vertreten. Die am stärksten erforschte Methode ist das Lippenlesen mittels Ultraschall. Aufgrund der Limitierungen dieser Methode ist diese nicht für den hier untersuchten Problemfall anwendbar und stellt somit keine Problematik für den Datenschutz von Koopango dar.

Da es keine perfekt anwendbare Methode für den Ultraschallbereich gibt, wurde eine Spracherkennungsmethode aus dem menschlichen Hörspektrum adaptiert und auf das Ultraschallspektrum angewendet. Es wurde untersucht, wie mittels Filtern Ultraschallsignale erstellt werden können, um diese in bereits existierende Spracherkennungsmethoden als Eingabe zu geben. Die daraus entstandenen Ultraschallsignale wurden in Systeme wie Whisper oder NVIDIA NeMo gegeben, um die Reaktion auf Ultraschallsignale dieser Systeme zu testen. Nach sehr vielen Tests waren diese Systeme nicht in der Lage, Sprachsignale aus diesen Aufnahmen zu entnehmen, da diese existierenden Systeme nicht darauf aufgebaut sind, Ultraschallsignale zu verarbeiten.

Auf Grund dessen wurde in dieser Arbeit ein eigenes System mit Hilfe von neuronalen Netzen gebaut, um Sprachsignale aus reinem Ultraschall zu erkennen. Um dieses System in den Rahmen einer Masterarbeit zu bringen, wurde die initiale Forschungsfrage etwas abstrahiert, um nur noch zu entscheiden, ob Sprache vorhanden ist oder nicht. Das fertig trainierte Modell des neuronalen Netzes ist in der Lage, mit einer hohen Genauigkeit zwischen den gegebenen Datensätzen zu unterscheiden. Leider scheint das Modell nicht die korrekten Merkmale zu lernen, da die Genauigkeit auf ungesehenen Datensätzen entweder 0% oder 100% ist. Einige Möglichkeiten für diesen Fehler könnten unter anderem die Mel Spektrogramme oder ein versteckter Moiré-Effekt sein.

An diesem Punkt wurde auch die zweite Forschungsfrage etwas abstrahiert, da diese ursprünglich auf die erste Frage aufgebaut hat. Somit wurde das Modell von dem bis jetzt erstelltem neuronalen Netz mit zusätzlichen Ortungsdaten von Koopango getestet. Es wurden die Testdatensätze mit den Ortungssignalen von Koopango überlappt und in das Modell eingegeben. Dabei kamen erstaunliche Ergebnisse heraus, da die überlappten Testdaten eine deutlich geringere Genauigkeit des Modelles aufweisen. Zusätzlich liegen

die Genauigkeiten im Rahmen des Möglichen. Es scheint, als würden die Ortungssignale von Koopango selbst bereits als eine Art Rauschen wirken, wodurch die Erkennung von Sprache schwieriger gemacht wird.

6.2 Ausblick

Diese Arbeit ist im Großen und Ganzen ein Anfangsschritt in die Richtung des Datenschutzes im Ultraschallspektrum. Es wurde eine Menge an Ergebnissen erzielt. Allerdings sind keine der Ergebnisse vollständig und benötigen mehr Zeit und Forschung. Da es keine existierenden Systeme gibt, die den Problemfall von Koopango ausnutzen können, war es schwierig, ein fertiges System in den Umfang dieser Arbeit zu bringen. Die Erstellung des neuronalen Netzes hat bei weitem die meiste Zeit beansprucht und hat auch die größte Menge an Potential für zukünftige Forschungen. Es muss untersucht werden, warum das Modell nicht die korrekten Merkmale lernt. Dafür müssen alle möglichen Fehlerquellen untersucht werden. Dazu gehört auch das Testen von wesentlich mehr Konfigurationen des Netzwerkmodelles. In der kurzen Zeit dieser Arbeit wurde kein Moiré-Effekt ausfindig gemacht. Dies heißt allerdings nicht, dass es nicht doch irgendwo unsichtbar geschieht. Der größte Punkt für zukünftige Forschung ist wahrscheinlich die Funktion der Mel Spektrogramme. Es wurde bereits in der Arbeit kurz angesprochen, dass die logarithmische Eigenschaft dieser Spektrogramme möglicherweise einen Nachteil mit sich bringt.

Als Motivation für die Lösung dieses Datenschutzproblems, wurde in der Einleitung die Vertraulichkeit des Wortes erwähnt. Es ist unklar, in wie weit dieses Gesetz auf den Ultraschallbereich anwendbar ist. Auch dies ist ein Thema, welches in Zukunft untersucht werden sollte.

Eine Idee, um diese Arbeit fortzuführen ist, das Erstellen eigener großer Datensätze. Die Daten für das neuronale Netz sind ebenfalls eine mögliche Fehlerquelle für das Lernen falscher Merkmale. Ebenso muss diese Form eines neuronalen Netzes in einer echten Testumgebung eingesetzt werden.

Diese Arbeit hat, wie bereits zu Beginn erwähnt, auf eine existierende Bachelorarbeit aufgebaut. In der Bachelorarbeit wurde für Koopango auf Konzeptebene die Trennung zwischen Hörschall und Ultraschall untersucht. In dieser darauf folgenden Arbeit hat sich herausgestellt, dass trotz der Trennung der Schallbereiche, immer noch Sprachinformationen aus dem Ultraschallbereich extrahiert werden können. Allerdings ist der Aufwand, um dies effektiv zu erreichen, sehr hoch. Einerseits ist dies ein gutes Zeichen für Koopango, weil es schwer umzusetzen ist. Andererseits scheint es immer noch möglich zu sein. Dies bietet eine Fortsetzung dieser Arbeit für Koopango an, indem untersucht wird, wie viel Aufwand nun wirklich für die Rekonstruktion ganzer Wörter oder sogar Sätze nötig ist. Zusätzlich kann dann auch wieder untersucht werden, wie viel Rauschen benötigt wird, um dies zu verhindern. Die Koopango Ortungssignale wirken bereits wie eine Form von Rauschen, um die Rekonstruktion zu erschweren. Jedoch könnte es sein,

dass ein neuronales Netz dies auch mit einberechnen kann. Wenn das alles erreicht wurde, kann ein Vorschlag an das World Wide Web Consortium (W3C) gestellt werden, um einen Standard für die Verarbeitung von Ultraschalldaten zu erstellen.

Literatur

[Aud] Audacity Team. *Audacity*. URL: https://www.audacityteam.org/ (besucht am 27. 09. 2023).

[Boo01] P. M. Boone. „NDT Techniques: Laser-based“. In: *Encyclopedia of Materials: Science and Technology*. Elsevier, 2001, S. 6018–6021. ISBN: 9780080431529. DOI: 10.1016/B0-08-043152-6/01059-7.

[Deu09] Deutsches Institut für Normung. *DIN 1320:2009-12, Akustik_- Begriffe*. Berlin, 2009. DOI: 10.31030/1544140.

[Din+22] Han Ding u. a. „UltraSpeech“. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.3 (2022), S. 1–25. DOI: 10.1145/3550303.

[Dos21] Ketan Doshi. *Audio Deep Learning Made Simple: Sound Classification, Step-by-Step: An end-to-end example and architecture for audio deep learning's foundational application scenario, in plain English.* 2021. URL: https://towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5 (besucht am 28. 07. 2023).

[Fon+20] Eduardo Fonseca u. a. *FSD50K: An Open Dataset of Human-Labeled Sound Events*. 2020. DOI: 10.48550/arXiv.2010.00475.

[Fu+22] Yongjian Fu u. a. „SVoice“. In: *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*. Hrsg. von Jeremy Gummeson u. a. New York, NY, USA: ACM, 2022, S. 622–636. ISBN: 9781450398862. DOI: 10.1145/3560905.3568530.

[Gao+20] Yang Gao u. a. „EchoWhisper“. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.3 (2020), S. 1–27. DOI: 10.1145/3411830.

[Gooa] Google LLC. *Gboard – die Google-Tastatur*. URL: https://apps.apple.com/de/app/gboard-die-google-tastatur/id1091700242 (besucht am 29. 06. 2023).

[Goob] Google LLC. *Speech-to-Text-Anfrage erstellen: Abtastraten*. URL: https://cloud.google.com/speech-to-text/docs/speech-to-text-requests?hl=de#sample-rates (besucht am 27. 07. 2023).

[Gul+20] Anmol Gulati u. a. „Conformer: Convolution-augmented Transformer for Speech Recognition“. In: *Interspeech 2020*. ISCA: ISCA, 2020, S. 5036–5040. DOI: 10.21437/Interspeech.2020-3015.

[Guo+22] Hanqing Guo u. a. „SUPERVOICE“. In: *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*. Hrsg. von Yuji Suga u. a. New York, NY, USA: ACM, 2022, S. 1019–1033. ISBN: 9781450391405. DOI: 10.1145/3488932.3517420.

[Han+20] Wei Han u. a. *ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context*. 2020. DOI: 10.48550/arXiv.2005.03191.

[Hun07] J. D. Hunter. „Matplotlib: A 2D graphics environment“. In: *Computing in Science & Engineering* 9.3 (2007), S. 90–95. DOI: 10.1109/MCSE.2007.55.

[Kooa] Koopango. *Koopango*. URL: https://koopango.com (besucht am 25. 09. 2023).

[Koob] Koopango. *Koopango - Klinik Navigation*. URL: https://koopango.com/wp-content/uploads/2023/01/MicrosoftTeams-image-11-1-440x440.png (besucht am 25. 09. 2023).

[Kri+20] Samuel Kriman u. a. „Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions“. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, S. 6124–6128. ISBN: 978-1-5090-6631-5. DOI: 10.1109/ICASSP40776.2020.9053889.

[Li+19] Jason Li u. a. *Jasper: An End-to-End Convolutional Neural Acoustic Model*. 2019. DOI: 10.48550/arXiv.1904.03288.

[lib] librosa. *librosa.stft: Short-time Fourier transform (STFT)*. URL: https://librosa.org/doc/latest/generated/librosa.stft.html (besucht am 20. 09. 2023).

[McF+23] Brian McFee u. a. *librosa/librosa: 0.10.0.post2*. 2023. DOI: 10.5281/zenodo.7746972.

[Nag+20] Arsha Nagrani u. a. „Voxceleb: Large-scale speaker verification in the wild“. In: *Computer Speech & Language* 60 (2020), S. 101027. ISSN: 08852308. DOI: 10.1016/j.csl.2019.101027.

[NVI] NVIDIA. *NVIDIA NeMo*. URL: https://github.com/NVIDIA/NeMo (besucht am 11. 09. 2023).

[Pan+15] Vassil Panayotov u. a. „Librispeech: An ASR corpus based on public domain audio books“. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, S. 5206–5210. ISBN: 978-1-4673-6997-8. DOI: 10.1109/ICASSP.2015.7178964.

[Par+19] Daniel S. Park u. a. „SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition“. In: (2019). DOI: 10.48550/arXiv.1904.08779.

[PC19] Daniel S. Park und William Chan. *SpecAugment: A New Data Augmentation Method for Automatic Speech Recognition*. 2019. URL: https://ai.googleblog.com/2019/04/specaugment-new-data-augmentation.html (besucht am 28. 07. 2023).

[Ped65] Paul Pedersen. „The Mel Scale“. In: *Journal of Music Theory* 9.2 (1965), S. 295. ISSN: 00222909. DOI: 10.2307/843164.

[Pic15] Karol J. Piczak. „Environmental sound classification with convolutional neural networks“. In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, S. 1–6. ISBN: 978-1-4673-7454-5. DOI: 10.1109/MLSP.2015.7324337.

[Rad+22] Alec Radford u. a. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. DOI: 10.48550/arXiv.2212.04356.

[RB18] Mirco Ravanelli und Yoshua Bengio. „Speaker Recognition from Raw Waveform with SincNet“. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, S. 1021–1028. ISBN: 978-1-5386-4334-1. DOI: 10.1109/SLT.2018.8639585.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri und Robert B. Dunn. „Speaker Verification Using Adapted Gaussian Mixture Models“. In: *Digital Signal Processing* 10.1-3 (2000), S. 19–41. ISSN: 10512004. DOI: 10.1006/dspr.1999.0361.

[SB17] Justin Salamon und Juan Pablo Bello. „Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification“. In: *IEEE Signal Processing Letters* 24.3 (2017), S. 279–283. ISSN: 1070-9908. DOI: 10.1109/LSP.2017.2657381.

[Sen] Sennheiser. *Sennheiser HD650 Spezifikationen*. URL: https://www.sennheiser-hearing.com/de-DE/p/hd-650/ (besucht am 11. 09. 2023).

[Sha49] C. E. Shannon. „Communication in the Presence of Noise“. In: *Proceedings of the IRE* 37.1 (1949), S. 10–21. ISSN: 0096-8390. DOI: 10.1109/JRPROC.1949.232969.

[Sil18] H. Ward Silver. „Filter Basics: Stop, Block, and Roll(off)“. In: (2018), S. 16–20. URL: https://nutsvolts.texterity.com/nutsvolts/201807/?folio=16&pg=16#pg16 (besucht am 06. 09. 2023).

[SJB14] Justin Salamon, Christopher Jacoby und Juan Pablo Bello. „A Dataset and Taxonomy for Urban Sound Research“. In: *Proceedings of the 22nd ACM international conference on Multimedia*. Hrsg. von Kien A. Hua u. a. New York, NY, USA: ACM, 2014, S. 1041–1044. ISBN: 9781450330633. DOI: 10.1145/2647868.2655045.

[Str] Max Strange. *AudioSegment*. URL: https://github.com/MaxStrange/AudioSegment (besucht am 20. 09. 2023).

[Tuo21] Jennifer P. Tuohy. *Some Echo speakers can now detect people*. 2021. URL: https://www.theverge.com/2021/11/12/22779044/amazon-echo-speakers-ultrasound-people-detection (besucht am 30. 08. 2023).

[Vir+20] Pauli Virtanen u. a. „SciPy 1.0: fundamental algorithms for scientific computing in Python“. In: *Nature methods* 17.3 (2020), S. 261–272. DOI: 10.1038/s41592-019-0686-2.

[Wan+18] Li Wan u. a. „Generalized End-to-End Loss for Speaker Verification“. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, S. 4879–4883. ISBN: 978-1-5386-4658-8. DOI: 10.1109/ICASSP.2018.8462665.

[Yan+21] Yao-Yuan Yang u. a. *TorchAudio: Building Blocks for Audio and Speech Processing*. 2021. DOI: 10.48550/arXiv.2110.15018.

[YVM19] Junichi Yamagishi, Christophe Veaux und Kirsten MacDonald. *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)*. 2019. DOI: 10.7488/ds/2645.

[Zac] ZachMan42. *Frequency Shifter Plugin Audacity*. URL: https://forum.audacityteam.org/t/frequency-shifter/66030 (besucht am 06. 07. 2023).

[Zha+21] Qian Zhang u. a. „SoundLip“. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.1 (2021), S. 1–28. DOI: 10.1145/3448087.

[Zha+22] Yu Zhang u. a. „BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition“. In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), S. 1519–1532. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2022.3182537.

Eidesstattliche Versicherung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, alle Ausführungen, die anderen Schriften wörtlich oder sinngemäß entnommen wurden, kenntlich gemacht sind und die Arbeit in gleicher oder ähnlicher Fassung noch nicht Bestandteil einer Studien- oder Prüfungsleistung war.

Rostock, 18. Oktober 2023

Jan Heisenberg