University of Erlangen-Nuremberg

Multimedia Communications and Signal Processing

Prof. Dr.-Ing. André Kaup

Titel

Directional noise suppression for robot audition

Thomas Heller

November 2015

Professor: Prof. Dr.-Ing. Walter Kellermann
Betreuer: M.Sc. Hendrik Barfuss

Inhaltsverzeichnis

1 Einleitung **3**

2 Grundlagen - Theorie **4**

2.1 Gerichteter Beamformer 4

2.2 Störunterdrückung G_f 6

2.3 Richtungsgewinn G_d 6

3 Hauptteil **7**

3.1 Akustisches Model 7

3.2 Erläuterungen zum MATLAB-Code 7

3.3 Einflussreiche Parameter 8

3.3.1 Entscheidungsgrenze des Richtungsgewinns 8

3.3.2 Unterschiedliche Mikrofonabstände 8

3.4 Evaluation 9

3.4.1 Antialiasing 9

3.4.2 Natürliche Phasendrehung 10

3.4.3 Wirkung unterschiedlicher Frequenzbereiche 11

3.4.4 Berechnete Gewinne 12

4 Zusammenfassung **15**

1 Einleitung

Das Ziel in der Entwicklung humanoider Roboter ist es, den Menschen möglichst echt abzubilden. Dafür ist eine flüssige akustische Kommunikation zwischen Mensch und Roboter unerlässlich. Dabei sind die komplexen Vorgänge im menschlichen Ohr und Gehirn schwer zu modellieren und zum Teil noch nicht ausreichend erforscht. Die Probleme, die sich im Bereich der akustischen Spracherkennung stellen, sind: der Hall des Raumes in dem der Roboter sich befindet, störende Quellen und Rauschen. Um diese Probleme ansatzweise zu lösen werden im Roboter üblicher Weise mehrere Mikrofone verwendet, die zusammengeschlossen ein sogenanntes Mikrofonarray bilden. Dadurch ist es möglich richtungsabhängige akustische Signale zu empfangen. Dazu werden zum Beispiel die Signale unterschiedlich gewichtet und verzögert addiert und ergeben den sogenannten Delay-Sum-Beamformer. Bei größeren Distanzen des Sprechers vom Roboter dominieren jedoch weiterhin die Probleme. Ein verbesserter Beamformer wird deshalb in [5] vorgestellt. In diesem Forschungspraktikum wird der dort vorgestellte Algorithmus am Beispiel des humanoiden Roboters NAO in MATLAB implementiert. Außerdem wird die Implementierung anhand realer Messwerte evaluiert.

2 Grundlagen - Theorie

2.1 Gerichteter Beamformer

In diesem Forschungspraktikum wird der Beamformer aus [5] implementiert und getestet.

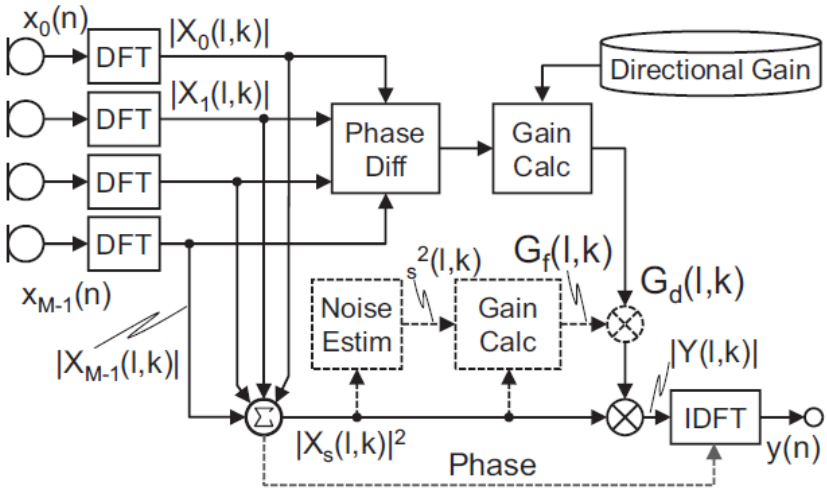

Abbildung 2.1: Richtungsabhängige Störunterdrückung nach [5]

Das Blockdiagramm des Beamformers ist in Abbildung 2.1 dargestellt. $x_{\mathrm{m}}(n)$ steht dabei für das Zeitsignal des $m-ten$ Mikrofons mit dem Zeitindex n. DFT steht für diskrete Fouriertransformation, deren Ausgangswerte die Spektren $X_{\mathrm{m}}(l,k)$ sind. Dabei ist l der zeitliche Blockindex und k ist der Frequenzindex im folgenden auch Frequenzbin genannt. Das Betragsquadrat der Summe der Spektren $|X_{\mathrm{s}}(l,k)|^2$ errechnet sich nach [5] mit (2.1).

$$|X_{\mathrm{s}}(l,k)|^2 = |\sum_{m=0}^{M-1} X_{\mathrm{m}}(l,k)|^2 \tag{2.1}$$

Aus der noise estimation zu deutsch Rauschschätzung geht die geschätzte Rauschleistungsdichte ${\sigma_{\mathrm{s}}}^2(l,k)$ hervor. Sie wird bei der Berechnung des spectral gains ($G_{\mathrm{f}}(l,k)$, dt. spektraler Gewinn) genutzt. Außerdem wird mithilfe der Phasendifferenzen der Spektren und einer Vergleichstabelle das directional gain ($G_{\mathrm{d}}(l,k)$, dt. Richtungsgewinn) errechnet. Das Ausgangsspektrum wird nach [5] mit Formel (2.2) berechnet.

$$|Y_{\mathrm{s}}(l,k)|^2 = G_{\mathrm{f}}(l,k) \cdot G_{\mathrm{d}}(l,k) \cdot |X_{\mathrm{s}}(l,k)|^2 \tag{2.2}$$

Um für die Spracherkennung ein Zeitsignal zu erhalten, wird die inverese diskrekte Fouriertransformation auf die Ausgangsspektren, multipliziert mit der Eingangsphase, angewandt.

Der Algorithmus besteht damit aus einem Sumbeamformer, mit dem eine einkanalige Rauschunterdrückung durchgeführt wird. Außerdem wird parallel mit Hilfe der Phasendifferenzen mehrerer Kanäle ein Richtungs-Beamformer realisiert. Dabei wird je Frequenzbin eine maximale Phasendifferenz definiert und bei einer Überschreitung wird das Frequenzbin unterdrückt. Die maximale Phasendifferenz wird in (2.3) aus der Frequenz f, dem Abstand zweier Mikrophone d, der Lichtgeschwindigkeit c und dem maximalen Ankunftswinkel ϕ errechnet. In [5] werden zwei verschiedene Ankunftswinkel verwendet um den Übergang zwischen gewollter und unterdrückter Richtung zu glätten. Ein Winkel ist die Passbandgrenze, bis zu der das Signal nicht verändert wird. Ein zweiter Winkel ist die Stopbandgrenze, ab der ein Signal unterdrückt wird. Zwischen den beiden Grenzen wird interpoliert.

$$\Delta\theta = 2\pi \cdot fd \cdot \sin\phi/c \tag{2.3}$$

Dadurch wird die Frequenzabhängigkeit der Hauptkeule kompensiert und der Beamformer ist für alle Frequenzen annähernd gleich. An den Randbereichen der Frequenz kommt es bei tiefen Frequenzen jedoch zu sehr starker Rauschsensitivität und bei ho-

hen Frequenzen tritt Aliasing auf, da die Phasendifferenz π überschreitet, wieder bei 0 anfängt und damit unter die maximale Phasendifferenz $\Delta\theta$ fällt.

2.2 Störunterdrückung G_f

Dieser Teil der Filterung ist vom Signal abhängig und ändert sich adaptiv. In [5] wird die Matrix, durch die die Störunterdrückung verwirklicht als spectral gain (G_f, dt. spektraler Gewinn) bezeichnet. Darin spiegelt sich wieder, dass je Frequenzbin geschätzt wird, ob eine Störung oder das Nutzsignal dominiert. Entsprechend wird dann ein gestörtes Frequenzbin unterdrückt. Die Richtung aus der die Störung kommt spielt dabei keine Rolle, weshalb die Algorithmen zur Errechnung des spektralen Gewinns einkanalig sind. In der Praxis käme hier ein einkanaliger Störunterdrücker zum Einsatz, zum Beispiel einer der in [5] vorgeschlagen Algorithmen aus [1, 3, 4, 6]. In diesem Forschungspraktikum wurde zur Vereinfachung der Algorithmus aus [2] implementiert. Dabei wird jedoch nicht mit einer Schätzung der Störung gearbeitet sondern mit der Originalstörung, die aufgrund des modellierten Versuchsaufbaus bekannt ist. Dadurch können weitere Untersuchungen unter der Annahme gemacht werden, dass G_f seine Funktion erfüllt und nicht der Ursprung von Fehlern ist.

2.3 Richtungsgewinn G_d

Dieser Teil des Algorithmus ist nicht vom Signal direkt abhängig, sondern von den Phasendifferenzen der DTFs der Eingangssignale. Diese Differenzen werden mit der Entscheidungsgrenze (2.3) verglichen und entsprechend unterdrückt oder nicht. Als Ergebnis entsteht eine Matrix, die als directioanl gain (G_d, dt. Richtungsgewinn) bezeichnet wird. Speziell G_d stellt die Neuerung im Paper [5] da und wird deshalb im weiteren Verlauf des Forschungspraktikums untersucht.

3 Hauptteil

3.1 Akustisches Model

Für den Test des implementieren Codes wird eine reale Situation verwendet. Allerdings werden Roboter und Störquellen nicht extra aufgebaut, sondern es wird auf eine gespeicherte Messung zurückgegriffen. Die Anzahl, Art und Position der Störer kann variiert werden. Dabei sind die Störquellen auf einem Kreis mit dem Radius $1,1\,m$ um den Roboter im $5\,^\circ$ Abstand aufgestellt. Zusätzlich befinden sich die Mikrofone im Kopf des Roboters, der diese um $0,74\,m$ erhöht. Die Geometrie der einzelnen Mikrofone im Roboterkopf ist aus dem vom Lehrstuhl erstellten MATLAB-Code zu entnehmen. Es wird mit fünf Mikrofonen gearbeitet. Dabei ist das dritte Mikrofon in der Mitte des Roboterkopfes und die anderen Mikrofone sind zu diesem dritten Mikrofon symmetrisch angeordnet. Sowohl das gewünschte Signal, als auch die Störung sind Sprache. Die Störung wird bei allen Messungen bei dem Winkel $55\,^\circ$ platziert. Das Nutzsignal liegt bei $0\,^\circ$ genau auf der mittleren Achse der Mikrofonanordnung. Die Raumimpulsantwort hat eine Länge von $190\,ms$.

3.2 Erläuterungen zum MATLAB-Code

Der Algorithmus in [5] beginnt mit den DFTs der Eingangssignale. Diese werden durch den vom Lehrstuhle zur Verfügung gestellten Code bereits erzeugt. Der in diesem Praktikum erstellte Code beginnt quasi nach den DFT Blöcken in Abbildung 2.1. Zuerst wird die Formel (2.1) implementiert. Um das Ergebnis der Implementierung testen

zu können, wird die Formel auch auf das reine Nutzsignal und die reine Störung angewandt. Als nächster Schritt wird G_f errechnet. Dabei wird ein Frequenzbin unterdrückt, wenn das SNR kleiner als $\gamma = 7$ ist. Danach werden die Phasendifferenzen aller Mikrofonkombinationen berechnet. Bevor G_d erstellt werden kann, werden die Mikrofonabstände und das mathematische Modell der Anordnung implementiert. Woraus die in 3.4.2 vorgestellte natürliche Phase errechnet wird. Dann werden Pass- und Stopbandgrenze initialisiert und für jede Mikrofonkombination G_d mit der Funktion $Gdir()$ errechnet. Danach wird aus den errechneten Richtungsgewinnen ein optimaler Richtungsgewinn zusammengestellt, dieser wird im Code G_{hybrid} genannt. Die auskommentierte Funktion $bestG()$ dient der schnellen Auswertungen der PESQ-Gewinne für alle Richtungsgewinne. Schließlich wird das Ausgangsspektrum berechnet und mit der Phase des Sumbeamformers versehen. Zum Schluss folgen die Plots und das Ergebnis wird evaluiert.

3.3 Einflussreiche Parameter

3.3.1 Entscheidungsgrenze des Richtungsgewinns

Die Entscheidung, ob ein Frequenzbin unterdrückt wird, hängt davon ab, ob die Phasendifferenz zweier DFT Werte höher als die Entscheidungsgrenze $\Delta\theta$ aus (2.3) ist. Dabei ist die Grenze in zwei Stufen aufgeteilt. Überschreitet die Phasendifferenz die Passbandgrenze wird das Frequenzbin abgeschwächt. Wird die Stopbandgrenze überschritten, führt dies zur vollständigen Unterdrückung des Frequenzbins. Diese beiden Grenzen werden in Abbildung 3.1 veranschaulicht. Zur Berechnung wurde $d = 7\,cm$ gewählt. Die Winkel für den Pass- und Stopbandbereich wurde bei 14° und 18° gewählt.

3.3.2 Unterschiedliche Mikrofonabstände

Um den Zusammenhang zwischen dem Mikrofonabstand, der Empfangsbereichsbreite und dem Aliasing zu verdeutlichen wurden in Abbildung A.1 mit verschiedene Mi-

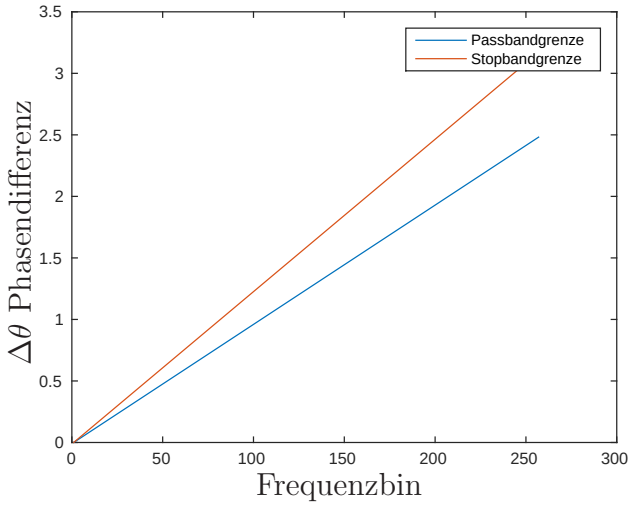

Abbildung 3.1: Pass- und Stopbandgrenze für Frequenzbins nach (2.3)

krofonabstände die Entscheidungsgrenzen und die Ankunftswinkel über der Frequenz dargestellt. Dabei entsteht viel Aliasing. Da die Mikrofonabstände nicht variiert werden können, wurde in Abbildung A.2 die Empfangsbereichsbreite verkleinert mit dem Ergebnis, dass weniger Aliasing auftritt. Zusammenfassend lässt sich sagen, dass der Abstand der Mikrofone beeinflusst, ab welchen Frequenzbin Aliasing beginnt. Die Empfangsbereichsbreite beeinflusst, wie viel Aliasing auftritt. Die im Roboter NAO verbauten Mikrofone haben einen relativ großen Abstand, so dass viel Aliasing auftritt. Beispielsweise wurde in [5] mit einem Mikrofonabstand von $4,5\,cm$ gemessen, der größte Abstand beim NAO liegt zwischen dem ersten und fünften Mikrofon bei $15\,cm$.

3.4 Evaluation

3.4.1 Antialiasing

Bei hohen Frequenzen kommt es durch den Abstand der Mikrofone zu Phasendifferenzen größer als π, wodurch die in (2.3) berechneten Entscheidungsgrenzen unterschritten werden. Zwei einfache Versuche Aliasing zu unterdrücken wurden untersucht. Sie sind in Abbildung A.3 dargestellt. Im ersten Versuch wurden die Grenzen ab dem Frequenzbin

193 zu einem Grenzband, das heißt, es wurde eine mindest-Phasendifferenz gefordert. Dadurch wird der Empfangsbereich von der Mitte her aufgespalten. Im zweiten Versuch wurden ab den Frequenzbin 193 die Pass- und Stopbandgrenze halbiert. Dadurch halbiert sich der Empfangsbereich. Da die Frequenzbins über 193 kaum relevante Informationen enthalten kann nicht gesagt werden, ob die Vorschläge den PESQ-Gewinn dauerhaft erhöhen können.

3.4.2 Natürliche Phasendrehung

Zur Auswertung wurden alle Mikrofonkombinationen verwendet um Richtungsgewinne zu erstellen. Mit jedem dieser Richtungsgewinne wurde der PESQ-Gewinn errechnet. Er ist in Tabelle 3.1 dargestellt. Dabei zeigt sich die größte Verbesserung, wenn der Richtungsgewinn aus dem zweiten und vierten Mikrofon errechnet wurde. Für die anderen Mikrofonkombinationen zeigen sich teilweise, wie zwischen dem ersten und dritten Mikrofon, nur sehr kleine PESQ-Gewinne, die sogar unter dem Niveau der Matrixdiagonalen liegen. Die Diagonale wird aus nur jeweils einem Mikrofon errechnet, so dass $G_\mathrm{d} = 1$ ist und der Gewinn allein durch $\mathrm{G_f}$ zustande kommt. Das heißt, nur wenn der PESQ-Gewinn in Tabelle 3.1 größer als $0,152$ ist, hat es sich gelohnt $\mathrm{G_d}$ zu verwenden. Der große Unterschied ist zunächst nicht erklärbar. Da die Geometrie der Mikrofonan-

0,152	0,118	0,077	0,092	0,168
0,118	0,152	0,161	0,213	0,122
0,077	0,161	0,152	0,155	0,132
0,092	0,213	0,155	0,152	0,133
0,168	0,122	0,132	0,133	0,152

Tabelle 3.1: PESQ-Gewinnmatrix; Zeile und Spalte als Indizes der Mikrofone

ordnung jedoch dreidimensional komplizierter ist, als bei einem linearen Mikrofonarray wurden ein mathematisches Modell des Aufbaus erzeugt, das bei der Phasenentscheidung nach Formel (2.3) einen sogenannten natürlichen Winkel ϕ_nat hinzuaddiert (3.1).

$$\Delta\theta = 2\pi \cdot fd \cdot \sin\phi/c + \phi_{\text{nat}} \tag{3.1}$$

Damit wird die räumliche Anordnung der Mikrofone kompensiert, wie in Tabelle 3.2 zu sehen ist. Es verschlechtert sich lediglich der Wert, der aus dem dritten und vierten Mikrofon gewonnen wurde. Die symmetrischen Mikrofone zwei, vier und eins, fünf verändern sich nicht, da die Symmetrie zu keiner natürliche Phasendrehung führt. Interessant ist außerdem der starke Unterschied zwischen dem Mikrofonpaaren zwei und drei sowie drei und vier. Testweise wurden alle berechneten Gewinne gemittelt und es ergab sich ein PESQ-Gewinn von $0,158$, der nur knapp über dem Wert der Diagonalen liegt. Der beste erzeuge PESQ-Gewinn ohne Veränderung des Algorithmus beträgt $0,213$ und wurde zwischen dem Mikrofonpaar 2 und 4 errechnet.

0,152	0,134	0,138	0,137	0,168
0,134	0,152	0,201	0,213	0,142
0,138	0,201	0,152	0,138	0,140
0,137	0,213	0,138	0,152	0,138
0,168	0,142	0,140	0,138	0,152

Tabelle 3.2: PESQ-Gewinnmatrix mit Phasenkorrektur; Zeile und Spalte als Indizes der Mikrofone

3.4.3 Wirkung unterschiedlicher Frequenzbereiche

Im tiefen Frequenzbereich zwischen dem Frequenzbin 1-15 tritt noch kein Aliasing auf, deshalb kann auf Gewinne zurückgegriffen werden die aus entfernten Mikrofonen errechnet wurden. Eine Kombination aus den Richtungsgewinnen vom ersten und fünften sowie dem zweiten und vierten Mikrofon ist optimal. In Verbindung mit dem Faktor $0,7$ kann der PESQ-Gewinn in (3.2) um $0,005$ gesteigert werden. Es handelt sich um eine Mittlung, wie sie in [5] vorgeschlagen wird.

$$G_{\text{d}}[1:15] = (G_{\text{d,1-5}}[1:15] + G_{\text{d,1-5}}[1:15])/2 * 0,7 \tag{3.2}$$

Die mittleren Frequenzen von Frequenzbin 15-49 haben hohen Einfluss auf den PESQ-Gewinn. Beispielsweise steigt er von $0,234$ auf $0,68$ nur durch eine Multiplikation mit dem Faktor 10.
Werden die hohen Frequenzbins 50-257 des Gewinns gleich $0,57$ gesetzt, verschlechtert sich der PESQ-Gewinn um lediglich $0,04$. Die hohen Frequenzbins im Gewinn enthalten folglich sehr wenig nützliche Informationen. Die hohen Frequenzbins werden optimaler Weise durch eine Mittlung der Richtungsgewinne aus den Mikrofonen zwei, vier und zwei,drei erzeugt.

3.4.4 Berechnete Gewinne

Der Gewinn durch die einkanalige Rauschunterdrückung G_f wird in Abbildung 3.2 gezeigt. Dabei steht die Farbe gelb für 1 und das blau für $0,1$. Allein durch die Berechnung würde 0 der niedrigste Wert von G_f sein und nicht $0,1$. Experimentell zeigt sich jedoch, dass eine komplette Unterdrückung zu schlechteren Ergebnissen führt. Das rote Audiosignal in Abbildung 3.2 stellt das gewollte Signal da, während das schwarze Audiosignal die Störung ist. Dominiert das rote Audiosignal ist das SNR hoch und in G_f dominiert die Farbe gelb.
Der, wie in 3.4.3 modifizierte, Richtungsgewinn wird in Abbildung 3.3 dargestellt.
Der gesamte Gewinn wird in Abbildung 3.4 dargestellt. Er resultiert aus der Multiplikation von G_d und G_f. Der damit erreichte PESQ-Gewinn liegt bei $0,685$.

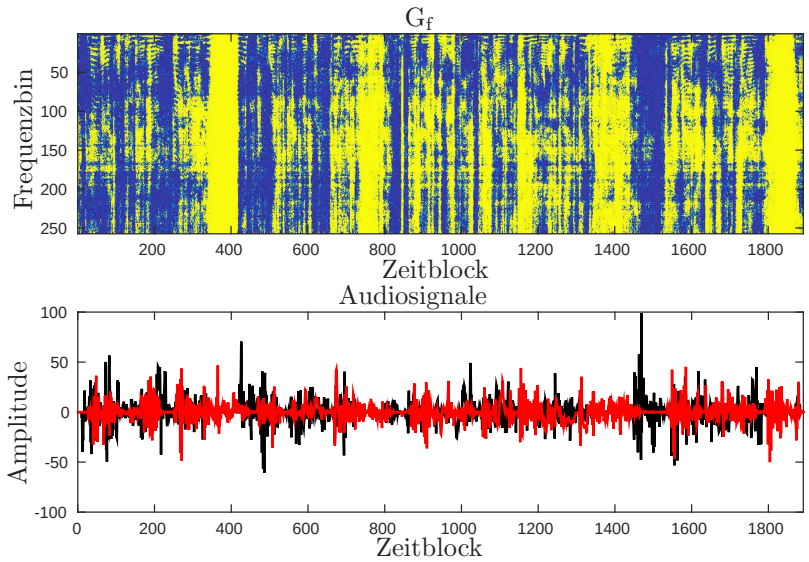

Abbildung 3.2: Audiosignale und resultierendes G_f gegenübergestellt; gelb = 1; blau = 0, 1; rot = Nutzsignal; schwarz = Störung

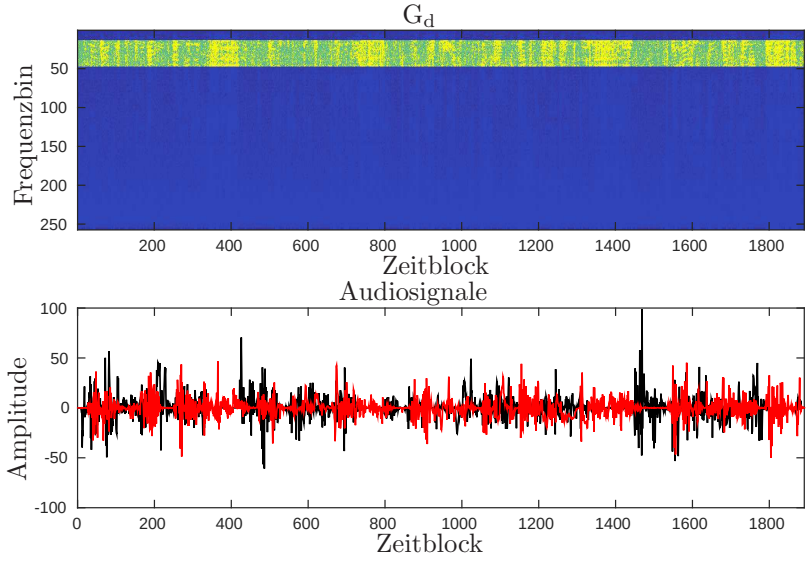

Abbildung 3.3: Audiosignale und resultierendes modifiziertes G_d gegenübergestellt; gelb = 1; blau = 0, 1; rot = Nutzsignal; schwarz = Störung

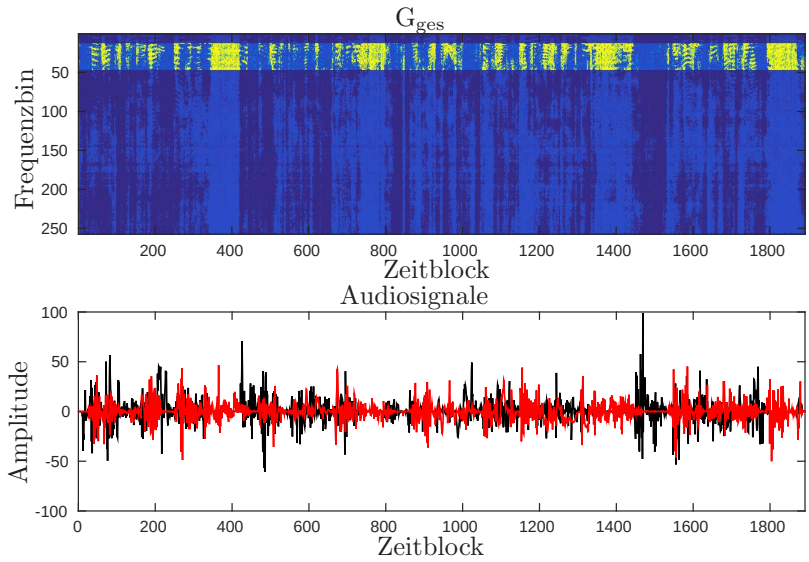

Abbildung 3.4: Audiosignale und resultierendes modifiziertes G_{ges} gegenübergestellt; gelb = 1; blau = 0,1; rot = Nutzsignal; schwarz = Störung

4 Zusammenfassung

In diesem Forschungspraktikum wurde der Beamformer aus [5] am Beispiel des humanoiden Roboters NAO implementiert und getestet. Es ergab sich ein PESQ-Gewinn. Der Aufbau ist deshalb geeignet die Spracherkennung zu verbessern. Anhand theoretischer Überlegungen wurde gezeigt, dass die Mikrofonabstände im Roboter zu groß sind um aus allen Mikrofonkombinationen einen funktionierenden Richtungsbeamformer zu bilden, da sonst zu viel Aliasing auftritt. Dafür wurden für verschiedene Parameter Plots erzeugt. Auch das Mitteln über alle Mikrofonkombinationen führte zu keine Verbesserung. Durch Kombination der symmetrischen Mikrofone und Gewichtung einzelner Frequenzbereiche konnte der PESQ-Gewinn auf bis zu $0,685$ erhöht werden.

Anhang

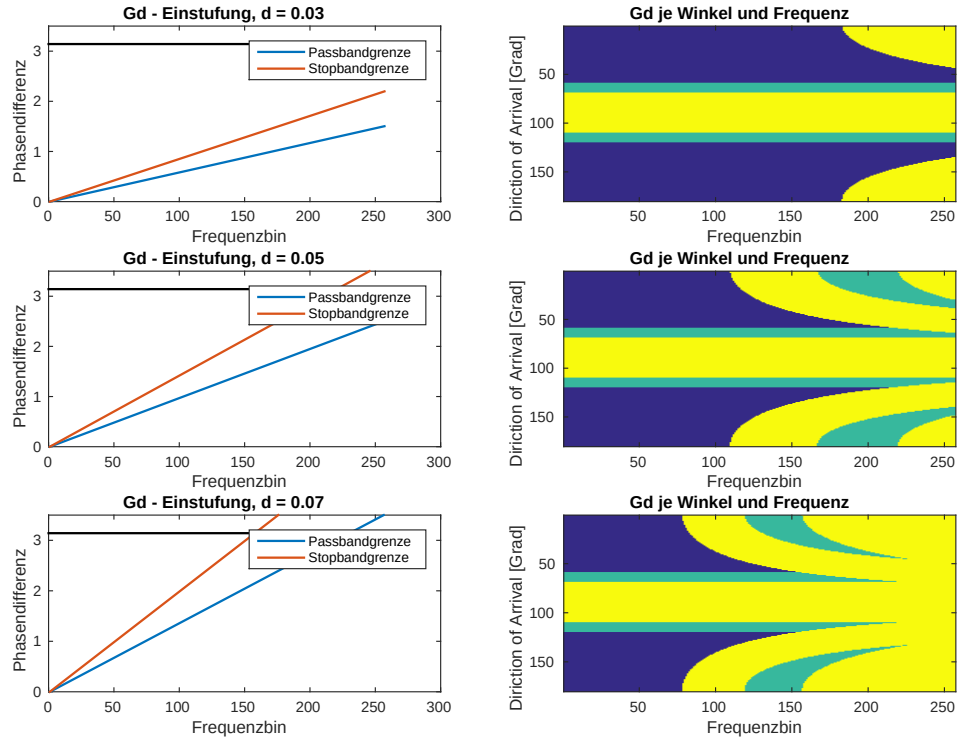

Abbildung A.1: Aliasing bei unterschiedlichen Mikrofonabständen; Passbandgrenze = 20 °; Stopbandgrenze = 30 °; schwarze Linie = π; blau = Unterdrückung; grün = teilweise Unterdrückung; gelb = Durchlassbereich

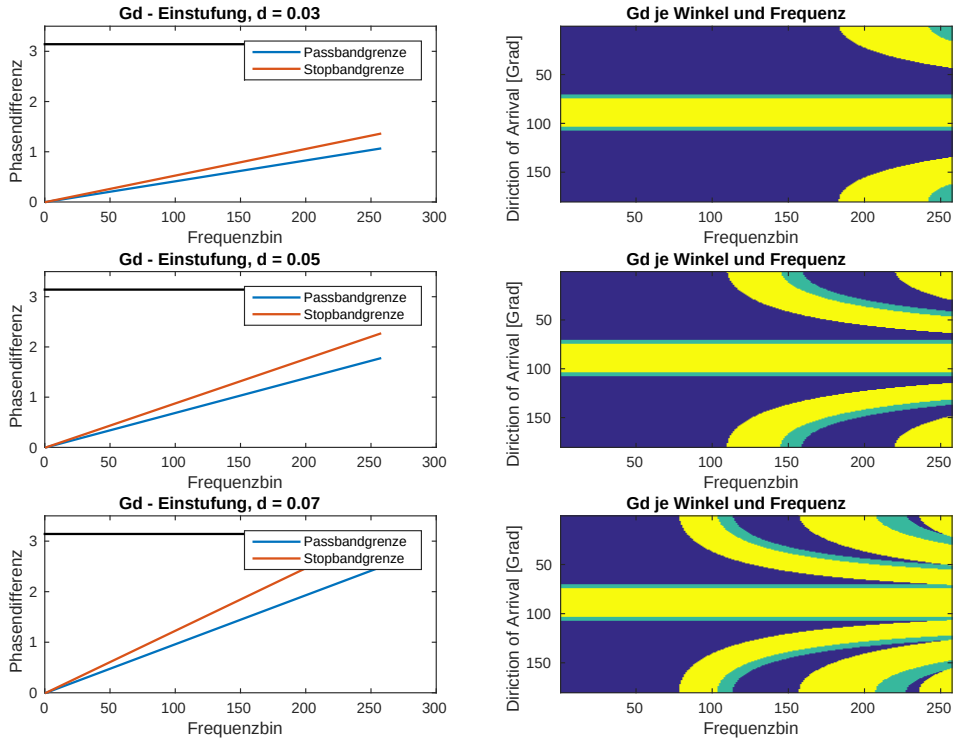

Abbildung A.2: Aliasing bei unterschiedlichen Mikrofonabständen; Passbandgrenze = 14 °; Stopbandgrenze = 18 °; schwarze Linie = π; blau = Unterdrückung; grün = teilweise Unterdrückung; gelb = Durchlassbereich

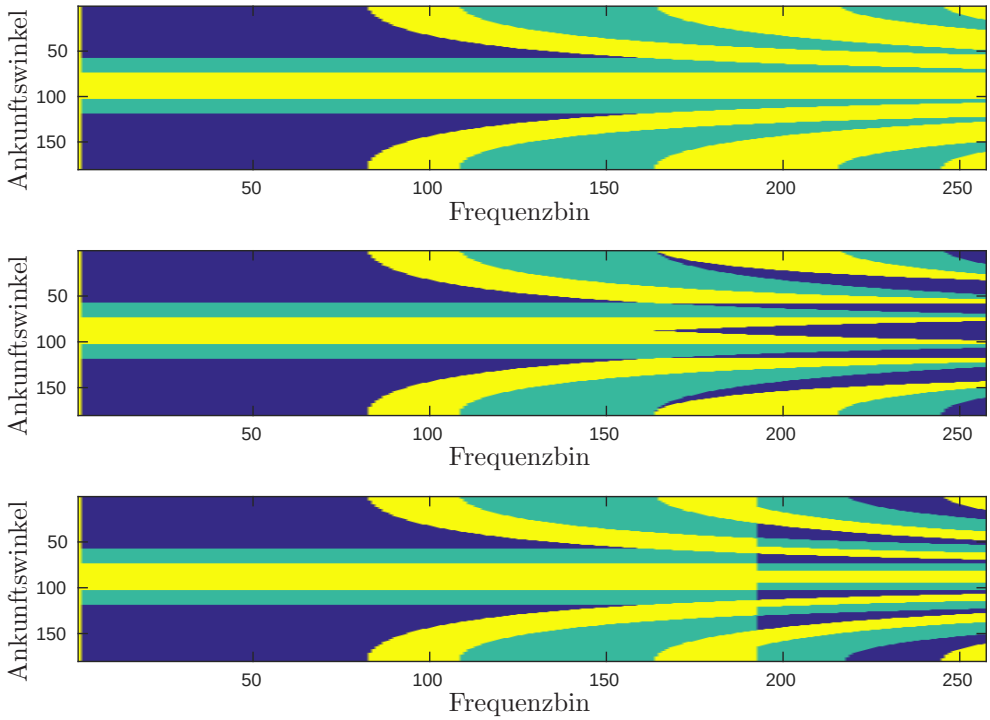

Abbildung A.3: Versuche Aliasing zu unterdrücken; blau = Unterdrückung; grün = teilweise Unterdrückung; gelb = Durchlassbereich

Literaturverzeichnis

[1] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2):113–120, Apr 1979.

[2] EricJ. Diethorn. Subband noise reduction methods for speech enhancement. In Yiteng Huang and Jacob Benesty, editors, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, pages 91–115. Springer US, 2004.

[3] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6):1109–1121, Dec 1984.

[4] Masanori Kato, Akihiko Sugiyama, and Masahiro Serizawa. Noise suppression with high speech quality based on weighted noise estimation and mmse stsa. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 89(2):43–53, 2006.

[5] A. Sugiyama and R. Miyahara. A directional noise suppressor with a specified beamwidth. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 524–528, April 2015.

[6] J. Taghia, J. Taghia, N. Mohammadiha, Jinqiu Sang, V. Bouse, and R. Martin. An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4640–4643, May 2011.