MEHRKANALIGES AKUSTISCHES FRONT-END FÜR SPRACHERKENNUNGSSYSTEME

Thomas Fehér, Rico Petrick, Rüdiger Hoffmann

Institut für Akustik und Sprachkommunikation, TU Dresden
thomas.feher@ias.et.tu-dresden.de

Kurzfassung: Mehrmikrofonanordnungen erweitern die Möglichkeiten der Vorverarbeitung von Audiosignalen für die Spracherkennung. Sie ermöglichen die Nutzung von Beamforming-Algorithmen, mit deren Hilfe Störgeräusche und Raumeinflüsse reduziert werden können. In diesem Beitrag werden verschiedene Beamformer vorgestellt und deren Vor- und Nachteile für die Spracherkennung diskutiert. Es wurden Messungen in praktisch relevanten Umgebungen vorgenommen, deren Auswirkung auf die Erkennungsrate hier gezeigt werden. Des Weiteren erfolgt die Vorstellung eines aktuellen Projektes des Institut für Akustik und Sprachkommunikation der TU Dresden mit mehreren Partnern zur Entwicklung eines public Terminals mit robuster Spracherkennung.

1 Einleitung

Spracherkennung unter realen akustischen Bedingungen ist derzeit noch ein ungelöstes Problem. Im Unterschied zu Laborbedingungen treten in realen Umgebungen die beiden Störarten Hall und Nebengeräusche auf. Die Behandlung von Störgeräuschen wird von der Wissenschaft seit ca. 30 Jahren intensiv untersucht [1]. Dieser Beitrag befasst sich mit dem noch jungen Forschungsgebiet der Unterdrückung raumakustischer Einflüsse.

Trotz hervorragender Ergebnisse der Klassifikatoren bei ungestörter Sprache (Erkennungsraten > 90 %) sinkt beispielsweise die Erkennungsrate in normalen Wohnräumen (Nachhallzeiten bei ca. 0,5 s) und einem Abstand von 2 m zwischen Sprecher und Mikrofon auf etwa 60 % (Abbildung 5 bzw. [2]).

Für die Spracherkennung existiert eine klassische Einteilung von Maßnahmen, die sich mit der Behandlung von akustischen Störungen befassen (sowohl Störgeräusch als auch Hall). Diese betrifft Maßnahmen auf Signalebene, auf Merkmalebene sowie auf Modellebene.

In diesem Artikel werden zwei relativ einfache Algorithmen zur mehrkanaligen Signalverarbeitung, mit Rücksicht auf ein in Echtzeit arbeitendes Front-End, untersucht: der Delay-and-Sum-Beamformer (DSB) und der Minimum-Variance-Distortionless-Response-Beamformer (MVDRB). Beide zählen zu den Maßnahmen auf Signalebene und wurden bislang vorwiegend für Geräuschstörungen untersucht. In der vorliegenden Arbeit werden beide Maßnahmen in Verbindung mit Hallstörungen umfangreich getestet. Derartige Untersuchungen sind in der Literatur vergleichsweise neu.

Die Hallstörungen wurden aus typischen akustischen Umgebungen im Wohn- und Büroumfeld ausgewählt. Dabei wurden Messungen bei Nachhallzeiten von 0,25 s, 0,45 s, 0,7 s und 0,9 s bei Sprecher-Mikrofon-Abständen (SMA) von 0,02 m bis 4 m durchgeführt. Die beiden untersuchten Algorithmen werden auf verhallte Sprache aus diesen Umgebungen angewandt und die Signalverbesserung wird mit einem Spracherkenner getestet. Es wird eine Verbesserung der Erkennungsraten erwartet, wobei der MVDRB deutlich bessere Ergebnisse liefern sollte, da dieser eine wesentlicher höhere Bündelung bei tiefen Frequenzen erreicht (Abbildung 4 - links).

2 Projekt public Terminals

Alle hier beschriebenen Versuche wurden im Rahmen des Projektes "Multimodales, personalisiertes Bedienkonzept für public Terminals" durchgeführt. Ziel dieses Projektes mit Partnern aus Forschung und Industrie ist ein Prototyp eines public Terminals, welcher per Sprache, Tastatur/Touchscreen oder mobilem Eingabegerät (Handy, PDA, etc.) bedient werden kann. Dieser Artikel geht auf Teilaspekte der Entwicklung des Front-Ends ein, der Robustheitssteigerung der Spracherkennung durch Mikrofonarrayalgorithmen.

Dieses Terminal kann von mehreren Personen gleichzeitig genutzt werden, da die Bedienung mit den mobilen Endgeräten per Bluetooth erfolgt. Es ist also möglich sich mit dem Mobiltelefon in die Nähe des Terminals zu stellen, und es genau so zu bedienen als würde man direkt mit dem Terminal arbeiten.

Die Bedienung per Spracheingabe direkt am Terminal soll möglichst nicht zu Lasten der Privatsphäre gehen. Das bedeutet für das Front-End: eine gute Spracherkennung zusammen mit einem guten Front-End (Mikrofonarray), wodurch die bedienende Person nicht unnötig laut sprechen muss, und eine gerichtete Schallabstrahlung für die Sprachsynthese, welche über mehrere Lautsprecher erfolgt und ebenfalls mit Hilfe des Beamforming-Prinzips gesteuert wird. Dadurch soll bei weiteren Personen, die sich in der Nähe des Terminals aufhalten, einen möglichst geringen Schalldruckpegel auftreten.

3 Algorithmen

3.1 Delay-and-Sum-Beamformer (DSB)

Der DSB stellt die einfachste Art der mehrkanaligen Verarbeitung dar. Alle Mikrofonsignale werden addiert, wodurch sich eine Richtwirkung einstellt, die von dem Verhältnis Wellenlänge zu Mikrofonabstand und der Anzahl an Mikrofonen abhängt. Bei der hier untersuchten Variante der Querzeile (broadside) sind Verzögerungsglieder nur für Schallquellen im Nahfeld des Arrays notwendig. Die Versuche wurden ohne Verzögerungsglieder durchgeführt, obwohl ein Teil der SMAs durchaus im Nahfeld liegt (siehe Auswertung).

Mehrmikrofonanordnungen realisieren eine räumliche Abtastung des Schallfeldes. Dadurch gelten prinzipiell die Gesetzmäßigkeiten des Abtasttheorems. Abbildung 1 zeigt das Richtdiagramm von 2 omnidirektionalen Mikrofonen, deren Signale einfach addiert werden, bei verschiedenen Wellenlänge- zu Abstand-Verhältnissen. Dabei werden die wichtigsten Einschränkungen des Beamformings sichtbar:

- Geringe Bündelung bei tiefen Frequenzen
- Aliasing bei hohen Frequenzen (Nebenkeulen)

In Abbildung 2 - links wird das Bündelungsmaß über die Frequenz für Anordnungen von 2, 4 und 8 Mikrofonen dargestellt.

Deutlich zu sehen ist die Erhöhung des Bündelungsmaßes um 3 dB bei einer Verdoppelung der Mikrofonanzahl für hohe Frequenzen. Für tiefe Frequenzen muss die Gesamtgröße des Arrays erhöht werden, wodurch bei Anwendungen für Sprache sehr große Maße erreicht werden. Daher gilt es immer einen Kompromiss zwischen Handhabbarkeit und Richtwirkung zu finden.

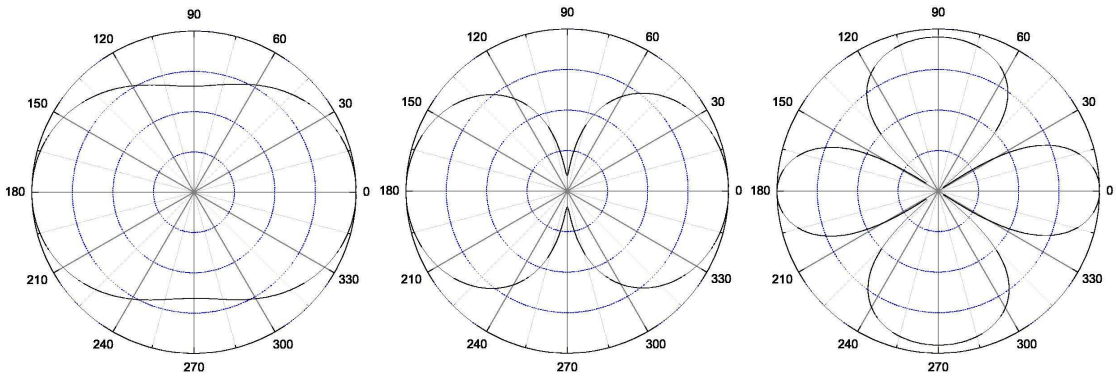

Abbildung 1 – Richtdiagramm für ein Array aus 2 omnidirektionalen Mikrofonen. *Links*: Wellenlänge groß gegenüber Mikrofonabstand (geringe Wirkung). *Mitte*: Wellenlänge gleich doppelter Mikrofonabstand (Idealfall). *Rechts*: Wellenlänge klein gegenüber Mikrofonabstand (räumliches Aliasing, erkennbar an den Nebenkeulen).

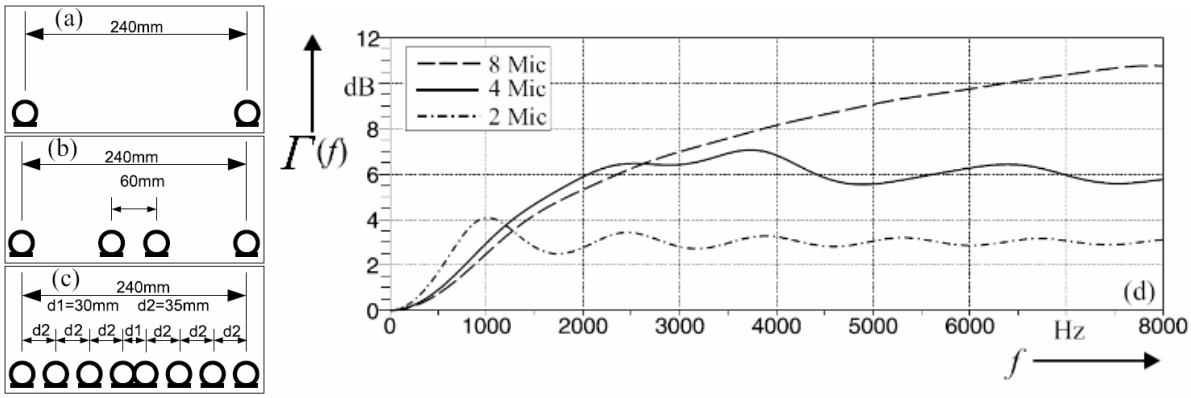

Abbildung 2 – (a) – (c): geometrische Anordnung des Arrays für 2, 4 und 8 Mikrofone. (d) Verlauf des Bündelungsmaßes eines DSB in Abhängigkeit der Frequenz und für verschiedene Anordnungen.

3.2 Minimum-Variance-Distortionless-Response-Beamformer (MVDRB)

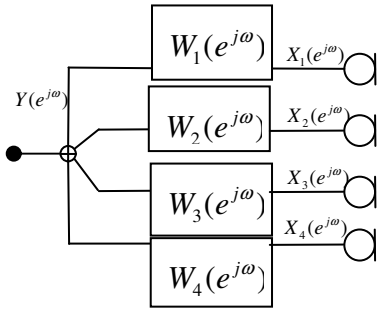

Abbildung 3 – Schematischer Aufbau des MVDRB.

Der MVDRB ist eine modifizierte Version des DSB. Das Verfahren wird in [3] ausführlich dargestellt. Die hier benutzten Notationen sowie die folgende Herleitung im Frequenzbereich beziehen sich ebenfalls auf die Ausführungen in [3]. Ausgangspunkt ist die Gleichung für das Ausgangssignal eines DSB mit zusätzlichen Wichtungsfaktoren W_i, für jeden Frequenzpunkt (Abbildung 3).

$$Y = \mathbf{W}^H \mathbf{X} \tag{1}$$

Das Ausgangssignal Y ergibt sich aus der Multiplikation des Vektors der Wichtungsfaktoren mit dem Vektor der Eingangssignale für jeden Frequenzpunkt $e^{j\omega}$ (die Abhängigkeit von $e^{j\omega}$ wird in den Gleichungen zur besseren Übersichtlichkeit nicht aufgeschrieben). Das hochgestellte H ist der Operator für die hermetisch transponierte Matrix.

Unter der Annahme, dass die einfache Addition nicht das optimale Ausgangssignal Y liefert und unter der Voraussetzung, dass der Beamformer richtig auf die Quelle ausgerichtet ist, ergibt sich folgende Optimierungsaufgabe: Die Varianz des Ausgangssignals Y soll minimiert werden unter der Bedingung, dass die Varianz des Nutzsignals gleich bleibt bzw. 1 wird.

$$\min_{\mathbf{W}}(\mathbf{W}^H \mathbf{\Phi}_{xx} \mathbf{W}) \text{ unter der Bedingung } \mathbf{W}^H \mathbf{d} = \mathbf{1} \tag{2}$$

$\mathbf{\Phi}_{xx}$ ist die Matrix der Leistungsdichtespektren (und Kreuzleistungsdichtespektren) der Eingangssignale ($\mathbf{\Phi}_{xx} = \mathbf{X} \cdot \mathbf{X}^H$). Für eine Quelle im Fernfeld und äquidistanten Mikrofonen ergibt sich $\mathbf{d}$, der Vektor der relativen Phasenlagen des Nutzsignals an den Mikrofonen, mit

$$\mathbf{d}^T = [1, \exp(-j\omega d_1 / c), \exp(-j\omega d_2 / c), \ldots] \tag{3}$$

In diesem Fall ist der Bezugspunkt das erste Mikrofon und d_i ist der Abstand von Mikrofon i zum Bezugspunkt, c entspricht der Schallgeschwindigkeit. Weil das Eingangssignal nicht verändert werden soll, kann $\mathbf{\Phi}_{xx}$ durch die Matrix der Leistungsdichtespektren der Störgeräusche $\mathbf{\Phi}_{nn}$ ersetzt werden.

Die Lösung für diese Aufgabe führt zu folgender Gleichung für den Vektor der komplexen Wichtungsfaktoren:

$$\mathbf{W} = \frac{\mathbf{\Phi}_{nn}^{-1} \mathbf{d}}{\mathbf{d}^H \mathbf{\Phi}_{nn}^{-1} \mathbf{d}} \tag{4}$$

Den Effekt dieses Algorithmus auf die Richtwirkung des Arrays zeigt Abbildung 4 - links.

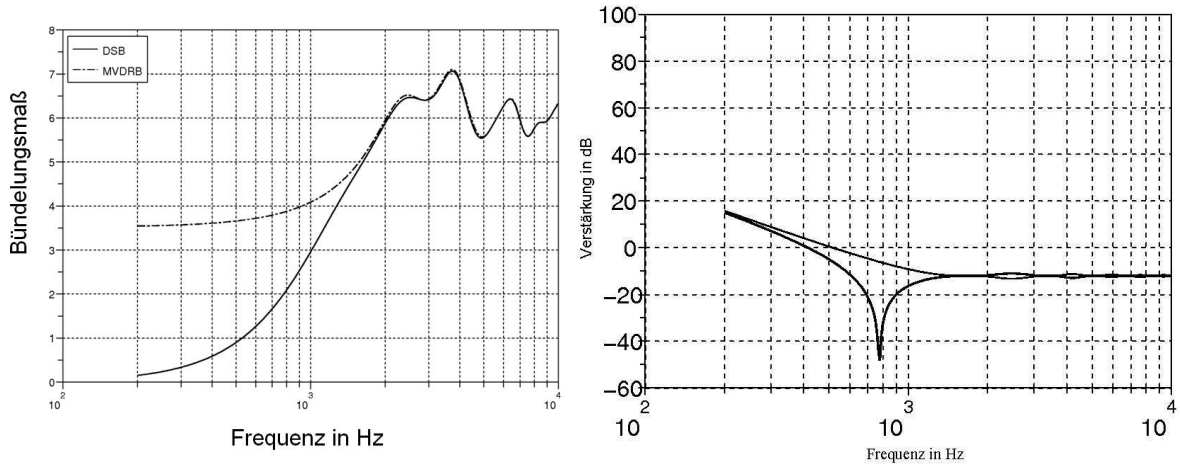

Abbildung 4 – *Links*: Verlauf des Bündelungsmaßes des DSB und MVDRB über die Frequenz. *Rechts*: Verlauf des Realteils von **W** (Verstärkung der Mikrofonsignale) über die Frequenz.

Sehr gut zu sehen ist das weit höhere Bündelungsmaß bei tiefen Frequenzen. Ab ca. 2 kHz sind die Verläufe des Bündelungsmaßes identisch. Auch die Wichtungsfaktoren sind denen des DSB gleich. In diesem Frequenzbereich scheint die einfache Addition optimal zu sein.

Der Nachteil, der dadurch entsteht, ist die Verstärkung des unkorrelierten Rauschens. Wie in Abbildung 4 - rechts zu sehen, erhöht der Algorithmus die Verstärkung (im Übrigen auch die Phasendrehung) wodurch das unkorrelierte Rauschen der Mikrofone angehoben wird. Eine Erhöhung um knapp 20 dB bei 200 Hz ist allerdings recht unproblematisch. Bei Arrays mit mehr Mikrofonen steigt der Verstärkungsfaktor jedoch wesentlich stärker (≈ 80 dB bei 200 Hz für 8 Mikrofone).

Bei dieser Anordnung muss demnach ein Kompromiss zwischen unkorreliertem Rauschen und Richtwirkung gefunden werden. Das ist wesentlich weniger problematisch als bei dem DSB, da die heutige Technologie äußerst rauscharme Mikrofone zur Verfügung stellt.

4 Experimente

4.1 Experimentierumgebung

4.1.1 Spracherkenner

Zur Untersuchung der Wirksamkeit der Algorithmen wurden Experimente mit einem Spracherkenner durchgeführt. Als Testkorpus wurden 1020 Phrasen (jeweils ca. 2 Sekunden Sprache) aus dem Apollo-Korpus [4] verwendet. Bei dem Erkenner handelt sich um den phonembasierten HMM-Erkenner des UASR Systems [5].

4.1.2 Räume

Die Räume wurden nach ihren akustischen Eigenschaften ausgewählt. Der erste Raum ist ein Tonstudio mit einer Nachhallzeit von 0,25 Sekunden. Der 2. Raum entspricht einem gewöhnlichen Wohnraum mit einer Nachhallzeit von 0,45 Sekunden. Raum 3 ist ein Büroraum mit 0,7 Sekunden Nachhallzeit und Raum 4 ist ein Laborraum mit einer Nachhallzeit von 0,95 Sekunden.

Damit sind die wichtigsten akustische Umgebungen im Wohn- und Bürobereich einbezogen. Andersartige Räume wie große Hallen, Kirchen, etc sind in die Betrachtungen nicht mit eingeflossen.

4.1.3 Mikrofonarray

Für die Experimente wurde das Mikrofonarray aus Abbildung 2 (b) ausgewählt. Die Geometrie wurde bewusst nicht gleichmäßig gewählt um Nebenkeulen zu verhindern [6]. Zum Einsatz kamen unidirektionale Kondensatormikrofone (Richtcharakteristik: Niere). Um eine gerichtete Aufnahme mit diesem Array zu simulieren, wurden Raumimpulsantworten (engl.: room impulse response – RIR) zwischen einer Schallquelle (Lautsprecher) und allen 4 Mikrofonen simultan für verschiedene SMAs gemessen. Dabei befand sich das Array am gleichen Ort, bewegt wurde nur der Lautsprecher zur Änderung des SMAs in Schritten von 20 cm, beginnend bei SMA = 0,2 m. Der maximale SMA beträgt 4 m. Die Messung erfolgte mit einer Sweepmessmethode. Mit den gemessenen RIRs werden durch Faltung mit dem Evaluationskorpus verhallte Sprachdaten für die betreffenden SMAs generiert. Die entstehenden vier Kanäle pro Position werden durch die beiden Mikrofonarrayalgorithmen je zu einem gerichteten Signal verarbeitet, das dem Spracherkenner zugeführt wird.

4.2 Ergebnisse

Abbildung 5 zeigt die Ergebnisse der Erkennungsexperimente. Zum Vergleich sind die Erkennungsraten für nur ein Mikrofon dargestellt. Beide Beamformer sind bei allen

Entfernungen und Räumen auf eine Quelle im Fernfeld (Entfernung zur Quelle wesentlich größer als Arrayabmessungen) eingestellt. Der MVDRB, als adaptiver Beamformer, wurde hier auf ein ideal diffuses Schallfeld optimiert und mit dieser festen Einstellung benutzt. Er arbeitet damit nicht adaptiv, dafür ist die benötigte Rechenleistung wesentlich geringer.

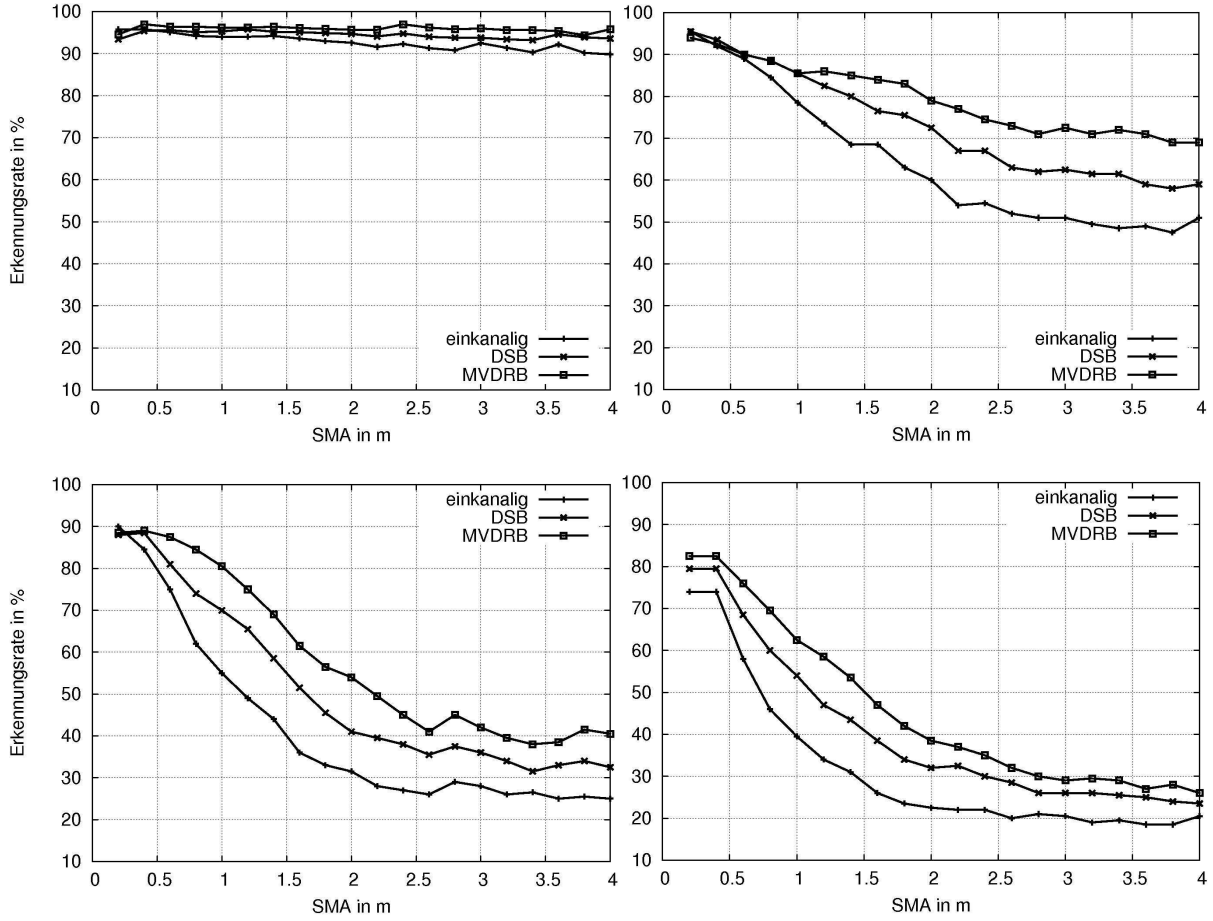

Abbildung 5 – Erkennungsraten des DSB und des MVDRB im Vergleich mit einkanaligen Aufnahmen bei verschiedenen Abständen und in verschiedenen Räumen. Nachhallzeiten: *links oben*: 0,25 s; *rechts oben*: 0,45 s; *links unten*: 0,7 s; *rechts unten*: 0,95 s.

Allen Experimenten ist gemeinsam, dass eine deutliche Verbesserung der Erkennungsraten durch den DSB und noch einmal eine weitere Verbesserung durch den MVDRB auftritt. Besonders bei dem Wohnraum (Nachhallzeit 0,45 s) ist die Leistungsfähigkeit beider Algorithmen erstaunlich hoch. Sie erhöhen die Erkennungsrate um ca. 10 % bzw. 20 %. Mit steigender Nachhallzeit nimmt der Grad der Verbesserung jedoch wieder ab. Zu beachten ist dabei, dass der zusätzliche Aufwand der durch den MVDRB entsteht sehr gering ist. Es ist lediglich eine FFT mit anschließender Multiplikation eines komplexen Faktors in jedem Frequenzpunkt sowie eine darauffolgende inverse FFT nötig. Diese Aufgabe ist für heutige digitale Signalprozessoren (DSP) in Echtzeit zu bewältigen. Deshalb empfehlen die Autoren die Anwendung dieses Algorithmus, wenn ein Mikrofonarray eingesetzt wird.

Ebenfalls charakteristisch für die beiden Beamformer ist der Abfall der Erkennungsrate bei sehr geringer Entfernung. Dies gilt für alle gemessenen Räume und ist besonders gut bei dem Büroraum (Nachhallzeit 0,7 s) zu sehen (Abbildung 5). Die Ursache dafür liegt in der Ausrichtung der Beamformer auf eine Quelle im Fernfeld. Dadurch ergibt sich eine Fehlanpassung und somit eine Verringerung der Signalqualität. Bei einem angepassten Beamformer tritt dieser Effekt nicht auf. Eine Anpassung könnte beispielsweise über eine

Quellenortung erreicht werden, die den Abstand der Quelle zum Array schätzt und die Arrayparameter entsprechend angepasst.

5 Ausblick

Die hier vorgestellten Algorithmen bilden die Grundlage der heutigen Arraytechnologie. In der Literatur ist eine Vielzahl spezieller Algorithmen für verschiedene Aufgaben zu finden, welche zum größten Teil auf dem DSB bzw. dem MVDRB beruhen. Die Untersuchung weiterer Arrayalgorithmen für Spracherkennungssysteme ist Bestandteil zukünftiger Arbeiten.

Der nächste Schritt wird die Modifikation des MVDRB zu einem adaptiven Beamformer beinhalten. Dabei wird die Matrix der Leistungsdichtespektren der Mikrofonsignale und die Wichtungsfaktoren nach Gleichung (4) neu berechnet. Daraus folgt, dass die Nullstellen im Richtdiagramm (Abbildung 1) genau auf die Winkel verlagert werden, aus denen Störschall eintrifft. Eine gezielte Ausblendung diskreter Störquellen ist das Ziel dieser Maßnahme.

Weiterhin sind Untersuchungen zur Verbesserung der Arraygeometrie geplant.

Parallel dazu erfolgt im Moment die Umsetzung der hier vorgestellten Algorithmen in einer Echtzeit-Evaluationsumgebung. Dadurch werden diese in praktischen Anwendungen wie dem oben genannten public Terminal nutzbar.

Literatur

[1] Boll, S. F.: Suppression of Acoustic Noise in Speech using Spectral Subtraction. IEEE Trans. on ASSP, Vol. 27(2), S. 113.120, 1979.

[2] Petrick, R.; Lohde, K.; Lorenz M.; Hoffmann R.: A new feature analysis method for robust ASR in reverberant environment based on the harmonic structure of speech. In: Proc. EUSIPCO2008.

[3] Bitzer, J.; Simmer, U.: Superdirective Microphone Arrays. In : Brandstein, M; Ward, D. (Eds.): Microphone Arrays. Springer-Verlag, ISBN 3-540-41953-5, 2001. S. 19 – 38.

[4] Maase, J.; Hirschfeld, D.; Koloska, U.; Westfeld, T. und Helbig, J.: Towards an evaluation standard for speech control concepts in real-world scenarios. Proc. EUROSPEECH 2003, Geneva, 2003. S. 1553 – 1556.

[5] Hoffmann, R.; Eichner, M; Wolff, M.: Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system. In: Esposito, A., et al: Verbal and Nonverbal Communication Behaviours. Berlin etc.: Springer, LNAI 4775, 2007. S. 200 – 218.

[6] Fehér, T.: Optimiertes Mikrofonarray für Studioanwendungen. In: Proc. ESSV 2006. S. 111 – 118.