SPRECHERLOKALISATION MIT EINEM 3D-MIKROFONFELD

Martin Birth

BTU Cottbus-Senftenberg
martin.birth@tu-cottbus.de

Kurzfassung: Die Signalkopplung zur Ausrichtung verschiedener Mikrofone auf eine spezifische Position im Raum ist einer der interessantesten Punkte bei der digitalen Verarbeitung von Audiosignalen. Das gewonnene fokussierte Signal wird dabei genutzt, um eine Raumabtastung nach möglichen Schallintensitätsmaxima hin zu untersuchen. Dabei spielen bei der Diskretisierung mit mehrdimensionalen Mikrofonfelder eine Reihe von Optimierungen eine wichtige Rolle, die zur Konsequenz eine Aufwandsreduktion haben. Solch eine technische Realisierung eines richtungsbezogenen Hörens wird im folgenden beschrieben.

1 Einleitung

Im Informationsaustausch zwischen Mensch und Maschine erfolgt immerwährend eine Verbesserung der Interaktion. Es ist zur Gewohnheit geworden, das Mikrofon in die Hand zu nehmen, es als Headset zu tragen oder gar das Mobiltelefon an das Ohr zu halten. Ungeachtet dessen, ist es durch den direkten Kontakt mit der Technik, keine rein natürliche Kommunikation. Die Entwicklung hat indessen Fortschritte gemacht und so gibt nun die Möglichkeit unmittelbare, technische Geräte komplett entfallen zu lassen. Die akustische Quellenortung durch Mikrofonfelder und die damit angeschlossene Verarbeitung von Audiosignalen sollen hierfür eine bedeutende Rolle spielen.

Als Ausgangspunkt für die Detektierung von akustischen Signalen dienen mehrkanalige Mikrofonsysteme, sogenannte Mikrofonarrays, die in einem Verbund synchronisiert werden. Solche Felder existieren für vielfältige Anwendungen in den unterschiedlichsten Anordnungen und Dimensionen. Grundlegend wird jedoch deren Arbeitsweise mit einem Hohlspiegel verglichen, bei dem die Schallquelle dem Brennpunkt entspricht [5]. Die Ausrichtung geschieht dabei unter Zuhilfenahme eines adäquaten Beamformers (dt. Strahlrichter), der es ermöglicht einen beliebigen Raumpunkt anzuvisieren, ohne physikalische Einflussnahme auf die Versuchsanordnung. Durch die rein passive Auslenkung der Sensoren erfolgt eine Verstärkung der Signale in „Blickrichtung“, unter gleichzeitiger Reduktion nebenstehender Geräusche und Nachhallelemente.

Das entstandene Gesamtausgangssignal wird im weiteren Verlauf dazu benutzt, den zu betrachtenden Raum systematisch abzutasten. Ein Vergleich der Energiegehalte der Abtastpunkte fördert auf diese Weise die Sprecherposition. Da die Detektierung rein digital erfolgt, können mehrere Beamformer parallel betrieben werden. Darüber hinaus wird der Standort des Sprechers gleichzeitig abgehört und das so gewonnene Mikrofonsignal für die weitere auditive Verarbeitung genutzt.

2 Versuchsaufbau

2.1 Mikrofonarraysystem

In einem aktuellen Projekt hat der Lehrstuhls für Kommunikationstechnik (BTU Cottbus-Senftenberg) das neue „SpeechLab“ mit einem 3D-Mikrofonarray ausgestattet, für die Erforschung der Interaktion mit multimedialen Komponenten. Der Proband soll dort eine möglichst naturgetreue Kommunikation erfahren, die für ihn eine intuitive Handhabung garantiert. Bei den Mikrofonarrays handelt es sich um zwei einzelne Felder in dreidimensionalen Anordnung, die orthogonal zueinander angebracht sind. Eine Arraykomposition, bestehend aus 32 Messmikrofonen, wurde um einen 72 Zoll Multitouchscreen angebracht, der sich an der Stirnseite des Raumes befindet. Das zweite Array, mit 32 weiteren Mikrofonen, befindet sich an der Raumdecke. Deren Komponenten sind spiralförmig in einer Plattenkonstruktion positioniert. Das vollständige Deckenarray lässt sich durch Schienenlagerung mit einem Schrittmotor in der Position verändern. Diese räumliche Anordnung hat einen entschiedenen Vorteil, gegenüber ein- oder zweidimensionalen Konstruktionen, da sich jeder Punkt im Raum anvisieren lässt. Demgegenüber steht die Tatsache, dass mit herkömmlichen Arraykompositionen nur, in einem von ihnen ausgehenden Halbkreis, die Quellenortung möglich ist.

Nicht nur die Kombination dieser zwei Felder ist ein großer Vorteil, auch die hohe Anzahl, der zum Einsatz kommenden Mikrofone. Durch sie wird die Genauigkeit des Beamformers gesteuert. Nach [4] kommt es zu einer Verdichtung der Hauptkeule, bei einer Zunahme der Anzahl der Mikrofonelemente. Gleichzeitig resultiert daraus eine Absenkung der Nebenkeulenmaxima.

2.2 Raumkoordinatensystem

Alle Mikrofone besitzen nur eine relative Position auf der jeweiligen konstruierten Platte. Die Sprecherlokalisation macht es jedoch notwendig, ein einheitlichen Raumkoordinatensystem zu schaffen. Für diesen Zweck wird ein festes Messfeld im Raum definiert (hier 4,40 x 4,40 x 2,20 m). Ihn diesem Rechtkant erhält jedes Mikrofon eine absolute Position. Darüber hinaus sind durch diese Einschränkung des Messfeldes auch alle möglichen Zielpositionen bekannt, da diese sich nur innerhalb des Feldes befinden dürfen. Um eine Aufwandsreduktion für die Diskretisierung zu erreichen, wurde der abzutastende Raum in Würfel mit einer Kantenlänge von 20 cm unterteilt. Dies sorgt in dieser projektbezogenen Anwendung für eine Ausreichend große Genauigkeit der Lokalisation.

Aus den absoluten Lagedaten der Mikrofone und den Daten für alle möglichen Sprecherpositionen lassen sich im nächsten Schritt alle Abstände und die daraus resultierenden Schallaufzeiten τ mit

$$\tau_i = \frac{1}{c_s} \left\| \vec{l_i} - \vec{l_p} \right\| \tag{1}$$

bestimmen. Hierbei wird c_s für die Schallgeschwindigkeit verwendet. Die Vektoren $\vec{l_i}$ und $\vec{l_p}$ stehen für die Position des i-ten Mikrofon $(i = 1,2,..,64)$ und Position des Sprechers. Diese Daten stellen nachfolgend die Grundlage für die Ortung einer möglichen Sprecherposition dar.

3 Sprecherlokalisation

3.1 Allgemeines

Durch die Festlegung eines Raumkoordinatensystems sind bereits alle möglichen Raumpunkte im Messfeld bekannt. Da die Bestimmung der Energiegehalte für jede Position einen erheblichen Aufwand beuten würde und im Zuge dessen auch eine schlechtere Erkennungsrate, ist es unerlässlich eine Optimierung der Abtastung einzurichten. Aus diesem Grund wurde das Messfeld in zwölf Ebenen (Abb. 1) unterteilt, die der Annahme zu Grunde liegen, dass ein einzelnen Sprecher, im Verlauf der Bedienung des Gesamtsystems, sich nicht in seiner Körpergröße ändert. Kleinere Bewegungen des Kopfes werden durch die Größe der Messpunkte durchaus toleriert.
Dies hat insgesamt zur Folge, dass eine Positionsabtastung nur stets auf einer Ebene erfolgen muss. Eine stetige Überprüfung der entsprechenden Höheneinheit ist somit nicht notwendig. Es ist folglich ausreichend, in einem kurzen Intervall nur die jeweilig betreffende Ebene abzuscannen und in einem parallel laufenden Beamformer mit einem größeren Zeitintervall nach der entsprechenden Höheneinheit zu suchen, um so auch eine neue Person mit einer abweichenden Körpergröße zu detektieren.

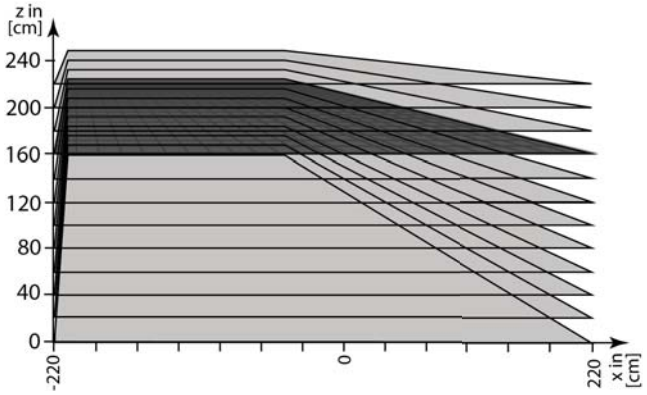

Abbildung 1 - Ebenendarstellung des Messfeldes.

3.2 Beamforming

Mit Hilfe des Beamformings im Zeitbereich wird unter Einsatz eines Mikrofonarrays ein Strahl durch räumlich, zeitliches Filtern der eintreffenden Signale gebildet. Dies geschieht unter Ausnutzung der Laufzeitunterschiede. Der so entstehende Fokuspunkt lässt sich beliebig auf eine Position lenken und kann somit für die Ortung einer beliebigen Schallquelle im Raum eingesetzt werden. Die Auslenkung ist dabei rein passiv und es wird kein Einfluss auf physikalische Anordnung der Mikrofone genommen. Für die exakte Synchronisation ist nach (1) eine Bestimmung der Laufzeiten für alle N Mikrofone eingangs notwendig. Dies ist begründet mit den differenzierenden Entfernungen und Einfallswinkel, wodurch der Schall auch ungleich an den Sensoren auftrifft.

Ausgehend davon, dass die Schallwellen nicht direkt aus der Querabrichtung auf ein Array mit N Mikrofone treffen, kommt es nun zu dem zeitlichen Versatz (engl. delay) der eintreffenden Signale. Die einzelnen Delayglieder τ_i $(i = 1,2,...,64)$ werden dazu genutzt die Verzögerung so auszugleichen, dass die anliegenden Signale $x_i(t)$ nach einer Verschiebung kohärent sind. Dazu erfolgt in einem weiteren Schritt eine Bestimmung der Verzögerungszeit des am weitesten entferntesten Mikrofons τ_{max} aus allen τ_i, im Verhältnis zu einer spezifischen Messposition. Aus

der Differenz dieser Werte lässt sich das Delay $d_{i/max}$ für jedes i-te Mikrofon bestimmen:

$$d_{i/max} = \tau_{max} - \tau_i. \tag{2}$$

Die Gesamtheit der Delayglieder für ein Mikrofonfeld zu einer spezifischen Messposition wird in dem Steeringvektor $\vec{\theta}$ zusammengefasst. Eine Wichtungsfunktion w_i sorgt im weiteren Verlauf für eine entsprechende Bewertung der unabhängigen Signale, um nachfolgend das Gesamtausgangsignal zu erhalten [3]. Das Ausgangssignal $y(t)$ des Beamformers lässt sich somit wie folgt beschrieben:

$$y(t) = \sum_{i=1}^{N} w_i \cdot x_i(t - \vartheta_i). \tag{3}$$

Die Vorgehensweise für diese Art des Beamformings ist charakteristisch, daher auch die Bezeichnung „Delay-and-Sum-Beamformer“ (kurz D&S-Beamformer). Durch die dynamische Anpassung der Verzögerungsglieder auf einen spezifischen Punkt, lässt sich der Beamformer in eine beliebige Richtung steuern. Um auch bewegende Ziele anvisieren zu können, muss die Anpassung des Steeringvektors $\vec{\vartheta}$ möglichst schnell geschehen, um keinen Verlust an Audiodaten zu erhalten.

3.3 Evaluation des Beamformers

Die räumliche Empfindlichkeit eines Mikrofonfelds wird als Richtcharakteristik bezeichnet. Zur Lokalisation des Sprechers müssen die Orte maximaler Empfindlichkeit („Haupt-“ und „Nebenkeulen “) bekannt sein.

Nach [1] ist die Fouriertransformierte eines einzelnen Mikrofonsignals

$$\underline{X}_i(\omega) = a_i \, \mathrm{e}^{-\mathrm{j}\omega\tau_i} \cdot \underline{S}(\omega), \tag{4}$$

wobei $\underline{S}(\omega)$ für die Fouriertransfomierte des Quellensignals (Sprecher) sowie a_i für die Dämpfung und τ_i für die Laufzeit des Quellensignals am Mikrofon stehen. Das Summensignal des Mikrofonfelds im Frequenzbereich ist:

$$\underline{Y}(\omega) = \sum_{i=1}^{N} w_i \mathrm{e}^{-\mathrm{j}\omega\vartheta_i} \cdot \underline{X}_i(\omega). \tag{5}$$

w_i steht dabei für die Verstärkung und und ϑ_i für die Verzögerung des Mikrofonsignals (steering vector). Einsetzen von (5) in (4) ergibt:

$$\underline{Y}(\omega) = \underbrace{\sum_{i=1}^{N} w_i a_i \exp\left[-j\omega\left(\vartheta_i + \tau_i\right)\right]}_{\underline{H}(\omega,\vec{l})} \cdot \underline{S}(\omega). \tag{6}$$

Die erhaltene räumlich-zeitliche Übertragungsfunktion $\underline{H}(\omega,\vec{l})$ wird im Anschluss dazu genutzt, um die Intensitätsverteilung zu einem beliebigen Raumpunkt $\vec{l}_p$ zu bestimmen. Dies erfolgt mit der Formel:

$$|H(\omega,\vec{l}_p)| = 20log \left| \sum_{i=1}^{N} w_i a_i \exp\left[-\mathrm{j}\omega\left(\vartheta_i + \frac{\left\|\vec{l}_i - \vec{l}_p\right\|}{c_s}\right)\right] \right|_{[dB]}. \tag{7}$$

Mit Hilfe der Gleichung (7) lässt sich folglich die Sensitivitätsverteilung für das betrachte Messfeld bestimmen. Für eine beispielhafte Darstellung 2, die einen horizontalen und vertikalen Querschnitt des Messfeldes zeigt, wurde für die Frequenz $f = 1000$ Hz, $l_p = [0;0;1,60]$ und $N = 64$ Mikrofonpositionen für das 3D-Mikrofonarray gewählt.

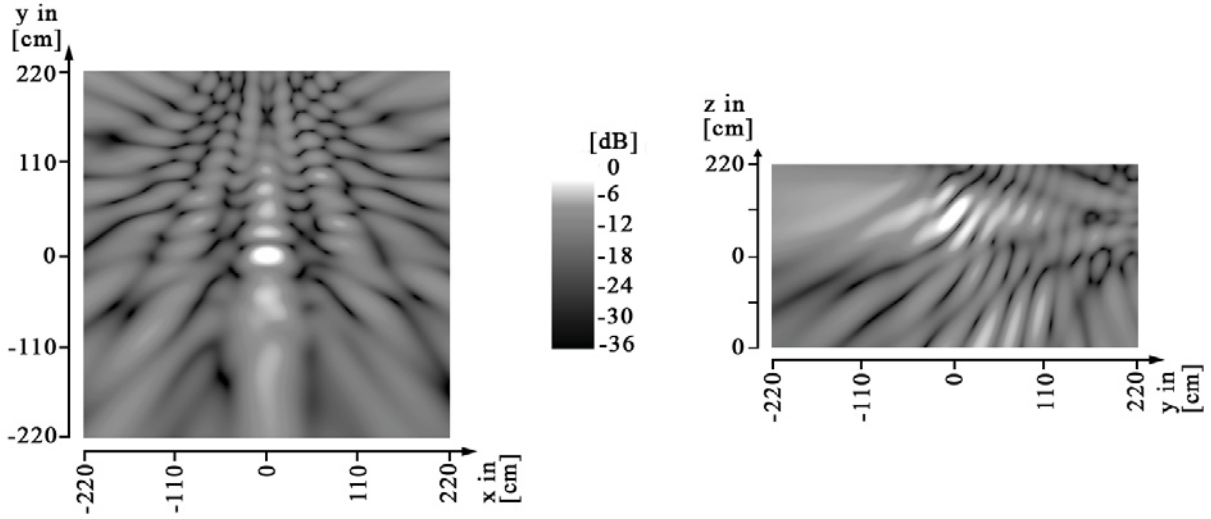

Abbildung 2 - Horizontalen und vertikaler Querschnitt der Sensitivitätsverteilung nach (7) für das Messfeld (f=1000 Hz, $\vec{l_p} = [0;0;1,60]$ und N=64 Mikrofone).

3.4 Sprecherlokalisation mittels Beamformern

Die Sprecherlokalisation baut in erster Linie auf dem gewählten Raumkoordinatensystem aus Abschnitt 2.2 auf. Zusätzlich zu diesen Daten wird noch die Laserposition des Deckenarrays ermittelt. Dies ist unverzichtbar durch die variable Schienenlagerung an der Raumdecke. Aus allen diesen Informationen heraus erfolgt daraufhin die Bestimmung der Steering-Vektoren. Durch eine stetige Überprüfung der Laserposition wird gewährleistet, dass keine fehlerhaften Abstandsvektoren in die Berechnungen des Beamformers mit einfließen.

Im Anschluss daran kann mit dem Einlesen der Audiodaten von allen 64 Mikrofonen begonnen werden. Um zu ermitteln auf welcher Ebene die genaue Positionsbestimmung erfolgen muss, wird in einem ersten Schritt auf jede Höheneinheit gezielt. Dies geschieht unter Zuhilfenahme eines unscharfen Beamformers, mit einer geringen Anzahl an Mikrofonen. Die Audiosignale werden zu diesem Zweck nach Formel (3) verschoben und summiert. Die Gewichtung richtet sich dabei nur nach der Anzahl der verwendeten Mikrofone $w_i = \frac{1}{N}$. Für jede Ebene wird somit der Energiegehalt vom Ausgangssignal des D&S-Beamformers bestimmt. Der resultierende Maximalwert dieser Überprüfung entspricht der Eingrenzung auf die zu untersuchende Höhe (Z-Wert), wie in Abbildung 1 beispielhaft dargestellt. Für eine Optimierung ist eine stetige Überprüfung auf die gesuchte Höheneinheit, bei einem nicht wechselnden Sprecher, nicht dringend notwendig.

In einem weiteren Schritt erfolgt die Untersuchung der genauen Sprecherposition auf den zuvor festgelegten Z-Wert. Dafür werden abermals die Audiosignale in einen D&S-Beamformer gegeben. Dabei wird die vollständige Anzahl an Mikrofonen verwendet, um so eine möglichst

hohe Genauigkeit zu erzielen. Es folgt daraufhin eine Abtastung aller 529 Messpunkte auf der Ebene und die Speicherung der jeweiligen resultierenden Energiewerte. In letzter Instanz werden diese Werte miteinander verglichen und das gefundene Schallintensitätsmaxima entspricht der Sprecherposition im Raum.

Für eine genaue Vorstellung des Ablaufes und der eigentlichen Implementierung ist im Algorithm 1 ein Pseudocode mit aufgeführt. Damit die hier gezeigte Vorgehensweise erfolgreich ist, muss sichergestellt werden, dass sich im Bereich des Messfeldes bzw. Raumes keine extrem lauten Nebengeräusche befinden. Diese könnten eine Erweiterung des Störfeld mit sich bringen. Die Nebengeräuschunterdrückung ist durch die Verwendung eines Beamformers zwar sehr hoch, es kann jedoch bei übermäßigen akustischen Störungen zu einer Verringerung der Erkennungsrate kommen.

```
input  : Audiodaten (64 Mics), Laserdaten
output: Sprecherposition
initialisierung();
berechneSteeringVektoren();
while Audiodaten!=NULL do
    einlesenAudiodaten();
    einlesenLaserdaten();                          // Position des Deckenarrays
    if alteLaserdaten != neueLaserdaten then
        alteLaserdaten ← neueLaserdaten;
        berechneNeueSteeringVektoren();
        gehe zu einlesenAudiodaten();
    end
    for i ← 1 to 12 do
        ebene[i] = GroberBeamformer(gesamtebene[i]);   // Geringe Mic-Anzahl
        neueEbene=getMaxOf(ebene[]);                   // Out: Index von ebene[]
        if scanEbene != neueEbene then
            scanEbene ← neueEbene;
        end
        for j ← 1 to 529 do
            werte[j] ← FeinerBeamformer(scanEbene.position[j]);
        end
        neueSprecherposition=getMaxOf(werte[]);  // Out: Index von werte[]
        if sprecherposition != neueSprecherposition then
            sprecherposition ← neueSprecherposition;
        end
    end
end
```

Algorithm 1: Pseudocode zum Ablauf der Erkennung der Sprecherlokalisation

4 Zusammenfassung und Ausblick

Die akustische Quellenortung und die damit in Verbindung stehende Ausgabe des Gesamtmikrofonarraysignals ist stets immer nur ein auditives Teilprojekt für die Verarbeitung von Sprachsignalen. Es dient überwiegend dazu, den entsprechenden Sprecher zu lokalisieren und ein scharfes Mikrofonsignal an etwaige Spracherkenner oder andere Verarbeitungsalgorithmen zu liefern. Das Beamforming ist in diesem Zusammenhang das Mittel der Wahl und offeriert neue Möglichkeiten zur Verarbeitung von digitalisierter Sprache. Somit wäre es nicht auszuschließen den Beamformer so zu modifizieren, dass auch parallel Sprecher herausgefiltert werden. Dies kann zum Beispiel durch die Anpassung der Wichtungsglieder mit adaptiven Transversalfilter geschehen [2]. Aber auch der Einsatz von superdirectiven oder anderen Filter-Beamformer (Wiener, Frost oder Kalman Filter) ist durchaus denkbar für die Weiterentwicklung des Systems.

Ein weiterer wichtiger Punkt für die Optimierung stellt der Suchalgorithmus an sich dar. Die Ortung der Sprecherposition erfolgt hinlänglich stabil und schnell, jedoch bedarf es noch Untersuchungen der genauen Fehlerraten für den Positionserkenner. Parallel dazu wird an einer Möglichkeit gearbeitet die Suchalgorithmen mittels *OpenGL* [6] direkt auf die Grafikkarte auszulagern. Der entschiedene Vorteil gegenüber der herkömmlichen Berechnung auf einer CPU, wäre die echt parallele Abarbeitung von 529 Messpunkten, die einer kompletten Messebene entsprechen. Dies würde eine massive Steigerung in der Geschwindigkeit der Erkennung mit sich bringen. Erste Tests beim Einsatz von *OpenGL* zur Berechnung der Sensitivitätsverteilung deuten auf eine erhebliche Steigerung der Rechengeschwindigkeit. Daher ist eine Ausweitung der Anwendung auf das Gesamtsystem durchaus vorstellbar.

Literatur

[1] BITZER, J. und K. SIMMER: *Superdirective Microphone Arrays*. In: *Microphone Arrays: Signal Processing Techniques and Applications*, Kap. 2, S. 20–23. Springer, 2001.

[2] DREWS, M.: *Mikrofonarrays und mehrkanalige Signalverarbeitung zur Verbesserung gestoerter Sprache*. Microfilm-Vorlesungsskripte, 1999.

[3] GOLDENBAUM, M.: *Time-Delay-and-Sum-Beamforming an den gemessenen Sensordaten eines Experimental-Sediment-Sonars*. Diplomarbeit, Hochschule Bremen, 2004.

[4] GREENSTED, A.: *The Lab Book Pages: Delay Sum Beamforming*. http://www.labbookpages.co.uk/audio/beamforming/delaySum.html, 2012. Zugriff: 03.06.2013.

[5] SARRADJ, E.: *Mikrofonarray mit Nahfeld-Beamforming*. Techn. Ber., Gesellschaft für Akustikforschung Dresden mbH, 2004.

[6] THE–KHRONOS–GROUP: *OpenGL*. http://www.opengl.org, 2012. Zugriff: 17.01.2014.