to appear in: Deutsche Gesellschaft für Akustik e.V. (DEGA), *Fortschritte der Akustik — DAGA98*, Oldenburg, 1998 (in press)

Blinde akustische Quellentrennung im Frequenzbereich

Jörn Anemüller und Tino Gramß (†)
Graduiertenkolleg Psychoakustik und AG Medizinische Physik
Carl von Ossietzky-Universität, D-26111 Oldenburg
email: ane@uni-oldenburg.de

Einführung

In vielen Anwendungen der Sprachverabeitung werden Überlagerungen mehrerer Signale gemessen, etwa die Überlagerung eines Nutzsprechers mit einem Störsprecher. Das menschliche Hörsystem besitzt die bemerkenswerte Eigenschaft, aus solchen vermischten Signalen die zugrundeliegenden Quellsignale zu rekonstruieren. Ein technischer Algorithmus, der eine ähnliche Fähigkeit besitzt, könnte etwa in der Spracherkennung nützlich sein, wo auch die besten Systeme bei Störschall versagen.

In diesem Beitrag wird ein Ansatz für einen solchen „Quellentrennungs"-Algorithmus beschrieben, der aus der Kenntnis mehrerer Überlagerungen die zugrundeliegenden Quellsignale rekonstruiert. Dabei wird als Kriterium zur Trennung benutzt, daß verschiedene Sprecher Signale aussenden, die voneinander statistisch unabhängig sind. Nimmt man Überlagerungen dieser Quellsignale mit mehreren Mikrofonen auf, so sind die an den verschiedenen Mikrofonen registrierten Signale korreliert. Das Ziel besteht nun darin, aus den Mikrofonsignalen wiederum mehrere statistisch unabhängige Signale zu rekonstruieren, die den Quellsignalen entsprechen. Da hierzu kein a priori Wissen über die räumliche Anordnung der Signalquellen benutzt wird, spricht man auch von „blinder" Quellentrennung [2].

Lineare Quellentrennung

Für den Fall einer instantanen linearen Überlagerung

$$m_i(t) \quad = \quad \sum_j A_{ij} s_j(t), \tag{1}$$

von N Quellsignalen $s_j(t)$ zu N Mikrofonsignalen $m_i(t)$ existieren mehrere Algorithmen zur blinden Quellentrennung; eine Übersicht findet sich etwa in [4]. Das ihnen gemeinsame Ziel besteht darin, eine möglichst gute Schätzung $\mathbf{W}$ für die Inverse $\mathbf{A}^{-1}$ der Mischmatrix $\mathbf{A} = [A_{ij}]$ zu finden, da dann die Quellensignale als $\vec{x}(t) = \mathbf{W}\vec{m}(t)$ aus den Mikrofonsignalen $\vec{m}(t)$ rekonstruiert werden können.

Der von uns benutzte Algorithmus basiert auf der Maximum-Likelihood Methode nach [3], erweitert für den Fall komplexwertiger Signale und Koeffizienten. $\mathbf{W}$ wird zu jedem Zeitpunkt t iterativ durch die Lernregel

$$\mathbf{W} \leftarrow \mathbf{W} + \eta(\mathbf{W} - 2\vec{u}(t)\vec{x}^H(t)\mathbf{W}) \tag{2}$$

adaptiert. Hierbei ist $u_i(t) = \tanh(|x_i(t)|)\, x_i(t)/|x_i(t)|$. $\vec{x}^H$ bezeichnet den zu $\vec{x}$ transponierten und komplex konjugierten Vektor. η steuert die Größe der Iterationsschritte.

Im Folgenden sind zwei prinzipielle Begrenzungen linearer Quellentrennung von Bedeutung. Die Rekonstruktion der Quellsignale ist nur bis auf eine Permutation möglich, so daß unbekannt ist, welcher Ausgabekanal des Algorithmus welcher Quelle entspricht („Permutationsinvarianz"). Außerdem kann jedes Quellsignal nur bis auf einen skalaren Faktor rekonstruiert werden („Skalierungsinvarianz").

†Dr. Tino Gramß, der diese Arbeit initiierte und betreute, verstarb im Januar 1998. Ein Nachruf findet sich im DEGA-Sprachrohr, Heft 16, S. 25f., Februar 1998.

Akustische Quellentrennung im Frequenzbereich

Für Schallquellen gilt die Annahme instantaner Überlagerung nicht mehr, da Schallaufzeiten und Echos berücksichtigt werden müssen. Daher wird die Multiplikation in Gl. 1 durch die Faltung von Quellsignal $s_j(t)$ mit der Impulsantwort $A_{ij}(\tau)$ des Raumes von Quelle j zu Mikrofon i ersetzt:

$$m_i(t) \quad = \quad \sum_j \sum_\tau A_{ij}(\tau) s_j(t-\tau). \tag{3}$$

Eleganter läßt sich die Filterung durch die Raumübertragungsfunktionen $\hat{A}_{ij}(f)$ im Frequenzbereich darstellen [1]. Berechnet man zu aufeinanderfolgenden Zeiten T jeweils die Kurzzeitspektren $\hat{m}_i(f,T)$ von $m_i(t)$ und $\hat{s}_j(f,T)$ von $s_j(t)$, so läßt sich Gl. 3 im Frequenzbereich in hinreichend guter Näherung schreiben als

$$\hat{m}_i(f,T) \quad = \quad \sum_j \hat{A}_{ij}(f) \hat{s}_j(f,T). \tag{4}$$

Mit Gl. 4 haben wir also das akustische Quellentrennungsproblem in voneinander unabhängige lineare Quellentrennungsprobleme für die Frequenzen $f_1, \ldots, f_N$ umformuliert, die mit den o.g. Algorithmen behandelt werden können.

Diese Eleganz wird dadurch „erkauft", daß im Allgemeinen die Quellsignale in den Ausgabekanälen des Algorithmus in reskalierter und permutierter Form auftreten, was eine Rekonstruktion im Zeitbereich ohne weitere Vorkehrungen unmöglich macht. Außerdem wird Quellentrennung nur langsam erreicht, da nach der Fouriertransformation nur noch relativ wenige Datenpunkte in jedem einzelnen Frequenzband vorhanden sind.

Quellentrennung bei δ-förmiger Impulsantwort

Wir betrachten zunächst die Quellentrennung in Räumen mit δ-förmiger Impulsantwort, wo also Laufzeit- und Pegelunterschiede, aber keine Echos auftreten. Dies entspricht der Schallausbreitung in einem idealen reflexionsarmen Raum. In diesem Fall ist die Überlagerung im Zeitbereich gegeben durch

$$m_i(t) \quad = \quad \sum_j A_{ij} s_j(t - \tau_{ij}). \tag{5}$$

Bei Transformation in den Frequenzbereich resultiert die Laufzeit τ_{ij} von Quelle j zu Mikrofon i in einer zur Frequenz proportionalen Drehung der Phase der komplexen Überlagerungskoeffizienten $A_{ij}e^{-2\pi i f \tau_{ij}}$, während deren Betrag $|A_{ij}|$ konstant bleibt:

$$\hat{m}_i(f,T) \quad = \quad \sum_j A_{ij} e^{-2\pi i f \tau_{ij}} \hat{s}_j(f,T). \tag{6}$$

Dieser Zusammenhang zwischen den Werten der Übertragungsfunktionen bei verschiedenen Frequenzen erlaubt es, den Trennungsalgorithmus frequenzübergreifend

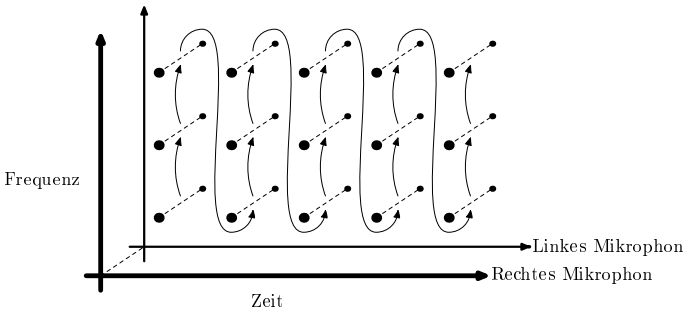

Abbildung 1: Schema zur Abtastung der Spektrogramme der Mikrofonsignale. Jeder Punkt symbolisiert einen Wert im Spektrogramm.

zu iterieren: Bei Abtastung der Spektrogramme der Mikrofonsignale analog zu Abb. 1 lassen sich Mehrdeutigkeiten in der Phase durch einen passenden unwrapping Algorithmus auflösen, was die korrekte Transformation der Überlagerungskoeffizienten zwischen den Frequenzen erlaubt.
Durch dieses Vorgehen werden die genannten Probleme der Quellentrennung im Frequenzbereich gelöst. Dadurch, daß die verschiedenen Frequenzen miteinander „verbunden“ werden, können keine Permutationen mehr zwischen den Ausgabekanälen des Algorithmus bei verschiedenen Frequenzen auftreten. Da außerdem zur Trennung zeitverzögerter Signale nur die „interauralen“ Differenzen der verschiedenen Quellen, also Laufzeitunterschiede und relative Pegel, benötigt werden, besteht auch das Skalierungsproblem nicht mehr. Eine Rekonstruktion der Quellsignale im Zeitbereich wird damit möglich. Schließlich wird durch die effektive Abtastung der Spektrogramme eine schnelle Konvergenz des Algorithmus erreicht.

Experimentelle Überprüfung

Mit im Oldenburger reflexionsarmen Raum gemachten Aufnahmen haben wir die Funktionsfähigkeit des Algorithmus überprüft. Dazu wurde Sprache über zwei im Abstand von 2,9m aufgestellte Lautsprecher abgespielt und die überlagerten Signale mit zwei Mikrofonen aufgenommen. Der Abstand der Mikrofone zueinander betrug 35cm; sie standen 1,5m vor den Lautsprechern, seitlich um 75cm von der Mitte versetzt.
Zur Trennung wurden Frequenzen von 33Hz bis 8kHz benutzt. Jeder Wert aus den Spektrogrammen der Mikrofonsignale wurde genau einmal zur Iteration benutzt. Die Initialisierung des Algorithmus erfolgte mit der Information, die Mikrofonsignale seien statistisch unabhängig; zu Beginn wurde $\mathbf{W}$ also als Einheitsmatrix vorgegeben.
Von diesem (natürlich unzutreffenden) Anfangszustand wurde Konvergenz zu den korrekten Werten innnerhalb von etwa 7s erreicht. Die wesentlichen, vom Algorithmus für jedes Quellsignal individuell geschätzten, Parameter sind die durch die jeweilige Quelle hervorgerufenen Pegel- und Laufzeitunterschiede zwischen den Mikrofonen. Der zeitliche Verlauf der Schätzung des Algorithmus für diese Parameter ist in Abb. 2 und Abb. 3 am Beispiel von Quelle 1 dargestellt. Insbesondere fällt auf, daß auch die in den Quellsignalen vorhandenen Sprachpausen vom Algorithmus bewältigt werden, ohne die Parameterschätzung wesentlich zu stören.

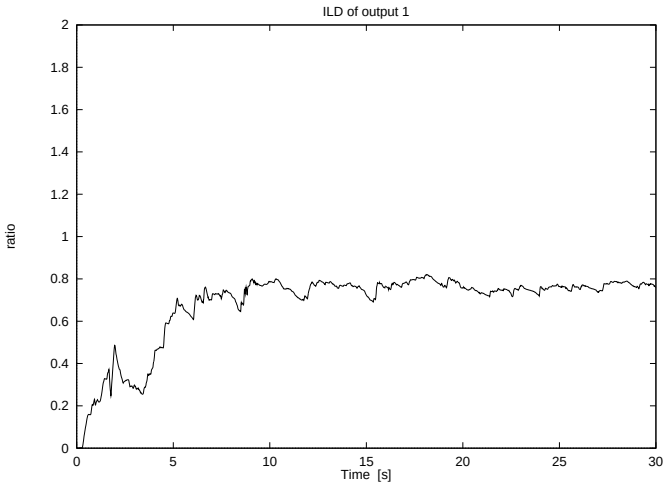

Abbildung 2: Zeitliche Entwicklung der Schätzung des Algorithmus für den durch Quellsignal 1 an den Mikrofonen hervorgerufenen Pegelunterschied.

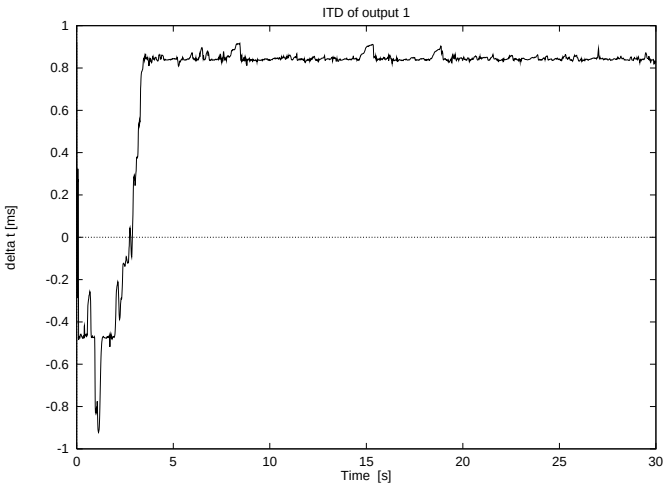

Abbildung 3: Geschätzte Laufzeitdifferenz in ms zwischen den Mikrofonen für Quellsignal 1.

Zusammenfassung und Ausblick

Es wurde ein Algorithmus vorgestellt, der durch Ausnutzung statistischer Verfahren mehrere Sprechersignale in einem reflexionsarmen Raum erfolgreich und in kurzer Zeit trennen kann.
Für die Zukunft ist die Trennung bewegter Quellen vorgesehen; erste Versuche hierzu sind erfolgversprechend verlaufen. Weiterhin erscheint die Trennung in beliebigen Räumen wünschenswert.

Literatur

1. T. Gramß. A neural model for the separation of acoustic signals. In J. Bower, editor, *Computational Neuroscience: Trends in Research 1995*, pages 191–195, New York, 1996. Academic Press. CNS 95, Monterey, California, July 1995.
2. C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.
3. D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis, draft 3.7. URL: ftp://wol.ra.phy.cam.ac.uk/pub/mackay/ica.ps.gz, Dec. 1996.
4. J.-P. Nadal and N. Parga. Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches. *Neural Computation*, 9:1421–1456, 1997.