Klassifikation von Korrelationsextrema zur Laufzeitdifferenzschätzung bei mehreren Sprachquellen

Jan Scheuing, Bin Yang
Lehrstuhl für Systemtheorie und Signalverarbeitung, Universität Stuttgart
Email: {jan.scheuing,bin.yang}@Lss.uni-stuttgart.de

Einleitung

Die Schätzung von Laufzeitdifferenzen ist eine häufige und wichtige Aufgabe bei Verarbeitung von Sprachsignalen, die mit einem Mikrofon-Array aufgenommen werden. Neben den klassischen Kreuzkorrelationsverfahren [1] werden derzeit verstärkt Methoden zur blinden Schätzung von Raumimpulsantworten verfolgt [2], die in halligen Umgebungen interessant sind, deren Ansatz jedoch vor einer einzelnen Quelle ausgeht.

Bei der Laufzeitdifferenzschätzung mehrerer gleichzeitig aktiver Sprecher besteht die Herausforderung unter Verwendung klassischer Verfahren in der durch mehrere Quellen hervorgerufenen Mehrdeutigkeit der Korrelationsmaxima, welche zu jenen Mehrdeutigkeiten hinzukommt, die durch Schallreflexionen im Raum und durch die Eigenschaften des Sprachsignals entstehen.

Dieser Beitrag beschreibt ein Verfahren zur Reduktion der Reflexionsmehrdeutigkeit unter Ausnutzung der Autokorrelierten. Anhand einer Rasterbedingung [3] werden die aus zwei Mikrofonsignalen geschätzten Laufzeitdifferenzen danach klassifiziert, ob sie durch Direktpfade oder durch Echopfade entstehen. Das Verwerfen von Echopfad-Laufzeitdifferenzen ist sowohl für Beamforming- als auch für Lokalisierungsanwendungen entscheidend, insbesondere im Hinblick auf eine geringe Komplexität bei der Zuordnung von Laufzeitdifferenzen mehrerer Mikrofonpaare zu verschiedenen Quellen.

Mehrsprecher-Echo-Modell

Betrachtet wird ein Raum mit N Sprechern und 2 Mikrofonen. Unter Vernachlässigung des Rauschens lassen sich die Mikrofonsignale $x_i(t)$ durch das lineare Modell

$$x_i(t) = \sum_{a=1}^{N} (h_{a,i} * s_a)(t), \quad i \in \{1,2\} \tag{1}$$

beschreiben, wobei $*$ den Faltungsoperator, $h_{a,i}(t)$ eine Raumimpulsantwort von Sprecher a zu Mikrofon i und $s_a(t)$ das Quellsignal darstellt. Es wird angenommen, dass jede Raumimpulsantwort nur $L_{a,i}$ signifikante Ausbreitungspfade $\mu \in \{0,\dots,L_{a,i}-1\}$ enthält, welche durch ihre Amplitude $h_{a,i,\mu}$ und ihre Laufzeit $\tau_{a,i,\mu}$ charakterisiert sind; der Resthall wird vernachlässigt.

$$x_i(t) = \sum_{a=1}^{N} \sum_{\mu=0}^{L_{a,i}-1} h_{a,i,\mu} s_a(t - \tau_{a,i,\mu}) \tag{2}$$

Die Pfade werden nach aufsteigender Laufzeit indiziert; es wird angenommen, dass sich unter den signifikanten Pfaden immer auch der Direktpfad $\mu = 0$ befindet.

Rasterbedingung

Das Signal $s(t)$ einer einzelnen weißen Rauschquelle[1] ruft in der Autokorrelierten $r_{ii}(t) = \mathrm{E}[x_i(t+t_0)\, x_i(t_0)]$ des Mikrofonsignals $x_i(t)$, $i \in \{1,2\}$ lokale Extrema an den Stellen $t_{ii,\mu\eta} = \tau_{i,\mu} - \tau_{i,\eta}$ und $-t_{ii,\mu\eta}$ hervor; die Kreuzkorrelierte $r_{12}(t) = \mathrm{E}[x_1(t+t_0)\, x_2(t_0)]$ weist an den Stellen $t_{12,\mu\nu} = \tau_{1,\mu} - \tau_{2,\nu}$ Maxima auf. Zwischen diesen gelten die in Abbildung 1 veranschaulichten Beziehungen

$$t_{11,\mu 0} = t_{12,\mu\eta} - t_{12,0\eta} \quad \text{und} \quad t_{22,\nu 0} = t_{12,\eta 0} - t_{12,\eta\nu}, \tag{3}$$

wobei irrelevant ist, auf welchem gemeinsamen Pfad η das Signal am jeweils anderen Sensor ankommt.

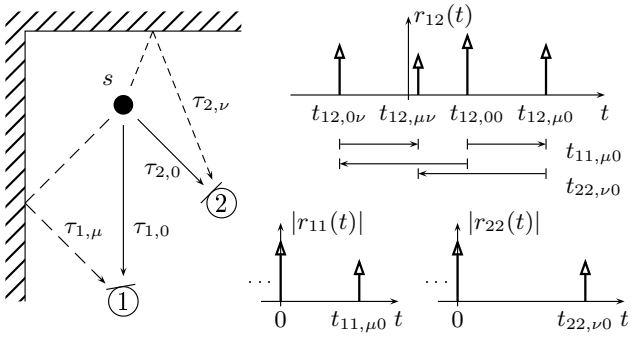

Abbildung 1: Raster der Extremstellen in Auto- und Kreuzkorrelierten

Da die Signalausbreitung entlang des Direktpfads immer die kürzeste Laufzeit hat, gilt in (3) zusätzlich $t_{11,\mu 0} > 0$ und $t_{22,\nu 0} > 0$[2]. Daraus folgt, dass bei einem Abstand $t_{11,\mu 0}$ die vom Echopfad stammende Maximumstelle in $r_{12}(t)$ weiter rechts, bei einem Abstand $t_{22,\nu 0}$ weiter links liegt. Die Direkt-Echopfad-Relation ist durch Anfänge und Spitzen der Zuordnungspfeile wiedergegeben.

In einem ersten Ansatz kann demnach die gesuchte Direktpfad-Laufzeitdifferenz $t_{12,00}$ als diejenige Maximumstelle in $r_{12}(t)$ identifiziert werden, welche keine Pfeilspitzen aufweist. Theoretisch genügt dies auch zur Identifizierung von Direktpfad-Laufzeitdifferenzen mehrerer unkorrelierter Quellen und bei mehreren Echopfaden.

Praktische Implementierung

Verursacht durch Signalfensterung, Abtastung und spektrale Glättung mittels GCC-PHAT [1] treten in der Praxis jedoch Ungenauigkeiten auf, die eine aufwendigere Klassifizierung der Maximumstellen erforderlich machen.

[1] Hier wird zunächst nur eine Quelle ($N = 1$) betrachtet und daher auf den Index a verzichtet.

[2] Nachfolgend wird $r_{ii}(t)$ nur noch für $t > 0$ ausgewertet.

In gemessenen Kreuzkorrelierten kommt es vor, dass das Direktpfad-Maximum einer Quelle mit dem Echopfad-Maximum einer anderen Quelle zusammenfällt (Abbildung 2). Im Vergleich zur Suche nach Maximumstellen ohne Pfeilspitzen ist eine Klassifikation unter Berücksichtigung der *Trefferhäufigkeit* von detektierten Direktpfaden (Pfeilanfänge) und Echopfaden (Pfeilspitzen) robuster gegenüber zufälligen Treffern. Betrachtet werden dazu die M größten lokalen Maximimstellen $t \in \mathbb{T}_M$ von $r_{12}(t)$.

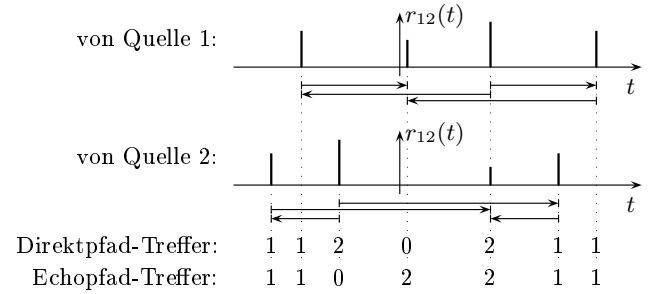

Abbildung 2: Trefferhäufigkeit für zwei Quellen bei ungünstiger Überlagerung: Ein Verwerfen aller Maximumstellen mit Echopfad-Treffern (Pfeilspitzen) würde die Direktpfad-Schätzung von Quelle 1 eliminieren.

Eine Überlagerung zweier Maxima unterschiedlichen Ursprungs an einer Stelle $t_\mu \in \mathbb{T}_M$ ist im *Qualitätsmaß*

$$q(t_\mu) = r_{12}(t_\mu) + \sum_{t_\eta^+} |r_{ii}(t_\eta^+)| - \sum_{t_\eta^-} |r_{ii}(t_\eta^-)| \qquad (4)$$

noch deutlicher zu erkennen; dieses wertet das Maximum $r_{12}(t_\mu)$ mit dem passenden Autokorrelierten-Maximum im Fall detektierter Direktpfade t_η^+ auf und im Fall von Echopfaden t_η^- ab.

Selbst nach Interpolation ist bei abgetasteten Signalen nicht damit zu rechnen, dass (3) exakt erfüllt ist. Eine Toleranzfunktion $w(t)$ mit einer Breite W von einigen Abtastwerten lässt Abweichungen bei der Rastersuche zu (Abbildung 3). Zugleich liefert die jeweilige Amplitude ein Maß für die *Treffergenauigkeit*; mit diesem wird die Auf- bzw. Abwertung in (4) zusätzlich gewichtet.

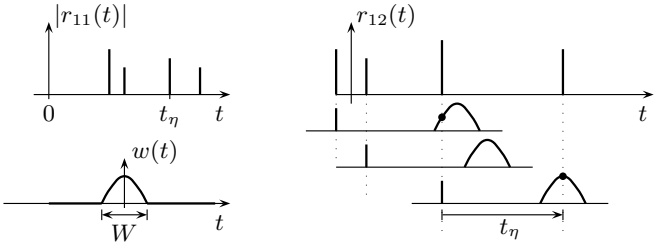

Abbildung 3: Suche nach Maxima in $r_{12}(t)$, deren Differenz näherungsweise zu einer Extremstelle t_η in $r_{11}(t)$ passt

Auch bei unbekannter Quellenanzahl kann nun anhand von $q(t_\mu)$ eine Klassifikation durchgeführt werden:

$$t_\mu \text{ ist } \begin{cases} \text{Direktpfad-Maximum, falls } q(t_\mu) > A \cdot \min\limits_{t_\nu \in \mathbb{T}_M} r_{12}(t_\nu) \\ \text{Echopfad-Maximum, sonst.} \end{cases}$$

Für $A \leq 1$ ist diese recht konservativ; jedes fehlerhafte Verwerfen von Laufzeitdifferenzen bedeutet Informationsverlust für die weiteren Verarbeitungsstufen.

Mess-Auswertung

Abbildung 4 zeigt den interessanten Ausschnitt der Kreuzkorrelierten (GCC-PHAT) eines Signalblocks mit 4096 Abtastwerten bei 96 kHz. Zwei Sprachquellen waren während der Messung in einem Raum mit $T_{60} \approx 300\,\text{ms}$ zeitgleich aktiv. Die tatsächlichen Laufzeitdifferenzen bei Wert 21 bzw. Wert 327 sind mit gepunkteten Linien markiert; nur die Laufzeitdifferenz der einen Quelle befindet sich unter den drei größten Maxima, markiert sind hier die 15 größten.

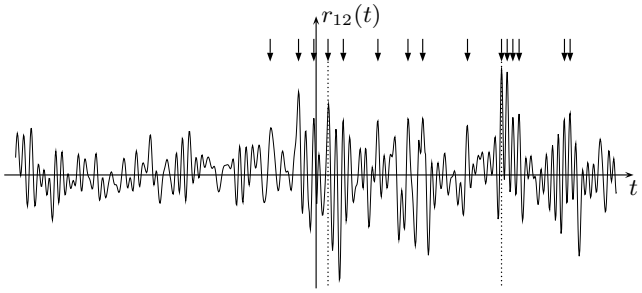

Abbildung 4: Kreuzkorrelierte eines gemessenen Signalblocks: Pfeile markieren die 15 größten Maximumstellen.

Die Autokorrelierten liefern Extremstellen bei 10, 21, 28 und 142 ($r_{11}(t)$), bzw. 35, 52, 79 und 224 ($r_{22}(t)$). Das entsprechend gefundene Raster (Toleranzbreite $W = 7$) ist in Abbildung 5 dargestellt. Durch zufällige Treffer ist Maximumstelle 438 die einzige ohne Pfeilspitze. Bei konservativer Klassifikation ($A = 1$) hingegen werden die Stellen $\mathbb{T}'_M = \{$-31, **21**, 48, 267, **327**, 337, 438$\}$ als potentielle Direktpfad-Laufzeitdifferenzen eingestuft. Gemessen über mehrere Signalblöcke reduziert das beschriebene Verfahren die Zahl betrachteter Maximumstellen im Durchschnitt um den Faktor 2 bis 4.

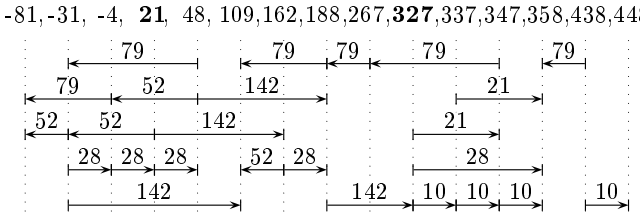

Abbildung 5: Gefundenes Raster: In der oberen Zeile stehen die quantisierten Maximumstellen aus Abbildung 4, die Zuordnungspfeile darunter sind mit näherungsweise passenden Extremstellen der Autokorrelierten beschriftet.

Literatur

[1] C. Knapp and G. Carter. The gereralized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech and Signal Proc.*, 24:320–327, 1976.

[2] Y. Huang, J. Benesty, and J. Chen. A blind channel identification-based two-stage approach to seperation and dereverberation of speech signals in a reverberant environment. *IEEE Trans. on Speech and Audio Processing*, 13:882–895, 2005.

[3] J. Scheuing and B. Yang. Disambiguation of TDOA estimates in multi-path multi-source environments (DATEMM). In *ICASSP*, 2006.