Binaurale Störgeräuschunterdrückung für digitale Hörgerätealgorithmen

V. Hohmann und G. Grimm

AG Medizinische Physik und Kompetenzzentrum „HörTech“
Carl-von-Ossietzky Universität Oldenburg, D–26111 Oldenburg

Ziel der vorliegenden Studie ist es, die Realisierbarkeit eines auf einer einfachen Quellentrennung basierten SNR-Schätzers (SNR: Signal-Rausch-Abstand) zu prüfen. Die Quellentrennung nutzt die interaurale Zeitdifferenz zur Lokalisation (Richtungsfilterung). Der darauf basierte SNR-Schätzer funktioniert auf sehr kurzen Zeitskalen und soll somit eine Unterdrückung von fluktuierenden Störgeräuschen ermöglichen.

I. Quellentrennung, Richtungsfilter

Die Aufgabe des Richtungsfilters ist es, Signale aus einer auserwählten Richtung zu erkennen. Diese *target*-Richtung ist in der Regel von vorne, aber auch andere Richtungen sind realisierbar.

Für eine Entscheidung, ob ein Frequenzbin zur *target*-Quelle gehört oder nicht, werden im wesentlichen vier Stufen benötigt:

1. Fourier-Transformation in beiden Mikrofonkanälen (binaurales Signal).

2. Berechnung der interauralen Phasendifferenz (IPD) für jedes Frequenzbin einzeln sowie Korrektur des eventuell frequenzabhängigen Phasenunterschiedes der Apparatur bei Signalen von vorne ohne Reflexionen.

3. Berechnung der interauralen Zeitdifferenz (ITD) für jedes Bin, also Division durch die zum Bin gehörende Winkelfrequenz . Diese Operation ist nur für Frequenzen unterhalb der Grenzfrequenz f_g eindeutig, welche durch die vom Mikrofonabstand gegebene maximale Zeitdifferenz bestimmt wird: $f_g = 1/t_{max}$. Für die weitere Berechnung können also nur Frequenzbins unterhalb dieser Grenzfrequenz verwendet werden. Der Abstand menschlicher Ohren entspricht etwa einer Grenzfrequenz von 900 Hz. Ein Histogramm dieses Parameters ist in Abb. 1 (links) zu sehen.

4. Binäre Entscheidung, ob ein Frequenzbin zur *target*-Quelle gehört oder nicht. In einer Tabelle sind die Wahrscheinlichkeiten für die Zugehörigkeit zur *target*-Quelle über der ITD gespeichert (Abb. 1, rechts). Liegt die Wahrscheinlichkeit über der Grenzwahrscheinlichkeit p_g (typischerweise 50%), so gehört das Frequenzbin zur *target*-Quelle (Bayes-Schätzung).

II. SNR-Schätzung in den Frequenzbändern

Für eine SNR-Schätzung in einem Frequenzband $(f_1, f_2]$ wird die Intensität der als *target*-Quelle erkannten Bins sowie der verbleibenden Bins getrennt im Frequenzband summiert. d_{targ} ist entweder die binäre Entscheidung, ob ein Bin zur *target*-Quelle gehört (1) oder nicht (0), oder alternativ die kontinuierliche Wahrscheinlichkeit, daß das Frequenzbin zur *target*-Quelle gehört.

$$I(t)_{targ} = \sum_{f_1 \le f < f_2} I(f,t) \cdot d_{targ} \quad (1)$$

$$I(t)_{noise} = \sum_{f_1 \le f < f_2} I(f,t) \cdot (1 - d_{targ}) \quad (2)$$

Der SNR in diesem Frequenzband ist dann

$$SNR(t) = \frac{I(t)_{targ}}{I(t)_{noise}}. \quad (3)$$

Entspricht die Breite des Frequenzbandes der Frequenzauflösung der FFT, so ist der SNR ist nur noch durch d_{targ} bestimmt.

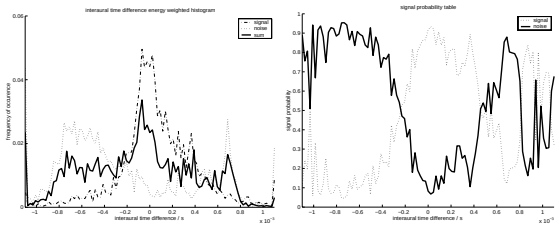

Abbildung 1: *Linkes Bild:* **Energiegewichtete Histogramme der interauralen Zeitdifferenzen (ITD) für Nutz-, Stör- und Summensignal. Es ist zu sehen, daß das Nutzsignal von vorne kommt (kleine interaurale Zeitdifferenzen), während das Störsignal überwiegend von links kommt (negative interaurale Zeitdifferenzen). Die Streuung des Parameters und das zweite Maximum bei pos. Differenzen werden von Reflexionen im Raum verursacht. In reflexionsarmen Umgebungen können sehr viel schärfere Histogramme erreicht werden.** *Rechtes Bild:* **Aus den ITD-Histogrammen abgeleitete Tabelle mit den Wahrscheinlichkeiten für die Entscheidung, ob ein Frequenzbin mit der interauralen Zeitdifferenz ITD zum Nutzsignal (target-Quelle) oder zum Störsignal gehört.**

III. Signaldatenbank und Apparatur

Für die Entwicklung und Bewertung des Algorithmus wurde eine Signaldatenbank zusammengestellt, welche aus möglichst repräsentativen akustischen Situationen besteht. Um die *a priori*-SNRs bestimmen zu können, wurden Nutz- und Störsignal jeweils getrennt aufgenommen und später gemischt. Dabei wurden drei Raumrichtungen verwendet: Richtung *m* entspricht 0° Azimuth aus 1,2 m Entfernung, Richtung *l* entspricht 60° Azimuth aus 1,2 m Entfernung, Richtung *r* entspricht $-60°$ Azimuth aus etwa 4 m Entfernung und indirekter Beschallung. Das Kürzel *l2* entspricht einer Richtung von etwa 60° Azimuth. Alle Aufnahmen aus Richtung *m*, *l* und *r* wurden in einem kleinen Wohnraum aufgenommen. Das Signal `12-sp-m-03-hall` wurde in einer stark verhallten Umgebung aufgenommen. Die Signale mit dem Kürzel *d* sind diffuse Störgeräusche aus unterschiedlichen Räumen. Es wurden Sprachsignale (*sp*), künstliche Störgeräusche (*art*) und natürliche Störgeräusche verwendet.

Alle Geräusche wurden mit einem Bügelmikrophon aufgenommen, dessen interauralen Eigenschaften (interaurale Zeitdifferenzen, Kopfabschattung) etwa denen eines beidseitig getragenem Hörgerätes entsprechen. Die Abtastrate beträgt 22050 Hz.

IV. Kriterien für die Machbarkeit der SNR-Schätzung

IV.1 Das Konzept der disjoined orthogonality

Das Kriterium der *disjoined orthogonality* (djo; [1] ist eine notwendige Bedingung an die SNR-Schätzung nach dem oben vorgestellten Verfahren.

Zwei Signale sind *disjoined orthogonal*, wenn das elementweise Produkt der Spektrogramme $s(t,f)$ überall verschwindet. Um ein Maß

für die Orthogonalität c_{djo} zu erhalten, wird das punktweise Produkt auf die Leistung $p(t,f)$ normiert:

$$s_{djo}(t,f) = |X_1(t,f) \cdot X_2(t,f)| \quad (4)$$

$$p(t,f) = \frac{1}{2}(|X_1(t,f)|^2 + |X_2(t,f)|^2) \quad (5)$$

Um die Anwendbarkeit des Algorithmus auf Signalpaare x_1 und x_2 zu testen, wird das globale djo-Kriterium

$$c_{djo} = \frac{\int\int s(t,f)\,\mathrm{d}f\,\mathrm{d}t}{\int\int p(t,f)\,\mathrm{d}f\,\mathrm{d}t} \quad (6)$$

berechnet (siehe Abb. 2).

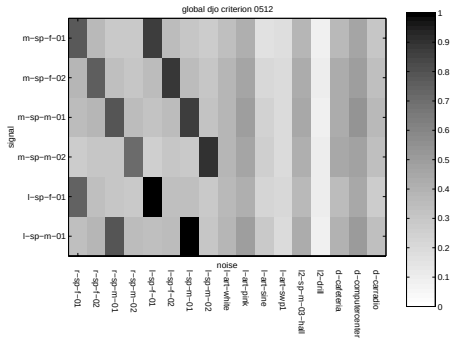

Abbildung 2: Globales disjoined orthogonality-Kriterium für einige ausgewählte Signalkombinationen. Hohe Werte dieses Kriteriums bedeuten, daß die Signalpaare nicht orthogonal sind und somit eine SNR-Schätzung nicht erfolgen kann. Besonders hohe Werte treten dann auf, wenn die Spektren der Signale nahezu identisch sind und die Signale sich nur durch die Einfallrichtung unterscheiden (linke Bildhälfte). Weiter führen auch breitbandige, diffuse Signale wie „computercenter" oder Cafeteria-Geräusch zu hohen Werten des djo-Kriteriums.

Um einen kleinen Fehler bei der SNR-Schätzung zu erreichen, müssen Nutz- und Störsignal möglichst orthogonal sein, d.h. der Wert des *disjoined orthogonality*-Kriterium c_{djo} muss möglichst klein sein. Dies ist beispielsweise dann erfüllt, wenn Nutz- und Störsignal vor allem aus tonalen Anteilen besteht. In den meisten getesteten Signal-Kombination ist die Überlappung jedoch groß (≥ 0.25). Abb. 3 zeigt dazu das ITD-Histogramm aus Abb. 1 parametrisiert mit dem c_{djo}-Maß. Für Werte größer als 0.25 läßt sich über die ITD aufgrund der Überlappung der Signale keine Trennung der frontalen Quelle erreichen. Die gilt gemäß Abb. 2 für fast alle Hörsituationen, so daß geschlossen werden muß, daß die SNR-Schätzung basierend auf Beobachtungen von Zeit- und Frequenzlokalisierten Signalparametern (Punkte im Spektrogramm) nicht möglich ist.

Die FFT-Länge hat einen wesentlichen Einfluß auf das *disjoined orthogonality*-Kriterium (hier nicht gezeigt). Je kürzer die FFT-Länge, desto geringer ist die spektrale Auflösung, wodurch die spektrale Überlappung steigt. Jedoch führen lange FFT-Längen zu verschlechterter Zeitauflösung und die Orthogonalität steigt für interessante Störsignale, wie etwa Cafeteria-Geräusch, nicht ausreichend.

IV.2 ITD Histogramme, Nutzsignalfehler

Die Zugehörigkeit eines Frequenz-Bins zum Nutzsignal wird anhand der interauralen Zeitdifferenz (ITD) bestimmt. Es ist daher für die Realisierbarkeit der SNR-Schätzung neben dem djo-Kriterium ferner auch von Bedeutung, inwiefern die ITD-Werte des Nutzsignals sich wirklich von den ITD des Störsignals unterscheiden. Dazu wurden energiegewichtete Histogramme der ITD sowohl für das Nutz- als auch für das Störsignal berechnet (hier nicht gezeigt). Die Schnittfläche dieser beiden Histogramme E_{itd} entspricht dem Anteil der nicht korrekt klassifizierbaren Energie, wobei hier zwischen den ITD-Bereichen des Nutz- und Störsignals unterschieden wird. Die Fläche E_{itd} im ITD-Bereich des Störsignals ist der *Nutzsignalfehler*, da dies der Energie des Nutzsignals entspricht, welche dem Störsignal zugeordnet wird. Die Fläche E_{itd} im ITD-Bereich des Nutzsignals entspricht umgekehrt dem Störsignalfehler.

In der realen Anwendung spielt der Nutzsignalfehler eine größere Rolle als der Störsignalfehler, da ein hoher Nutzsignalfehler unmittelbar zu Informationsverlust führt, während ein hoher Störsignalfehler nur einem reduziertem SNR entspricht.

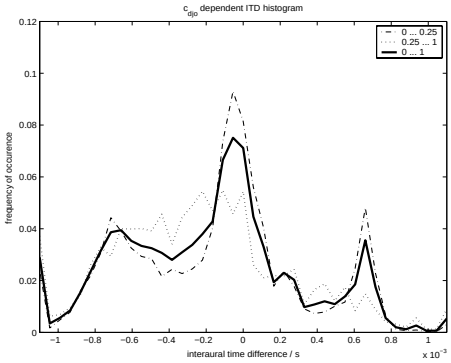

Abbildung 3: Energiegewichtete Histogramme der interauralen Zeitdifferenzen (ITD) für Nutz-, Stör- und Summensignal parametrisiert nach dem c_{djo}-Maß. Für c_{djo}-Werte kleiner 0.25 sind Nutz- und Störsignal mittels der ITD trennbar. Für Werte größer 0.25 ergibt sich jedoch nur eine eingipflige, breite Verteilung, so daß die binaurale Information nicht zur Trennung ausreicht.

V. Diskussion

In dieser Studie wurden die Anforderungen an eine binaurale Störgeräuschbefreiung gezeigt. Eine binaurale Störgeräuschschätzung ist grundsätzlich nur dann möglich, wenn alle hier gezeigten Kriterien erfüllt sind. Das wichtigste Kriterium ist dabei das der *disjoined orthogonality*. Dieses Kriterium ist für ein Gemisch von wenigen Sprechern erfüllt. Bei breitbandigen Störgeräuschen wie beispielsweise Cafeteria-Lärm ist es nicht mehr erfüllt. Aufgrund der hier gefundenen starken Überlappung von Nutzsignal und Störsignal im Spektrogramm erscheint es notwendig, binaurale Information über die Frequenz zu einem Zeitpunkt zu integrieren, um robuste Richtungsinformation zu erhalten. Damit ist es aber nicht möglich den SNR frequenzabhängig zu schätzen, so daß in diesem Fall auf andere Methoden der Störgeräuschunterdrückung zurückgegriffen werden muss. Möglich wäre etwa die Steuerung eine adaptiven Richtmikrofons zur Unterdrückung der Störquelle mit dem aktuell höchsten Pegel. Für diese Steuerung ist keine frequenzabhängige, sondern nur eine globale Pegelschätzung notwendig.

Die hier verarbeiteten Signale wurden mit einem Mikrofonabstand von etwa 22 cm aufgenommen. Die dadurch bedingten maximalen Laufzeitdifferenzen betragen etwa 1 ms, oberhalb der dadurch gegebenen Grenzfrequenz von etwa 900 Hz ist die Zuordnung von Phase zu interauraler Zeitdifferenz nicht mehr eindeutig, es kann keine SNR-Schätzung mehr erfolgen. Ein kleinerer Mikrofonabstand würde diese Grenzfrequenz nach oben verschieben und somit den schätzbaren Frequenzbereich vergrößern. Eine weitere Möglichkeit wäre, oberhalb der Grenzfrequenz nicht die interauralen Phasendifferenzen des Signales, sondern die der Einhüllenden zu berechnen. Diese wären wiederum eindeutig und können einer interauralen Zeitdifferenz zugeordnet werden.

[1] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separa- tion of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In Proceedings of the 2000 IEEE Con- ference on Acoustics, Speech, and Signal Processing (ICASSP2000), volume 5, pages 2985–2988, Istanbul, Turkey, 2000.

Unterstützt durch BMBF-Kompetenzzentrum 'HörTech'