Schätzung des Restechos mit Hilfe eines Mikrofonarrays

Markus Kallinger, Jörg Bitzer, Karl-Dirk Kammeyer
Universität Bremen, kallinger@ant.uni-bremen.de

ABSTRACT

Der Echokompensator stellt im systemtheoretischen Sinn die optimale Lösung zur Unterdrückung akustischer Echos in Freisprecheinrichtungen dar. In einer realen Anwendung und unter Berücksichtigung von Störungen können aber zusätzliche Maßnahmen zur Dämpfung der Echos notwendig werden. Eine dieser Möglichkeiten besteht in der Verwendung adaptiver Post-Filter. Diese werden meistens zusammen mit einem Echokompensator eingesetzt. Um ein Post-Filter richtig zu entwerfen, muss das Auto-Leistungsdichtespektrum des Restechos, das am Ausgang des Echokompensators verbleibt, genau geschätzt werden. In diesem Beitrag wird zu diesem Zweck ein neues Schätzverfahren vorgestellt, das auf einem Mikrofonarray basiert. Die Schätzwerte können vor allem bei dem Auftreten von additiven Interferenzen, zu denen ein naher Sprecher zählt, starke Abweichungen aufweisen. Unter Ausnutzung des bekannten Minimum Statistik-Verfahrens sowie räumlicher Informationen kann jedoch die Robustheit der Schätzwerte deutlich erhöht werden. Auch in Situationen mit starkem Gegensprechen kann das Restecho zuverlässig geschätzt werden.

1. EINLEITUNG

Die Kombination von Echokompensatoren mit Beamformern bietet sich in Freisprecheinrichtungen an, wenn eine sehr hohe Qualität des zu übertragenden Sprachsignals gefordert ist. Dabei können mehrere Echokompensatoren, die auf einem Referenzkanal beruhen, dem Beamformer vorgeschaltet sein. Die Alternative besteht darin, einen Echokompensator am Ausgang des Beamformers zu betreiben. Allerdings müssten dann mehrere Echokompensatoren parallel für mehrere „diskrete“ Blickrichtungen des Beamformers betrieben werden [1]. In beiden Fällen steigt der Rechenaufwand um ein Vielfaches.

Abhilfe kann geschaffen werden, indem man die Echokompensatoren verkürzt und dafür zusätzlich ein Post-Filter einsetzt (die hier betrachtete Verschaltung ist in Bild 1 skizziert). Für den Entwurf eines Post-Filters nach Wiener muss jedoch das Auto-Leistungsdichtespektrum (ALDS) des Restechos nach dem Beamformer bekannt sein [2]. Das Restecho im $i$ten Mikrofonkanal lässt sich anhand des be-

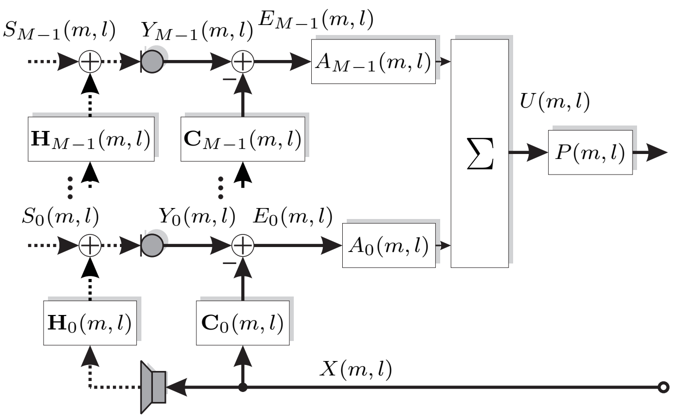

Bild 1. Signalmodell im Frequenzbereich mit M Echokompensatoren vor einem Beamformer und einem Post-Filter.

trachteten Signalmodells ausdrücken als $\Xi_i(m,l) = E_i(m,l) - S_i(m,l)$. m ist der diskrete Frequenzindex und l der Blockindex in zeitlicher Richtung. Das Restecho

$$\Xi_B(m,l) = \sum_{i=0}^{M-1} A_i(m,l)\Xi_i(m,l) \tag{1}$$

als Anteil des Signals $U(m,l)$ muss am Ausgang des Beamformers geschätzt werden, um das Post-Filter zu entwerfen. Um den unterschiedlichen Ordnungen der betrachteten Raumimpulsantworten und Echokompensatoren Rechnung zu tragen, wird entsprechend der partitionierten schnellen Faltung [3] für deren Übertragungsfunktionen $\mathbf{H}_i(m,l)$, bzw. $\mathbf{C}_i(m,l)$ eine Schreibweise mit den Vektoren

$$\begin{aligned}
\mathbf{H}_i(m,l) &= \left[H_{i,0}(m,l) \quad \cdots \quad H_{i,L'_H-1}(m,l)\right], && (2)\\
\mathbf{C}_i(m,l) &= \Big[C_{i,0}(m,l) \quad \cdots \quad C_{i,L'_{EK}-1}(m,l) \\
&\qquad 0 \quad \cdots \quad 0\Big], && (3)\\
\mathbf{D}_i(m,l) &= \mathbf{H}_i(m,l) - \mathbf{C}_i(m,l), && (4)\\
\mathbf{X}(m,l) &= \left[X(m,l) \quad \cdots \quad X(m,l-L'_H+1)\right]^T. && (5)
\end{aligned}$$

verwendet. $L_H = L'_H L_{DFT}$ und $L_{EK} = L'_{EK} L_{DFT}$ sind die Längen der modellierten Raumimpulsantworten, bzw. Echokompensatoren. L_{DFT} ist die verwendete DFT-Länge. Nachdem das Restecho $\Xi_B(m,l)$ linear mit dem Referenzsignal $X(m,l)$ verknüpft ist, muss schließlich die partitionierte Übertragungsfunktion des gesamten Systemfehlers

$$\mathbf{D}_B(m,l) = \sum_{i=0}^{M-1} \left(\mathbf{H}_i(m,l) - \mathbf{C}_i(m,l)\right) A_i(m,l) \tag{6}$$

geschätzt werden, um das Restecho zu bestimmen.

2. SCHÄTZUNG DES SYSTEMFEHLERS

Die Übertragungsfunktion des Systemfehlers kann mit Hilfe der Wiener-Hopf Gleichung bestimmt werden.

$$\hat{\mathbf{D}}_B(m,l) = \hat{\mathbf{\Phi}}_{XU}(m,l) \otimes \hat{\mathbf{\Phi}}_{XX}^{-1}(m,l); \tag{7}$$

$\otimes$ kennzeichnet die elementweise Vektor-Multiplikation. Der Vektor

$$\hat{\mathbf{\Phi}}_{XX}^{-1}(m,l) = \left[\hat{\Phi}_{XX}^{-1}(m,l) \quad \cdots \quad \hat{\Phi}_{XX}^{-1}(m,l-L'_{SFS}+1)\right] \tag{8}$$

mit den inversen geschätzten Auto-Leistungsdichtespektren kann nur unter der Annahme eines unkorrelierten Referenzsignals $X(m,l)$ verwendet werden. Mit Hilfe von L'_{SFS} kann eingestellt werden, auf welcher Länge der Systemfehler geschätzt werden soll. Im Allgemeinen sollte die Beziehung $L'_{EK} < L'_{SFS} < L'_H$ gelten. Der Vektor $\hat{\mathbf{\Phi}}_{XU}(m,l)$ ist ähnlich definiert. Der Unterschied besteht darin, dass in seinem $j$en Element $U(m,l)$ mit $X(m,l-j+1)$ korreliert wird. Alle geschätzten ALDS werden mit Hilfe der Welch-Methode und einer rekursiven Glättung berechnet.

2.1. Robuste Verfahren

Im Beamformer-Ausgangsignal $U(m,l)$ ist das Sprachsignal $S(m,l)$ enthalten, welches Fehlschätzungen verursacht. Aufgrund der Instationarität von Sprache betreffen diese Störungen lediglich einzelne Teilbänder und dies i.d.R. nicht länger als 400 ms. Deshalb kann das Schätzverfahren auf Grundlage der Wiener-Hopf Gleichung in den betroffenen Teilbändern angehalten werden, während ein naher Sprecher aktiv ist.

Ein Verfahren zur Ermittlung gestörter Frequenzbänder stellt ein einkanaliges Verfahren dar, das auf der bekannten Minimum Statistik basiert und in [4] vorgestellt wurde.

Gestörte Frequenzbänder können auch unter Ausnutzung räumlicher Informationen gefunden werden [4]. Dabei soll zunächst das *Array Gain* am Beamformer betrachtet werden, das nach

$$G_A(m,l) = \frac{\mathrm{SNR}_{\mathrm{Array}}(m,l)}{\mathrm{SNR}_{\mathrm{Microphone}}(m,l)} \approx \frac{\bar{\Phi}_{\Xi\Xi}(m,l)}{\Phi_{\Xi_B\Xi_B}(m,l)} \tag{9}$$

definiert ist. Bei der Vereinfachung wurde vorausgesetzt, dass das Sprachsignal kohärent ist und daher durch den Beamformer nicht beeinflusst wird. $\bar{\Phi}_{\Xi\Xi}(m,l)$ ist das linear über alle Mikrofonkanäle gemittelte ALDS des Restechos vor dem Beamformer. Wenn nun weiterhin angenommen wird, dass das Geräuschfeld, das durch das Restecho erzeugt wird, diffus ist, so ergibt sich das Array Gain zum so genannten *Directivity Factor* $\mathrm{DF}(m)$ [5], der nur von den Beamformer Koeffizienten $A_i(m,l)$ abhängt. Es entsteht der Zusammenhang

$$\Phi_{\Xi_B\Xi_B}(m,l) = \mathrm{DF}^{-1}(m)\bar{\Phi}_{\Xi\Xi}(m,l). \tag{10}$$

In einem realen System kann vor und nach dem Beamformer nur auf die Signale $E_i(m,l)$ und $U(m,l)$ zugegriffen werden. Bildet man den Quotienten aus den entsprechenden Auto-Leistungsdichtespektren, so lässt sich die Umformung

$$\begin{aligned}\frac{\Phi_{UU}(m,l)}{\bar{\Phi}_{EE}(m,l)} &= \frac{\Phi_{\Xi_B\Xi_B}(m,l)+\Phi_{SS}(m,l)}{\bar{\Phi}_{\Xi\Xi}(m,l)+\Phi_{SS}(m,l)} \\ &= \frac{\mathrm{DF}^{-1}(m)\bar{\Phi}_{\Xi\Xi}(m,l)+\Phi_{SS}(m,l)}{\bar{\Phi}_{\Xi\Xi}(m,l)+\Phi_{SS}(m,l)} \\ &= \frac{\mathrm{DF}^{-1}(m)+\mathrm{SRER}(m,l)}{1+\mathrm{SRER}(m,l)} \\ &> \mathcal{T}_{DF} \end{aligned} \tag{11}$$

vornehmen. $\mathrm{SRER}(m,l)$ ist das Signal-zu-Restecho Verhältnis (*signal-to-residual echo ratio*)

$$\mathrm{SRER}(m,l) = \frac{\Phi_{SS}(m,l)}{\bar{\Phi}_{\Xi\Xi}(m,l)}. \tag{12}$$

Für große SRERs erreicht der Quotient Werte nahe 1 und überschreitet dabei den Schwellwert $\mathcal{T}_{DF}$. Ein solcher Fall tritt ein, sobald ein naher Sprecher aktiv ist. Ist nur der ferne Sprecher aktiv, so nimmt der Quotient Werte an, die nahe beim inversen Directivity Factor $\mathrm{DF}^{-1}(m)$ liegen. Ein Problem kann sich bei tiefen Frequenzen ergeben, wenn sich der Directivity Factor an 1 annähert. Der Schwellwert $\mathcal{T}_{DF}$ kann in Abhängigkeit des Directivity Factors sowie eines zu wählenden SRERs angegeben werden (siehe Gleichung (11)).

Mit Hilfe der zuvor beschriebenen Fallunterscheidung können Teilbänder $m_{a,j}^{DF}$ gefunden werden, in denen die Schätzung des Systemfehlers nach der Wiener-Hopf Gleichung anzuhalten ist. Wird nun parallel das auf Minimum Statistik beruhende Verfahren angewandt, lassen sich ebenso Teilbänder $m_{a,j}^{MS}$ finden. Beide Verfahren können verbunden werden, indem die Vereinigungsmenge aller gestörten Teilbänder gebildet wird

$$\mathbb{M}_{a,MS} = \left\{ m_{a,1}^{MS},\ \ldots,\ m_{a,q}^{MS} \right\} \tag{13}$$

$$\mathbb{M}_{a,DF} = \left\{ m_{a,1}^{DF},\ \ldots,\ m_{a,r}^{DF} \right\} \tag{14}$$

$$\mathbb{M}_{a} = \mathbb{M}_{a,MS} \bigcup \mathbb{M}_{a,DF}. \tag{15}$$

3. SIMULATIONSERGEBNISSE

Für die nachfolgend dargestellten Simulationsergebnisse wurde ein superdirektiver Beamformer mit 4 Mikrofonen in Endfire-Anordnung verwendet, um einen hinreichend großen Directivity Factor bei tiefen Frequenzen zu erhalten. In diesem Fall handelt es sich um einen Directivity Factor von 4 bei 200 Hz. Beim Entwurf des superdirektiven Beamformers wurde ein Leistungsverhältnis des Signals zum Sensorrauschen von 30 dB angenommen. Die mit Hilfe der Spiegelbild-Methode simulierten Raumimpulsantworten haben eine Länge von 4096 Abtastwerten und eine Nachhallzeit von $\tau_{60} = 400$ ms. Wie in Bild 1 angedeutet befindet sich hinter jedem Mikrofon ein Echokompensator. Es kamen Filter der Länge 512 zum Einsatz, die mit Hilfe eines Affinen Projektions-Algorithmus der Ordnung 3 adaptiert werden. In Bild 2 sind geschätzte Signalleistungen des Restechos für

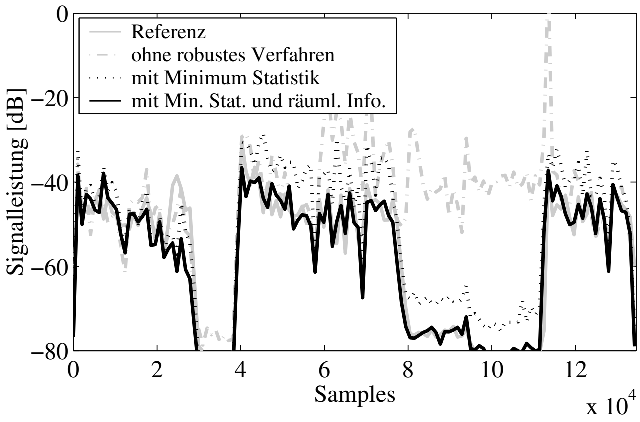

Bild 2. Geschätzte Breitband-Signalleistung des Restechos $\Xi_B(m,l)$. Sowohl für den nahen als auch für den fernen Sprecher wurden Sprachsignale verwendet. Der nahe Sprecher ist zwischen Sample 30.000 und 40.000 sowie zwischen Sample 58.000 und 122.000 aktiv.

mehrere Verfahren aufgetragen. Besonders in der Phase mit Gegensprechen zwischen Sample 58.000 und 122.000 lässt sich die Wirksamkeit des neuen robusten Schätzverfahrens demonstrieren. Nur unter Verwendung räumlicher und statistischer Informationen lässt sich bei Gegensprechen ein starker Bias vermeiden.

4. ZUSAMMENFASSUNG

In diesem Beitrag wurde ein neues Verfahren zur Schätzung des Restechos für den Entwurf eines Post-Filters vorgestellt. Das Verfahren stützt sich sowohl auf statistische als auch auf räumliche Informationen. In Phasen mit „Double-Talk“ werden robuste Schätzergebnisse erzielt, die den Entwurf von Post-Filern mit einer sehr guten Sprachqualität ermöglichen (s. www.ant.uni-bremen.de/research/speech/).

5. LITERATUR

[1] W. L. Kellermann, “Acoustic Echo Cancellation for Beamforming Microphone Arrays,” in *Microphone Arrays: Signal Processing Techniques and Applications* (M. S. Brandstein and D. Ward, eds.), ch. 13, pp. 281–306, Springer-Verlag, 2001.

[2] G. Enzner, R. Martin, and P. Vary, “Unbiased Residual Echo Power Estimation for Hands–Free Telephony,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, (Orlando, Florida, USA), May 2002.

[3] J.-S. Soo and K. Pang, “Multidelay Block Frequency Domain Adaptive Filter,” *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 38, pp. 373–376, Feb 1990.

[4] M. Kallinger, J. Bitzer, and K. D. Kammeyer, “Multi-Microphone Residual Echo Estimation,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, (Hong Kong, China), Apr 2003. accepted.

[5] J. Bitzer and K. U. Simmer, “Superdirective microphone arrays,” in *Microphone Arrays: Signal Processing Techniques and Applications* (M. S. Brandstein and D. Ward, eds.), ch. 2, pp. 19–38, Springer-Verlag, 2001.