ARRAYTECHNIKEN ZUR GERÄUSCHREDUKTION BEI SPRACHÜBERTRAGUNG

Kristian Kroschel

Institut für Automation und Robotik, Universität Karlsruhe
Fraunhofer Institut für Informations– und Datenverarbeitung
IITB, Karlsruhe

1. Einleitung

Bei der Sprachübertragung spielen Geräusche in der Umgebung des Sprechers immer dann eine Rolle, wenn der Kommunikationspartner dadurch Verständigungsprobleme hat. Daß dieses Thema erst in neuerer Zeit aktuell geworden ist, liegt daran, daß man Anwendungsbereiche der Kommunikation erschließt, bei denen das Signal–zu–Rauschverhältnis am Mikrophonausgang nicht dadurch in einen akzeptablen Bereich gebracht wird, daß man z.B. einen Handapparat wie beim Telefon oder ein Kopfgeschirr wie im Cockpit eines Flugzeuges verwenden kann oder akzeptiert. Beispiele für diese Anwendungsfälle sind Freisprecheinrichtungen beim Autotelefon oder bei Auskunftsystemen. Dieselbe Problematik stellt sich bei Spracheingabesystemen, bei denen durch die Geräuschumgebung die Erkennungsrate ohne Maßnahmen zur Geräuschreduktion auf unakzeptable Werte abfällt.

Es gibt mehrere Möglichkeiten zur Geräuschreduktion. Interpretiert man die Trennung von Sprache und Geräusch als Schätzaufgabe für den Sprachsignalprozeß, so stellt das Wiener–Filter als Schätzsystem die ideale Lösung [1] dar. Paßt man das Wiener–Filter an die Realisierungsmöglichkeiten an, gelangt man zur Spektralsubtraktion [2]. Dieses Verfahren ist recht robust, benötigt wenig Aufwand an Hardware, hat allerdings den Nachteil, daß die Sprache hörbar verzerrt wird und die Gewinne an Signal–zu–Rauschverhältnis mit etwa 5 dB recht gering sind [3]. Ein erheblich besseres Ergebnis läßt die Geräuschkompensation erwarten, bei der ein zweites Mikrophon als Referenz für die Störungen verwendet wird. Mit Hilfe des Signals im Referenzkanal wird die Störung im Sprachkanal z.B. mit Hilfe des LMS–Algorithmus geschätzt [4] und vom Signal im Sprachkanal subtrahiert. In der Praxis sind die Forderungen dieses Verfahrens, daß kein Sprachanteil im Referenzkanal vorhanden und die Störanteile in den beiden Kanälen stark miteinander korreliert sind, kaum zu realisieren. Beachtet man noch, daß der Aufwand an Hardware erheblich höher als bei der Spektralsubtraktion ist und daß die Gewinne an Signal–zu–Rauschverhältnis nur in Sonderfällen diejenigen der Spektralsubtraktion übersteigen, so empfiehlt sich der Einsatz der Geräuschkompensation nur in Sonderfällen. Auch eine Kombination der Verfahren Spektralsubtraktion und Geräuschkompensation zeigt keine wesentlichen Vorteile [5].

Neuere Ansätze zur Geräuschreduktion verwenden deshalb entweder andere Signalverarbeitungsverfahren, so z.B. neuronale Netze [6], oder mehrkanalige Eingabe in Form von Mikrofonarrays [7]. Beide Ansätze sind natürlich auch kombinierbar. Eine umfassende Analyse des Fragenkomplexes der Spracheingabe in Geräuschumgebung wird in dem ESPRIT–Projekt 2101 "Adverse Environment Recognition of Speech" der Europäischen Gemeinschaft angestrebt, was die Aktualität des Themas der Spracheingabe in Geräuschumgebung unterstreicht.

In den folgenden Abschnitten sollen einige Mikrophonarraysysteme näher vorgestellt werden, wobei die Zahl der verwendeten Mikrophone von 2 [8] bis 8 [7] variiert.

2. Arrays mit zwei Mikrophonen

Die Grundidee zu diesem Verfahren [9] von Allen et al. wurde dazu verwendet, Nachhall in Räumen zu reduzieren: das Sprachsignal wird in halligen Räumen von dessen Reflexionen überlagert, die man als Störung interpretieren kann, da sie die Verständlichkeit der

Sprache verschlechtern. Zur Trennung von Nutzschall und Reflexionen dient das in Bild 1 gezeigte System.

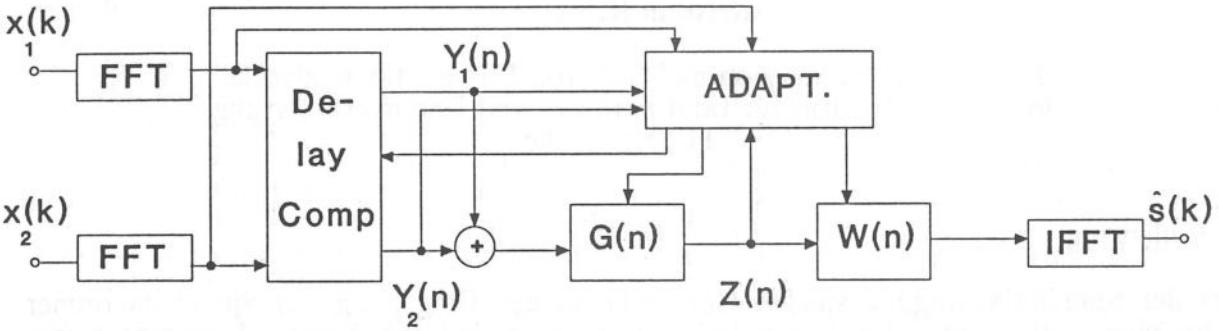

Bild 1 System zur Geräuschreduktion mit zwei Eingängen

An den zwei Eingängen stehen zwei gestörte Versionen $x_1(k)$ und $x_2(k)$ des mit der Frequenz f_a abgetasteten Sprachsignals $s(k)$ zur Verfügung, wobei

$$x_1(k) = s(k-k_1) + n_1(k) \tag{1a}$$

$$x_2(k) = s(k-k_2) + n_2(k) \tag{1b}$$

gilt, d.h. die Nutzanteile in den Kanälen sind um eine bestimmte Laufzeit k_i verzögert und durch Störkomponenten $n_i(k)$ verfälscht. Zur weiteren Signalverarbeitung werden die Signale in Blöcke zerlegt und der mit FFT im Blockschaltbild gekennzeichneten schnellen Fourier–Transformation unterworfen, wobei die einzelnen Blöcke vor der Transformation mit einem Hamming–Fenster [10] gewichtet und mit einer Überlappung von 1/2 oder 3/4 verarbeitet werden. Block für Block und damit adaptiv werden der Laufzeitausgleich und die Übertragungsfunktion G(n) berechnet, wobei der Laufzeitausgleich durch einen Allpaß mit der Übertragungsfunktion A(n) erfolgt, während G(n) einen frequenzabhängigen Verstärkungsfaktor darstellt. Für die Allpaßübertragungsfunktion wählt man

$$A(n) = \frac{X_1(n,i)\cdot X_2^*(n,i)}{|X_1(n,i)|\cdot|X_2(n,i)|}, \tag{2}$$

wobei $X_j(n,i)$ die FFT des i–ten Datenblocks der Länge N von $x_j(k)$ ist und n die diskrete Frequenzvariable $f\cdot N/f_a$ ist. Der Betrag von A(n) ist gleich eins, die Phase entspricht dem momentanen Laufzeitunterschied von $x_1(k)$ und $x_2(k)$.

Der frequenzabhängige Verstärkungsfaktor ist durch

$$G(n) = \frac{|\hat{S}_{y_1y_2}(n,i)|}{\hat{S}_{y_1y_1}(n,i)+\hat{S}_{y_2y_2}(n,i)} \tag{3}$$

gegeben, wobei der Zähler der rekursiv berechnete Schätzwert für die Kreuzleistungsdichte der laufzeitkorrigierten Prozesse $y_1(k)$ und $y_2(k)$

$$\hat{S}_{y_1y_2}(n,i) = \alpha\cdot\hat{S}_{y_1y_2}(n,i-1) + (1-\alpha)\cdot Y_1(n,i)\cdot Y_2^*(n,i) \tag{4}$$

ist und entsprechend im Nenner die Schätzwerte der Autoleistungsdichten von x_1 und x_2

stehen. Dabei ist α eine Glättungskonstante, die in Abhängigkeit von der Stationarität der Signalprozesse zu wählen ist. Die Wirkung des frequenzabhängigen Verstärkungsfaktors besteht darin, diejenigen Spektralkomponenten, für die $x_1(k)$ und $x_2(k)$ auf Grund der in ihnen enthaltenen Spektralkomponenten eine hohe Korrelation besitzen, zu verstärken, während diejenigen mit geringer Korrelation, die durch Störungen verursacht werden, zu dämpfen. Die Berechnung von A(n) für den Laufzeitausgleich nach (2) und von G(n) nach (3) erfolgt in dem mit ADAPT bezeichneten Block. Durch die Addition der Komponenten $Y_1(n)$ und $Y_2(n)$ werden die zeitlich synchronen Sprachanteile verstärkt und die unkorrelierten Störanteile geschwächt. Schließlich dient die mit IFFT bezeichnete inverse schnelle Fourier–Transformation und die Verkettung der Datenblöcke dazu, einen Schätzwert für das Sprachsignal zu gewinnen.

Eine Modifikation dieses Verfahrens wurde in [8] von Kaneda und Tohyama beschrieben. Die Eingangsstufe des Systems ist identisch mit der beim System von Allen. Der Laufzeitausgleich erfolgt hier jedoch nicht durch Ermitteln der momentanen Phase nach (2), vielmehr wird hier eine Glättungsoperation durch Berechnung der Kreuzkorrelationsfunktion vorgenommen, um den Einfluß der Störungen zu reduzieren. Die Lage des Maximums der geschätzten Kreuzkorrelationsfunktion entspricht dem Laufzeitunterschied der beiden Signale. Um diese genauer zu ermitteln, kann man neben dem Hauptmaximum die Nebenmaxima der Kreuzkorrelationsfunktion heranziehen. Die Kreuzkorrelationsfunktion ist bei fehlender Störung mit der verschobenen Autokorrelationsfunktion des Sprachsignalprozesses identisch, so daß die Maxima in einem bestimmten, vom Sprachsignalprozeß abhängigen Abstand aufeinanderfolgen. Durch Mittelung der Abstände der korrespondierenden Maxima der Kreuzkorrelationsfunktion und der Autokorrelationsfunktionen von x_1 und x_2 erhält man eine genauere Schätzung der Laufzeit. Aus Aufwandsgründen erfolgt die Berechnung im Frequenzbereich.

Ein zweiter Unterschied der Systeme von Allen und Kaneda betrifft die Berechnung der Filterkoeffizienten von G(n). Die Übertragungsfunktion G(n) ist hier der Quotient aus dem Realteil der geschätzten Kreuzleistungsdichte von x_1 und x_2 und der Wurzel aus dem Produkt der Betragsquadrate der geschätzten Autoleistungsdichten von x_1 und x_2, sofern dieser positiv ist. Im andern Fall wird dieser Wert für G(n) zu null. Die Schätzung der Auto– bzw. Kreuzleistungsdichten erfolgt wieder nach (4). Diese Festlegung von G(n) wird damit begründet, daß im Falle unkorrelierter Störungen $n_1(k)$ und $n_2(k)$ in den beiden Kanälen und bei korrektem Laufzeitausgleich der Realteil der Kreuzleistungsdichte positiv sein muß. Durch die endliche Mittelungsdauer für den Schätzwert nach (4) und Fehler bei der Laufzeitkorrektur trifft dies aber nicht zu, weswegen die Spektralkomponenten, die negativ sind, zu null gesetzt werden. Die weitere Signalverarbeitung ist wie im System nach Bild 1.

Die hier genannten Verfahren lassen sich mit der spektralen Subtraktion [1] kombinieren, die ebenfalls im Frequenzbereich arbeitet. Dazu ist nach dem Filter mit der Übertragungsfunktion G(n) ein weiteres mit der Übertragungsfunktion

$$W(n) = \begin{cases} [\hat{S}_{zz}(n,i) - a \cdot \hat{S}_{nn}(n,i)]/\hat{S}_{zz}(n,i) \; ; & \hat{S}_{zz}(n,i) > a \cdot \hat{S}_{nn}(n,i) \\ 0 & \text{sonst} \end{cases} \tag{5}$$

einzufügen. Der Faktor a liegt im Bereich von $2 \leq a \leq 4$ und dient zur Überkompensation der Störungen. Die Leistungsdichten des gestörten Sprachsignalprozesses $z(k)$ und des Störprozesses $n(k)$ werden entsprechend (4) rekursiv geschätzt. Man benötigt einen Sprachpausendetektor, der z.B. ein Leistungskriterium verwendet, um die rekursive Schätzung der Leistungsdichte des Störprozesses $n(k) = z(k)$ in den Sprachpausen durchführen zu können.

Es wurden Untersuchungen zur Störunterdrückung mit diesen Verfahren vorgenommen. Als Gütekriterium SNRE wurde der Quotient aus den Signal–zu–Rauschverhältnissen am Ausgang und am Eingang der Systeme gewählt. Das schlechteste Ergebnis wurde mit dem System von Allen ohne zusätzliche Spektralsubtraktion mit SNRE $\approx$ 5 dB erreicht, Spektralsubtraktion allein erzielte SNRE $\approx$ 8 dB. Das beste Ergebnis wurde mit dem System von Allen, zusätzlicher Spektralsubtraktion und einer Laufzeitkompensation über die aus der Kreuzkorrelationsfunktion durch Mittelung gewonne Laufzeit

erreicht und betrug SNRE ≈ 13,5 dB. Das System nach Kaneda hat dagegen den Vorteil, daß es auch ohne Laufzeitmittelung einen Wert von SNRE ≈ 9 dB liefert, der durch Spektralsubtraktion nur wenig auf SNRE ≈ 10 dB gesteigert werden kann.

3. Mikrophonarray mit nachgeschaltetem Wiener–Filter

Das von Zelinski [11] vorgestellte Verfahren bedient sich eines Mikrophonarrays von 4 Mikrophonen, um zum einen durch eine höhere Richtwirkung den Störanteil um etwa 6 dB zu senken und zum anderen eine Referenz für die Störung zu erhalten. Dies wird durch Addition und Subtraktion der laufzeitkompensierten Eingangskanäle erreicht, so daß Schätzwerte für die Leistungsdichten von Sprachsignal– und Störprozeß zur Verfügung stehen, die zur weiteren Signalschätzung des Sprachsignals mit Hilfe eines Wiener–Filters [1] herangezogen werden können. Das zugehörige System zeigt Bild 2.

Die erste Stufe, der Laufzeitausgleich, entspricht dem der Verfahren nach Allen und Kaneda. Entweder wird die Phase aus den Meßsignalen selbst oder nach Mittelung über die Kreuzkorrelationsfunktion bestimmt. Das Summensignal r(k), in dem sich das Sprachsignal zeitsynchron überlagert, während sich die unkorrelierten Störanteile gegenseitig um ca. 6 dB abschwächen, wird im Wiener–Filter W(n) weiter verarbeitet. Die Übertragungsfunktion des Wiener–Filters, das sich in zweifacher Ausführung mit identischer Übertragungsfunktion im System befindet, ist durch [1]

$$W(n) = \frac{S_{ss}(n)}{S_{rr}(n)} \approx \frac{S_{s_1 s_1}(n)}{S_{r_1 r_1}(n)} \tag{6}$$

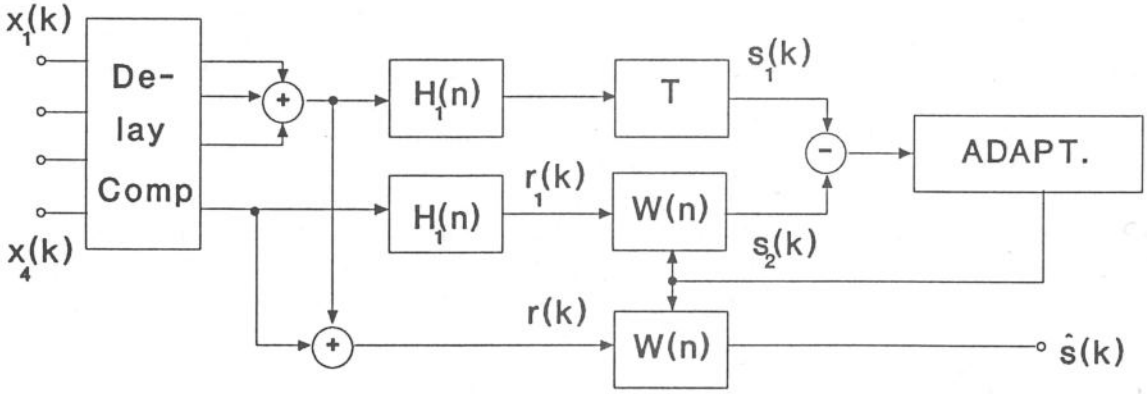

Bild 2 Mikrophonarray mit Wiener–Filter zur Geräuschreduktion

gegeben, wobei die Leistungsdichten des Sprachsignalprozesses $s(k)$ selbst und des gestörten Sprachsignalprozesses $r(k)$ das untere, die Leistungsdichten von $s_1(k)$ und $r_1(k)$ das obere Wiener–Filter bestimmen. Dabei entsteht $s_1(k)$ durch Mittelung dreier laufzeitkompensierter Mikrophonsignale und stellt damit eine Approximation des ungestörten Sprachsignals dar, während $r_1(k)$ das gestörte Sprachsignal eines Mikrophonzweiges ist. Das Ausgangssignal $s_2(k)$ des Wiener–Filters im oberen Zweig ist damit aber der Schätzwert von $s_1(k)$, der durch Filterung des gestörten Sprachsignals $r_1(k)$ gewonnen wird. Das Wiener–Filter wird durch ein Transversalfilter der Länge L approximiert

$$W(n) = \sum_{m=0}^{L} w_m z^{-m} \Big|_{z=e^{j2\pi n/N}} = \sum_{m=0}^{L} w_m e^{-j2\pi nm/N} \tag{7}$$

und die Filterkoeffizienten w_m so bestimmt, daß der mittlere quadratische Fehler

$$E\{e^2(k)\} = E\{[s_1(k) - s_2(k)]^2\} = E\{[s_1(k) - w(k) * r_1(k)]^2\}$$

$$= E\{[s_1^2(k)\} - 2\cdot E\{[s_1(k)\, \mathbf{r}_1^T(k)\}\cdot \mathbf{w}(k) + \mathbf{w}^T(k)\, \mathbf{S}_{r_1 r_1}(k)\, \mathbf{w}(k) \tag{8}$$

mit

$$\mathbf{r}_1(k) = \begin{bmatrix} r_1(k) \\ \vdots \\ r_1(k-L) \end{bmatrix} ;\ \mathbf{S}_{r_1 r_1}(k) = E\{\mathbf{r}_1 \mathbf{r}_1^T\} ;\ \mathbf{w}(k) = \begin{bmatrix} w_0(k) \\ \vdots \\ w_L(k) \end{bmatrix} \tag{9}$$

zum Minimum wird. Die im Vektor $\mathbf{w}(k)$ zusammengefaßten Filterkoeffizienten werden rekursiv mit Hilfe des Least–Mean–Square– oder LMS–Algorithmus [4] berechnet

$$\begin{aligned} \mathbf{w}(k+1) &= \mathbf{w}(k) + 2\,\mu(k)\cdot E\{[e(k)\, \mathbf{r}_1(k)\} \\ &\approx \mathbf{w}(k) + 2\,\mu(k)\cdot s_2(k)\cdot \mathbf{r}_1(k) \quad . \end{aligned} \tag{10}$$

Der Adaptionsfaktor $\mu(k)$ ist im optimalen Fall gleich dem Kehrwert des größten Eigenwerts der Korrelationsmatrix von $\mathbf{r}_1(k)$ nach (9). Da dessen Berechnung sehr aufwendig ist, verwendet man eine geeignete Näherung [11]. Der Algorithmus (10) wird in dem mit ADAPT bezeichneten Block in Bild 2 ausgeführt. Die Filter $H_l(n)$ sind Hochpässe und dienen zur Preemphase. Das mit T bezeichnete Element dient zur Kompensation der Laufzeit des Wiener–Filters $W(n)$. In [12] wird eine Modifikation des bisher beschriebenen Systems vorgeschlagen, indem die Realisierung mit Hilfe der FFT im Frequenzbereich erfolgt und die Leistungsdichten in (6) geschätzt werden. Die Autoleistungsdichte des Sprachsignalprozesses $s(k)$ wird durch Mittelung der Kreuzleistungsdichten der laufzeitkorrigierten Signalprozesse $y_i(k)$ geschätzt

$$\hat{S}_{ss}(n) = \frac{1}{6} \sum_{i=1}^{3} \sum_{j=i+1}^{4} Y_i(n)\cdot Y_j^*(n) \quad , \tag{11}$$

wobei $Y_i(n)$ die FFT von $y_i(k)$ ist. Da in jedem $y_i(k)$ das laufzeitkorrigierte Sprachsignal und eine unkorrelierte Störkomponente enthalten ist, handelt es sich um eine Schätzung der Leistungsdichte des ungestörten Sprachsignals. Die Leistungsdichte des gestörten Sprachsignals wird entsprechend durch

$$\hat{S}_{rr}(n) = \frac{1}{4} \sum_{i=1}^{4} Y_i(n)\cdot Y_i^*(n) \tag{12}$$

geschätzt. Der Schätzwert der Leistungsdichte des Sprachsignals nach (11) muß reell sein, weil die zugehörige Korrelationsfunktion symmetrisch ist. Deshalb wird in [12] nur der Realteil von (11) verwendet, da der Imaginärteil nur durch Störungen verursacht sein kann. Schließlich werden die Spektralkomponenten in (11), die stark gestört sind, mit einem niedrigeren Gewicht belegt als die mit geringer Störung.

Es wird über eine Verbesserung des Signal–zu–Rauschverhältnisses um 28 dB berichtet, wobei die zuletzt diskutierten Modifikationen einen Anteil von 10 dB haben und keine genauen Angaben über den Rauschabstand am Systemeingang und das Berechnungeverfahren für diesen Wert gemacht werden.

4. Mikrophonarray mit Spektralsubtraktion und Nachfilterung

Dieses in [13] publizierte Verfahren verwendet Ansätze zur Geräuschreduktion, die von den bisher beschriebenen Verfahren her bekannt sind, so die Bildung von Summen– und Differenzkanälen des Verfahrens von Zelinski oder die Kombination mit der Spektralsubtraktion, die als Erweiterung der Verfahren von Allen und Kaneda vorgeschlagen

wurde. Daneben werden aber auch Verbesserungen benutzt, um z.B. die bei Spektralsubtraktion auftretenden störenden "musical tones" zu reduzieren oder eine in Bezug auf die Störungen robuste Laufzeitkompensation zu realisieren. Das Gesamtsystem besteht aus vier Stufen, die aus dem Aufbau nach Bild 3 erkennbar werden: eine Einheit zur Laufzeitkompensation, die Bildung von Summen– und Differenzkanälen aus jeweils zwei laufzeitkompensierten Eingangskanälen, der Spektralsubtraktionsstufe einschließlich einer Transformation der Differenzkanäle in die Summenkanäle, sowie einer mit POST bezeichneten Nachverarbeitung, um die Reststörungen in Form der "musical tones" zu reduzieren. Alle diese Operationen werden im Frequenzbereich ausgeführt, so daß am Eingang eine FFT und am Ausgang eine inverse FFT oder IFFT ausgeführt wird.

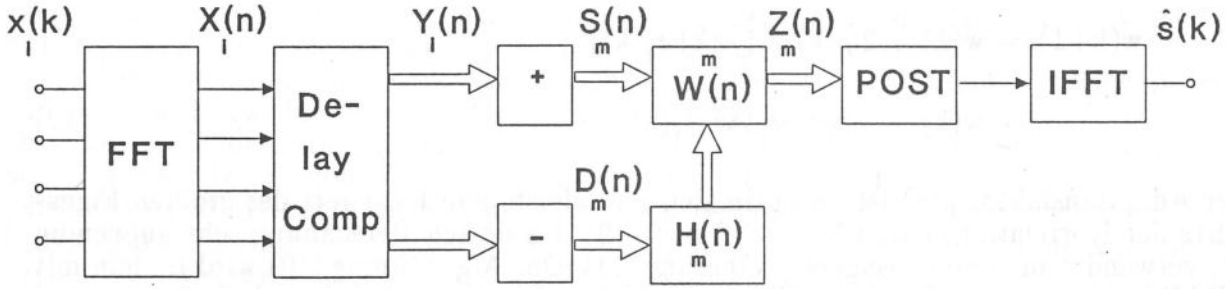

Bild 3 Geräuschredukton mit Mikrophonarray, Spektralsubtraktion und Nachfilterung

Zur Laufzeitkompensation der vier Eingangssignale wird die Fehlerfunktion

$$f(\tau) = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \sum_{n=n_u}^{n_o} |X_i(n)\, e^{j2\pi\tau_i n/N} - X_j(n)\, e^{-j2\pi\tau_j n/N}|^2 \tag{13}$$

als Funktion der im Vektor τ zusammengefaßten, auf die Abtastzeit T normierten Laufzeiten τ_i zum Minimum gemacht. Dabei sind die $X_i(n)$ die Werte der FFT der Eingangssignale $x_i(k)$ mit der Blocklänge N. Durch numerische Optimierungsverfahren, z.B. das Fletcher–Reeves–Verfahren [14], das mit dem aus (13) leicht ableitbaren Gradienten arbeitet, wird das absolute Minimum gesucht, das im Falle von Sprachaktivität gut ausgeprägt ist, während es in Sprachpausen nicht erkennbar wird, sofern keine punktförmige, sondern nur flächig abstrahlende Störquellen vorhanden sind. Auf Grund dieser Eigenschaft läßt sich $f(\tau)$ auch zur Sprachpausendetektion verwenden. Die Summation in (13) erfolgt nur über die Spektralkomponenten n, die einen besonders hohen Anteil des Sprachsignals erwarten lassen, d.h. im Bereich um 1 kHz. Als Startwert wird der Frequenzbereich von 312 Hz bis 2 kHz verwendet, was bei einer Abtastfrequenz $f_a = 8$ kHz und $N = 256$ Stützstellen auf $n_u = 10$ und $n_o = 64$ führt. Dadurch wird vermieden, daß Störer hoher Leistung, die in einem anderen Spektralbereich als das Sprachsignal angesiedelt sind, das Minimum der Fehlerfunktion verfälschen. Dieser Spektralbereich wird adaptiv an das aktuelle Sprachsignal angepaßt.

In der nächsten Stufe werden paarweise Summen und Differenzen der Spektren der laufzeitkompensierten Signale $y_i(k)$ gebildet

$$S_m(n) = \frac{1}{2}(Y_i(n) + Y_j(n))\ ;\ 1 \le m \le M \tag{14a}$$

$$D_m(n) = \frac{1}{2}(Y_i(n) - Y_j(n))\ ;\ 1 \le m \le M\ , \tag{14b}$$

wobei sich unter Beachtung von $i \neq j$ insgesamt 6 Summen– und Differenzsignale ergeben. Bei fehlerfreier Laufzeitkompensation verstärkt sich in $S_m(n)$ der Sprachsignalanteil, in

$D_m(n)$ löscht er sich aus. Während der Summenkanal Tiefpaßcharakter hat, besitzt der Differenzkanal Hochpaßcharakter. Weil die Differenzkanäle im wesentlichen Störungen enthalten, die denen in den entsprechenden Summenkanälen zuzuordnen sind, versucht man die Störung $n_s(k)$ im Summenkanal durch Transformation der Störung $n_d(k)$ im Differenzkanal zu schätzen und durch Subtraktion zu kompensieren. Die Leistungsdichte der Störungen im Summenkanal wird über die noch zu bestimmende Übertragunsfunktion $H_m(n)$ aus der Leistungsdichte des nur Störungen enthaltenden Differenzkanals geschätzt

$$\hat{S}^m_{n_s n_s}(n) = |H_m(n)|^2 \cdot \hat{S}^m_{n_d n_d}(n) \quad . \tag{15}$$

Die Übertragungsfunktion $H_m(n)$ bzw. deren Betrag läßt sich in Sprachpausen, die z.B. durch Auswertung von (13) bestimmt werden, als Quotient aus den geschätzten Leistungsdichten nach (15) berechnen. Diese Übertragungsfunktion bleibt bei Sprachaktivität unverändert, während sie in Sprachpausen adaptiert wird. Eine Adaption ist erforderlich, da der Differenzkanal sowohl instationäre Störungen enthält wie auch Fehleranteile, die durch unpräzise Laufzeitkorrektur der Sprachsignale entstehen. Die Berechnung der geschätzten Leistungsdichten erfolgt wie bei (4).

Nach der Transformation der Störleistung in den Summenkanal kann dort die spektrale Subtraktion ausgeführt werden. Im Prinzip entspricht sie der Filterung mit $W(n)$ nach (5), wobei die Schätzung der Leistungsdichte der Störungen im Summenkanal aber mit Hilfe der transformierten Störleistungsdichte nach (15) erfolgt. Ferner erfolgt diese Operation parallel für alle M Summen– und Differenzkanäle, so daß mit $1 \leq m \leq M$

$$W^m(n) = \begin{cases} (\hat{S}^m_{ss}(n) - a \cdot |H_m(n)|^2 \, \hat{S}_{n_d n_d}(n)) / \hat{S}^m_{ss}(n) \\ b \cdot \hat{S}^m_{ss}(n) \end{cases} \tag{16}$$

gilt, wobei $\hat{S}^m_{ss}(n)$ die geschätzte Leistungsdichte im Summenkanal ist und nur für

$$\hat{S}^m_{ss}(n) > a \cdot |H_m(n)|^2 \, \hat{S}_{n_d n_d}(n) \tag{16a}$$

die obere Beziehung verwendet wird, damit die Differenz, der Schätzwert für die Leistungsdichte des Sprachsignals, nicht negativ wird, was definitionsgemäß nicht möglich ist. Ist (16a) nicht erfüllt, gilt die untere Beziehung.
In (16) bzw. (16a) ist a der aus (5) bekannte Überschätzfaktor, während b ein Dämpfungsfaktor in der Größe von 0,1 ist, durch den ein Rest der Störgeräusche übertragen wird. Damit soll vermieden werden, daß dem Benutzer die Übertragungsstrecke als unterbrochen erscheint. Im Gegensatz zu (5) wurde die Abhängigkeit vom i–ten verarbeiteten Datenblock nicht explizit aufgeführt.

Bei der Spektralsubtraktion treten die bereits erwähnten störenden "musical tones" auf. Sie werden durch isolierte, von Datenblock zu Datenblock in der Frequenzlage sich ändernde Spektrallinien hervorgerufen. Bei Sprachaktivität sind die durch Spektralsubtraktion gewonnenen M Spektren $Z_m(n)$ weitgehend identisch, während sie in Sprachpausen deutliche Unterschiede aufweisen. Deswegen werden sie bei Sparchaktivität in der mit POST bezeichneten Nachverarbeitungsstufe gemittelt, während man bei Sprachpausen zur Synthese des Ausgangsspektrums jeweils die kleinste Spektrallinie der M Spektren $Z_m(n)$ verwendet.

Das System wurde mit mehreren Signal–zu–Rauschverhältnissen am Eingang und verschiedenen Störern getestet. Verbesserungen SNRE des Signal–zu–Rauschverhältnisses lagen bei einem Rauschabstand von 0 dB am Eingang bei 9 dB $\leq$ SNRE $\leq$ 14 dB. Der Gewinn durch die Mittelung nach (14a) lag im Bereich von 1,4 dB bis 2,5 dB, die Spektralsubtraktion nach (16) lieferte einen Beitrag von 7,2 dB bis 9,8 dB und die Nachverarbeitung ergab eine weitere Verbesserung von 0,5 dB bis 1,8 dB. Mit zunehmendem Rauschabstand am Systemeingang verringern sich die Gewinne, was aber unkritisch ist, da die Störeinflüsse absolut gesehen abnehmen.

Die Gewinne SNRE des Systems wurden mit denen anderer Ansätze verglichen.

So liefert die einkanalige Spektralsubtraktion nach (5) Verbesserungen von 0 dB $\leq$ SNRE $\leq$ 9 dB, der Delay–and–Sum Beamformer von 2 dB $\leq$ SNRE $\leq$ 5 dB, wobei dieser Ansatz der Mittelung von vier laufzeitkorrigierten Signalen nach (14a) entspricht. Daraus folgt, daß die Gewinne bei anderen Ansätzen entweder erheblich niedriger sind oder erheblich stärker von der Störquelle abhängen. Allerdings ist der Realisierungsaufwand auch erheblich niedriger.

Danksagung

Beiträge zu diesem Aufsatz wurden in der von Herrn Dr.–Ing. K. Linhard vom Daimler Benz Forschungsinstitut, Ulm, betreuten Diplomarbeit von Herrn Dipl.–Ing. H. Rehmann und im Rahmen eines von Herrn Dr.–Ing. S. Gierl bearbeiteten Forschungsauftrags der Deutschen Bundespost Telecom gewonnen.

Schrifttum

[1] Kroschel, K.: Statistische Nachrichtentheorie. Zweiter Teil: Signalschätzung. 2. Auflage. Berlin 1988, Springer–Verlag

[2] Lim, J.S.: Speech Enhancement. Signal Processing Series. Englewood Cliffs 1983: Prentice Hall

[3] Kroschel, K.; Reich, W.: A Comparison for Noise Reduction Systems for Speech Transmission. Proc. of the European Conf. on Circent Theory and Design, ECCTD–85, S. 565–568

[4] Widrow, B. et al.: Adaptive Noise Concelling: Principles and Applications. Proc. IEEE, Vol. 63, No. 12, 1975, S. 1692–1716

[5] Kroschel, K.; Linhard, K.: Combined Methods for Adaptive Noise Concellation. Proc. Europ. Signal Processing Conference, EUSIPCO–88, S. 417–420

[6] Tamura, S.; Waibel, A.: Noise Reduction Using Connectionist Models. Proc. Intl. Conference on Acoustic, Speech, and Signal Processing, ICASSP–88, 1988, S. 553–556

[7] Grenier, K.; Xu, M.: An Adaptive Microphone Array for Speech Input in Cars. Proc. 22nd. Intl. Symposium on Automative Technology and Automation. ISATA –90, 1990, S. 485–492

[8] Kaneda, Y; Tohyama, M.: Noise Suppression Using 2–Point Received Signals. Electronis and Communications in Japan. Vol. 67–A, No. 12, 1984, S. 19–28

[9] Allen, J.B. et al.: Multimicrophone Signal–Processing Technique to Remove Room Reverberation from Speech Signals. The J. of the Acoustical Soc. of America, Vol. 62, No. 4, 1979, S. 912–915

[10] Kammeyer, K.D.; Kroschel, K.: Digitale Signalverarbeitung. Filterung und Spektralanalyse. Stuttgart 1989: Teubner

[11] Zelinski, R.: Ein Geräuschunterdrückungssystem mit zweidimensionaler Mikrofongruppe und nachgeschalteter adaptiver Wiener Filterung. 6. Aachener Symposium Signaltheorie, ASST–87, 1987, S. 372–375

[12] Zelinski, R.: A Microphone Array with Adaptive Postfiltering for Noise Reduction in Reverberant Rooms. Proc. Intl. Conf. on Acoustic, Speech, an Signal Processing, ICASSP–88, 1988, S. 2578–2581

[13] Gierl, S.: Noise Reduction for Speech Input Systems Using an Adaptive Microphone Array. Proc. 22nd Intl. Symposium on Automotive Technology and Automation, ISATA–90, 1990, S. 517–524

[14] Fletcher, R.; Reeves, C.M.: Function Minimization by Conjugate Gradients. The Computer Journal, Vol. 7, 1964, S. 149–154