Signalunterraum-Verfahren zur Störgeräuschbefreiung von Sprachsignalen mit Stimmhaft/Stimmlos-Unterscheidung

J. Schultz, Volkswagen AG
D. Ronneberger, Universität Göttingen
K. Kroschel, Universität Karlsruhe
K. Schaaf, Volkswagen AG

Einleitung

Verfahren zur Unterdrückung von Störgeräuschen bei Sprachsignalen finden insbesondere im Sprechfunk, bei Telefon-Freisprecheinrichtungen und als Vorverarbeitung für Spracherkennungssysteme im Kraftfahrzeug eine steigende Bedeutung. Die größte Verbreitung haben Verfahren, die auf der spektralen Subtraktion basieren [1]. Die Anwendung von Signalunterraum-Verfahren ist vergleichsweise neu. Dabei werden die Eigenwerte und Eigenvektoren einer Kovarianzmatrix analysiert. Unkorrelierte Signalanteile, die mit Rauschen identifiziert werden, können auf diese Weise unterdrückt werden. Dieser Ansatz ist für stimmlose Sprache nur eingeschränkt gültig. Es wird gezeigt, wie anhand der Eigenwertverteilung stimmhafte von stimmloser Sprache unterschieden werden kann. Dies erlaubt die getrennte Optimierung des Verfahrens für beide Fälle.

Signal-Unterraum-Verfahren

Kern des Verfahrens ist die $K \times K$-dimensionale Kovarianzmatrix, die zur Zeit k gemäß

$$R_k = \frac{1}{N} \sum_{j=-(N/2)+1}^{N/2} \vec{z}_{k+j} \vec{z}^{\,T}_{k+j}$$

wobei der Vektor $\vec{z} = (z_k, z_{k+1}, \cdots, z_{k+K-1})^T$ aus dem verrauschten Sprachsignal z_k gebildet wird [2]. Liegt weißes Rauschen w_k vor, so ergibt sich für $N \to \infty$ die mit σ_w^2 multiplizierte Einheitsmatrix, d.h. es liegt eine konstante Eigenwertverteilung vor, wobei σ_w^2 die Varianz des weißen Rauschens ist. Ein sinusförmiges Signal führt zu zwei Eigenwerten größer Null. In Gegenwart von weißem Rauschen und einem sinusförmigen Signal ergeben sich zwei Eigenwerte, die größer als σ_w^2 sind, während die übrigen Eigenwerte den Wert σ_w^2 haben. Die zu den großen Eigenwerten gehörenden Eigenvektoren spannen den zweidimensionalen Signal-plus-Rausch-Unterraum auf, der meist kurz als „Signalunterraum“ bezeichnet wird.

Die Störgeräuschbefreiung wird dadurch erreicht, dass das verrauschte Signals auf die Eigenvektoren der Kovarianzmatrix projiziert wird. Die Projektionsanteile werden entsprechend dem Wert des zugehörigen Eigenwerts gewichtet. Damit ergibt sich das störgeräuschbefreite Signal zu

$$\vec{x}_k = UGU^T \vec{z}_k \,.$$

U ist die Matrix, die in den Spalten die Eigenvektoren von R enthält. $G = diag(g(i))$ ist eine Diagonalmatrix, dessen Diagonalelemente im einfachsten Fall 1 betragen, wenn der zugehörige Eigenwert größer σ_w^2 ist, und sonst Null (least-square-(LS-)Schätzer).

Gewichtungsfunktionen

Der LS-Schätzer führt zu tonalem Störgeräusch. Besser ist ein gleitender Übergang. Die Forderung, dass die Energie des verbleibenden Restrauschens pro Komponente unter einer vorgegebenen Schwelle bleiben soll, führt zum „spectral-domain-constrained“-(SDC-)-Schätzer mit dem Parameter γ [2]:

$$g(i) = \begin{cases} \left(\dfrac{\lambda_z(i) - \sigma_w^2}{\lambda_z(i)} \right)^{\gamma/2} & \lambda_z > \sigma_w^2 \\ 0 & sonst \end{cases}$$

$\lambda_z(i)$ ist der i-te Eigenwert der Kovarianzmatrix, die aus dem verrauschten Sprachsignal gebildet wurde. Es zeigt sich, dass ein raues Klangbild weitestgehend vermieden werden kann, wenn alle Projektionen berücksichtigt werden. Eine mögliche Wahl besteht darin, die Gewichtungsfunktion um σ_w^2 nach links zu verschieben. Führt man noch analog zum Oversubtraction-Faktor bei der spektralen Subtraktion die Möglichkeit ein, mit einem um den Faktor $\ddot{u}$ erhöhten Schätzwert für den Rauschanteil zu arbeiten, ergibt sich:

$$g(i) = \begin{cases} \left(\dfrac{\lambda_z(i)}{\lambda_z(i) + \ddot{u} \cdot \sigma_w^2} \right)^{\gamma/2} & \lambda_z > \sigma_w^2 \\ 0 & sonst \end{cases}$$

Abbildung 1 zeigt die Graphen der Gewichtungsfunktion für einige Werte von γ.

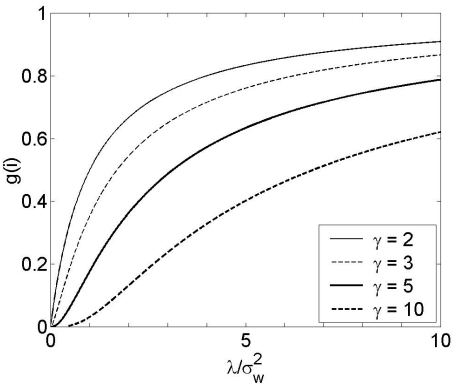

Abbildung 1: Graph der Gewichtungsfunktion für verschiedene Werte von γ und ü=1.

Effektive Frequenz eines Eigenvektors

Um das Verfahren für hohe und tiefe Frequenzanteile optimieren zu können, kann die effektive Frequenz eines Eigenvektors durch

$$f(i) = \frac{\sum_{j=1}^{K-1} u_i(j) \cdot u_i(j+1)}{\sum_{j=1}^{K} u_i(j) \cdot u_i(j)}$$

definiert werden. Zu tiefen Frequenzen liegt zunehmend konstanter Verlauf vor, so dass sich $f(i)$ dem Wert 1 nähert. Nahe der Nyquistfrequenz ergibt sich –1.

Vokal-Konsonant-Unterscheidung

Korrelierte Signale führen zu wenigen großen Eigenwerten. Die Projektion des verrauschten Sprachsignals auf diese wenigen Eigenvektoren erlaubt eine gute Unterdrückung des Störgeräuschs. Bei stimmloser Sprache kann die Unterdrückung der Eigenvektoren, die zu kleinen Eigenvektoren gehören, schneller zu eine Verzerrung führen.
Stimmhafte Sprache kann durch wenige Formanten gut beschrieben werden und zeichnet sich durch wenige große Eigenwerte aus, während stimmlose Sprache zu einer breiten Eigenwertverteilung führt. Abbildung 2 zeigt die Eigenwertverteilung des Wortes „schade“. Der größte Eigenwert ist auf 1 normiert. Die stimmhaften Vokale führen zu einer deutlich schnelleren Abfall der Eigenwertverteilung. Als Maß für die Stimmhaftigkeit kann der Schwerpunkt der Eigenwertverteilung gemäß

$$c_k = \frac{\sum_{i=1}^{N} i \cdot \lambda_k(i)}{\sum_{i=1}^{N} \lambda_k(i)}$$

verwendet werden.

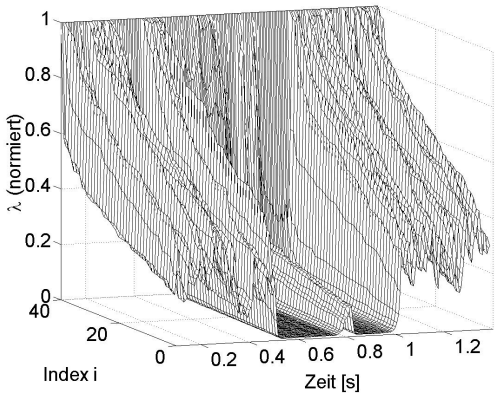

Abbildung 2: Eigenwertverteilung des Wortes „schade“. Der Vokal „a“ ist von 0,4 bis 0,8s und der Vokal „e“ von 0,82 bis 1,0s an der Konzentration auf wenige Eigenwerte zu erkennen. Die Eigenwerte sind auf den jeweils größten Eigenwert normiert.

Abbildung 3 zeigt den zeitlichen Verlauf des Schwerpunkts bei verschiedenen SNR-Werten des weißen Rauschens. Auch bei starken Rauschpegeln ist eine gute Trennung von stimmhafter und stimmloser Sprache möglich.
Ist das Störgeräusch nicht weiß, so ist ein Prewhitening notwendig. Nur dann stellt die Basis der Kovarianzmatrix des Signals auch eine Basis der Kovarianzmatrix des Störgeräuschs dar. Durch das Prewhitening werden auch stimmlose Laute eingefärbt. Bei einem Störgeräusch mit einer spektralen Leistungsdichte $\propto \omega^{-4}$, wie es z.B. für das Innengeräusch eines Fahrzeugs bei nicht zu tiefen Frequenzen typisch ist, führt das Prewhitening auch bei stimmlosen Lauten zu einer Korrelation, die eine Stimmhaft-/Stimmlos-Unterscheidung nicht mehr zulässt (Abbildung 3).

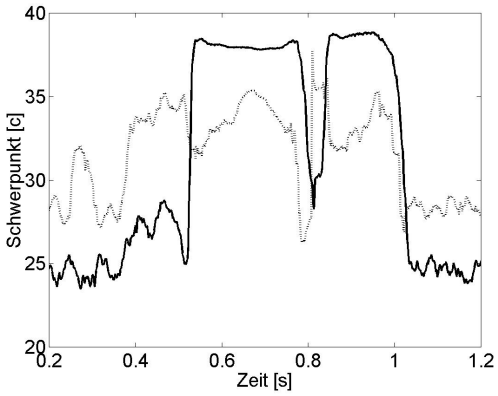

Abbildung 3: Zeitlicher Verlauf des Schwerpunkts c des Wortes „schade“ in Gegenwart von weißem Rauschen (durchgezogene Linie) und bei Rauschen mit einer spektralen Leistungsdichte $\propto \omega^{-4}$ (gepunktete Linie).

Ergebnisse

Um die Frage zu beantworten, welche Parametereinstellungen als optimal empfunden werden, wurde eine graphische Oberfläche erstellt, bei der die Parameter γ und $\ddot{u}$ für stimmhafte und stimmlose und jeweils für hohe und tiefe Frequenzen getrennt eingestellt werden konnten. Bei den Probanden handelte es sich um Personen im Alter zwischen 23 und 30 Jahren, die keine Vorerfahrung mit Störgeräuschbefreiungsverfahren hatten. Die Probanden wurden darauf hingewiesen, dass stimmhafte und stimmlose Laute und jeweils die hohen und tiefen Frequenzanteile eingestellt werden können. Es sollte besonders auf die Parametereinstellung von stimmhafter und stimmloser Laute geachtet werden. Tabelle 1 zeigt eine repräsentative Auswahl der gewählten Parameter. In 11 von 13 Einstellungen wurde für die stimmlosen Laute eine „sanftere“ Einstellung gewählt, d.h. es wurde ein kleinerer Wert für γ und/oder ein kleinerer Wert für $\ddot{u}$ gewählt (Tabelle 1). Die dadurch erreichbaren Verbesserungen stellen allerdings nur eine geringfügige Verbesserung dar. Bei einer – eigentlich wünschenswerten – deutlich verschiedenen Parameterwahl würde sich der Pegel des verbleibenden Restrauschens zu stark verändern, was als störend empfunden wird. Die durch die getrennte Parametrisierung erreichbare Verbesserung ist zwar gering aber eindeutig.

stimmhaft				stimmlos			
tieffrequent		hochfrequent		tieffrequent		hochfrequent	
ü	γ	ü	γ	ü	γ	ü	γ
1,5	5	1,5	5	1,25	5	1,25	3
1,5	5	1,5	3	1,25	3	1,25	3
1,25	7	1,25	7	1,25	7	1,25	5

Tabelle 1: Typische Parametrisierung für stimmhafte und stimmlose sowie jeweils hohe und tiefe Frequenzanteile.

Literaturangaben

[1] Boll S.F.: *Suppression of acoustic noise using spectral subtraction.* IEEE Trans. On ASSP, vol. 27, no. 2, S. 113-120, April 1979

[2] Ephraim Y. Van Trees H.L.: *A signal subspace approach for speech enhancement.* IEEE Trans. on speech and audio processing, vol. 3, no. 4, S.251-266, Juli 1995