Zweistufige Sprache/Pause-Detektion in stark gestörter Umgebung

Ernst Warsitz, Reinhold Häb-Umbach, Jörg Schmalenströer

Universität Paderborn, Inst. für Elektrotechnik und Informationstechnik, Fachgebiet Nachrichtentechnik, 33098 Paderborn

Email: {warsitz,haeb,schmalen}@nt.uni-paderborn.de

Einleitung

Bei der Verarbeitung von Sprachsignalen in Freisprechanlagen ist üblicherweise eine Störgeräuschreduktion notwendig. Dazu bedarf es einer möglichst guten Schätzung des Rauschspektrums, welches zu entfernen ist. Grundsätzlich können die Verfahren hierbei in zwei Klassen eingeteilt werden: Zum einen kann z. B. mit Hilfe der so genannten Minimum-Statistik-Methode [1, 2] eine kontinuierliche Schätzung erfolgen und zum anderen ist eine abschnittsweise Mittelung des Rauschens in Sprachpausen durch Verwendung einer Sprache/Pause-Detektion möglich. Letzteres Vorgehen hat den Vorteil für weitere Verarbeitungsschritte die Information über Sprachaktivität bereit zu stellen. Problematisch wird jedoch eine solche Detektion bei einem sehr niedrigen Signal-zu-Rausch-Verhältnis, wie z. B. in einem Kraftfahrzeug [3].

Hier soll ein robustes zweistufiges Verfahren zur Detektion von Sprachaktivität (engl. Voice Activity Detection, VAD) vorgestellt werden, welches in einem ersten Schritt eine grobe Entstörung des Sprachsignals mittels Wiener Filterung und mit Hilfe der Minimum-Statistik-Methode vornimmt. In einem zweiten Schritt erfolgt dann der Einsatz einer effizienten Sprache/Pause-Detektion basierend auf einem Energiekriterium. Die Leistungsfähigkeit der vorgeschlagenen zweistufigen Methode wird im Vergleich zu anderen Verfahren zur Sprache/Pause-Detektion untersucht [4].

Energiebasierte VAD

Für eine energiebasierte VAD ist es notwendig, die Energie des Eingangssignals für unterschiedlich lange Zeitabschnitte zu mitteln bzw. zu glätten. Hierfür sind unterschiedliche Strategien denkbar, wobei sich das im folgenden beschriebene Vorgehen in vielen praktischen Anwendungen als sehr wirkungsvoll erwiesen hat. Dabei wird das mit $f_a = 1/T$ abgetastete zeitdiskrete Eingangssignal $y(n)$ blockweise betrachtet, wobei die Blocklänge mit B und der Blockindex mit m bezeichnet werden soll. Zunächst wird die instantane Energie $P(m)$ eines Signalabschnittes berechnet

$$P(m) = \frac{1}{B} \sum_{n=mB}^{(m+1)B-1} y^2(n) \tag{1}$$

und rekursiv zur Ermittelung der Kurzzeitenergie $P_K(m)$ mit einer kleinen Glättungskonstante α exponentiell geglättet

$$P_K(m) = \alpha P_K(m-1) + (1-\alpha)P(m). \tag{2}$$

Die Langzeitenergie $P_L(m)$ ergibt sich aus der gemittelten Energie über D Blöcke

$$P_L(m) = \frac{1}{DB} \sum_{n=(m+1-D)B}^{(m+1)B-1} y^2(n), \tag{3}$$

wobei dann die minimale Energie $P_N(m)$ eines der letzten M Blöcke der Langzeitenergie gesucht wird

$$P_M(m) = \min_{\tilde{m}\in\{m-N,m\}} \{P_L(\tilde{m})\}. \tag{4}$$

Die Entscheidung für Sprachaktivität im Block m wird getroffen, wenn die Kurzzeitenergie über einem bestimmten Vielfachen η der minimalen Energie plus einem festen Wert δ liegt:

$$P_K(m) > \eta P_M(m) + \delta. \tag{5}$$

Minimum Statistik und Wiener Filter

Das oben beschriebene Verfahren stellt eine sehr effiziente Methode zur Sprache/Pause-Detektion dar. Mit sinkendem Eingangs-SNR nimmt jedoch die Erkennungsrate der energiebasierten Sprachdetektion deutlich ab. Daher wird in einer ersten Stufe der eigentlichen VAD ein Wiener Filter (WF) in Form einer spektralen Subtraktion vorgeschaltet. Die Schätzung der spektralen Leistungsdichte der Störung erfolgt dabei mit der sogenannten Minimum Statistik (MS) [1, 2]. Mit dem so entrauschten Signal wird dann die VAD als nachgeschaltete Stufe betrieben, siehe Abb. 1. Die Parameter für

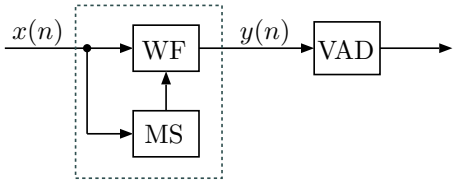

Abbildung 1: Schematische Darstellung der zweistufigen Sprache/Pause-Detektion.

die Störgeräuschreduktion werden derart eingestellt, dass ein hohes SNR am Ausgang $y(n)$ zu erwarten ist. Daher wird z. B. keine Filterung der Gewichtungskoeffizienten des Wiener Filters in Frequenzrichtung (Glättung, psychoakustisch motivierte Filterbänke) vorgenommen und ein hoher Übersubtraktionsfaktor eingestellt. Das resultierende Ausgangssignal $y(n)$ weist somit erhebliche Verzerrungen und so genannte musical tones auf, bildet jedoch eine gute Grundlage für die anschließende VAD. Abb. 2 zeigt beispielhaft die Zeitverläufe eines reinen Sprachsignals $s(n)$, das mit weißem Rauschen

überlagerte Sprachsignal $x(n)$ und das gefilterte Signal $y(n)$. Zusätzlich sind zum einen die optimale und zum anderen die energiebasierte VAD-Entscheidung dargestellt.

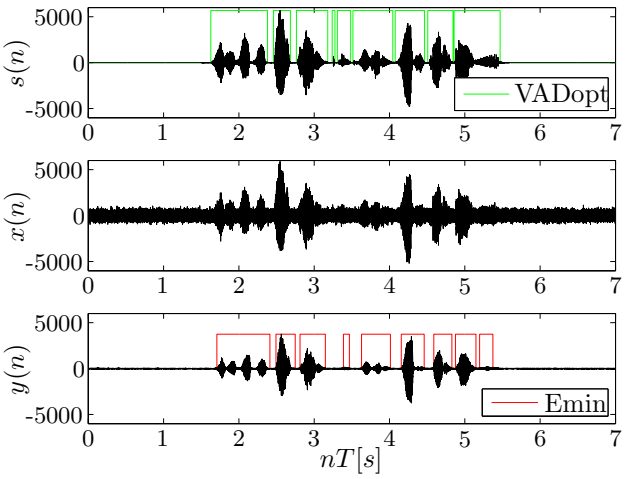

Abbildung 2: Oben: Reines Sprachsignal $s(n)$ und optimale Sprachdetektion (VADopt). Mitte: Verrauschtes Signal $x(n)$ mit einem SNR von 5 dB. Unten: Signal $y(n)$ nach Störgeräuschreduktion und das Ergebnis der energiebasierten VAD (Emin).

Experimentelle Ergebnisse

Um die Leistungsfähigkeit des beschriebenen Verfahrens zur Sprachdetektion zu evaluieren wurden mit $f_a = 16$ kHz abgetastete Sprachsignale in unterschiedlichen Fahrsituationen bei 60, 80, 100 und 120 km/h in einem Kraftfahrzeug (BMW E46/2) aufgenommen [3]. Für jede der 4 Geschwindigkeiten wurden jeweils 4 Sprachsequenzen von dem Fahrer und Beifahrer mit einer durchschnittlichen Länge von 10 s aufgezeichnet.

Die verarbeitete Blocklänge der VAD betrug $B = 256$, die Glättungskonstante hatte einen Wert von $\alpha = 0,45$ und weitere Parameter waren $D = 62$, $N = 310$ und $\delta = 1000$. Die Wahrscheinlichkeit einer korrekten Sprachdetektion $\mathrm{P}(\mathit{Sprache}|\mathit{Sprache})$ ist in Abb. 3 über der Wahrscheinlichkeit eines Detektionsfehlers $\mathrm{P}(\mathit{Sprache}|\mathit{Pause})$ als so genannte receiver operating characteristic für die energibasierte VAD (Emin) aufgetragen. Zum Vergleich sind Ergebnisse für das maximum likelihood (ML) und das decision-directed (DD) Verfahren der VAD aus [4] mit optimal eingestellten Parametern dargestellt. Für die drei gezeigten Verfahren wurden jeweils für unterschiedliche Schwellwerte η die Ergebnisse über alle Audiobeispiele gemittelt.

In einem weiteren Experiment wurden 10 reine Sprachsignale mit weißem Rauschen bei einem SNR von 5 dB überlagert. Die Abb. 4 zeigt die entsprechenden Ergebnisse der Detektionsgenauigkeiten für die drei Verfahren als receiver operating characteristic.

Beide Experimente (Abb. 3 und 4) zeigen deutlich die Funktionsfähigkeit des vorgestellten Verfahrens, bei dem in zwei separaten Stufen erst eine Störgeräuschreduktion und anschließend die eigentliche Sprachdetektion statt

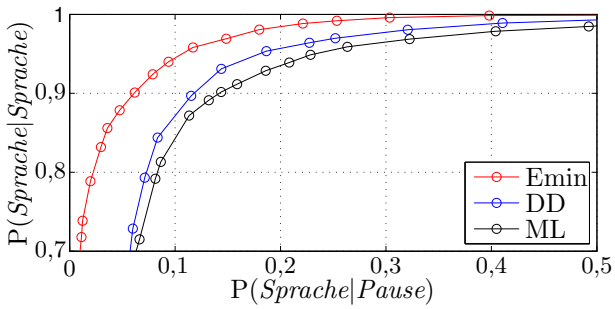

Abbildung 3: Receiver operating characteristic der energiebasierten VAD (Emin), der maximum likelihood (ML) und der decision-directed (DD) basierten VAD für Sprachdaten in einem Kraftfahrzeug.

findet. Außerdem wurden höhere Detektionsgenauigkeiten im Vergleich zu den statistisch motivierten Verfahren ML und DD erzielt, bei denen die Schätzung der spektralen Eigenschaften der Sprache und des Rauschens *direkt* mit der Entscheidungsvariablen verknüpft sind. Hierbei wirken sich unterschiedliche Entscheidungsschwellen η rekursiv auf die Spektralschätzung aus, die wiederum die Entscheidungsregel beeinflußt. Diese enge Koppelung führt zu einer hohen Empfindlichkeit gegenüber der Entscheidungsschwelle, was insbesondere an den Schwankungen der entsprechenden Kurven in Abb. 4 zu erkennen ist.

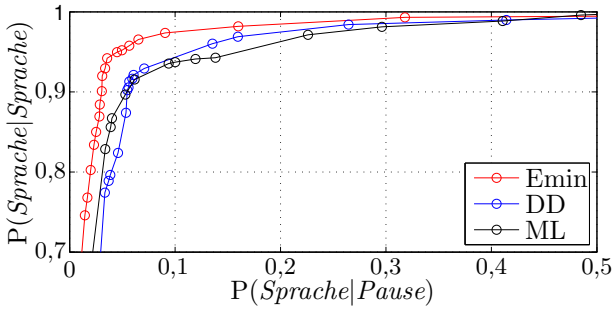

Abbildung 4: Receiver operating characteristic der energiebasierten VAD (Emin), der maximum likelihood (ML) und der decision-directed (DD) basierten VAD für Sprachsignale überlagert mit weißem Rauschen von 5 dB.

Literatur

[1] Martin, R.: Spectral subtraction based on minimum statistics, European Signal Processing Conference (EUSIPCO), 1994, pp. 1182-1185

[2] Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Transactions on Speech and Audio Processing, vol. 108 July 2001, pp. 504-512

[3] Warsitz, E. und Häb-Umbach, R.: Mehrkanalige Sprachsignalverarbeitung durch adaptives Eigenbeamforming für Freisprecheinrichtungen im Kraftfahrzeug, 32. Deutsche Jahrestagung für Akustik (DAGA), 2006, pp. 49-50

[4] Sohn, J. and Kim, N. and Sung, W.: A statistical model-based voice activity detection, IEEE Signal Processing Letters, vol. 6 Jan. 1999, pp. 1-3