Einkanalige Störgeräuschunterdrückung zur Steigerung der Worterkennungsrate eines Spracherkenners

S. Wittenberg, R. Petrick, M. Wolff und R. Hoffmann

Technische Universität Dresden, Institut für Akustik und Sprachkommunikation
soeren.wittenberg@ias.et.tu-dresden.de

Abstract

Diese Arbeit zeigt die Integration einer einkanaligen Störgeräuschunterdrückung in das Frontend eines internetbasierten, verteilten Spracherkennungssystems. Das als Java-Applet realisierte Frontend ermöglicht durch seine Einbettung in einen Web-Auftritt die Navigation auf den angebotenen Webseiten mittels Spracheingabe. Die daraus resultierende Loslösung von einer speziellen Zielplattform stellt besondere Forderungen an die Vorverarbeitung. Erstens kann die zur Verfügung stehende Rechenleistung des Endgerätes gering sein und zweitens ist der Einsatz in beliebigen Störgeräuschumgebungen möglich. Erschwerend kommt hinzu, dass gewöhnlich nur ein Audioaufnahmekanal vorhanden ist. Durch die Anwendung der wohluntersuchten Spektralen Subtraktion, gepaart mit Erweiterungen für die Spracherkennung, konnte ein Kompromiss zwischen der Steigerung der Worterkennungsleistung in gestörter Umgebung und der dafür notwendigen Rechenzeit gefunden werden. Vergleichend dazu wurden die in MOS-Hörtests besser bewerteten Verfahren nach Ephraim und Malah betrachtet, erwiesen sie sich allerdings als weniger geeignet.

1 Einführung

Infolge des stetigen Ausbaus der Datennetze und der Verringerung der Antwortzeiten wurden im Bereich der Sprachdialogsysteme neue Anwendungen denkbar. So konnte am *Institut für Akustik und Sprachkommunikation* der *Technischen Universität Dresden* ein internetfähiges Sprachdialogsystem entwickelt werden. Das System ist als typische Client-Server-Architektur realisiert. Der serverseitige Spracherkenner ist in der Programmiersprache C bzw. C++ entwickelt worden und bedarf so eines speziell für das Serverbetriebssystem übersetzten Programmcodes. Da es beim Client keine Einschränkung auf ein Betriebssystem geben sollte, wurde die plattformunabhängige Programmiersprache Java verwendet. Dadurch ist keine Installation von Spezialsoftware auf dem Client notwendig, wenn dieser im Sprachdialogsystem z.B. durch eine sprachsteuerbare Internetseite im HTML-Format und ein zugehöriges Java-Applet repräsentiert und in einem javafähigen Browser ausgeführt wird. Als Client sind aber auch eigenständige Anwendungen in Java denkbar, welche mittels Sprachsteuerung bedienbar sind. Dies stellt besonders für PDAs einen Eingabekanal für völlig neue Anwendungen bereit.

Für die notwendige Kommunikation zwischen dem sogenannten *jLab Client* (Abb. 1) und dem *UASR* [8] als Spracherkenner gibt es eine auf Java basierende Middleware, den *jLab*

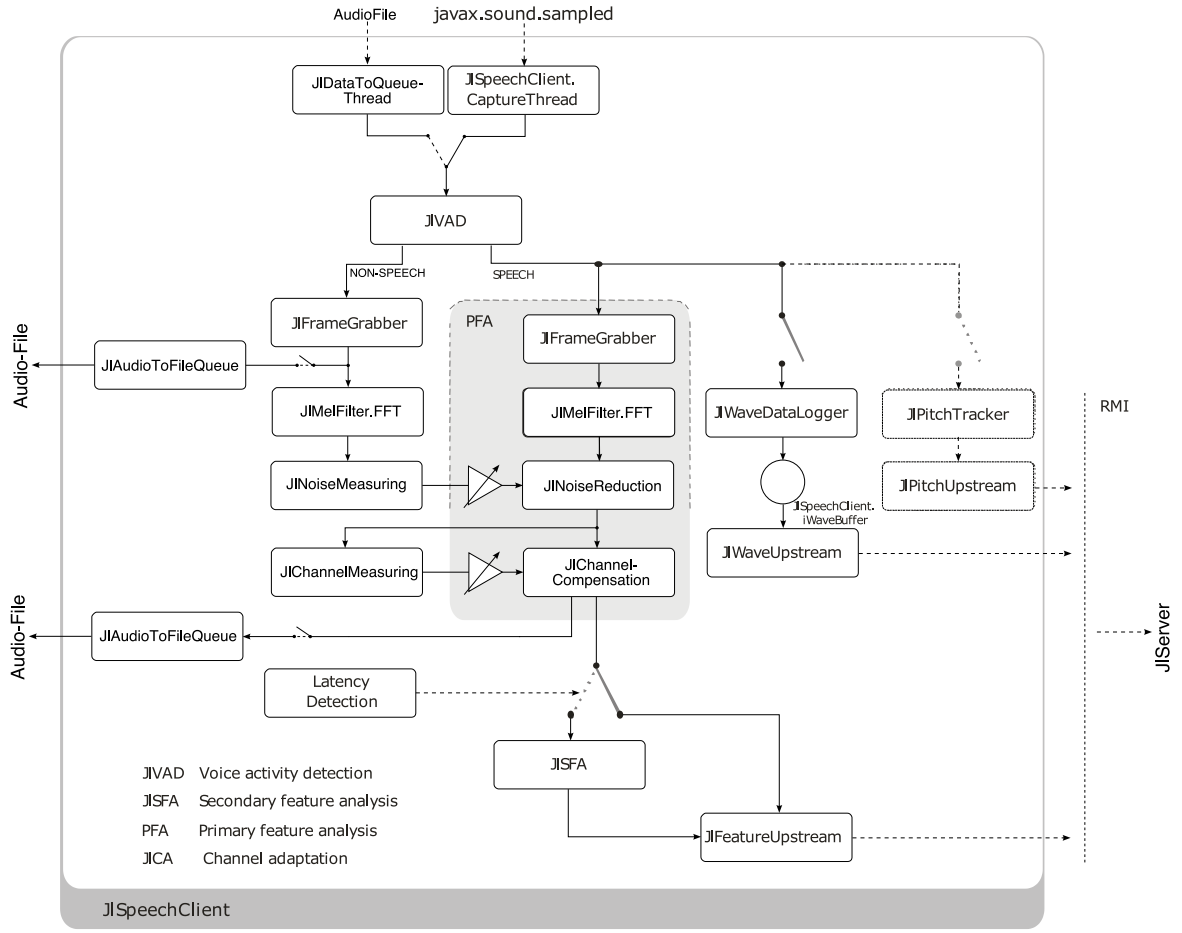


Abbildung 1. Struktur des *jLab Client*

Speech Server. Der entscheidende Vorteil dieser Architektur liegt in der Vorverarbeitung und Merkmalextraktion auf dem Endgerät, was einerseits die zu übertragende Datenmenge reduziert und andererseits den Server entlastet, was sich bei einer großen Anzahl gleichzeitiger Sitzungen vorteilhaft auswirkt. Um tragbaren Systemen mit geringer Prozessorleistung (z.B. PDAs) die Nutzung dieses Dienstes anzubieten, muß die Vorverarbeitung allerdings möglichst Ressourcen sparend durchgeführt werden oder bei zu geringer Rechenleistung wieder auf den Server auslagerbar sein.

2 Einkanalige Verfahren zur Störgeräuschunterdrückung

Der Entwurf der Algorithmen zur Reduzierung von Störgeräuscheinflüssen erfolgte unter dem Gesichtspunkt, dass nur ein Aufnahmekanal zur Verfügung steht und keine Angaben zum Einsatzort und den sich damit ergebenden Störumfeld verfügbar sind. Sekundär wurde die algorithmische Komplexität gering gehalten, um so auch den Einsatz auf mobilen Endgeräten zu ermöglichen.

2.1 Spektrale Subtraktion

Unter der Annahme, dass das Sprachsignal (engl. *speech*) s und Störsignal (engl. *noise*) n additiv überlagert und unkorreliert sind, kann das Nutzsignal durch die Subtraktion des Störsignals vom gestörten Signal am Mikrofon (engl. *microphone*) m wiedergewonnen werden [3]. Bei blockweiser Verarbeitung des gestörten Sprachsignals ergibt sich das

diskrete Kurzzeitspektrum eines ungestörten Signalabschnittes i zu

$$\underline{S}_i(n) = \underline{M}_i(n) - \underline{N}_i(n) \tag{1}$$

Der Zugriff auf das Störsignalspektrum $\underline{N}_i(n)$ ist nicht direkt möglich, es ist bei stationärem Störsignal aber in Sprachpausen schätzbar. Dazu werden die Betragsspektren der letzten $L \in \mathbb{N}$ als Sprachpausen (Sprachpausendetektion notwendig) erkannten Signalabschnitte gemittelt. Die Mittelung gemäß Gleichung (2) verläuft adaptiv mit einem Adaptionskoeffizienten $0.9 < \rho < 1$ der die Streuungen der einzelnen Kurzzeitgeräuschspektrallinien um ihren Mittelwert glättet. Für langsamere Änderungen des Mittelwertes des Rauschens (Veränderungen der Geräuschcharakteristik) passt sich das geschätzte Störgeräuschspektrum mit einer Adaptionsgeschwindigkeit abhängig von ρ an.

$$\overline{|\underline{N}_i(n)|} = \rho\overline{\left|\underline{N}_{i-1}(n)\right|} + (1-\rho)\,|\underline{M}_i(n)| \tag{2}$$

Das geschätzte Störsignalspektrum wird vom aktuellen Kurzzeitbetragspektrum des gestörten Signals abgezogen. Dabei ist sicherzustellen, dass das resultierende Kurzzeitbetragsspektrum positiv bleibt. Anders als der Betrag wird das Argument nicht korrigiert. Durch die Einführung des Parameters γ besteht die Möglichkeit die Subtraktion auf Basis der Amplituden- ($\gamma = 1$) oder Leistungsspektren ($\gamma = 2$) durchzuführen.

Die adaptive Gewichtung des gestörten Kurzzeitspektrum entsprechend Gleichung (3) ergibt das geschätzte Kurzzeitspektrum des ungestörten Signalabschnittes i.

$$\left|\hat{\underline{S}}_i(n)\right|^\gamma = max\left\{1 - \alpha\frac{\overline{|\underline{N}_i(n)|}^\gamma}{|\underline{M}_i(n)|^\gamma} \quad ; \quad 0\right\} \cdot |\underline{M}_i(n)|^\gamma \tag{3}$$

Die Schätzung des Störsignals führt in der Anwendung zu hörbaren Tönen - den für dieses Verfahren typischen *musical tones*. Zur Verminderung kann der (ggf. auch frequenzselektive) Überschätzungsfaktors α einführt werden, da die *musical tones* aber hauptsächlich in den Sprachpausen auftreten und hier isolierte Spektrallinien bilden, ist die Verwendung nicht notwendig. Die Einflüsse während der Sprachsegmente können durch die Einführung eines *spectral floor* vermindert werden.

Der *spectral floor* ist ein spektrales Grundrauschen, welches die *musical tones* verdeckt. Dies kann durch die Begrenzung der Dämpfung (mit Hilfe eines s.g. *Flooringfaktors* β) des gestörten Signals auf einen unteren Wert realisiert werden. Aus Gleichung (3) ergibt sich

$$\left|\hat{\underline{S}}_i(n)\right|^\gamma = max\left\{1 - \alpha\frac{\overline{|\underline{N}_i(n)|}^\gamma}{|\underline{M}_i(n)|^\gamma} \quad ; \quad \beta(SNR_i)^\gamma\right\} \cdot |\underline{M}_i(n)|^\gamma \tag{4}$$

Auf diese Art und Weise wird eine vollständige Auslöschung der spektralen Bestandteile des gestörten Signals verhindert. Für die Wahl von β empfiehlt [4] einen Wert größer $0,2$. Gemäß Untersuchungen mit verschiedenen Werten für β (Abschnitt 3.1) ergab sich ein vom Signal-Rausch-Abstand in Grenzen linear abhängiger *Flooringfaktor*, der aber für Signal-Rausch-Abstände größer 0 dB der Empfehlung von [4] folgt. Es empfiehlt sich, eine dynamische Anpassung von β zur Laufzeit entsprechend Abbildung 2a vorzunehmen.

Neben der Variante mit dem von der Amplitude des gestörten Eingangsignals ($|\underline{M}_i(n)|^\gamma$) abhängigen Floors (Gleichung (4)), die zur Verbesserung des Höreindrucks des entstörten Sprachsignal entwickelt wurde , schlägt [9] bzw. [10] für die Spracherkennung eine Variante mit konstantem Noisefloor (Gleichung (5)) vor. Der konstante Noisefloor belässt stets

ein um den Faktor β reduziertes Rauschen im Spektrum. Auch hier empfiehlt sich eine dynamischen Anpassung von β zur Laufzeit.

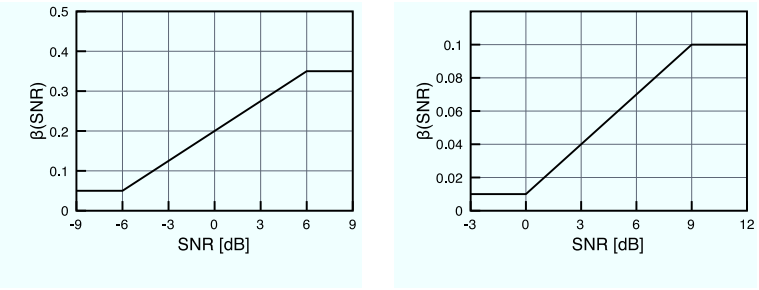
$$\left|\underline{\hat{S}}_i(n)\right|^\gamma = \max\left\{\left(1-\alpha\frac{\overline{|\underline{N}_i(n)|}^\gamma}{|\underline{M}_i(n)|^\gamma}\right)\cdot|\underline{M}_i(n)|^\gamma \quad ; \quad \left(\beta(SNR_i)\cdot\overline{|\underline{N}_i(n)|}\right)^\gamma\right\} \tag{5}$$

Zur Realisierung der sich dynamisch anpassenden Dämpfungsbegrenzung ist der auf den aktuell zu verarbeitenden Signalblock bezogene SNR nach Gleichung (6) notwendig.

$$SNR_i = 10\cdot\left(log_{10}\left(\sum_{n=0}^{N-1}|\underline{M}_i(n)|^2 - \overline{|\underline{N}_i(n)|}^2\right) - log_{10}\left(\sum_{n=0}^{N-1}\overline{|\underline{N}_i(n)|}^2\right)\right)\ dB \tag{6}$$

Da dieser SNR von Segment zu Segment große Sprünge machen kann, wird er zusätzlich entsprechend Gleichung (7), beeinflusst vom Parameter κ, stark geglättet.

$$SNR_i = (1-\kappa)\cdot SNR_{i-1} + \kappa\cdot SNR_i \quad 0\le\kappa\le 1 \tag{7}$$



(a) für Gleichung (4); $\gamma = 1$ (b) für MMSE-STSA

Abbildung 2. Dynamische Dämpfungsbegrenzung in Abhängigkeit vom SNR

2.2 Spektrale Subtraktion nach Ephraim and Malah

Zu Vermeidung von *musical tones* griff Olivier Cappé die in [1] und [2] formulierten Ansätze von Yariv Ephraim und David Malah auf und stellte in [7] ein Verfahren zur Minimierung des mittleren quadratischen Fehlers der spektralen Kurzzeitamplituden (MMSE-STSA) vor.

Ausgangspunkt ist der adaptive spektrale Gewichtungsfaktor (Gleichung (8)), der auf jedes Kurzzeitspektrum $\underline{M}_i(n)$ anzuwenden ist.

$$\begin{aligned} G_i(n) &= \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+R_{post}(n,i)}\right)\left(\frac{R_{prio}(n,i)}{1+R_{prio}(n,i)}\right)} \\ &\quad\cdot Z\left[(1+R_{post}(n,i))\left(\frac{R_{prio}(n,i)}{1+R_{prio}(n,i)}\right)\right] \end{aligned} \tag{8}$$

mit

$$Z[\theta] = e^{\left(-\frac{\theta}{2}\right)}\cdot\left[(1+\theta)I_0\left(\frac{\theta}{2}\right)+\theta I_1\left(\frac{\theta}{2}\right)\right] \tag{9}$$

$$R_{post}(n,i) = \frac{|\underline{M}_i(n)|^2}{\overline{|\underline{N}(n)|}^2} - 1 \tag{10}$$

$$R_{prio}(n,i) = max\Big\{ \quad (1-\alpha)\cdot max\left\{R_{post}(n,i),0\right\} + \alpha\cdot\frac{|G_{i-1}(n)\,\underline{M}_{i-1}(n)|^2}{\overline{|\underline{N}(n)|}^2}\ ,\ \beta(SNR)\Big\} \quad mit\ 0<\beta(SNR)\leq 1 \tag{11}$$

Die beiden Funktionen I_0 und I_1 beschreiben modifizierte BESSEL-Funktionen erster Art nullter und erster Ordnung, $R_{post}(n,i)$ beschreibt den *a-posteriori Signal-Rausch-Abstand* des aktuellen Kurzzeitanalyseblocks i und $R_{prio}(n,i)$ den *a-priori Signal-Rausch-Abstand*. Die Konstante α dient der Beeinflussung der glättenden Wirkung von R_{prio}.

Wie bereits bei der Spektralen Subtraktion wurde auch hier die Möglichkeit zur dynamischen Begrenzung der Dämpfung integriert (Abbildung 2b).

3 Experimente

Zur Durchführung der Experimente wurde der javabasierte Client für die Vorverarbeitung und der HMM-basierte Kommandophrasenerkenner des UASR für die Spracherkennung benutzt. Das Training des Spracherkenners erfolgte mit der Verbmobil Datenbasis. Während des Trainings waren die Verfahren zur Störgeräuschreduzierung deaktiviert. Im Abbildung 3 ist die Vorverarbeitung im Client dargestellt. Für die entwickelten Algorith-

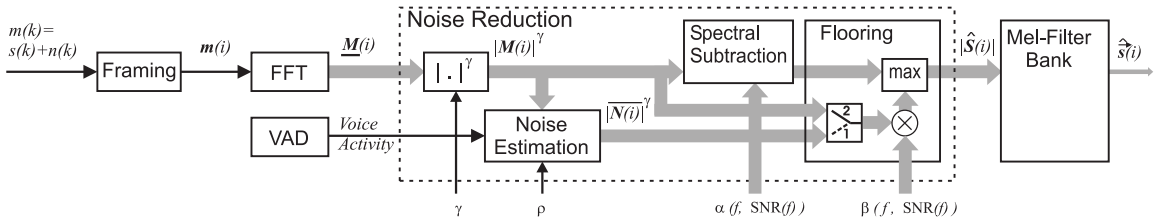


Abbildung 3. Blockschaltbild der Geräuschunterdrückung; Schalterstellung: 1 realisiert Gleichung (5) und 2 realisiert Gleichung (4)

men wurden die Verläufe der Erkennungsrate bei verschiedenen Signal-Rausch-Abständen unter idealisierten und realen Bedingungen ermittelt. Die erste Versuchsreihe schloss eine eventuelle Beeinflussung der Erkennung durch die Sprach-Pausen-Detektion (engl. *Voice Activity Detection - VAD*) aus. Dazu wurde die VAD deaktiviert und eine ideale VAD durch getrennte Einleitung der Sprach- bzw. Pausensignale in die Verarbeitung simuliert. Hierfür wurden für jede ungestörte Aufnahme Start- und Stopmarkierungen generiert, die anschließend zur Einspeisung der gestörten Sprachaufnahmen in das System dienten. Dabei wurde großzügig zugunsten der Sprache entschieden, um möglichst keine Signalabschnitte abzuschneiden. Die verwendeten Aufnahmen (Set I) von 20 Sprechern entstammen der Apollo Datenbasis [5]. Sie bestehen jeweils aus ca. 3 s Pause gefolgt von der entsprechenden Kommandophrase (ca. 2 s) und wiederum ca. 1 s Pause. Für jeden Sprecher konnten jeweils 3 Realisierungen der 17 Kommandos zur Bedienung einer Dunstabzugshaube (insgesamt 1020) klassifiziert werden.

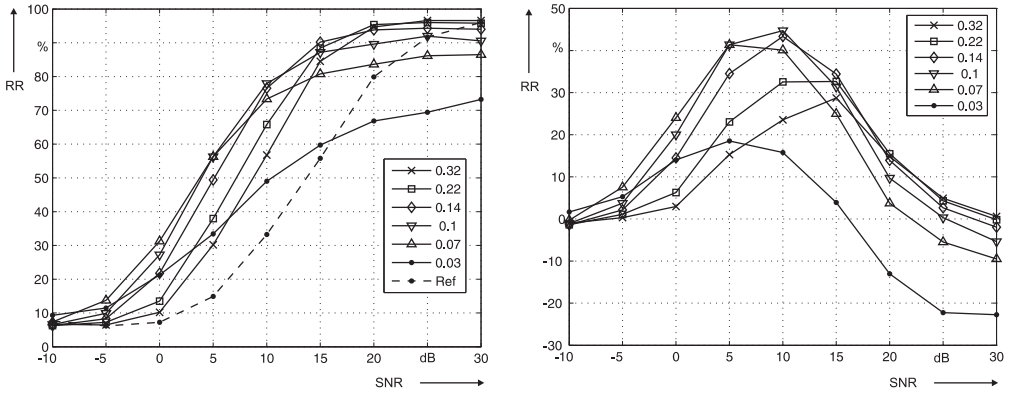
Der zweite Versuch fand unter realen Bedingungen statt und umfasste die Möglichkeit das System an einen Sprecher zu adaptieren. Da für die verschiedenen Sprecher aus Set I zu wenige Daten für ein Adaptierung des Spracherkenners an einen Sprecher vorhanden waren, wurde ein weiteres Testset entworfen. Set II bestand aus insgesamt 1239 Realisierungen eines Sprechers von 18 Kommandos zur Bedienung eines Diktiersystems. Evaluiert

wurde mit aktiver VAD, allerdings wurden die Audiodaten nicht wie üblich über den Mikrofonkanal an die Sprach-Pausen-Erkennung, sondern direkt aus den Dateien zugeführt.

Als Störsignale dienten die Aufnahmen eines Staubsaugers (Störsignal I) und ein künstlich erzeugtes weißes Rauschen (Störsignal V). Diese Signale wurden in ihren Intensitäten manipuliert und additiv mit den ungestörten Testdaten (Set I und II) überlagert. Bei den Untersuchungen wurde bewusst auf die Wirkung des Lombard-Effektes verzichtet, der durch die nachträgliche additive Überlagerung der ungestörten Sprachsignales mit einem Störsignal so nicht zum Tragen kommt.

3.1 Spektrale Subtraktion mit festen Flooringfaktor β

Um günstigste Einstellungen für die Geräuschunterdrückung zu finden, wurden umfangreiche Experimente durchgeführt. Dabei wurden für die beiden Möglichkeiten des Floorings (Gleichungen (4) bzw. (5)) jeweils die Fälle $\gamma = 1$ (Amplitudenspektrum) und $\gamma = 2$ (Leistungsspektrum) unter Variation des Parameters β durchgeführt. Eine Kurvenschar ist beispielhaft für die Spektrale Subtraktion nach Gleichung (5) auf Basis der Amplitudenspektren in Abbildung 4 dargestellt. Die Ergebnisse der anderen drei Variationen der Gleichungen (4, 5) und γ ähneln Abbildung 4, unterschieden sich jedoch in einer günstigen Wahl von β. Keine der vier Varianten hat sich in der Verbesserung der Erkennungsrate markant hervorgehoben. Die beste Einstellung von β hängt jeweils vom SNR ab, weshalb eine adaptive, SNR-abhängige Einstellung von β sinnvoll erscheint.



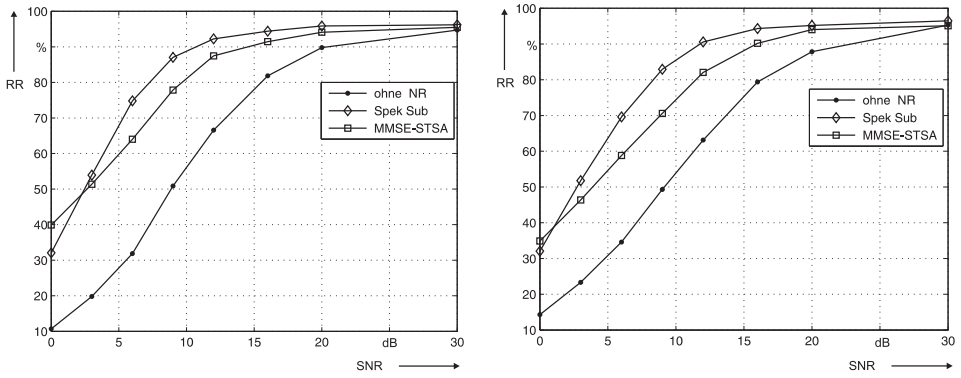
(a) Erkennungsraten für verschiedene β und ohne Geräuschunterdrückung (*Ref*)

(b) Steigerung der Erkennungsrate gegenüber der Referenz bei unterschiedlichen β Werten

Abbildung 4. Optimierung von β für Glg. (5) (Set I, Störsignal V, ideale VAD, $\alpha = 1.2$, $\gamma = 1$)

3.2 Spektrale Subtraktion mit dynamischen Flooringfaktor β

Aus den Erkenntnissen für die Einstellung eines festen Flooringfaktors entsprechend dem vorangegangenen Abschnitt wurden Erkennungsexperimente mit einem sich dynamisch anpassenden β durchgeführt. Die Experimente basieren auf Gleichung (4). In Abbildung 5 sind die Erkennungsraten für die Kommandophrasen aus dem Testset I mit überlagertem Störsignal I und V für verschiedene Signal-Rausch-Abstände dargestellt. Die Untersuchungen wurden mit idealer VAD sowohl bei in- als auch bei aktiver Störgeräuschreduzierung durchgeführt. Dabei konnte die Spektrale Subtraktion mit SNR-abhängiger Anpassung von β die Erkennungsrate sowohl in stark, als auch in schwach gestörter Umgebung - im Vergleich zu den Versuchen ohne Störgeräuschreduzierung - deutlich erhöhen.



(a) Störsignal I (b) Störsignal V

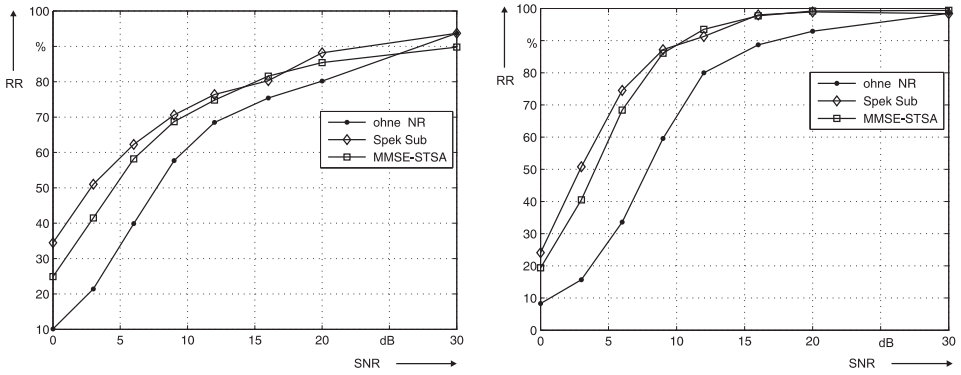
Abbildung 5. Worterkennungsrate mit und ohne Störgeräuschunterdrückung für Set I; ideale VAD

Nach den Untersuchungen unter idealisierten Bedingungen wurde das verteilte Spracherkennungssystem im realen Einsatz mit aktivierter Sprach-Pausen-Erkennung (rein signalbasierte VAD), unter Verwendung des Testset II mit Störsignal I, getestet. Abbildung 6a zeigt, dass die Anwendung der Spektralen Subtraktion in realen Umgebungen mit näherungsweise stationären Störsignalen eine Steigerung der Kommandophrasenerkennungsrate um bis zu 30 % ermöglicht.

3.3 Spektrale Subtraktion nach Ephraim und Malah

Die in Abschnitt 2.2 diskutierte Spektrale Subtraktion mit dem Verfahren zur Minimierung des mittleren quadratischen Fehlers der spektralen Kurzzeitamplituden (MMSE-STSA) konnte ebenfalls erfolgreich getestet werden (Abbildungen 5 und 6). Allerdings blieb die Erkennungsrate trotz der ebenfalls in MOS-Hörversuchen ([6]) bestätigten geringeren Signalverzerrungen unter der mit der normalen Spektralen Subtraktion zurück. Der Einsatz dieses algorithmisch komplexeren Verfahrens ist nicht gerechtfertigt, da die Beseitigung der *musical tones* nur einen Gewinn für die akustische Wiedergabe des bereinigten Sprachsignals darstellt.

3.4 Sprecheradaptierung



(a) unadaptiert (b) adaptiert

Abbildung 6. Worterkennungsrate mit und ohne Sprecheradaptierung (Set II, Störsignal I)

Ein Schritt zur Steigerung der Erkennungsrate ist die zusätzliche Adaption des Systems an den Sprecher. Abbildung 6b zeigt, dass eine Adaptierung an den Sprechers, und so auch

an das Verfahren zur Störgeräuschreduzierung, eine deutlich Steigerung der Erkennungsleistung ermöglicht (bis 40 %). Dabei ist zu beachten, dass der Spracherkenner ohne die Verfahren zur Störgeräuschreduzierung trainiert wurde und nachträglich adaptiert wird. Die zur Adaptierung herangezogenen Aufnahmen besitzen ein SNR von 30 dB.

4 Zusammenfassung

Es konnte gezeigt werden, dass sowohl die Spektrale Subtraktion als auch die Spektrale Subtraktion mit Minimierung des mittleren quadratischen Fehlers der spektralen Kurzzeitamplituden eine Steigerung der Erkennungsleistung erzielt, wenn sich der Client des internetbasierten Sprachdialogsystems in Umgebungen mit stationären Störsignalen befindet. Besonders hervorgehoben hat sich die Spektrale Subtraktion mit sich dynamische anpassenden *spectral floor*, da so unabhängig vom Signal-Rausch-Abstand die größtmögliche Steigerung der Erkennungsrate erzielt werden könnte.

Literatur

[1] EPHRAIM Yariv; MALAH David;. Speech enhancement using optimal non-linear spectral amplitude estimation. *IEEE International Conference Acoustic Speech Signal Processing (Boston)*, pages 1118–1121, 1983.

[2] EPHRAIM Yariv; MALAH David;. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions Acoustic Speech Signal Processing*, ASSP-33(2):443–445, 1985.

[3] BOLL S. F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):pp. 113–120, 1979.

[4] PUDER Henning. Single channel noise reduction using time-frequency dependent voice activity detection. Technical report, Darmstadt University of Technology, Signal Theory, 1999.

[5] MAASE J., HIRSCHFELD D., KOLOSKA U., WESTFELD T., and HELBIG J. Towards an evaluation standard for speech control concepts in real-world scenarios. *EUROSPEECH-2003*, pages 1553–1556, 2003.

[6] PETERS Mike. *Psychoakustische Signalverbesserung und Geräuschreduktion in Kraftfahrzeugen.* PhD thesis, Universität Kaiserslautern - Fachbereich Elektrotechnik und Informationstechnik, April 2002.

[7] CAPPÉ Olivier. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349, 1994.

[8] HOFFMANN R., EICHNER M., WERNER S., and WOLFF M. The project uasr (unified approach for speech synthesis and recognition) - a progress report). pages 17–24, 2003.

[9] PETRICK R., HIRSCHFELD D., HOFFMANN R., and JOKISCH O. Verbkey – a dsp based speech control for the automotive environment. In *Workshop on DSP in Mobile and Vehicular Systems*, 2003.

[10] PETRICK R., KINAST G., and HIRSCHFELD D. Influence of a single channel and a multi channel noise reduction on the recognition of noisy speech. *Studientexte zur Sprachkommunikation - Proc. of the 16. Conf. of Electronic Speech Signal Processing ESSP*, 30:159–166, 2005.