Robuste Sprachverbesserung durch statistische Signalverarbeitung und maschinelles Lernen[1]

Robert Rehr[2]

Abstract: In vielen Anwendungen, z. B. für Hörgeräte, zur Interaktion zwischen Mensch und Maschine, oder für Telekommunikation, spielen Sprachsignale eine besondere Rolle. In Umgebungen, in denen neben dem Sprachsignal auch weitere Quellen präsent sind, nehmen die Mikrofone nicht nur das gewünschte Sprachsignal sondern auch weitere Störgeräusche auf. Hiedurch reduziert sich die Qualität und potentiell auch die Verständlichkeit des aufgenommenen Sprachsignals. Zur Reduktion der Hintergrundgeräusche werden Sprachverbesserungsalgorithmen eingesetzt. Diese Arbeit konzentriert sich hierbei auf einkanalige Verfahren, bei denen Synergien zwischen Verfahren, die auf maschinellem Lernen basieren, und Verfahren der statistische Signalverarbeitung genutzt werden.

1 Einleitung

Sprache ist eine der natürlichsten Formen der Kommunikation für Menschen und wird zum Austausch von Informationen, Ideen und Gefühlen genutzt. Begünstigt durch die Verfügbarkeit von leistungsfähigen, elektronischen Geräten spielt Sprache eine immer wichtigere Rolle in vielen Anwendungen, z. B. in der mobilen Telekommunikation und in Hörhilfen. Neben der zwischenmenschlichen Kommunikation ist Sprache auch ein wichtiger Bestandteil für die Interaktion zwischen Mensch und Maschine, z. B. mit Robotern oder virtuellen persönlichen Assistenten.

Die für die Umsetzung solcher Anwendungen eingesetzten Geräte werden oft in Umgebungen verwendet, in denen neben dem Zielsignal auch weitere Geräusche auftreten. In solchen Situationen nehmen die Mikrofone nicht nur das gewünschte Sprachsignal sondern zusätzlich auch ungewünschte Signale auf. Dies verschlechtert die wahrgenommene Qualität des Sprachsignals und wirkt sich potentiell auch negativ auf die Verständlichkeit aus. Außerdem erhöht sich auch die Leistung automatischer Spracherkennungsalgorithmen. Um die Qualität und, wenn möglich, auch die Verständlichkeit der gestörten Sprache wiederherzustellen, werden Sprachverbesserungsalgorithmen eingesetzt.

In dieser Arbeit werden einkanalige Sprachverbesserungsalgorithmen betrachtet, die entweder das Signal eines einzelnen Mikrofons oder den Ausgang eines mehrkanaligen räumlichen Filters, d. h. eines Mikrofon-Arrays, verarbeiten. Viele solcher Verfahren transformieren zur Verbesserung das verrauschte Zeitsignal in eine Zeit-Frequenz Repräsentation, wobei häufig eine Kurzzeit Fourier-Transformation (short-time Fourier transform, STFT) verwendet wird. Zeitfrequenzpunkte, die hauptsächlich dem Geräusch zugeordnet werden,

[1] Robust Speech Enhancement Using Statistical Signal Processing and Maching Learning

[2] Universität Hamburg, robert.rehr@uni-hamburg.de

werden anschließend mit einer sogenannten Gewichtungsfunktion unterdrückt und auf Werte nahe Null gesetzt. Abschließend wird das Signal zurück in den Zeitbereich transformiert. Einkanalige Sprachverbesserungsalgorithmen lassen sich grob in zwei Kategorien einteilen:

1. Verfahren basierend auf statistischer Modellierung Bei solchen Ansätzen werden die Gewichtungsfunktionen, die auf die STFT-Koeffizienten angewendet werden, in einem statistischen Rahmenwerk hergeleitet. Hierzu werden die Koeffizienten der unverrauschten Sprache und des Rauschens durch parametrische Wahrscheinlichkeitsdichtefunktionen (probability density function, PDFs) modelliert. Die Parameter sind im Allgemeinen durch die Leistungsdichtespektren (power spectral density, PSDs) von Sprache und Rauschen gegeben, die blind aus den verrauschten Beobachtungen geschätzt werden. Hierzu wird häufig angenommen, dass sich das Geräusch über die Zeit weniger stark verändert als Sprache.

2. Verfahren basierend auf maschinellem Lernen (ML) Im Gegensatz zu dem konventionellen Ansatz nutzen ML-basierte Algorithmen repräsentative Beispiele, um die statistischen Eigenschaften der Sprache und des Rauschens zu lernen. Häufig sind ML-basierte Ansätze dadurch motiviert, dass konventionelle Ansätze nicht in der Lage sind, hochinstationären, d. h. sich schnell ändernden Geräuschtypen, zu folgen. Im Gegensatz dazu besteht bei ML-basierten Ansätzen allerdings die potentielle Gefahr, dass Eingangsdaten, die stark von den Trainingsdaten abweichen, nicht korrekt verarbeitet werden.

Das Ziel dieser Arbeit ist es, die Robustheit einkanaliger Verfahren zur Sprachverbesserung zu erhöhen. Um dieses Ziel zu erreichen, werden die beiden oben beschriebenen Ansätze betrachtet und Synergien zwischen beiden Ansätzen genutzt. Die sieben Forschungsbeiträge, auf denen die Arbeit [Re19] basiert, sind hierzu in drei Teile gegliedert und präsentieren Verbesserungen für konventionelle, nicht-ML-basierte Ansätze, ML-basierte Ansätze sowie Kombinationen aus beiden Verfahren. Die folgenden Abschnitte dieser Kurzzusammenfassung geben eine Übersicht über die drei Teile der Dissertation.

2 Geräusch-PSD Schätzer basierend auf adaptiver Glättung

Das Bestimmen der Geräusch-PSD ist äquivalent zur Bestimmung der Varianz der komplexwertigen spektralen Koeffizienten des Geräuschs $N_{k,\ell}$. Hierbei symbolisiert $N_{k,\ell}$ die STFT des Geräuschs, wobei k der Frequenzindex ist und ℓ der Zeitindex. Die Fouriertransformation erlaubt es, die spektralen Koeffizienten durch eine mittelwertfreie Verteilung zu beschreiben. Aufgrund dessen kann die Varianz des Geräuschkoeffizienten $\Lambda_{k,\ell}^{(\mathrm{n})}$ eines spektralen Koeffizienten durch den Erwartungswert $\Lambda_{k,\ell}^{(\mathrm{n})} = \mathbb{E}\{|N_{k,\ell}|^2\}$ definiert werden. In praktischen Anwendungen wird die Berechnung des Erwartungswert häufig durch eine Mittelung der Betragsquadrate $|N_{k,\ell}|^2$ über der Zeit ℓ bestimmt. Die Mittelung findet dabei für jedes Frequenzband k separat statt. In der Signalverarbeitung werden hierzu häufig rekursive Glättungsfilter erster Ordnung verwendet. Solche Filter ermöglichen es, Änderungen entlang der Zeit zu verfolgen und zusätzlich lässt sich zeigen, dass diese Filter erwartungstreue Schätzer sind. Im Kontext der Geräusch-PSD-Schätzung ließe sich ein

solches Filter durch

$$\hat{\Lambda}_{k,\ell}^{(\mathrm{n})} = (1-\alpha)|Y_{k,\ell}|^2 + \alpha\hat{\Lambda}_{k,\ell-1}^{(\mathrm{n})} \tag{1}$$

beschreiben, wobei $\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}$ die Schätzung der Geräusch-PSD $\Lambda_{k,\ell}^{(\mathrm{n})}$ ist. Hierbei ist $0 < \alpha < 1$ die Glättungskonstante, welche die Stärke der Glättung kontrolliert. Zusätzlich ist $Y_{k,\ell}$ die STFT des verrauschten Sprachsignals, d. h. $Y_{k,\ell} = S_{k,\ell} + N_{k,\ell}$ und $S_{k,\ell}$ ist die STFT des unverrauschten Sprachsignals.

Aufgrund des zusätzlichen Sprachsignals würde eine einfache Anwendung des Filters in (1) allerdings dazu führen, dass auch Sprachanteile in die Schätzung $\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}$ einfließen. Um dies zu verhindern, tauschen die Geräusch-PSD-Schätzer in [GH11] und [HS04, Kap. 14.1.3] den festen Parameter α gegen einen zeitlich veränderlichen aus. Hierzu wird in Abschnitten, in denen Sprache anwesend ist, der Glättungsparameter auf einen größeren Wert gesetzt, wodurch die Aktualisierung der Schätzung verlangsamt wird. In Abschnitten, in denen nur das Hintergrundgeräusch präsent ist, wird der Glättungsparameter auf einen kleineren Wert zurückgestellt. Die Wahl des Glättungsparameters hängt bei den Verfahren in [GH11, HS04] von dem Verhältnis $|Y_{k,\ell}|^2/\hat{\Lambda}_{k,\ell-1}^{(\mathrm{n})}$ ab. In der vorliegenden Arbeit konnten wir zeigen, dass die Einführung des zeitlich veränderlichen Glättungsparameters allerdings dazu führt, dass das rekursive Filter nicht länger ein erwartungstreuer Schätzer ist. Das bedeutet, dass die Größe entweder unterschätzt wird, was zu einer verringerten Geräuschreduktion führt, oder überschätzt wird, was zu einer Verzerrung des Sprachsignals bei der Verbesserung führt. In [Re19, Kap. 3 und Kap. 4] werden die Geräusch-PSD-Schätzer, welche auf einer solchen adaptiven rekursiven Glättung basieren, analysiert. Darüber hinaus werden in dieser Arbeit Methoden zur Bestimmung des Fehlers und zur Kompensation des Fehlers vorgestellt. Hierzu werden die in [GH11, HS04] vorgestellten Geräusch-PSD-Schätzer als Beispiele verwendet.

In [Re19, S. 45ff] wird gezeigt, dass die betrachteten Geräusch-PSD-Schätzer skalierungsinvariant sind. Wenn das Signal nur Geräusch enthält, lässt sich der systematische Fehler aufgrund dieser Eigenschaft durch die Multiplikation eines einzelnen Faktors $\mathscr{C}$ korrigieren. Der Faktor kann hierzu auf das Eingangs- oder das Ausgangssignal angewendet werden. In [Re19, Kap. 4] wird eine weitere Methode zur Korrektur des Schätzfehlers vorgestellt, bei der ebenfalls nur ein einzelner Faktor $\mathscr{C}^{(\mathrm{a})}$ zur Korrektur des Fehlers notwendig ist, wenn das Signal nur das Geräusch enthält. Im Gegensatz zu der in [Re19, Kap. 3] vorgestellten Methode findet die Korrektur an einer anderen Stelle der rekursiven Struktur statt, sodass im Allgemeinen $\mathscr{C} \neq \mathscr{C}^{(\mathrm{a})}$ gilt.

Aufgrund der Rekursion und der Nichtlinearität der adaptiven Glättungsfilter, ist die analytische Bestimmung von $\mathscr{C}$ als auch $\mathscr{C}^{(\mathrm{a})}$ eine besondere Herausforderung. In [Re19, Kap. 3 und 4] werden analytische Lösungen vorgestellt, mit denen beide Korrekturfaktoren näherungsweise bestimmt werden können. Beide Ansätze verfolgen hierzu das Ziel den Erwartungswert des Geräuschschätzers, d. h. $\mathbb{E}\{\hat{\Lambda}_{k,\ell}^{(\mathrm{n})}\}$, zu bestimmen. Hieraus lässt sich die Abweichung von der wahren Geräusch-PSD bestimmen, woraus sich der benötigte Korrekturfaktor ableiten lässt. Bei der Anwendung der festen Korrekturfaktoren $\mathscr{C}$ und $\mathscr{C}^{(\mathrm{a})}$ wird nicht berücksichtigt, dass das verrauschte Eingangssignal auch Sprache enthält. Daher

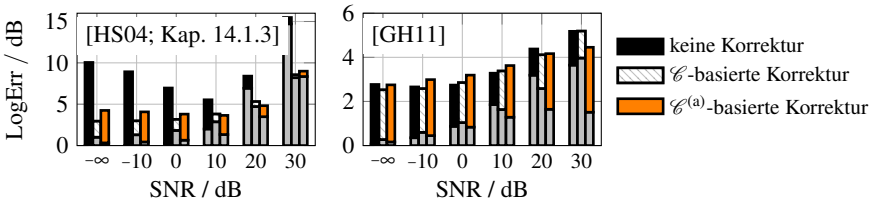


Abb. 1: Überschätzung (unterer, grauer Teil) und Unterschätzung (oberer, farbiger Teil) der Geräusch-PSD für die in [HS04, Kap. 14] und [GH11] beschriebenen Schätzer mit und ohne die in [Re19, Kap. 3 und Kap. 4] vorgestellte Korrektur im Cafeteria-Geräusch.

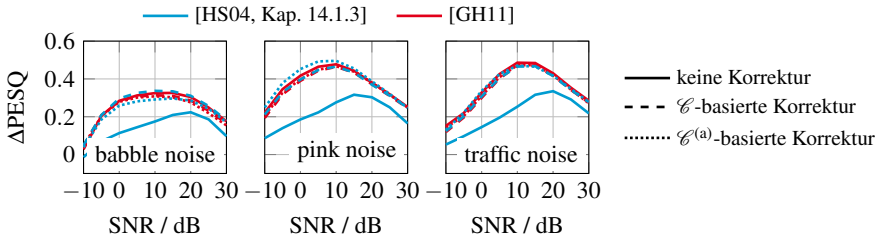


Abb. 2: Vergleich der in [Re19, Kap. 3 und 4] vorgestellten Korrekturmethoden für die Geräusch-PSD-Schätzer in [GH11] und [HS04, Kap. 14.1.3] in Bezug auf PESQ-Verbesserungen.

werden in in [Re19, Kap. 3.5] und [Re19, Kap. 4.2] Erweiterungen vorgeschlagen, um das Sprachsignal bei der Korrektur mit einzubeziehen.

Abbildung 1 zeigt die logarithmische Abweichung für beide in [Re19, Kap. 3 und Kap. 4] untersuchten Geräusch-PSD-Schätzer in einem Cafeteria-Hintergrundgeräusch mit und ohne die vorgeschlagenen Korrekturansätze. Das Fehlermaß erlaubt es die Über- und Unterschätzung der Geräusch-PSD zu quantifizieren. Bei dem in [GH11] vorgestellten Geräusch-PSD-Schätzer ist der Fehler im allgemeinen relativ gering. Aufgrund dessen haben die vorgeschlagenen Korrekturmethoden hier nur eine geringe Auswirkung. Für das in [HS04, Kap. 14.1.3] vorgestellte Verfahren lässt sich hingegen erkennen, dass bei hohen Eingangs-Signal-zu-Rausch Verhältniss (signal-to-noise ratio, SNRs), d. h. wenn das Sprachsignal lauter ist als das Geräusch, die Geräusch-PSD ohne Korrektur überschätzt wird. Im Gegensatz dazu wird die Geräusch-PSD bei niedrigen SNRs, also wenn Sprache deutlich leiser ist als das Geräusch, unterschätzt. Beide Fehlschätzungen lassen sich mit den vorgestellten Korrekturmethoden effektiv verringern.

Zusätzlich wurden die vorgeschlagenen Korrekturverfahren mit Perceptual Evaluation of Speech Quality (PESQ) [P.01] untersucht, einem instrumentellen Maß, dass die Sprachqualität der verbesserten Sprachsignale algorithmisch vorhersagt. Abbildung 2 zeigt die Verbesserung von PESQ gegenüber dem verrauschten Signal, wobei höhere Werte eine bessere Qualität widerspiegeln. Ähnlich zur in Abbildung 1 dargestellten logarithmischen Abweichung hat die Korrektur nur einen geringen Einfluss auf den in [GH11] vorgestellten Geräusch-PSD-Schätzer, während sich für das Verfahren in [HS04, Kap. 14.1.3] ein klarer positiver Effekt beobachten lässt. Die $\mathscr{C}$-basierte und die $\mathscr{C}^{(a)}$-basierte Korrekturmethode liefern ähnliche Verbesserungen.

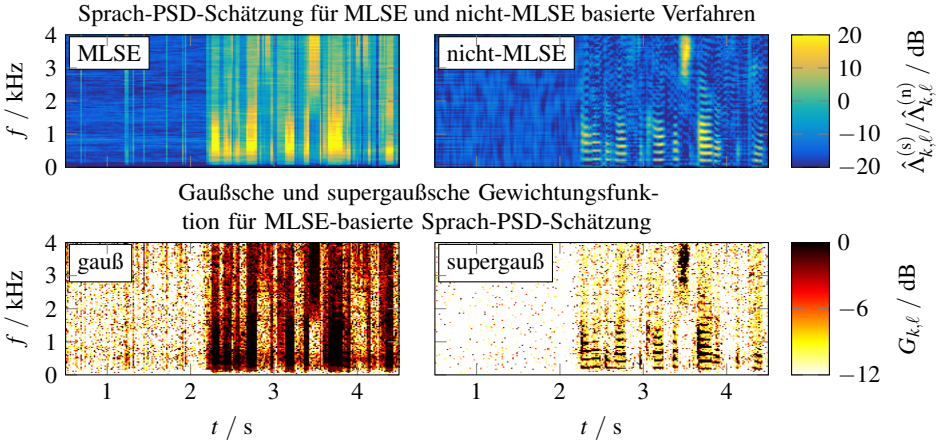


Abb. 3: Oben: Verhältnis der Sprach-PSD und der Rausch-PSD eines nicht-MLSEs und eines MLSEs. Unten: Gewichtungsfunktion eines gaußschen und eines supergaußschen Sprachschätzers für den Fall, dass die MLSE-basierte Sprach-PSD Schätzung verwendet wird. Gleiches Sprachsignal in allen vier Fällen.

3 ML-basierte Spracheinhüllendeverfahren

Im zweiten Abschnitt der Arbeit [Re19, Kap. 5 – 7], werden ML-basierte Sprachverbesserungsansätze adressiert, die typische Strukturen der Sprache erlernen. Für die Verarbeitung wird anschließend das Modell selektiert, das die Beobachtung am besten erklärt. Das Besondere an den in diesem Teil betrachteten Methoden ist, dass die betrachteten ML-basierten Ansätze nur eine grobe spektrale Einhüllende der Sprache repräsentieren. Diese Art von Verfahren werden im folgenden als ML-basierte Spracheinhüllendenverfahren (machine-learning spectral envelope, MLSE) bezeichnet und werden unter anderem zur Sprachverbesserung eingesetzt [CGG16]. Die Form der spektralen Einhüllende von Sprache resultiert aus den Resonanzen, die durch den Vokaltrakt des Menschens erzeugt werden, und ermöglicht es verschiedene Phoneme, z. B. die Vokale, akustisch voneinander zu unterscheiden. Allerdings enthält die spektrale Einhüllende nicht die spektrale Feinstruktur, die dem Grundton und deren Harmonischen entspricht, die durch das Schwingen der Stimmlippen in stimmhaften Lauten erzeugt werden. MLSE haben zum einen den Vorteil, dass sie sehr gut generalisieren und zu recheneffizienten Lösungen führen. Allerdings überschätzen MLSE-Ansätze die Sprach-PSD zwischen den spektralen Harmonischen der Sprache. Dadurch wird das Geräusch zwischen diesen Harmonischen typischerweise nicht unterdrückt. Infolgedessen ist die Geräuschreduktion in sprachaktiven Segmenten begrenzt und führt zu hörbaren Aktivierungen des Geräuschs in solchen Abschnitten. Der obere Teil von Abbildung 3 zeigt das Verhältnis der geschätzten Sprach-PSD $\hat{\Lambda}^{(s)}_{k,\ell}$ und der Geräusch-PSD $\hat{\Lambda}^{(n)}_{k,\ell}$ für ein MLSE- und ein nicht-MLSE-Verfahren. Das Hintergrundgeräusch ist ein stationäres rosa Rauschen, das mit einem SNR von 5 dB dem Sprachsignal hinzugefügt wurde. Während sich für das nicht-MLSE-Verfahren klar die harmonischen Strukturen von Sprache erkennen lassen, gehen diese bei dem hier betrachteten MLSE-Verfahren verloren.

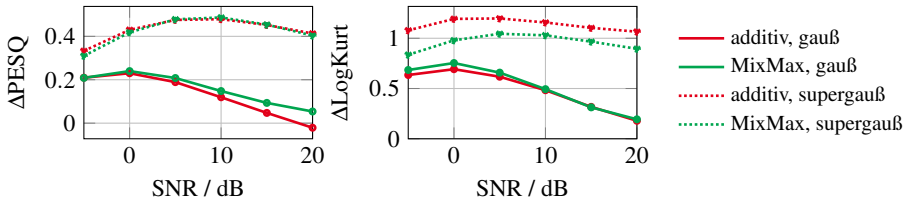


Abb. 4: Vergleich der Sprachqualität und der Artefakte von gaußschen und supergaußschen Sprachschätzern, die jeweils unter dem additiven Modell und dem MixMax-Modell hergeleitet wurden, bei MLSE-Verfahren.

In [Re19] zeigen wir, dass Sprachschätzer, die auf einer supergaußschen Annahme zur Modellierung der Sprachkoeffizienten basieren, das Problem der Unterdrückung des Störgeräuschs ohne heuristische Nachverarbeitung lösen. Die Untersuchungen zeigen, dass Beobachtungen $|Y_{k,\ell}|^2$, die einen ähnlichen Wert haben wie die geschätzte Geräusch-PSD $\hat{\Lambda}^{(\mathrm{n})}_{k,\ell}$, mit einem supergaußschen Schätzer stärker unterdrückt werden. Dies gilt insbesondere auch für den Fall, wenn die Sprach-PSD $\Lambda^{(\mathrm{s})}_{k,\ell}$ deutlich größer als die Geräusch-PSD $\Lambda^{(\mathrm{n})}$ ist, was bei einer Überschätzung der Fall ist. Der untere Teil von Abbildung 3 stellt die Gewichtungsfunktion $G_{k,\ell}$ eines gaußschen und eines supergaußschen Sprachschätzer, wenn eine MLSE-basierte Sprach-PSD-Schätzung verwendet wird. Das Ergebnis zeigt, dass ein Schätzer basierend auf der supergaußschen Annahme die Feinstruktur des Sprachsignals wiederherstellen kann, was mit einem gaußschen Schätzer hingegen nicht möglich ist.

In [Re19] untersuchen wir neben dem additiven Signalmodell auch Sprachschätzer, die unter dem MixMax-Modell hergeleitet wurden. Das MixMax-Modell arbeitet im log-spektralen Bereich, der durch $\log(|Y_{k,\ell}|^2)$ definiert ist, und modelliert das verrauschte Log-Spektrum durch $\log(|Y_{k,\ell}|^2) = \max(\log(|S_{k,\ell}|^2), \log(|N_{k,\ell}|^2))$. Log-spektrale Repräsentationen sind gut geeignet, um generalisierende Sprachmodelle zu trainieren und werden daher häufig auch zur Spracherkennung eingesetzt. Durch die nichtlineare Transformation ist allerdings der sonst üblicherweise genutzte additive Zusammenhang zwischen Sprache und Rauschen mathematisch nicht länger handhabbar. Hierbei wird das MixMax-Modell zur mathematischen Vereinfachung genutzt [CGG16]. In [GM09] wird gezeigt, dass Spektralkoeffizienten, die einer supergaußschen Verteilung folgen, eine höhere Varianz im Log-Spektralbereich aufweisen als gaußverteile Spektralkoeffizienten. Mit diesem Zusammenhang wird gezeigt [Re19, Kap. 6], dass sich der Vorteil von supergaußschen Schätzverfahren auch für MixMax-basierte Sprachschätzer übertragen lässt. Die Verwendung des MixMax-Modells hat dabei den Vorteil, dass weniger wahrnehmbare Prozessierungsartefakte auftreten. Dieses Ergebnis wurde instrumentell mit dem Log-Kurtosis Verhältnis verifiziert, das in Abbildung 4 neben dem Sprachqualitätsmaß PESQ dargestellt ist. Höhere Werte des Log-Kurtosis-Verhältnis entsprechen dabei einem häufigeren Auftreten von Prozessierungsartefakten. Das supergaußsche Modell unter dem additiven und dem MixMax-Modell liefern ähnliche Ergebnisse im Bezug auf die Sprachqualität, die durch PESQ vorhergesagt wird. Allerdings ist Log-Kurtosis-Verhältnis für den MixMax-basierten sußergaußschen Schätzer deutlich kleiner, was auf weniger Artefakte hindeutet. Zusätzlich wurde die Effektivität von supergaußschen Schätzern in einem Hörversuch verifiziert [Re19, Kap. 5.6].

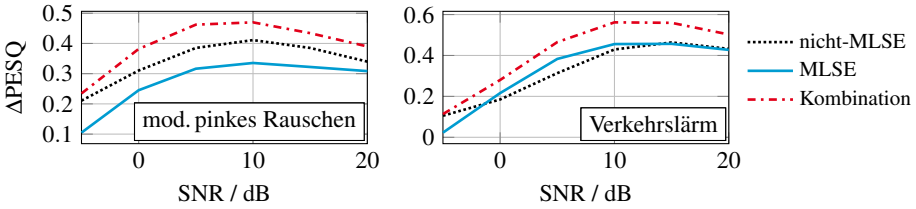


Abb. 5: Verbesserungen in PESQ für die statistische Kombination von konventionellen Sprachverbesserungsalgorithmen und MLSE-Verfahren.

In [Re19, Kap. 7] wird ein Ansatz verfolgt, bei dem mit Hilfe einer Kombination von MLSE-Verfahren und konventionellen Verfahren die Feinstruktur wiederhergestellt wird. Die Kombination der beiden Verfahren wird mit Hilfe eines statistischen Rahmenwerks realisiert. Das Resultat der Kombination ist, dass die konventionellen Algorithmen das Geräusch zwischen spektralen Harmonischen unterdrücken, während das MLSE-Verfahren hauptsächlich die Unterdrückung von Koeffizienten verhindert, die von Sprache dominiert sind. Dementsprechend wird bei einer Geräuschreduktion, die vergleichbar mit konventionellen Ansätzen ist, weniger Sprache verzerrt. Abbildung 5 zeigt die Verbesserungen der vorgeschlagenen Kombination von konventionellen und MLSE-Ansätzen gegenüber reinen nicht-MLSE- und MLSE-Ansätzen in Form von PESQ-Verbesserungen.

4 Generaliserung bei DNN-basierter Sprachverbesserung

Im dritten Teil der Arbeit [Re19, Kap. 8] wird die Generalisierung von ungesehenen akustische Umgebungen im Zusammenhang mit ML-basierten Verbesserungsverfahren, die auf tiefen neuronalen Netzwerken (deep neural networks, DNNs) basieren, untersucht. Bei DNNs handelt es sich um eine Methode des MLs, die es ermöglicht beliebige Funktionen auf einem beschränkten Raum zu approximieren. Solche Verfahren wurden kürzlich mit vielversprechenden Ergebnissen für die Sprachverbesserung eingesetzt [Xu15]. Ein übliches Problem, dass für viele Verfahren des MLs gilt, trifft auch auf DNNs zu und zwar, wie gut das gelernte Modell ungesehene Eingangsdaten generalisieren kann. Diese Fragestellung ist auch für DNN-basierte Sprachverbesserungssysteme untersucht worden [Xu14, KTJ17], wobei in [Xu14] das geräuschbasierte Training (noise-aware training, NAT) vorgeschlagen wurde, um eine erhöhte Robustheit von DNN-basierter Sprachverbesserung zu erzielen. Hierbei wird eine Schätzung der Geräusch-PSD an die Merkmale, die aus dem verrauschten Eingangssignal extrahiert werden, angehängt. Die Geräusch-PSD kann mit konventionellen Verfahren, z. B. [GH11], erfolgen, die im Allgemeinen den Vorteil haben, unabhängig von der akustischen Umgebung zu sein.

In [Re19, Kap. 8] stellen wir eine neuartige Methode zur Erhöhung der Robustheit von DNN-basierten Sprachverbesserungsalgorithmen vor. Hierzu werden Schätzungen der Sprach- und der Geräusch-PSD als Eingangsmerkmale für ein DNN-basiertes Sprachverbesserungsverfahren verwendet, die aufgrund der Robustheit mit einem konventionellen bestimmt werden. Im Gegensatz zum NAT werden aber SNRs als Eingangsmerkmale verwendet. Zum einen wird das *a priori* SNR, d. h. das Verhältnis zwischen Sprach- und

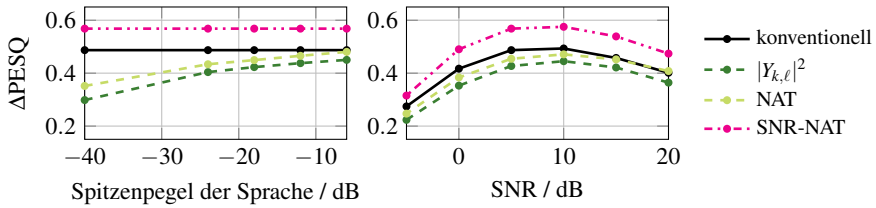


Abb. 6: Vergleich der verschiedener Eingangsmerkmale für DNN-basierte Sprachverbesserungsverfahren mittels PESQ. Zum Vergleich sind auch die Ergebnisse eines konventionellen Ansatzes dargestellt.

Geräusch-PSD $\Lambda_{k,\ell}^{(\mathrm{s})}/\Lambda_{k,\ell}^{(\mathrm{n})}$, vorgeschlagen. Zum anderen wird auch das *a posteriori* SNR, d. h. das Verhältnis der verrauschten Eingangsperiodogramm $|Y_{k,\ell}|^2$ und der Geräusch-PSD $\Lambda_{k,\ell}^{(\mathrm{n})}$, genutzt. Die beiden SNRs können separat verwendet werden oder durch Aneinanderhängen kombiniert werden. Letzteres wird als SNR-basiertes NAT (SNR-based NAT, SNR-NAT) bezeichnet wird.

Die NAT- und SNR-NAT-Merkmale werden in [Re19] experimentell mit der Methode der Kreuzvalidierung verglichen. Hierzu wird eine Menge von neun verschiedene Geräuschtypen, mit denen Sprachsignale künstlich verrauscht werden, betrachtet. Jedes Sprachsignal wird mit allen verfügbaren Störsignalen verrauscht und für jedes Sprachsignal wird das SNR als auch der Gesamtpegel zufällig anders gewählt, um das DNN lernen zu lassen, auf solche Variationen richtig zu reagieren. Für jedes der untersuchten Eingangsmerkmale, werden neun verschiedene Modelle trainiert, wobei alle verfügbaren Geräusche bis auf eins für das Training verwendet werden. Bei jedem dieser neun Modelle gibt es daher ein Geräusch, das nicht während des Trainings gesehen wurde. Außerdem wurden zusätzliche Geräuschtypen in dem Trainingsset aller Modelle eingefügt, um die Robustheit des DNN-basierten Verfahrens allgemein zu stärken. Die Größe des Trainingssatzes für jedes Modell entspricht dabei ungefähr 20 Stunden Audiomaterial, die genutzt werden, um ein Feedforward-Netzwerk mit drei versteckten Ebenen zu trainieren.

Die in Abbildung 6 dargestellten Ergebnissen zeigen den Mittelwert über alle ausgewerteten Geräuschtypen, d. h. ohne die für das Training zusätzlich eingefügten Geräuschtypen. Die Ergebnisse zeigen, dass das vorgeschlagene SNR-NAT gegenüber NAT zwei wesentliche Vorteile hat: Im linken Teil von Abbildung 6 sind die PESQ-Verbesserung für verschiedene Spitzenpegel der Sprache, die im direkten Zusammenhang mit dem Gesamtpegel des Signals stehen, dargestellt. Das Resultat zeigt, dass das DNN-basierte Verfahren mit der Verwendung des vorgeschlagenen SNR-NAT über den gesamten Pegelbereich nahezu identische Ergebnisse liefert, während die Qualität von NAT abhängig vom Gesamtpegel ist. Das bedeutet, dass das vorgeschlagenen SNR-NAT nicht durch den Gesamtpegel des Eingangssignals beeinflusst wird. Zweitens führt das vorgeschlagene SNR-NAT zu robusteren Modellen als NAT, wie der rechte Teil in Abbildung 6 zeigt. Dies lässt sich an den PESQ-Werten für das SNR-NAT-Verfahren erkennen, die über einen weiten SNR-Bereich über den verglichenen Verfahren liegen. Dies gilt insbesondere für den Fall, wenn wie im durchgeführten Experiment wenige Trainingsdaten zur Verfügung stehen.

Die Ergebnisse sind mit Hilfe eines Hörversuchs mit elf Teilnehmern verifiziert worden, bei dem die Qualität der verbesserten Signale bewertet wurde. Hierfür wurden zwei Hintergrundgeräusche untersucht und zwar, Fabriklärm und Verkehrslärm, die beide für die Verbesserungsverfahren nicht während des Trainings zur Verfügung standen. Die Resultate des Hörversuchs zeigen, dass die vorgeschlagenen SNR-NAT Merkmale signifikant besser bewertet werden als NAT, wenn das Geräusch nicht aus dem Training bekannt ist.

5 Zusammenfassung

Die Dissertation betrachtet verschiedene Aspekte der einkanaligen Sprachverbesserung und nutzt Synergien zwischen konventionellen Sprachverbesserungsalgorithmen und ML-basierten Verfahren aus. In diesem Rahmen sind verschiedene Beiträge zur Verbesserung konventioneller und ML-basierter Verfahren entstanden.

Zum einen wurden Verbesserungen für konventionelle Verfahren zur Schätzung der Geräusch-PSD vorgeschlagen, um systematische Schätzfehler zu vermeiden. Hierzu wurden verschiedene analytische Lösungen vorgestellt, die es ermöglichen den Schätzfehler approximativ zu bestimmen. Zusätzlich sind Methoden vorgestellt worden, mit denen der systematische Schätzfehler kompensiert werden kann. Die Evaluation zeigt, dass insbesondere Geräusch-PSD-Schätzer, die einen großen Schätzfehler aufweisen, von der Korrektur des Fehlers profitieren.

Zusätzlich wurden supergaußsche Sprachschätzer zur Verbesserung von MLSE-basierten Sprachverbesserungsalgorithmen vorgeschlagen. MLSE-basierte Verfahren verwenden trainierte Sprachmodelle, bei denen aber nur die spektrale Einhüllende des Sprachsignals abgebildet wird. Im Gegensatz zu Methoden, bei denen eine heuristische Nachverarbeitung eingesetzt wird, um die spektrale Feinstruktur wiederherzustellen, ermöglichen es supergaußsche Schätzer die Feinstruktur ohne diesen Umweg zu rekonstruieren. Die Effektivität dieser Methode ist durch die Evaluationen mit instrumentellen Maßen als auch durch Hörversuche verifiziert worden. Neben dem Einsatz von supergaußschen Schätzern ist außerdem eine Kombination von konventionellen Methoden und MLSE-Methoden vorgeschlagen worden, die in einem statistischen Rahmenwerk eingebettet wurde. Die Effektivität der Methode konnte auch hier mittels instrumenteller Maße verifiziert werden.

Zuletzt wurde die Generalisierung DNN-basierter Sprachverbesserungsverfahren betrachtet. Hierzu wurden SNR-basierte Merkmale vorgestellt, wobei für die Schätzung der Sprach- und Geräusch-PSD konventionelle Verfahren eingesetzt werden. In den Auswertungen konnte gezeigt werden, dass das vorgeschlagene SNR-NAT zwei Vorteile gegenüber dem zuvor vorgestellten NAT hat. Zum einen sind die Merkmale skalierungsinvariant, sodass die Leistung des Sprachverbesserungsalgorithmus nicht länger vom Gesamtpegel des Eingangssignals abhängen. Des Weiteren zeigen die Evaluationen mit instrumentellen Maßen und Hörversuche, dass diese Merkmale außerdem zu robusteren Modellen im Bezug auf den Einsatz in ungesehenen akustischen Umgebung führen.

Literaturverzeichnis

[CGG16] Chazan, S. E.; Goldberger, J.; Gannot, S.: A Hybrid Approach for Speech Enhancement Using MoG Model and Neural Network Phoneme Classifier. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(12):2516–2530, Dezember 2016.

[GH11] Gerkmann, Timo; Hendriks, Richard. C.: Noise Power Estimation Based on the Probability of Speech Presence. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY, USA, S. 145–148, 2011.

[GM09] Gerkmann, Timo; Martin, Richard: On the Statistics of Spectral Amplitudes after Variance Reduction by Temporal Cepstrum Smoothing and Cepstral Nulling. IEEE Transactions on Signal Processing, 57(11):4165–4174, 2009.

[HS04] Hänsler, Eberhard; Schmidt, Gerhard: Acoustic Echo and Noise Control: A Practical Approach. Adaptive and Learning Systems for Signal Processing, Communication and Control. Wiley & Sons, 2004.

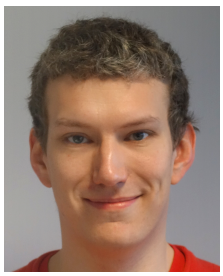
[KTJ17] Kolbæk, M.; Tan, Z. H.; Jensen, J.: Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1):149–163, Januar 2017.

[P.01] : P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. ITU-T recommendation, International Telecommunication Union, Januar 2001.

[Re19] Rehr, Robert: Robust Speech Enhancement Using Statistical Signal Processing and Machine-Learning. Dissertation, Universität Hamburg, Hamburg, Januar 2019.

[Xu14] Xu, Yong; Du, Jun; Dai, Li-Rong; Lee, Chin-Hui: Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks. In: Interspeech. Singapore, Singapore, S. 2670–2674, September 2014.

[Xu15] Xu, Y.; Du, J.; Dai, L. R.; Lee, C. H.: A Regression Approach to Speech Enhancement Based on Deep Neural Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1):7–19, Januar 2015.



Robert Rehr hat Hörtechnik und Audiologie studiert und seinen B.Eng. an der Jade Hochschule in Oldenburg 2011 abgeschlossen und seinen M.Sc. an der Carl-von-Ossietzky Universität in Oldenburg 2013 erhalten. Von 2013 bis 2016 hat er in der Speech Signal Processing Group der Carl-von-Ossietzky Universität, Oldenburg seine Doktorarbeit begonnen, wo er von Sep. 2014 bis Feb. 2015 an einem Projekt mit „Sivantos – the hearing company" zusammengearbeitet hat. Er hat seine Doktorarbeit von 2017 bis. 2018 in der Signal Processing Group der Universität Hamburg abgeschlossen und war dort bis 2019 als wissenschaftlicher Mitarbeiter tätig. Er arbeitet seit Mai 2019 bei Oticon an Geräuschreduktionsalgorithmen für Hörgeräte.