SCHÄTZUNG DER SPEKTRALEN EINHÜLLENDEN - EIN VERGLEICH VON TIEFEN NEURONALEN NETZEN UND CODEBÜCHERN

Christopher Seitz, Mohammed Krini

Hochschule für angewandte Wissenschaften Aschaffenburg
christopher.seitz@t-online.de

Kurzfassung: Viele Methoden der Sprachsignalverbesserung, wie die Bandbreitenerweiterung oder die Geräuschreduktion, sind modellbasiert und basieren auf dem sogenannten Quelle-Filter-Modell der Vokalproduktion. Mit Bezug auf das Modell lässt sich Sprache in zwei Komponenten aufteilen. Die Stimmbänder generieren einen Laut, mit eigener spektralen Form und Struktur (Quelle-Teil), der anschließend von den Resonanzeigenschaften des Vokaltrakts gefiltert wird (Filter-Teil). Daher sind zuverlässige Methoden zur Schätzung des Filter-Teils (spektrale Einhüllende) und des Quelle-Teils (Anregungssignal) notwendig. Der Fokus dieser Arbeit liegt auf der Schätzung der spektralen Einhüllenden, da diese sehr wichtig ist für die Rekonstruktion von stark gestörten Sprachsignalen. Konventionelle Methoden liefern keine zuverlässige Schätzung der spektralen Einhüllenden mit hohem Geräuschanteil. Daher werden zwei verschiedene Verfahren vorgestellt und analysiert. Zum einen ein Codebuch, das ungestörte spektrale Einhüllenden enthält. Das Codebuch wird mittels *k-Means* Algorithmus in unterschiedlichen Größen trainiert. Die Einhüllenden des Codebuch-Verfahrens werden mit der Schätzung der Einhüllenden eines tiefen rekurrenten neuronalen Netzes (RNN) verglichen. Weiterhin werden verschiedene Optimierungen bei dem tiefen RNN angewendet, um die Mappingfunktion zwischen gestörter und ungestörter spektralen Einhüllenden zu verbessern. Die Qualität der erhaltenen Modelle wird durch ein objektives Maß analysiert.

1 Einführung

In den vergangenen Jahren wurde der Forschung in der Sprachsignalverarbeitung immer mehr Aufmerksamkeit gewidmet. Viele Applikationen, wie zum Beispiel Hörgeräte, Mobiltelefone und Freisprecheinrichtungen, benötigen eine gute Verarbeitung des Sprachsignals. Es gibt verschiedene Techniken der Geräuschreduktion von Mikrofonsignalen, die in der Literatur beschrieben sind. Die Verfahren lassen sich, je nach Anzahl der verwendeten Mikrofone, in zwei Klassen unterscheiden: einkanalige und mehrkanalige Geräuschreduktion. Mehrkanalige Verfahren nutzen die räumliche Verteilung von Sprache und Geräusch aus [1]. Einkanalige Verfahren, wie die spektrale Gewichtung, sind heutzutage weit verbreitet und finden ihren Einsatz in vielen Applikationen. Das Wiener-Filter basiert auf einer statistischen Optimierung [2]. Dabei wird eine Übertragungsfunktion ermittelt, die das gestörte Sprachsignal durch den mittleren quadratischen Fehler von dem Geräusch befreit. All diese konventionellen Methoden verbessern die Qualität des Sprachsignals für hohe und mittlere Signal- zu Geräuschabständen (SNR). Bei starken Störungen kann keine zuverlässige Verbesserung erzielt werden. Dort können Sprachverzerrungen und durch die Geräuschreduktion verursachte Artefakte auftreten. Das Ziel der Sprachverbesserung ist es, die Verständlichkeit und Qualität von stark gestörter Sprache zu verbessern.

Basierend auf dem Quelle-Filter-Modell der Vokalproduktion lässt sich ein Sprachsignal in zwei Komponenten aufteilen. Der Quelle-Teil beschreibt das Signal, das von den Stimmbändern erzeugt wird. Dieses Signal wird anschließend von den Resonanzeigenschaften des Vokaltrakts gefiltert. Dadurch lässt sich jedes beliebige Sprachsignal durch ein Anregungssignal (Quelle-Teil) und eine spektrale Einhüllende (Filter-Teil) beschreiben [3]. Wie schon zuvor erwähnt liefern konventionelle Methoden für ein niedriges SNR keine zuverlässige Schätzungen des ungestörten Sprachsignals. Daher können Sprachrekonstruktionsverfahren [4], basierend auf dem Quelle-Filter-Modell, eingesetzt werden. Für eine erfolgreiche Rekonstruktion ist eine zuverlässige Schätzung der spektralen Einhüllenden von großer Bedeutung. In diesem Beitrag wird der Augenmerk auf den Filter-Teil des Signals gelegt.

Es werden zwei Verfahren, zum einen das Codebuch-Verfahren und zum anderen ein neuronales Netz, für die Einhüllendenschätzung gegenüber gestellt und analysiert. Bei dem Codebuch-Ansatz wird ein Trainingssatz von ungestörten spektralen Einhüllenden mittels *k-Means* Algorithmus trainiert. Zur Gegenüberstellung wird ein rekurrentes neuronales Netz (RNN) mit gestörten und ungestörten Einhüllenden trainiert. Das RNN nutzt die zeitliche Korrelation von Sprache aus und wird mit unterschiedlichen Parametern trainiert. Dabei werden unter anderem die Neuronenanzahl, die Schichten und die Zeitschritte variiert und unterschiedliche Optimierungen durchgeführt.

2 Schätzung der spektralen Einhüllenden

2.1 Analyse-Filterbank

Es wird angenommen, dass das Mikrofonsignal $y(n)$ sich additiv aus dem Sprachsignal $s(n)$ und dem Geräusch $b(n)$ zusammensetzt:

$$y(n) = s(n) + b(n). \tag{1}$$

Zunächst wird das Sprachsignal mit einer Analyse-Filterbank verarbeitet. Diese Filterbank teilt das Signal in überlappende Blöcke auf. Anschließend wird jeder Block mit einer Fensterfunktion $h_{\text{ana},k}$ gewichtet und in den Frequenzbereich transformiert, mittels einer diskreten Fourier Transformation (DFT):

$$Y_\mu(n) = \sum_{k=0}^{N-1} y(nr-k)\; h_{\text{ana},k}\; e^{-j\Omega_\mu k}, \tag{2}$$

wobei r den Versatz zwischen den Blöcken beschreibt. Verwendet wird eine DFT-Ordnung von $N = 512$, eine Überlappung von 75 % und eine Abtastfrequenz von $f_s = 16000$ Hz. Dies ergibt eine Blocklänge von 32 ms. Die resultierenden Teilbandsignale $Y_\mu(n)$ setzen sich additiv aus dem Nutzsignal und dem Störsignal zusammen. Der Index n beschreibt den aktuellen Block. Die Größe μ kennzeichnet die normierten, diskreten Frequenzstützstellen mit

$$\mu \in \{0, 1, ..., N-1\}. \tag{3}$$

2.2 IIR-Glättung des Betragsspektrums

Eine einfache Methode zum Schätzen der spektralen Einhüllenden ist die Frequenzglättung des Sprektrums in Vorwärts- und Rückwärtsrichtung. Diese bidirektionale Glättung wird druch einen *Infinite Impulse Response* (IIR)-Filter erster Ordnung realisiert. Als Eingangsgröße wird

das Betragsspektrum $|Y_\mu(n)|$ genutzt, und es kann eine Frequenzglättung in positiver Frequenzrichtung folgendermaßen beschrieben werden:

$$\overline{Y'}_\mu(n) = \begin{cases} |Y_\mu(n)|, & \text{für } \mu = 0, \\ \lambda_f \overline{Y'}_{\mu-1}(n) + (1-\lambda_f)|Y_\mu(n)|, & \text{für } \mu \in \{1,...,N-1\}. \end{cases} \tag{4}$$

Daraufhin wird eine Glättung in negativer Frequenzrichtung vorgenommen. Dies wird, wie anschließend beschrieben, durchgeführt:

$$\overline{Y''}_\mu(n) = \begin{cases} \overline{Y'}_\mu(n), & \text{für } \mu = N-1, \\ \lambda_f \overline{Y''}_{\mu+1}(n) + (1-\lambda_f)\overline{Y'}_\mu(n), & \text{für } \mu \in \{1,...,N-2\}. \end{cases} \tag{5}$$

Die Zeitkonstante λ_f gibt die Geschwindigkeit der Frequenzglättung an und wurde wie folgt festgelegt:

$$\lambda_f = 0,8. \tag{6}$$

2.3 Sprachaktivitätsdetektor

Zur Sprachaktivitätsdetektion (SAD) wird das Signal-Rausch-Verhältnis (SNR) von dem Eingangssignal im Frequenzbereich bestimmt, vgl. [5]. Für die Extraktion des SNR's wird das Verhältnis aus einem entstörten Kurzzeitspektrum $S_{\hat{s}\hat{s}}(\mu,n)$ und einer geschätzten Geräuschleistung $S_{\hat{b}\hat{b}}(\mu,n)$ gebildet:

$$\widetilde{SNR}(\mu,n) = \frac{S_{\hat{s}\hat{s}}(\mu,n)}{S_{\hat{b}\hat{b}}(\mu,n)}. \tag{7}$$

Das geräuschreduzierte Leistungsdichtespektrum wird durch Subtraktion der geschätzten Geräuschleistung vom Betragsquadrat des momentanen Eingangsspekrums geschätzt:

$$S_{\hat{s}\hat{s}}(\mu,n) = |Y_\mu(n)|^2 - S_{\hat{b}\hat{b}}(\mu,n). \tag{8}$$

An dieser Stelle sei erwähnt, dass durch Schätzfehler der Geräuschleistung negative Werte für das SNR resultieren können. Durch eine Maximumbildung wird dem Problem entgegengewirkt:

$$SNR(\mu,n) = \max\{0, \widetilde{SNR}(\mu,n)\}. \tag{9}$$

Das mittlere SNR wird bestimmt um die Sprachaktivität zu detektieren:

$$\overline{SNR}(n) = \frac{1}{\mu_1 - \mu_0 + 1} \sum_{\mu=\mu_0}^{\mu_1} SNR(\mu,n). \tag{10}$$

Die Frequenzstützstellen μ_0 und μ_1 entsprechen den realen Frequenzen 100 Hz und 4000 Hz. Schließlich wird der Sprachaktivitätsdetektor wie folgt bestimmt:

$$c_{\text{SAD}}(n) = \begin{cases} 1, & \text{für } \overline{SNR}(n) > 0.3, \\ 0, & \text{sonst.} \end{cases} \tag{11}$$

Dabei beschreibt $c_{\text{SAD}}(n)$ in welchem Block eine Sprachaktivität vorhanden ist.

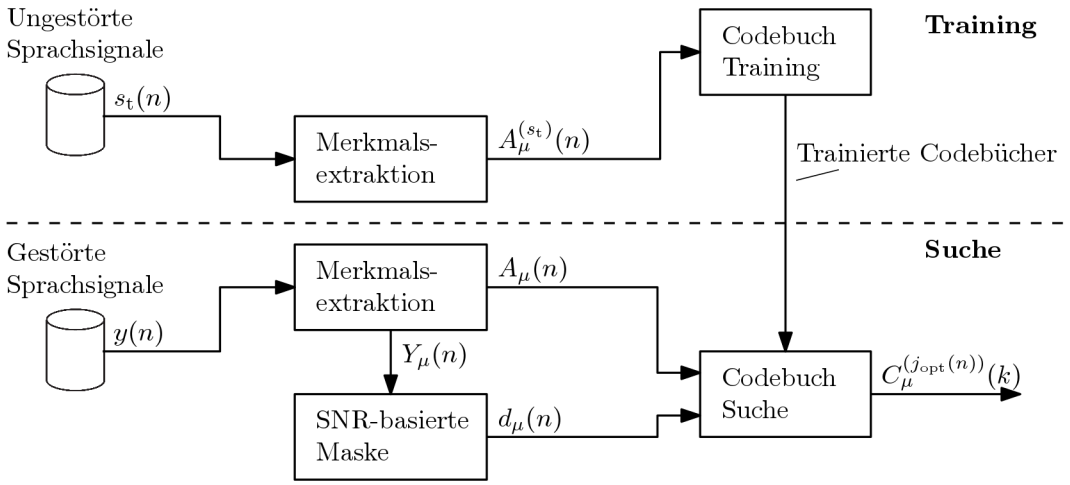

Abbildung 1 – Funktionsübersicht für das Codebuch-Training (oben) sowie die Suche (unten) der spektralen Einhüllenden.

2.4 Codebuch-Verfahren zur Einhüllendenschätzung

Die Sprachdaten für das Codebuch sowie für das RNN stammen aus dem CSTR VCTK Korpus [6]. Die ungestörten Sprachdaten aus dem Korpus wurden mit einen Geräusch beaufschlagt. Das verwendete Geräusch wurde in einem fahrenden Auto bei einer Geschwindigkeit von 140 km/h aufgenommen. Durch variieren der Lautstärke des Geräuschs und anschließender Addition des Sprachsignals mit dem Geräuschsignal können unterschiedliche SNRs erzeugt werden. Abb. 1 zeigt die Funktionsübersicht für das Codebuch-Verfahren. Im oberen Teil ist der Trainingsablauf dargestellt. Für das Training werden ungestörte Sprachsignale $s_t(n)$ verwendet. Diese Sprachsignale durchlaufen die Merkmalsextraktion und es werden die logarithmierten und mittelwertbefreiten Einhüllenden $A_\mu^{(s_t)}(n)$ für das Codebuch-Training bereitgestellt. Das Training nutzt den *k-Means* Algorithmus [7]. Ziel des *k-Means*-Verfahren ist es, die Eingangsdaten $A_\mu^{(s_t)}(n)$ in $N_{cb} \in \{64, 256, 1024\}$ Cluster aufzuteilen. Dabei soll die Summe der quadratischen Abweichungen von dem Clustermittelpunkt minimal sein. Die folgende Kostenfunktion wird iterativ minimiert, damit die optimalen Cluster berechnet werden können:

$$K\left(A_\mu^{(s_t)}(n), C_\mu^{(j)}(k)\right) = \sum_{n=0}^{N_a-1} \sum_{k=0}^{N_{cb}-1} \sum_{\mu=0}^{N/2-1} \varphi_{nk}^{(j)} \left(A_\mu^{(s_t)}(n) - C_\mu^{(j)}(k)\right)^2. \tag{12}$$

Dabei beschreibt $C_\mu^{(j)}(k)$ die Zentroiden der Cluster zum Iterationszeitpunkt j und $C_\mu^{(j_{opt}(n))}(k)$ die optimalen Prototyp-Einhüllenden im letzten Iterationsschritt. Die binäre Maske $\varphi_{nk}^{(j)} \in \{0,1\}$ dient dazu, die Eingangsmerkmale $A_\mu^{(s_t)}(n)$ einem k-ten Cluster zuzuweisen. N_a beschreibt die Anzahl der verwendeten spektralen Einhüllenden für das Training.

Im unteren Teil der Abb. 1 wird die Funktionsübersicht zur Codebuch-Suche dargestellt. Für die Merkmalsextraktion werden gestörte Sprachsignale mit einem SNR aus dem Bereich 0–20 dB genutzt. Danach werden in der SNR-basierten Maske $d_\mu(n)$ für die Codebuch-Suche die Teilbänder ermittelt, die ein hohes SNR aufweisen:

$$d_\mu(n) = \begin{cases} 1, & \text{für } SNR(\mu, n) > \lambda, \\ 0, & \text{sonst,} \end{cases} \tag{13}$$

mit

$$\lambda = 0{,}5. \tag{14}$$

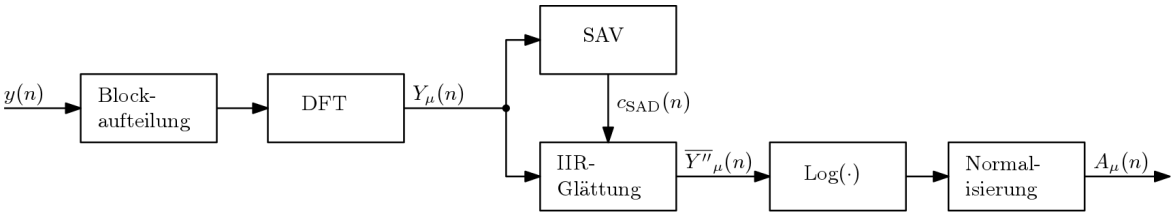

Abbildung 2 – Das genaue Vorgehen bei der Merkmalsextraktion: Für die Normalisierung wird beim Codebuch-Verfahren die sprecherspezifische Lautstärke entfernt. Bei dem RNN wird zusätzlich eine Mittelwertbefreiung über die Frequenz mit anschließender Normierung in der Varianz durchgeführt.

Anschließend wird der beste Codebucheintrag, basierend auf dem quadratischen Abstand, ausgesucht:

$$k_{\mathrm{opt}}(n) = \underset{0\le k\le N_{\mathrm{cb}}-1}{\operatorname{argmin}} \sum_{\mu=0}^{N/2} d_\mu(n)\left(A_\mu(n) - C_\mu^{(j_{\mathrm{opt}}(n))}(k)\right)^2. \tag{15}$$

Der erhaltene Eintrag $C_\mu^{(j_{\mathrm{opt}}(n))}(k)$ beschreibt die Prototyp-Einhüllende, die der beste Ersatz der geräuschbehafteten Einhüllenden $A_\mu(n)$ ist.

Die Merkmalsextraktion ist in Abb. 2 genauer dargestellt. Die Sprachsignale $y(n)$ werden zuerst mit einer Analyse-Filterbank verarbeitet und in den Frequenzbereich transformiert, vgl. Abschnitt 2.1. Anschließend findet für das Codebuch-Verfahren eine Sprachaktivitätsdetektion statt. Es wird nur in den Blöcken eine Einhüllende geschätzt, in denen Sprachaktivität detektiert wurde. Danach werden die Einhüllenden $\overline{Y''}_\mu(n)$ logarithmiert und von der sprecherspezifischen Lautstärke, durch Mittelwertbildung, befreit.

2.5 Rekurrentes Neuronles Netz zur Einhüllendenschätzung

Wie auch beim Codebuch-Ansatz werden bei dem Training des RNNs Sprachsignale aus dem gleichen Korpus verwendet. Dabei werden für das Training sowohl die Einhüllenden während Sprachaktivität sowie die Einhüllenenden während Sprachpausen genutzt. In Abb. 3 ist die Funktionsübersicht dargestellt. Im oberen Teil ist das Vorgehen für das Training zu sehen. Dabei werden ungestörte und gestörte Sprachsignale verarbeitet, da ein überwachtes Lernen stattfindet. Es werden Einhüllendenpaare mit verschiedenen SNRs aus dem Bereich 0–20 dB erzeugt und RNN-Modelle mit unerschiedlichen Parametern trainiert. Die Parameter, welche variiert wurden sind unter anderen die Anzahl der Schichten, Neuronen und Zeitschritte. Die Zeitschritte beschreiben die Anzahl der gestörten Einhüllenden $(A_\mu(n), A_\mu(n-1), ...)$, die für die rekursive Berechnung, genutzt werden. Des Weiteren wurden Modelle mit der *Leaky ReLU*-Aktivierungsfunktion [8] erstellt sowie mit der *ELU*-Aktivierungsfunktion [9]. Es wurde bei allen Modellen die *Long Short-Term Memory* (LSTM)-Zelle verwendet, dadurch konnte der, zuvor aufgetretene, *Exploding Gradient* vermieden werden [10]. Weitergehend wurde die He-Initialisierung und L2-Regularisierung bei allen Modellen angewendet [10]. Eine Normalisierung [11] der Trainingsdaten, sowie Dropout [12] und eine Layernormalisierung [13] kam nur bei einigen Modellen zum Einsatz. Durch die folgende Kostenfunktion wurden die Gewichte in den Modellen optimiert:

$$F\left(A_{\mathrm{est},\mu}(n), A_\mu^{(s_{\mathrm{t}})}(n)\right) = \frac{1}{N_{\mathrm{bs}}}\frac{1}{N/2}\sum_{n=0}^{N_{\mathrm{bs}}-1}\sum_{\mu=0}^{N/2}\left(A_{\mathrm{est},\mu}(n) - A_\mu^{(s_{\mathrm{t}})}(n)\right)^2. \tag{16}$$

Dabei beschreibt $A_{\mathrm{est},\mu}(n)$ die geschätzte Einhüllende des Netzes in der Trainingsphase. N_{bs} beschreibt die Größe der verwendeten Minibatch [10]. Im unteren Teil der Abb. 3 wird der Ablauf der Schätzung, nach Optimierung der Gewichte durch Formel 16, gezeigt. Dabei werden in der

Merkmalsextraktion nur die Blöcke betrachtet, die Sprachaktivität aufweisen, und aus diesen Spektren wird die Einhüllende geschätzt. Danach schätzt das RNN die ungestörte Einhüllende $A_{\mathrm{RNN},\mu}(n)$ aus der gestörten Einhüllenden $A_\mu(n)$.

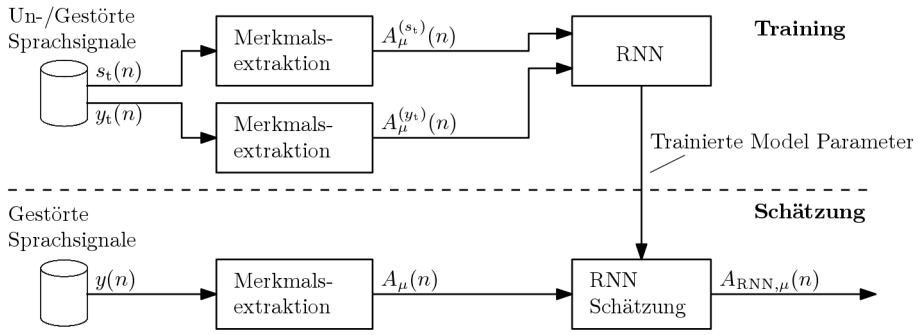

Abbildung 3 – Funktionsübersicht für das rekurrente neuronale Netz. Im oberen Teil ist der Trainingsablauf dargestellt und im unteren Teil die Schätzung der spektralen Einhüllenden.

3 Analyse der beiden Verfahren

Für den Vergleich der beiden Verfahren wurde eine Vielzahl an Sprachsignalen mit unterschiedlichem SNR, aus dem Bereich 0–15 dB, erzeugt. Die Einhüllenden wurden im niederfrequenten Bereich 0–2000 Hz geschätzt, da sich das Geräusch im Fahrzeug am stärksten in dem Bereich auswirkt. Abb. 4 zeigt die mittlere Abweichung der verschiedenen RNN Modelle. Die Berechnung wurde, wie folgt, durchgeführt:

$$K(A_\mu^{(s_t)}(n), A_\mu(n)) = \frac{1}{N_{\mathrm{TD}}} \sum_{n=0}^{N_{\mathrm{TD}}-1} \sqrt{\frac{1}{N_g(n)} \sum_{\mu=0}^{\mu_g-1} d'_\mu(n) \left(A_\mu^{(s_t)}(n) - A_\mu(n)\right)^2}, \tag{17}$$

Der Parameter N_{TD} beschreibt die Anzahl der Blöcke aller Testdaten. $A_\mu(n)$ beschreibt die geschätzte Einhüllende des jeweiligen RNN-Modells bzw. die der verschiedenen Codebücher. Der Parameter μ_g beschreibt die Grenzfrequenz von 2000 Hz. $N_g(n) = \sum_{\mu=0}^{\mu_g-1} d'_\mu(n)$ ist die Summe

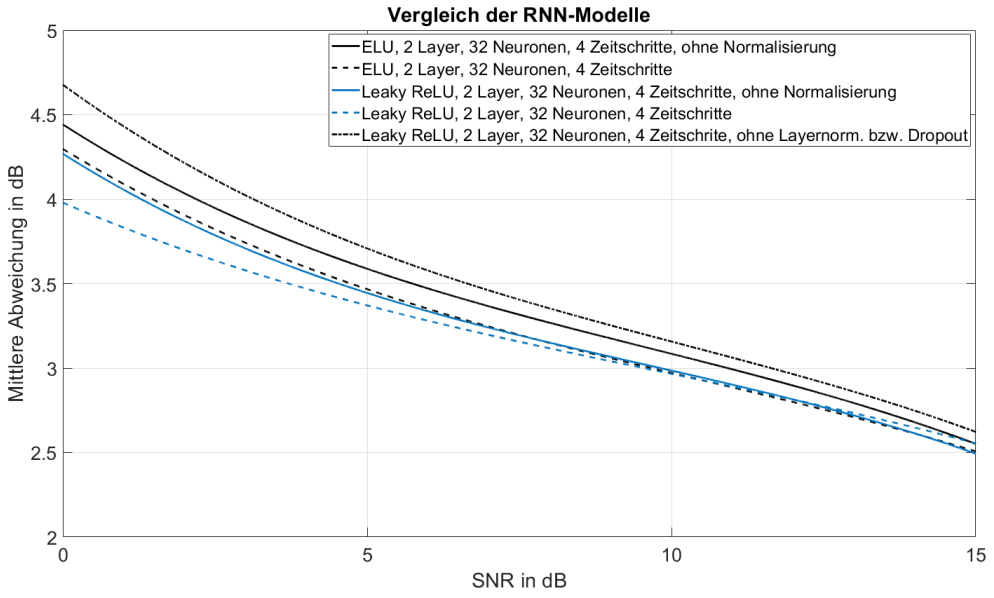

Abbildung 4 – Mittlere Abweichungen der RNN-Modelle mit und ohne Optimierungen.

der binären Maske $d'_\mu(n)$. Die Maske gewichtet jede Frequenzstützstelle mit einer 1 für ein niedrieges SNR und mit einer 0 für ein hohes SNR. Das Modell, das am besten abgeschnitten hat, nutzte die *Leaky ReLU*-Aktivierungsfunktion mit allen Optimierungen des RNNs. Aus vorhergehenden Versuchen haben sich zwei Layer mit jeweils 32 Neuronen und 4 Zeitschritte als günstig erwiesen. In Abb. 5 werden die drei verwendeten Verfahren gegenübergestellt. Die schwarze gestrichelte Linie beschreibt das einfache Verfahren der IIR-Glättung. Es ist zu sehen, dass es bei hohem Geräusch schlechter abschneidet als die anderen Verfahren. Darunter ist das Codebuch-Verfahren mit verschiedenen Größen zu sehen. Bei hohem SNR wird die mittlere Abweichung mit zunehmender Codebuchgröße geringer. Bei niedrigem SNR hat der Such-Algorithmus Schwierigkeiten den passenden Codebucheintrag auszuwählen. Dies kommt dadurch zustande, dass weniger Frequenzstützstellen mit einem zufriedenstellenden SNR für die Codebuch-Suche zur Verfügung stehen. Des Weiteren zeigt sich, dass die geschätzten Einhüllenden des RNNs am besten abschneiden. Das verwendete Modell nutzte die *Leaky ReLU*-Aktivierungsfunktion, die Normalisierung, 2 Layer, 32 Neuronen und 4 Zeitschritte.

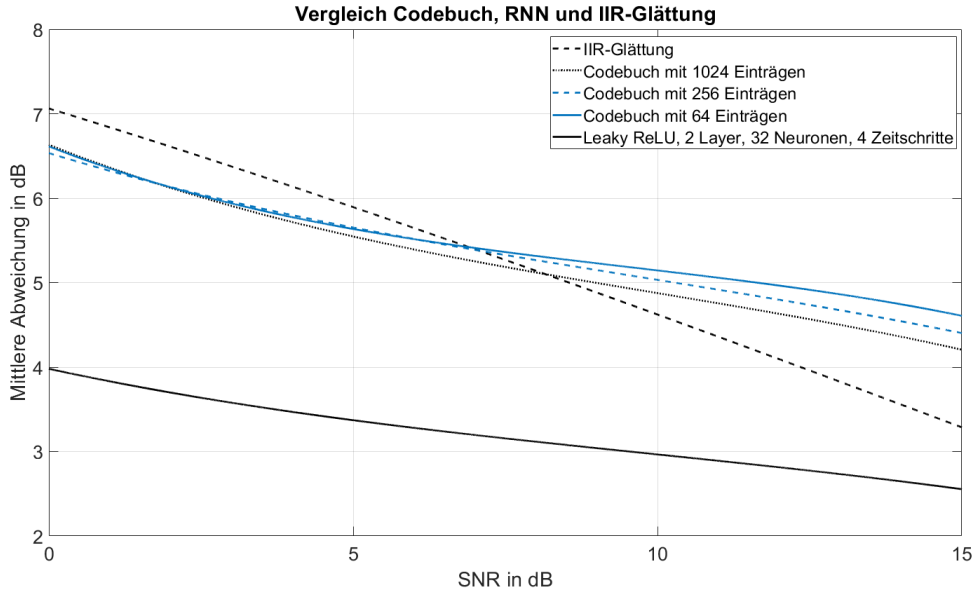

Abbildung 5 – Mittlere Abweichung der verwendeten Verfahren: IIR-Glättung, Codebücher mit verschiedenen Größen und das RNN-Modell mit dem besten Ergebnis.

4 Zusammenfassung und Ausblick

In dieser Arbeit wurden zwei Verfahren zur verbesserten Schätzung von spektralen Einhüllenden vorgestellt und analysiert. Das Codebuch-Verfahren nutzt den *k-Means*-Algorithmus zum trainieren der Prototyp-Einhüllenden. Zur Gegenüberstellung wurde ein rekurrentes neuronales Netz trainiert. Dazu wurden verschiedene Modelle trainiert, um aus der gestörten Einhüllenden die ungestörte Einhüllende zu schätzen. Des Weiteren wird die zeitliche Korrelation der Sprache ausgenutzt, in dem die vergangenen Einhüllenden berücksichtigt werden. Dies wird über die zeitliche Entfaltung des RNNs realisiert. Am besten hat das Modell mit der *Leaky ReLU*-Aktivierungsfunktion, zwei Layer mit jeweils 32 Neuronen und 4 Zeitschritten abgeschnitten, sowohl im Vergleich mit dem Codebuch als auch mit der IIR-Glättung. Die geschätzten Einhüllenden der beiden Verfahren können genutzt, um stark gestörte Frequenzstützstellen zu ersetzen.

Darüber hinaus gibt es viele Möglichkeiten die beiden Verfahren weiterhin zu optimieren. Zum einen kann der Sprachaktivitätsdetektor verbessert werden, damit die Codebuch-Suche

passendere Prototyp-Einhüllenden liefert. Zum anderen können die Trainingsdaten vergrößert und die Anzahl der Cluster erhöht werden. Eine Klassifizierung durch ein RNN könnte auch die Schätzung der ungestörten Einhüllenden verbessern. Dabei können durch einen Clustering-Algorithmus zunächst Prototyp-Einhüllenden trainiert und diese dem RNN für die Klassifikation zur Verfügung gestellt werden.

Literatur

[1] KAJALA, M. und M. HÄMÄLÄINEN: *Filter-and-Sum Beamformer with Adjustable Filter Characteristics*. Bd. 5, S. 2917 – 2920 vol.5. 2001. doi:10.1109/ICASSP.2001.940257.

[2] WIENER, N.: *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. MIT Press, 1949.

[3] DELLER, J., J. HANSEN, und J. PROAKIS: *Discrete-Time Processing of Speech Signals*. IEEE Press, 2000.

[4] HANNON, P., M. KRINI, und I. SCHALK-SCHUPP: *Advanced Speech Enhancement with Partial Speech Reconstruction*. In *21st European Signal Processing Conference (EUSIPCO 2013)*, S. 1–5. 2013.

[5] COHEN, I.: *Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging*. Bd. 11. IEEE Transactions on Speech and Audio Precessing, 2003.

[6] VEAUX, C., J. YAMAGISHI, und K. MACDONALD: *CSTR VCTK Corpus*. University of Edinburgh, 2012.

[7] LLOYD, S. P.: *Least Squares Quantization in PCM*. In *IEEE Transactions on Information Theory*, Bd. 28, S. 129 – 137. 1982.

[8] MAAS, A. L., A. Y. HANNUN, und A. Y. NG: *Rectifier Nonlinearities Improve Neural Network Acoustic Models*. *Proc. ICML*, 2013. URL `https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf`.

[9] CLEVERT, D., T. UNTERTHINER, und S. HOCHREITER: *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. *CoRR*, abs/1511.07289, 2015. URL `http://arxiv.org/abs/1511.07289`. `1511.07289`.

[10] GÉRON, A.: *Hands-On Machine Learning with Scikit-learn and Tensorflow*. O'Reilly Media, 2017.

[11] JAIN, P. und H. HERMANSKY: *Improved Mean and Variance Normalization for Robust Speech Recognition*. In *Proc. ICASSP*, Bd. 6, S. 4012 – 4015. 2001.

[12] GAL, Y. und Z. GHAHRAMANI: *A Theoretically Grounded Application of Dropout in Recurrent Neural Networks*. *arXiv e-prints*, arXiv:1512.05287, 2015. `1512.05287`.

[13] LEI BA, J., J. R. KIROS, und G. E. HINTON: *Layer Normalization*. *arXiv e-prints*, arXiv:1607.06450, 2016. `1607.06450`.