Einfluss von Entrauschungsverfahren auf die automatische Segmentierung mit WebMAUS

Lorenz Gutscher, Nicola Klingler, Michael Pucher

Institut für Schallforschung, Österreichische Akademie der Wissenschaften
lorenz.gutscher@oeaw.ac.at

Kurzfassung: Archivierte Audioaufnahmen von Sprecher*innen ermöglichen es, Sprachentwicklungen über viele Jahrzehnte hinweg zu untersuchen. Aufgrund der technischen Beschaffenheit früherer Aufnahmegeräte, wie beispielsweise eines Phonographen, weisen solche Aufnahmen ein niedriges Signal-Rausch-Verhältnis (SRV) auf und sind geprägt von Nebengeräuschen und Verzerrungen. Für phonetische Untersuchungen stellt das ein großes Problem dar, da wichtige Informationen maskiert werden. Entrauschungsmethoden bieten die Möglichkeit, das Nutzsignal teilweise wiederherzustellen, was jedoch mit Verlusten einhergeht. Während spektrale Analysen (wie Formantanalysen) bei stark verrauschten Daten dementsprechend schwierig sind, können zeitliche Informationen sinnvoller ausgewertet werden. In dieser Studie wird untersucht, ob Entrauschungsmethoden auf stark verrauschtes Archivmaterial eines Phonographen aus dem Jahr 1910 angewendet werden können und welche der drei angewandten Tools (Audacity [1], iZotope [2] und Praat [3]) sinnvolle Ergebnisse liefern, wenn auf den Output eine automatische Segmentierung angewandt wird. Für einen Vergleich der Ergebnisse wird von einer Linguistin eine manuell segmentierte Version angefertigt, die als Referenz dient. Damit können zeitliche Abweichungen zwischen manuell gesetzten und automatisch gesetzten Segmentgrenzen gemessen und objektiv evaluiert werden. Die Studie zeigt, dass sowohl Audacity als auch iZotope gute Ergebnisse erreichen und eine sinnvolle automatische Segmentierung ermöglichen. Bei der mit Praat entrauschten Version kommt es dagegen zu Fehlern bei der automatischen Segmentierung und damit zu großen Abweichungen von der Referenz.

1 Einführung

Aufnahmen in Laborqualität bilden die Grundlage phonetischer Forschung. Doch wie sollte mit archivierten Daten umgegangen werden, die nicht in Laborqualität vorliegen? Eine scheinbar gute Lösung stellen Methoden zur Rauschverminderung oder auch Entrauschungsverfahren dar, die Hintergrundgeräusche reduzieren und das Signal-Rausch-Verhältnis (SRV) verbessern. Das zu lösende Problem ist hierbei eine Trennung zwischen Nutzsignal und Nebengeräuschen (Speech Separation) – eine Aufgabe, die das menschliche Gehör sehr gut lösen kann. Ein bekanntes Beispiel dafür ist der Cocktailparty-Effekt, der beschreibt, dass es durch selektive Wahrnehmung möglich ist, trotz Überlagerung von mehreren Stimmen und Geräuschen eine einzelne Stimme herauszuhören. Während klassische Entrauschungsverfahren auf Algorithmen basieren, welche die spektrale Information des Hintergrundgeräuschs von dem Originalsignal subtrahieren, verfolgen neuere Verfahren komplexere Ansätze, wie die Verwendung von tiefen neuronalen Netzwerken, oder methodenspezifischen Techniken, die das Ergebnis der Verarbeitung natürlicher klingen lassen sollen (z. B. durch das Hinzufügen von harmonischen Obertönen des Nutzsignals [2]).

Dass Entrauschungsverfahren einen großen Einfluss auf spektrale Analysen haben können (z. B. durch den Informationsverlust bei der spektralen Subtraktion), scheint einleuchtend. Weniger gesichert ist der Einfluss von Entrauschungsverfahren auf temporale Analysen, wie beispielsweise Dauermessungen von Vokalen und Konsonanten. In dem hier vorgestellten Vorhaben werden mehrere Entrauschungsverfahren auf deren Anwendbarkeit auf historische Daten verglichen. Die Untersuchungen werden anhand eines Korpus aus dem Jahr 1910 durchgeführt, welches von Anton Pfalz mit Sprechern aus der Region des Marchfelds erhoben wurde. Das Korpus besteht aus Aufnahmen von drei männlichen Sprechern (Alter: 18, 38, 54 Jahre), die in Deutsch-Wagram aufgenommen wurden und den dort ortstypischen ostmittelbairischen Dialekt gesprochen haben.

Die Aufnahmen wurden mit einem „Archiv-Phonographen“ [4] aufgezeichnet und weisen aus technischen Gründen ein niedriges SRV auf. Die Technik des Archiv-Phonographen ist vergleichbar mit der eines Edisonapparats. Die Membran des für die hier analysierten Aufnahmen verwendeten Archiv-Phonographen waren aus Glas, der Träger aus Wachs [5, 6]. Die Digitalisierung der Daten wurde vom Phonogrammarchiv der Österreichischen Akademie der Wissenschaften durchgeführt: Die Digitalisierung fand im Zuge des sogenannten „Re-Recordings“, unter Verwendung eines Technics-Plattenspielers, eines Vorverstärkers und eines externen A/D-Wandlers statt. Das Ergebnis ist ein Waveform-Audio-Dateiformat im Archivstandard mit 96 kHz (24 Bit), dem gängigen Standardformat (siehe „IASA Guidelines on the Production and Preservation of Digital Audio Objects“[1]).

2 Methode

2.1 Entrauschungsverfahren

Für die Analyse werden die Entrauschungsverfahren von Praat, iZotope RX 8 und Audacity auf mehrere Sätze (Wenkersätze) aus dem Korpus angewandt. Bei der Auswahl der verwendeten Tools wurde darauf geachtet Vertreter von freier und in der Sprachwissenschaft häufig verwendeter Software (Praat und Audacity) als auch Vertreter von kommerzieller Software (iZotope) zu repräsentieren. Die Methoden von Praat und iZotope benötigen einen kurzen Ausschnitt des Audiosignals, bei dem nur das zu entfernende Rauschen auftritt, um ein sogenanntes Rauschprofil davon zu erstellen. Diese Methoden sind dadurch auf ein annähernd stationäres Rauschen angewiesen. Für die ermittelten Rauschprofile werden in dieser Studie, wenn es die Methode erlaubt bzw. erfordert, immer dieselben Zeitpunkte im Originalsignal genutzt. Bei der kommerziellen Variante mit iZotope entfällt dieser Schritt, da hier der „Adaptive Modus“ verwendet wird, welcher automatisch ein Rauschprofil ermittelt. Details zu den gewählten Einstellungen sind in Abbildung 1 angeführt. Bei der Wahl der Parameter wird versucht, möglichst viel Rauschen zu entfernen, ohne das Nutzsignal stark zu beeinflussen. Die einzustellenden Parameter werden anhand eines Satzes aus dem Korpus einmalig optimiert. Daraufhin wird jeder Satz mit den festgelegten Einstellungen verarbeitet.

Die genannten Softwarepakete verwenden unterschiedliche zugrunde liegende Methoden zur Entrauschung: Audacity verwendet ein Verfahren, bei dem einzelne Frequenzbänder abgeschwächt werden, wenn diese unter einen im Rauschprofil definierten Schwellenwert fallen. Dies wird allgemein oft als *Spectral Noise Gating* bezeichnet. Praat greift auf die Methode der einfachen spektralen Subtraktion zurück, bei der vom Rauschen der Mittelwert des Leistungsspektrums berechnet wird. Damit kann das ermittelte Leistungsspektrum des Rauschens im Frequenzbereich von dem Nutzsignal subtrahiert werden [7]. IZotope bietet die Auswahl zwischen den Algorithmen „Simple, Advanced, Advanced + Extreme“, wobei in dieser Studie

[1] `https://www.iasa-web.org/tc04/audio-preservation`

nur die Methode „Advanced + Extreme“ angewendet wird. Diese Methode kombiniert einen *Non-Local Means Algorithmus* [8] – eine Methode aus der Bildverarbeitung – mit anschließender Nachbearbeitung mittels *Discrete Fourier Transform Thresholding* (DFTT) zur Entfernung von akustischen Artefakten [9].

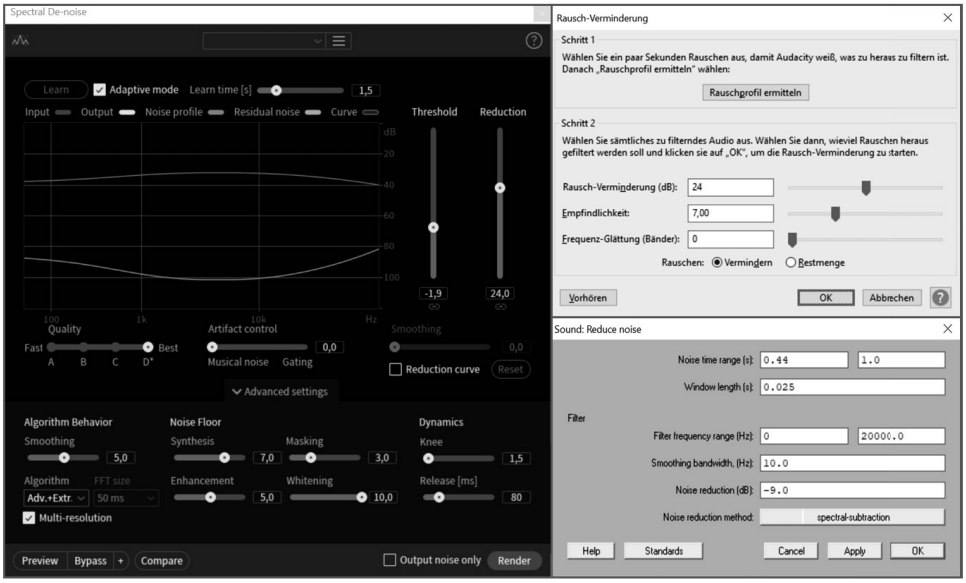


Abbildung 1 – Screenshots der gewählten Einstellungen für die drei Entrauschungsverfahren (iZotope: links, Audacity: oben rechts, Praat: unten rechts).

2.2 Klick-Entferner

Da auf den Aufnahmen gerätetypische Klickgeräusche zu hören sind, wird untersucht, ob ein zusätzlicher Klick-Entferner vor dem Entrauschen das Ergebnis verbessern kann. Dafür wird die Audiodatei mit der Funktion *Klick-Entferner* in Audacity bearbeitet. Die einzugebenden Parameter wurden nach subjektivem Höreindruck des Erstautors ausgewählt, sodass möglichst viele Klicks entfernt werden ohne das Nutzsignal zu stark zu beeinflussen (Schwellenwert: 110, bzw. minimale Spitzenbreite: 40). Die resultierende Audiodatei wird anschließend, parallel zum Ablauf in Unterabschnitt 2.1, entrauscht und mit WebMAUS automatisch segmentiert.

2.3 Evaluationsmethode

Eine objektive Beurteilung der Entrauschungsverfahren durch den Vergleich des SRV zwischen den Methoden wäre für eine objektive Evaluation interessant, ist jedoch mit den verwendeten Daten nicht möglich, da das reine Nutzsignal nicht isoliert zur Verfügung steht. Eine weitere Möglichkeit zur objektiven Evaluation bieten Spracherkennungssysteme. Die Anzahl der korrekt erkannten Wörter wird dabei mit einer *Word Error Rate (WER)* gemessen. Diese Methode ist allerdings für die verwendeten Daten ebenfalls nicht anwendbar, da die entrauschten Audiodateien immer noch ein hohes Maß an Rauschen beinhalten und keine sinnvollen Worte oder Phone von der Spracherkennung erkannt werden. Neben dem geringen SRV ist dafür der in den Aufnahmen gesprochene Dialekt verantwortlich, der eine besondere Schwierigkeit darstellt. Aus diesem Grund wird zur Evaluation die Genauigkeit einer automatischen Segmentie-

rung mit *Forced Alignment* herangezogen und mit einer handsegmentierten Version verglichen (siehe Abbildung 3). Dafür wird für jeden Laut die zeitliche Abweichung zwischen automatischer und manueller Segmentgrenze betrachtet und für die Berechnung des Root Mean Square Error (RMSE) herangezogen:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i^{manuell} - t_i^{automatisch})^2}$$

3 Experiment

Die Originaldaten sind einem bairischen Dialekt zuzuordnen, weshalb die Transkription von einer Linguistin als SAMPA-Umschrift angefertigt wurde, um diese als Input für WebMAUS zu verwenden (siehe Abbildung 2). Die SAMPA-Umschrift enthält sowohl spektrale als auch temporale Informationen über die einzelnen Laute, da phonologisch langen und kurzen Lauten unterschiedliche Symbole zugeordnet werden. Somit gilt, je präziser die SAMPA-Umschrift ist, desto mehr sinnvolle Information erhält WebMAUS als Input. Aus dieser Umschrift wird ein Template erstellt und in WebMAUS eingespeist, damit jedes entrauschte Audiofile mit derselben Transkriptionsgrundlage kombiniert wird. In WebMAUS wird der MAUS Modus „Forced alignment to input transcript“ gewählt. Ziel der Entrauschung ist eine möglichst prä-

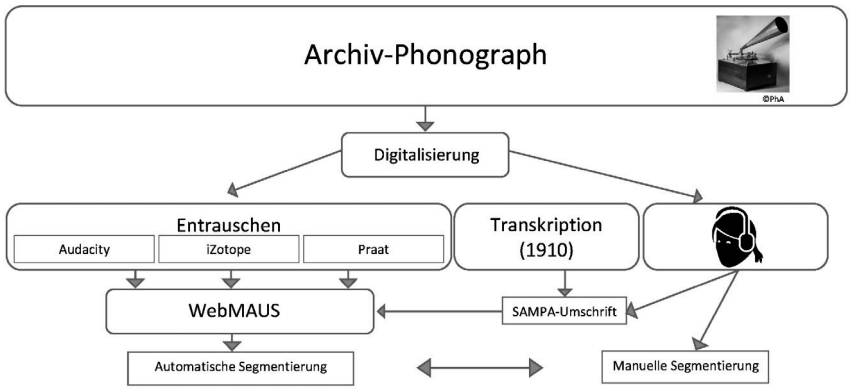


Abbildung 2 – Schematische Darstellung der Datenverarbeitung. Nach der (externen) Digitalisierung der Daten werden drei Entrauschungsverfahren angewendet. Parallel dazu wird eine manuelle Segmentierung der Daten angefertigt. Die entrauschten Daten werden gemeinsam mit der SAMPA-Umschrift in WebMAUS eingepflegt und der Output mit der manuellen Segmentierung verglichen.

zise Segmentierung durch WebMAUS im Vergleich zu einer manuell segmentierten Version. WebMAUS führt bei vorgegebener Abfolge von Segmenten ein sogenanntes *Forced Alignment* durch, um Segmentgrenzen zu eruieren. Die von WebMAUS erstellten Segmentierungen werden als TextGrid Dateien ausgegeben. Anhand dieser werden die drei Entrauschungsverfahren mit dem manuell erstellten TextGrid verglichen. Dafür werden die Segmentgrenzen der automatischen und manuellen Segmentierung zueinander in Beziehung gesetzt, indem die zeitliche Abweichung zum manuell transkribierten Signal berechnet wird (siehe Abbildung 3). Als Maß wird hierfür der RMSE verwendet.

Für den Vergleich wurde sichergestellt, dass WebMAUS und die manuelle Segmentierung die gleiche Anzahl an Segmenten hatten. Abweichungen traten innerhalb von Worten nur für Di-

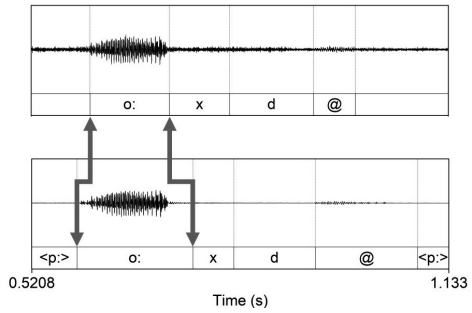
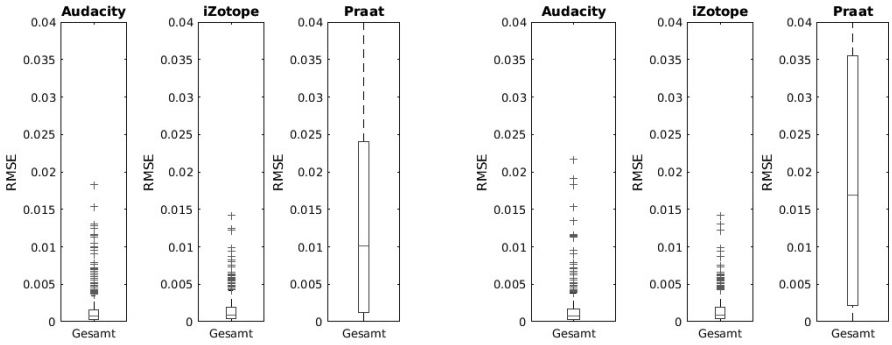


Abbildung 3 – Vergleich der Segmentgrenzen der automatischen und manuellen Segmentierung. Der Root Mean Square Error (RMSE) wird auf die Abweichung der Grenzen zwischen automatischer und manueller Segmentierung berechnet.

phthonge auf, die von WebMAUS als einzelne Laute segmentiert wurden. Dies wurde händisch korrigiert, indem die mittlere Grenze zwischen zwei Diphthongsegmenten gelöscht wurde. Andere Veränderungen wurden an den gesetzten Grenzen nicht vorgenommen.

3.1 Ergebnisse der entrauschten Dateien

Die Ergebnisse in Abbildung 4a zeigen, dass die Entrauschungsverfahren von Audacity und iZotope ähnlich niedrige RMSE Werte aufweisen, was auf eine geringe Abweichung zwischen manueller und automatischer Segmentierung hinweist. Der Entrauschungsfilter von Praat weist einen hohen RMSE Wert auf, was auf eine größere Abweichung zwischen manueller und automatischer Segmentierung mit dieser Methode schließen lässt. Die hohen RMSE Werte bei Praat sind darauf zurückzuführen, dass die automatische Segmentierung hier schlechter funktioniert. Da Praat eine sehr starke Rauschreduktion des gemittelten Spektrums durchführt, bleiben einzelne Artefakte im Spektrogramm isoliert übrig und werden von WebMAUS als Sprachsignal gewertet. Dies führt zu fehlerhaftem Alignment und damit zu großen Fehlern beim Vergleich der Segmentgrenzen. Aus den Ergebnissen lässt sich schließen, dass die Entrauschungsverfah-



(a) Ohne Vorverarbeitung.

(b) Vorverarbeitung mit Klick-Entferner.

Abbildung 4 – Boxplot des RMSE zwischen automatischer und manueller Segmentierung mit drei unterschiedlichen Entrauschungsverfahren.

ren von Audacity und iZotope trotz stark verrauschter Aufnahmen eine automatische Segmen-

tierung ermöglichen und eine deutliche Arbeitserleichterung bieten.

3.2 Ergebnisse der entrauschten Dateien mit vorheriger Klick-Entfernung

Die Vorverarbeitung mit zusätzlichem Klick-Entferner verringert die Abweichung der automatisch gesetzten Segmentgrenzen für die entrauschte Version mit Praat nicht, sondern erhöht für diese den RMSE sogar von $0,0248$ auf $0,0324$ (vgl. Abbildung 4b). Für iZotope und Audacity erhält man sehr ähnliche Werte wie ohne Klick-Entfernung. Das zeigt, dass der zusätzliche Schritt keine Verbesserung für die automatische Segmentierung darstellt, da die Entrauschungsverfahren die Klicks ebenfalls gut entfernen.

Tabelle 1 – RMSE gemittelt über alle Sätze für die entrauschte Version, sowie auf die entrauschte Version mit vorheriger Klick-Entfernung.

Verfahren / Tool	RMSE Entrauscht	RMSE Klick-Entfernt + Entrauscht
Audacity	0,0028	0,0028
iZotope	0,0025	0,0024
Praat	0,0248	0,0324

4 Zusammenfassung

In dieser Studie wird der Einfluss von drei Entrauschungsverfahren (Audacity, iZotope, Praat) auf die automatische Segmentierung (mit WebMAUS) von stark verrauschten Daten untersucht. Es wird gezeigt, dass archivierte Aufnahmen, trotz niedrigem Signal-Rausch-Verhältnis, dank aktueller Methoden für phonetische Fragestellungen herangezogen werden können. Die Beurteilung der Verfahren beschränkt sich dabei auf die Abweichung der automatischen Segmentgrenzen zu einer manuell segmentierten Version ohne spektrale Unterschiede der entrauschten Dateien im Detail zu betrachten.

Es zeigt sich, dass vor allem die Entrauschungsverfahren von Audacity und iZotope angewendet werden können, um automatische Segmentierung zu nutzen und somit den Arbeitsaufwand (d. i. händisches Segmentieren) zu reduzieren. Die Entrauschungsmethode von Praat funktioniert ohne zusätzliche Nachbearbeitung für die geforderte Aufgabe nur schlecht, da im Spektrum übrig gebliebene Artefakte bei der automatischen Segmentierung mit WebMAUS als Sprachsignale identifiziert werden. Für diese Methode müssten daher nachträglich weitere Bearbeitungen durchgeführt werden, bevor die automatische Segmentierung sinnvoll durchgeführt werden kann.

5 Zitate, Literaturangaben

Die Entrauschungsverfahren wurden auf Phonogramme (Archivnummern: Ph 921, Ph 925 und Ph 929) angewendet, die vom Phonogrammarchiv der Österreichischen Akademie der Wissenschaften zur Verfügung gestellt wurden.

Die Autoren stehen weder in wirtschaftlicher noch in persönlicher Verbindung zu der verwendeten Software.

Literatur

[1] AUDACITY TEAM: *Audacity(r): Free audio editor and recorder [computer programm]*. 2021. URL `https://audacityteam.org/`.

[2] IZOTOPE, INC.: *iZotope RX 8 [computer programm]*. 2020. URL `https://www.izotope.com/`.

[3] BOERSMA, P. AND WEENINK, D.: *Praat: doing phonetics by computer [computer programm]*. 2021. URL `http://www.praat.org/`.

[4] GRAF, W.: *Aus der Geschichte des Phonogrammarchivs der Österreichischen Akademie der Wissenschaften. Bulletin phonographique 6*, S. 9–39, 1964.

[5] EXNER, S.: *II. Bericht über den Stand der Arbeiten der Phonogramm-Archivs Commission. Anzeiger der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Klasse 39, Beilage*, S. 1–31, 1902.

[6] HAUSER, F.: *Über einige Verbesserungen am Archivphonographen. Sitzungsbericht der mathematisch-naturwissenschaftliche Klasse der Österreichischen Akademie der Wissenschaften IIa, 112*, S. 1397–1406, 1903.

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

[8] BUADES, A., B. COLL, und J. MOREL: *Image denoising by non-local averaging*. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Bd. 2, S. ii/25–ii/28 Vol. 2. 2005.

[9] LUKIN, A. und J. G. TODD: *Suppression of musical noise artifacts in audio noise reduction by adaptive 2-d filtering. Journal of The Audio Engineering Society*, 2007.