Detektion von Schnarchen in Audiosignalen mit Convolutional Neural Networks

Matthias Stammler, Andreas Rommel, Bernhard Wirnitzer, Stefan Feldes

Institut für Digitale Signalverarbeitung, Hochschule Mannheim, E-Mail: matthias.stammler@hotmail.de

Abstract

Dieser Beitrag untersucht die Eignung von Convolutional Neural Networks (CNN) für den Zweck der Schnarchdetektion in Audiosignalen. Dazu werden spektrale Merkmale des Audiosignals in ihrem zeitlichen Verlauf als Input für die Netze in ein Bildformat mit festen Dimensionen gewandelt. Am Netzausgang werden zur Differenzierung des Schnarchens gegenüber Nebengeräuschen und Stille drei entsprechende Klassen angesetzt. Zum Training der CNNs wird umfangreiches Trainingsmaterial aus diversen Quellen (klinisches sowie häusliches Umfeld) herangezogen, dessen Zusammensetzung diskutiert wird. Experimente werden insb. zur Dimensionierung der CNN-Architekturen bzgl. Filtergröße und Filteranzahl durchgeführt, sowie zur Datenaugmentierung, die hier durch Überlagern von Rauschen sowie Aufprägen von unterschiedlicher Raumakustik realisiert wird. Nach Optimierung wurden Tests mit separat aufgezeichneten Übernachtaufnahmen aus häuslichem Umfeld durchgeführt. Im Ergebnis konnte eine Korrektklassifikationsrate von 95,5 % in bekanntem Umfeld und von ca. 80 % in unbekanntem Umfeld und auf unbekannte Personen erzielt werden.

Einleitung

Motivation

Die mit starkem Schnarchen einhergehende Verengung der Atemwege mit ihren vielfältigen negativen Auswirkungen stellt ein ernstes gesundheitliches Problem dar, das bis zum Krankheitsbild der Schlafapnoe führen kann. Diagnose, Behandlung und Verlaufskontrolle von Schnarchen und Schlafapnoe sind häufig apparativ aufwendig und mit Einschränkungen (z.B. Verkabelung) für die Patienten verbunden. Zur Unterstützung der Diagnose und Kontrolle des Therapieverlaufs ist eine Überwachung der nächtlichen Schnarchaktivität im häuslichen Umfeld des Patienten sehr hilfreich. Diese Überwachung durch akustische Detektion von Schnarchgeräuschen zu realisieren, vermeidet störende Verkabelungen, etc. und wäre bspw. über Smartphone leicht realisierbar. Sog. Schnarch-Apps sind bereits verfügbar, so z.B. „SnoreApp“ oder „SnoreLab“ in Googles Play Store. Im häuslichen Umfeld jedoch müssen Schnarchgeräusche gegenüber anderen Geräuschen (und Grundrauschen) differenziert und dies robust gegenüber unterschiedlichen raumakustischen Wirkungen (Hall, etc.) auf das entfernte Mikrofon ausgelegt werden. So wird dann eine korrekte quantitative Erfassung der Schnarchaktivität ermöglicht.

Stand der Technik

Bereits 1994 ermöglicht ein Künstliches Neuronales Netz (KNN) die automatische Erkennung von Schnarchen aus den Luftdruckdaten einer Beatmungsmaschine mit einer Erkennungsrate von 75% (Accuracy) [1]. Neuere Arbeiten nutzen bspw. Matratzen, die die Liegeposition der Schlafenden detektieren, sowie EKG-Signale, um zu entscheiden, ob die Betroffenen schnarchen. Hierbei werden Erkennungsraten von bis zu 94,5 % erzielt [2].

Die Akustik des Schnarchens wird seit mehreren Jahrzehnten in der Schlafmedizin untersucht [3]. Die dabei verwendeten charakteristischen Merkmale sind z.B. die Lautstärke, der Root-Mean-Square-Wert (RMS), der Crest-Faktor (Maximalwert/RMS-Wert), die Zero-Crossing-Rate und spektrale Merkmale wie Linear-Predictive-Coding-Koeffizienten (LPC) oder Mel-Frequenz-Cepstrum-Koeffizienten (MFCC). Die Selektion der relevanten Merkmale führt zu sogenannten akustischen Biomarkern [4]. Ähnliche Merkmale sind auch die Basis für die automatische Erkennung von Schnarchen und die Klassifikation von krankhaftem Schnarchen aus Audiosignalen mit Hilfe von KKNs [4]. Auf der Basis von MFCC-Merkmalen wird in [5] mit Hilfe von Hidden-Markov-Modellen eine Differenzierung von Schnarchen gegenüber häuslichen Nebengeräuschen realisiert. Innerhalb des trainierten Umfelds kann so das Schnarchen mit 96% korrekt klassifiziert werden. In [6] wird ein aufwandsarmes Verfahren zur Schnarchdetektion für Embedded Systeme vorgestellt, das auf einer LPC-Analyse und Merkmalen wie Grundfrequenz der Schnarchvibration, Energiegehalt etc. basiert.

Bei der Verwendung eines CNN wird die schwierige Merkmalsselektion implizit von den Faltungsschichten (Convolutional Layer) des Netzes übernommen. So kann ein CNN z.B. aus Spektrogrammen nach geeignetem Training selbständig die relevanten Merkmale extrahieren. Mit einem derartigen CNN kann aus Audiosignalen der Entstehungsort des Schnarchens entsprechend der sogenannten vier Klassen V,O,T,E mit einem „Unweighted Average Recall" von 67 % bestimmt werden [7] [8].

Dieser Beitrag untersucht die Eignung eines CNN für den Zweck der Schnarchdetektion in Audiosignalen. Dazu werden spektrale Merkmale des Audiosignals in ihrem zeitlichen Verlauf in Form von Spektrogrammen als Input für das Netz verwendet. Im Sinne der Alltagstauglichkeit und Unempfindlichkeit gegen variierende Raumakustik werden die Trainingsdaten erstmals mit unterschiedlichen Raumimpulsantworten augmentiert.

CNN Aufbau und Spektrogrammerzeugung

Für die Untersuchungen wird eine bewährte Grundstruktur von CNNs gemäß Abbildung 2 gewählt. Diese Netzarchitektur wird im Training auf vier Parameter hin optimiert: (1) Anzahl der Feature Extraction Layers (jeweils bestehend aus Convolutional Layer, Batch Normalization Layer, ReLu Layer und MaxPooling Layer); (2) Anzahl der Fully Connected Layers; (3) die Größe der Filter und (4)

Anzahl der Filter in den Convolutional Layers. Um die Robustheit des CNNs insb. bzgl. Abweichungen der raumakustischen Eigenschaften und Störgeräuschen zwischen Trainingsmaterial und realem Einsatz zu erhöhen, wird zudem noch die Wirksamkeit von Maßnahmen zur Generalisierung untersucht.

Ausgangspunkt für die Klassifikation durch das CNN sind, wie Abbildung 2 zeigt, Spektrogramme, die zuvor aus dem Audiosignal errechnet werden. Der typische zeitliche Verlauf des Schnarchens wird in Abschnitten von 1s Länge gut erfasst. Da CNNs an ihrem Eingang meist Bilder konstanter Größe erwarten, werden die Audioaufnahmen entsprechend Abb.1 aufbereitet.

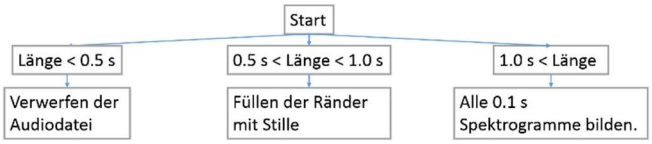

Abbildung 1 Erzeugung von Spektrogrammen der festen Länge 1 s aus einem Korpus mit Audioaufnahmen variierender Länge

In der Klassifikationsschicht am Ausgang des CNNs werden zur Differenzierung des Schnarchens gegenüber Nebengeräuschen und Stille drei entsprechende Klassen angesetzt.

Datensatzaufbau

Zum Training und zur Architekturoptimierung wurden sowohl eigene Aufnahmen im häuslichen Umfeld erhoben als auch verfügbare Datensätze aus anderen Quellen herangezogen. So wurden Schnarchaufnahmen aus dem Schlaflabor des Uniklinikums Mannheim sowie des Munic Passau Sound Snore Corpus [8] genutzt, die eine breite Verteilung unterschiedlicher Schnarchcharakteristika auch bzgl. des Schnarchentstehungsortes abbilden. Um den Mangel an Nebengeräuschen im Gesamtkorpus auszugleichen, wurden Wörter und Nebengeräusche aus der Tensorflow Speech Commands Datenbank [9] verwendet, sowie eigene Aufnahmen von Straßengeräuschen hinzugefügt. Stillepassagen konnten hauptsächlich aus den häuslichen Aufnahmen gewonnen werden. Die Dateien des Gesamtkorpus wurden wie folgt aufgeteilt: 70 % zum Training, 15 % zur Validierung und 15 % zum Test.

Tabelle 1: Anzahl (Gesamtdauer) der Audiodateien aus hochschulinternen und fremden Datenbanken

Kategorie	Anzahl (Gesamtdauer) der Dateien aus häuslichem Bereich	Anzahl (Gesamtdauer) der Dateien aus fremden Datenbanken
Schnarchen	373 (~10 min)	1759 (~47 min)
Stille	1218 (~45 min)	1 (1 min)
Nebengeräusch	1325 (~26 min)	1003 (~19 min)

Experimente und Ergebnisse

Die Erzeugung der Spektrogramme erfolgte mit folgender Parametrierung: Abtastfrequenz 8 kHz (ggf. Downsampling der Audioaufnahme), Hanning-Fensterung mit Fensterbreite von 16 ms (128 Abtastwerte) und 8 ms (64 Abtastwerte) Überlapp, Berechnung der FFT mit 128 Punkten in dB. Das Training der Netze wurde unter Matlab R2018b durchgeführt. Es wurde über 16 Epochen trainiert. Für jede Architektur-Konfiguration wurde das Training mit jeweils zufälliger Initialisierung dreimal wiederholt, um das beste lokale Optimum zu finden.

Um die Leistungsfähigkeit der verschiedenen Architektur-Konfigurationen zu vergleichen, wurde jeweils die Korrektklassifikationsrate (Accuracy) der Testproben bestimmt. Sie errechnet sich als das Verhältnis der Anzahl korrekt klassifizierter Schnarch-, Nebengeräusch- und Stilleproben zur Gesamtzahl der Testproben.

Netzwerkoptimierung

Wie die Abbildungen 3, 4 und 5 zeigen, hängt die erzielte Accuracy nur in einem relativ geringen Maß von der Anzahl der Feature Extraction Layer, der Größe der Filter in diesen Schichten und der Anzahl der nachgeschalteten Fully Connected Layer ab. Die Veränderungen liegen im Bereich weniger Prozent und damit im Bereich der Schwankungen bei mehrfachem Training. Die Gründe für die gewählten besten Parameter sind: 4 bis 5 Feature Extraction Layer (Abbildung 3), da später eine Filtergröße von 5 bis 7 gewählt wird (Abbildung 4). Auch bzgl. der Anzahl der Fully Connected Layer ist kein Trend erkennbar, sowohl bei einer wie bei zwei Schichten liegt die durchschnittliche Accuracy bei 92,5 %. Dies deutet darauf hin, dass die Merkmalsextraktion die Komplexität des Problems so weit reduziert, dass sich die extrahierten Merkmale durch einfachere Strukturen klassifizieren lassen.

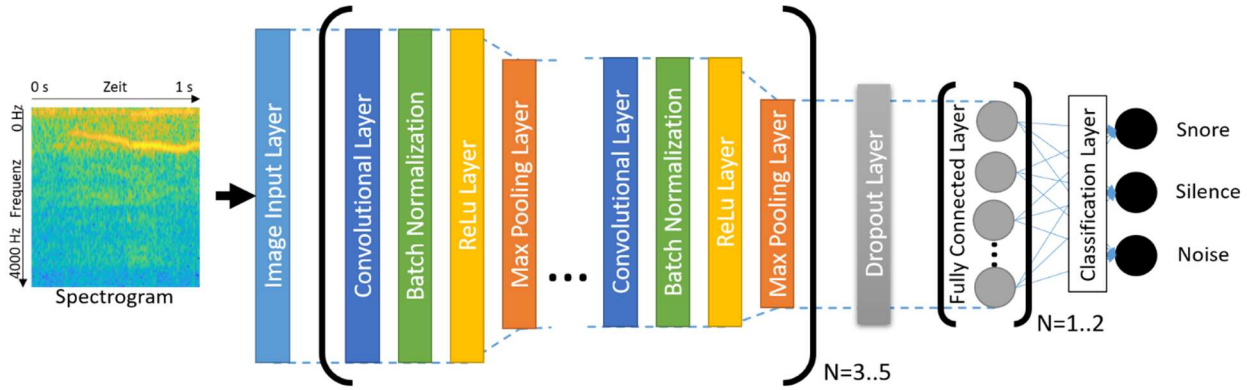

Abbildung 2 Schematischer Aufbau der Netzarchitekturen mit zu optimierenden Parametern

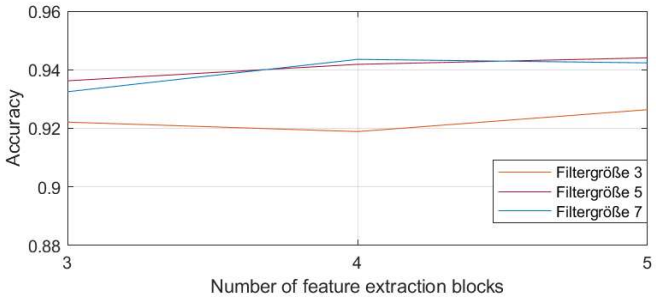

Abbildung 3 Accuracy bei Variation der Anzahl der Feature Extraction Layer und der Filtergröße

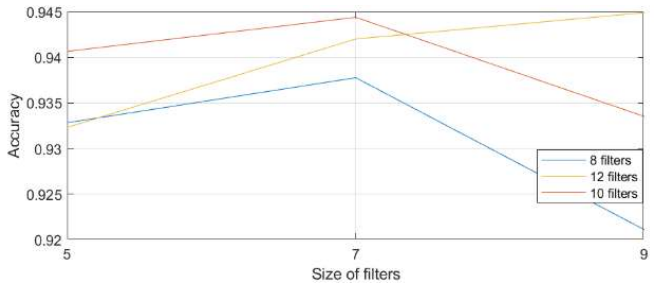

Abbildung 4 Accuracy bei Variation der Filtergröße und -Anzahl

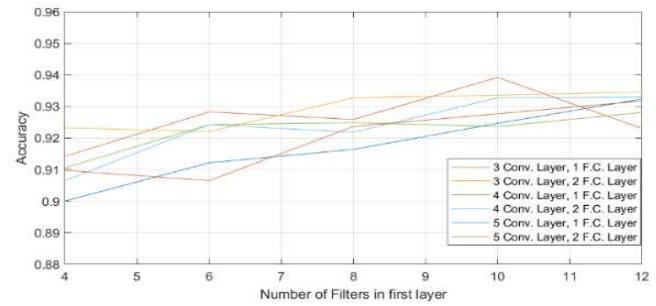

Abbildung 5 Accuracy bei Variation der Anzahl der Filter sowie Anzahl der Convolutional und Fully Connected Layer

Die Anzahl der Filter in den Convolutional Layer sollte zwischen 8 und 14 gewählt werden. Weniger Filter lassen es nicht zu, dass das CNN alle für die Schnarcherkennung notwendigen Merkmale erlernt. Dies zeigt sich darin, dass bei wiederholtem Training und gleichbleibender Architektur immer andere Filterkerne erlernt werden. Das beste Netz erlernt sowohl 2D-Tiefpass- als auch 2D-Hochpassfilter, sowie Ecken- und Kantendetektoren (Abbildung 6). Auch mit einer geringeren Anzahl von Filtern lassen sich hohe Accuracys erzielen.

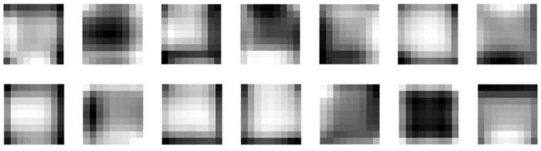

Abbildung 6 Die erlernten 14 Filter des besten Netzes

Generalisierungsmaßnahmen

Um die Generalisierungsfähigkeit des CNNs im Sinne einer robusten Erkennung im Alltagsbetrieb zu erhöhen, wurden zweimal 15% der Trainingsdaten augmentiert. Dies wurde zum einen durch Falten der Audiosignale mit zwei Raumimpulsantworten aus der Aachener AIR Datenbank [10] realisiert, und zum anderen durch Addieren von normalisiertem weißem Rauschen auf das normalisierte Audiosignal. Die verwendeten Impulsantworten aus der AIR Datenbank sind mit einem Mikrofonabstand von 1 m bzw. 2 m in einem Büroraum aufgenommen worden. Die RT60-Zeit beträgt 0,37 s bzw. 0,44 s.

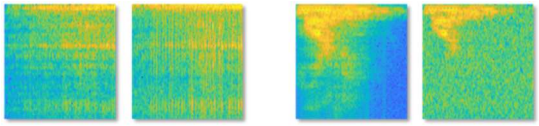

Abbildung 7 Spektrogramm eines verhallten Audiosignals (links) und eines verrauschten Audiosignals (rechts) mit Spektrogramm des jeweiligen unveränderten Audiosignals

In den Spektrogrammen (Abbildung 7) zeigen sich beide Maßnahmen visuell, die gedämpft-periodische Wiederholung von spektralen Informationen aufgrund des aufgeprägten Halls ebenso wie das Rauschen.

Zum Test der Generalisierungsmaßnahmen wurden nochmals separate Übernachtaufnahmen in unterschiedlichen häuslichen Umfeldern und mit unbekannten Personen erhoben. Für diese unbekannten Testdaten wurde mit verschiedenen Netzarchitekturen die Accuracy jeweils für ein Training mit und ohne Augmentierung bestimmt. Abbildung 8 zeigt die Verteilung der erzielten Accuracys.

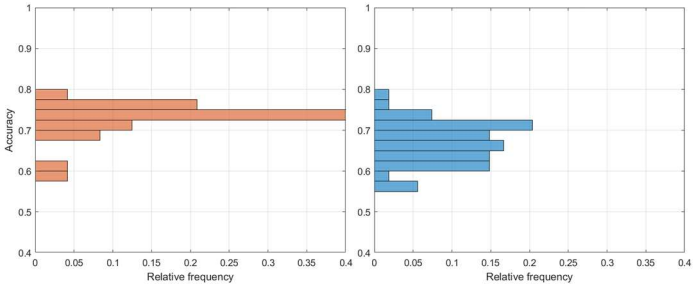

Abbildung 8 Accuracy auf Testdaten aus unbekannter Umgebung und von unbekannten Personen mit (links) und ohne (rechts) Generalisierungsmaßnahmen

Deutlich ist der Erfolg dieser Maßnahme zu erkennen. Die mittlere Accuracy wird durch die Datenaugmentierung von 71,6 ± 6,0 % auf 75,4 ± 4,1 %, also um ca. 4% gesteigert.

Beste Netzarchitektur

Die folgende Tabelle zeigt die Parametrierung der besten Netzarchitekturen für bekanntes bzw. unbekanntes Umfeld.

Tabelle 2: Netzarchitekturen mit jeweils höchster Accuracy für bekanntes bzw. unbekanntes Umfeld

Architekturparameter	Bekanntes Umfeld	Unbekanntes Umfeld
Anz. Feature Extraction Layer	5	4
Anz. Fully Connected Layer	2	2
Filtergröße	5	7
Filteranzahl	14	8

Auffällig ist, dass ein Netz mit vergleichsweise wenigen, aber größeren Filter besser in der Lage ist, das Erlernte auf unbekannte Daten zu generalisieren. Durch eine geringere

Anzahl von Filtern wird das Netz offenbar gezwungen, sich auf die wesentlichen Merkmale zu fokussieren, sodass die Gefahr des Übertrainierens reduziert wird. Eine Feature Extraction Layer weniger führt zudem dazu, dass die Komplexität der erlernten Merkmale reduziert wird.

Abbildung 9 zeigt die Verwechslungsmatrizen der jeweils besten Netze für bekanntes und unbekanntes Umfeld. Der Vergleich zeigt, dass der Recall im unbekannten Umfeld nur geringfügig sinkt, die Precision dagegen deutlich. Folglich wird tatsächliches Schnarchen weiter zuverlässig erkannt, jedoch werden zusätzlich unbekannte Nebengeräusche fälschlicherweise auch als Schnarchen klassifiziert. Anzumerken ist, dass sich die häufigste Verwechslung, nämlich dass Stille fälschlicherweise als Schnarchen erkannt wird, mit einfachen Mitteln vermeiden ließe; bspw. durch einen akustischen Eventdetektor, der das Signal erst nach Überschreiten einer gewissen Schwelle auf das CNN zur Klassifikation durchlässt.

Bekanntes Umfeld

True Class \ Predicted Class	Snore	Silence	Noise
Snore	1797	0	42
Silence	0	1662	54
Noise	63	164	2933

Accuracy: 95,2 % Recall: 97,7 % Precision: 96,6 %

Unbekanntes Umfeld, unbekannte Personen

True Class \ Predicted Class	Snore	Silence	Noise
Snore	484	0	23
Silence	204*	467	16
Noise	40	8	234

Accuracy: 80,3 % Recall: 95,5 % Precision:66,5 %

Abbildung 9 Verwechslungsmatrizen der jeweils besten Netze für Testdaten aus bekanntem (oben) und unbekanntem (unten) Umfeld. *Verwechslungen von Stille können mit Eventdetektor vermieden werden.

Zusammenfassung und Ausblick

Die Experimente zeigen, dass CNNs basierend auf den Spektrogrammen der Audiosignale eine robuste Schnarcherkennung ermöglichen. Die Architektur der Netze hat dabei weniger Einfluss auf die erreichbare Genauigkeit, als die Aufbereitung der Datensätze. Durch Datenaugmentierung, insb. Aufprägen verschiedener Raumakustiken, kann die Leistungsfähigkeit der Netze auf Daten aus unbekanntem Umfeld und von unbekannten Personen deutlich verbessert werden. Nach Optimierung erreichen die besten Netze auf Testdaten aus bekanntem Umfeld eine Korrektklassifikationsrate von 95,2 %, sowie aus unbekanntem Umfeld von 80,3 %.

Im unbekannten Umfeld ließe sich die Genauigkeit durch einen akustischen Eventdetektor nochmals steigern, da der häufigste Fehlerfall fälschlicherweise als Schnarchen erkannte Stille ist.

Es bleibt zu klären, wie mit konstant hohen Pegeln umzugehen ist, bspw. bei offenem Fenster oder lauter Heizung. Um ein Netz auf die jeweilige häusliche Umgebung anzupassen, wäre ein Nachtrainieren vor Ort eine Option. Weitere Arbeiten sind nötig, um zwischen dem harmlosen Schnarchen und dem Schnarchen in Verbindung mit Schlafapnoe zu unterscheiden, bspw. durch Detektion längerer Pausen zwischen Schnarchereignissen. Auch die Zeitpunktbestimmung einer Elektrostimulation bspw. der Zunge des Schnarchenden könnte mithilfe dieser Methode zuverlässig umgesetzt werden.

Literaturverzeichnis

[1] F. Lopez, k. Behbehani und F. Kamangar, „An artifical neural network based snore detector," in *Preoceedings of 16th Annual Conference of the IEEE Engineering in Medicine and Biology Studies*, Baltimore, 1994.

[2] J. M. Perez-Macias, S. Adavanne, J. Viik, A. Värri, S.-L. Himanen und M. Tenhunen, „Assessment of support vector machines and convolutional neural networks to detect snoring using Emfit mattress," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Seogwipo, 2017.

[3] D. Pevernagie, R. M. Aarts und M. D. Meyer, „The acoustics of snoring," *Sleep Medicine Reviews,* Nr. 14, pp. 131-144, 2010.

[4] T. Kim, J.-W. Kim und K. Lee, „Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques," *BioMedical Engineering OnLine,* Bd. 17, Nr. 16, 2018.

[5] B. Kraus und S. Feldes, „Schnarcherkennung mit diskreten Hidden-Markov-Modellen," in *Fortschritte der Akustik - DAGA 2014*, Oldenburg, 2014.

[6] M. Mousa, S. Feldes und K.-H. Krauß, „Eingebettetes System zur Schnarch-Erkennung und Schnarch-Unterbindung," *FuE-Profil - Forschungsbericht der HAW Mannheim ,* pp. 12 - 17, 2013.

[7] S. Amiparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird und B. Schuller, „Snore Sound Classification Using Image-based Deep Spectrum Features," in *INTERSPEECH 2017*, Stockholm, 2017.

[8] C. Janott und M. Schmitt, „Snoring classified: The Munich-Passau Snore Sound Corpus," *Computers in Biology and Medicine,* pp. 106-118, 2018.

[9] P. Warden, „Speech Commands: A Dataset for Limited-Vocabulary Speech," Google Brain, Montain View, 2017.

[10] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant und P. Vary, „Do We Need Dereverberation for Hand-Held Telephony?," in *Proceedings of 20th International Congress on Acoustics*, Sydney, Australia, 2010.