C. Janott[1] · M. Schmitt[2] · C. Heiser[3] · W. Hohenhorst[4] · M. Herzog[5] · M. Carrasco Llatas[6] · W. Hemmert[1] · B. Schuller[2,7]

[1] Munich School of BioEngineering, Technische Universität München, Garching, Deutschland

[2] ZD.B Lehrstuhl für Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Augsburg, Deutschland

[3] Hals-Nasen-Ohrenklinik und Poliklinik, Klinikum rechts der Isar, Technische Universität München, München, Deutschland

[4] Klinik für HNO-Heilkunde, Kopf- und Hals-Chirurgie, Alfried Krupp Krankenhaus, Essen, Deutschland

[5] Klinik für HNO-Krankheiten, Kopf- und Halschirurgie, Carl-Thiem-Klinikum Cottbus, Cottbus, Deutschland

[6] Servicio de Otorrinolaringología, Hospital Universitario Doctor Peset, Valencia, Spanien

[7] Department of Computing, Imperial College London, London, Großbritannien

VOTE versus ACLTE: Vergleich zweier Schnarchgeräusch-klassifikationen mit Methoden des maschinellen Lernens

Für die Behandlung des Schnarchens gibt es konservative und operative Therapien. Diagnostische Maßnahmen wie die Polygraphie oder Polysomnographie können die Häufigkeit des Schnarchens und eine ggf. vorliegende obstruktive Schlafapnoe messen, lassen aber keine Rückschlüsse auf deren Entstehungsmechanismen zu. Die akustische Analyse von Schnarchgeräuschen mittels maschinellem Lernen kann dabei unterstützen, verschiedene Orte und Arten der Schnarchschallentstehung automatisiert zu unterscheiden und so eine Unterstützung bei der Wahl der individuell besten Therapie bieten.

Hintergrund

Für eine möglichst gezielte individuelle Behandlung des Schnarchens ist die Kenntnis der zugrunde liegenden Entstehungsmechanismen wichtig. Die medikamenteninduzierte Schlafvideoendoskopie (MISE) [1] wird zunehmend und mit Erfolg für eine differenzierte Diagnostik eingesetzt [2–4], sie ist jedoch kosten- und zeitaufwendig. Zudem kann sie für den Patienten unangenehm und belastend sein.

Methoden der akustischen Analyse von Schnarchgeräuschen hingegen sind nichtinvasiv und beeinflussen den natürlichen Schlaf des Patienten nicht [5]. Zudem können sie mit vergleichsweise geringem Aufwand, im häuslichen Umfeld des Patienten und auch über mehrere Nächte oder regelmäßig über einen längeren Zeitraum eingesetzt werden.

Jüngst ist die Verwendung von Methoden der künstlichen Intelligenz (KI) zur Schnarchgeräuschanalyse verstärkt in den wissenschaftlichen Fokus gerückt, mit teilweise vielversprechenden Ergebnissen [6–10]. Die überwiegende Anzahl der publizierten Untersuchungen basiert auf Methoden des überwachten maschinellen Lernens („supervised machine learning“). Dabei wird ein maschineller Klassifikator trainiert, Audioereignisse anhand ihrer akustischen Eigenschaften einer von mehreren vorher definierten Klassen zuzuordnen. Während des Trainingsvorgangs lernt der Klassifikator, anhand eines Trainingsdatensatzes und der Information, zu welcher Klasse die Ereignisse jeweils gehören, charakteristische Muster in den Daten zu erkennen. Das als Ergebnis des Trainings entstehende Klassifikationsmodell (umgangssprachlich manchmal auch als „Algorithmus“ bezeichnet) kann dann von den Trainingsdaten unabhängige Testdaten selbstständig den definierten Klassen zuordnen.

Je mehr unterschiedliche Klassen ein maschinelles Lernsystem unterscheiden soll, desto höher ist die Anforderung an die Größe des Trainingsdatensatzes. Außerdem ist die zu erwartende Erkennungsleistung eines maschinellen Klassifikators bei gleicher Größe des Trainingsdatensatzes umso höher, je weniger Klassen unterschieden werden müssen.

Die Aufgabe liegt nun darin, Klassifikationsschemata für Schnarchgeräusche zu entwickeln, welche einerseits eine möglichst aussagekräftige Unterstützung bei der Wahl der therapeutischen Maßnahmen bieten, andererseits aber aus möglichst wenigen Klassen bestehen.

Material und Methoden

Klassifikationsschemata

Ein häufig referenziertes Schema für die standardisierte Klassifikation von MISE-Untersuchungsbefunden bei Patienten mit obstruktiver Schlafapnoe (OSA) ist

Tab. 1 Vibrationsorte und -muster entsprechend der VOTE-Klassifikation

Ort	Muster		
	Anterior-posterior	*Lateral*	*Konzentrisch*
Velum	V – a-p	V – l	V – c
Oropharynx	O – a-p	O – l	O – c
Zungengrund	T – a-p	T – l	T – c
Epiglottis	E – a-p	E – l	E – c

V Velum, *O* Oropharynx, *T* Zungengrund, *E* Epiglottis

Tab. 2 Definierte Klassen, s-VOTE-Schema („simplified VOTE")

Ort	Muster		
	Anterior-posterior	*Lateral*	*Konzentrisch*
V		V	
O		O	
T		T	
E		E	

V Velum, *O* Oropharynx, *T* Zungengrund, *E* Epiglottis

Tab. 3 Definierte Klassen ACLTE-Schema(5 Klassen, Kombinationen aus Vibrationsort und -orientierung)

Ort	Muster		
	Anterior-posterior	*Lateral*	*Konzentrisch*
V	A	–	C
O	–	L	
T	T		
E	E		

Die mit einem Strich (–) gekennzeichneten Kombinationen von Ort und Muster kommen nicht vor
V Velum, *O* Oropharynx, *T* Zungengrund, *E* Epiglottis

die VOTE-Klassifikation [11]. Sie erlaubt eine standardisierte Beschreibung der Obstruktionen bezüglich ihres Orts (V – Velum, O – Oropharynx, T – Zungengrund und E – Epiglottis), ihres Musters (a-p – anterior-posterior, l – lateral, c – konzentrisch) und des Ausprägungsgrads (0 = keine Obstruktion; 1 = partielle Obstruktion, 2 = komplette Obstruktion). Die **Tab. 1** zeigt die theoretisch möglichen Kombinationen aus Vibrationsort und Muster entsprechend der VOTE-Klassifikation. Definitionsgemäß liegt einem Schnarchereignis eine Vibration von Weichgewebe zugrunde, weswegen im Folgenden Ort und Muster hörbarer Vibrationen unterschieden werden.

Ausgehend von der VOTE-Klassifikation wurden 2 Schemata zur Schnarchgeräuschklassifikation definiert. Die **Tab. 2 und 3** zeigen, welche Orte und Muster die Klassen umfassen. In den **Abb. 1a, b** sind die entsprechenden anatomischen Bereiche in den oberen Atemwegen anhand einer schematischen Darstellung eines Sagittalschnitts durch den Kopf dargestellt.

Im ersten Schema (**Tab. 2**; **Abb. 1a**) wurde nur die Vibrationsebene berücksichtigt. Dieses 4-Klassen-Schema wird im Folgenden als „s-VOTE" („simplified VOTE") bezeichnet, die Klassen entsprechen genau den in der VOTE-Klassifikation im Detail beschriebenen Ebenen [11]. Im Gegensatz dazu werden im zweiten Schema (**Tab. 3**; **Abb. 1b**) verschiedene Kombinationen aus Vibrationsort und -orientierung unterschieden. Das Schema umfasst 5 Klassen und wird im Folgenden als „ACLTE" bezeichnet. Dabei sind die Klassen folgendermaßen definiert:

- A: anterior-posteriore Vibration des weichen Gaumens und/oder der Uvula
- C: konzentrische Vibration auf Ebene des weichen Gaumens, der Uvula oder im Bereich des Oropharynx
- L: laterale Vibrationen auf Ebene des Oropharynx
- T: Vibration auf Höhe des Zungengrundes
- E: Vibration im Bereich der Epiglottis

Bei den Klassen T und E wird die Orientierung der Vibration nicht berücksichtigt.

Schnarchdatenbanken

Ausgangsmaterial für die Erstellung der Schnarchdatenbanken waren Audio- und Videoaufzeichnungen von MISE-Untersuchungen, die zwischen 2006 und 2016 an den HNO-Kliniken 4 klinischer Zentren durchgeführt wurden:

- Klinikum rechts der Isar, München
- Alfried Krupp Krankenhaus, Essen
- Universitätsklinik Halle/Saale, Halle/Saale
- Hospital Dr. Peset, Valencia, Spanien

Zur Audioaufnahme wurden an den einzelnen Zentren unterschiedliche Mikrofone und Mikrofonplatzierungen verwendet. Dies ist für eine Maschinenlernaufgabe prinzipiell wünschenswert. Durch verschiedene Hintergrundgeräuschcharakteristika, Raumeigenschaften und Mikrofonpositionen werden die trainierten Modelle von akustischen Rahmenbedingungen unabhängig und dadurch die Klassifikation von Schnarchsignalen aus unterschiedlichen akustischen Umgebungssituationen prinzipiell robuster. Da alle Schnarchereignisse vor der Merkmalsextraktion und der Klassifikation pegelnormalisiert, d. h. im Lautstärkepegel einander angeglichen, wurden, ist das resultierende Klassifikatormodell von dem absoluten Schallpegel und damit auch von dem Mikrofonabstand zur Schallquelle unabhängig.

Im Einzelnen wurde das folgende Equipment für die Audioaufzeichnungen verwendet.

- Klinikum rechts der Isar: Storz Telepack X Aufnahmesystem (Fa. Storz, Tuttlingen, Deutschland), Headsetmikrofon Sennheiser ME3 (Fa. Sennheiser, Wedemark, Deutschland), Aufnahmeposition 5–10 cm seitlich vom Mund des Patienten.
- Alfried Krupp Krankenhaus: Rehder/Partner rpSzene Aufnahmesystem (Fa. Rehder und Partner, Hamburg, Deutschland), bei Verwendung des handgehaltenen Mikrofons Aufnahmeabstand ca. 1 m vor dem Mund des Patienten, alternativ mit Stirnmikrofon des Systems Aufnahmeabstand ca. 30 cm oberhalb des Patientenmundes.
- Universitätsklinik Halle/Saale: Videosystem AIDA (Fa. Storz, Tuttlingen, Deutschland), externes Kondensatormikrofon NT3 (Fa. RODE, Silverwater, Australien); Aufnah-

meabstand 30 cm vor dem Mund des Patienten.
- Hospital Dr Peset: AverMedia C285 Capture Box (Fa. AverMedia, New Taipei City, Taiwan), Aufnahme mit Lavaliermikrofon, Befestigung auf Höhe des Schlüsselbeins, Aufnahmeabstand ca. 20 cm vom Mund des Patienten.

Insgesamt wurden MISE-Aufzeichnungen von über 2500 Patienten analysiert.

Zunächst wurden Schnarchereignisse in den Audiospuren der MISE-Aufzeichnungen mit einem semiautomatischen Verfahren identifiziert, dann in separaten Audiodateien (amplitudenpegelnormalisiert, Format wav, Auflösung 16 bit, Samplerate 16 kHz) gespeichert und die Zeitpunkte des Auftretens im Video notiert. Schnarchereignisse, die nichtstationäre Störsignale (z. B. Warntöne von Geräten im Untersuchungsraum) enthielten, sowie verzerrte Audiosignale wurden verworfen.

Damit der maschinelle Klassifikator „weiß", zu welcher Klasse die einzelnen Schnarchereignisse gehören, wurde jedes Ereignis mit einem Label versehen, das die jeweilige Klasse enthält. Als objektiver Referenzwert (Ground Truth) für das Label diente der endoskopische Befund basierend auf dem aufgezeichneten Video der MISE-Untersuchung. Jedes Schnarchereignis wurde durch 2 erfahrene und verblindete Untersucher anhand des Videobefunds bewertet und manuell einer der vorgegebenen Klassen zugeordnet (annotiert). Über eine Tabelle wurde diese Information den jeweiligen Audiodateien zugeordnet (gelabelt). Nur Schnarchereignisse, bei denen eine eindeutige Klasse von beiden Untersuchern übereinstimmend erkannt wurde, wurden in die Datenbank aufgenommen. Ereignisse mit mehreren Vibrationsorten oder Ereignisse, bei denen die Untersucher zu unterschiedlichen Ergebnissen gekommen sind, wurden verworfen.

Die Annotation nach den beiden Klassifikationsschemata wurde unabhängig voneinander zu getrennten Zeiten durchgeführt. Die Annotation nach s-VOTE entstand im Zeitraum von 2015 bis 2017, die Annotation nach ACLTE zwischen 2017 und 2018.

C. Janott · M. Schmitt · C. Heiser · W. Hohenhorst · M. Herzog · M. Carrasco Llatas · W. Hemmert · B. Schuller

VOTE versus ACLTE: Vergleich zweier Schnarchgeräuschklassifikationen mit Methoden des maschinellen Lernens

Zusammenfassung

Hintergrund. Die akustische Analyse von Schnarchgeräuschen ist eine nichtinvasive Methode für die Diagnose von Entstehungsmechanismen des Schnarchens, die während des natürlichen Schlafs durchgeführt werden kann. Ziel der Arbeit ist die Entwicklung und Bewertung von Klassifikationsschemata für Schnarchgeräusche, die eine möglichst aussagekräftige Diagnoseunterstützung ermöglichen.
Material und Methoden. Basierend auf 2 annotierten Schnarchgeräuschdatenbanken mit unterschiedlicher Klassifikation (s-VOTE – 4 Klassen versus ACLTE – 5 Klassen) wurden identisch aufgebaute maschinelle Klassifikationssysteme trainiert. Der Merkmalsextraktor openSMILE wurde in Kombination mit einer linearen Support-Vektor-Maschine zur Klassifikation eingesetzt.
Ergebnisse. Mit einem ungewichteten Average Recall (UAR) von 55,4 % für das s-VOTE-Modell und 49,1 % für das ACLTE liegen die Ergebnisse auf ähnlichem Niveau. In beiden Modellen gelingt die beste Differenzierung für Epiglottisschnarchen, während velares und oropharyngeales Schnarchen häufiger verwechselt werden.
Schlussfolgerung. Automatisierte akustische Verfahren können bei der Diagnose von Schlafatmungsstörungen unterstützen. Einschränkungen in der Erkennungsleistung sind u. a. durch die begrenzte Größe der Trainingsdatensätze bedingt.

Schlüsselwörter

Respiratorische Symptome · Intrinsische Schlafstörungen · Obstruktive Schlafapnoe · Datenanalyse · Schlafvideoendoskopie

VOTE versus ACLTE: comparison of two snoring noise classifications using machine learning methods

Abstract

Background. Acoustic snoring sound analysis is a noninvasive method for diagnosis of the mechanical mechanisms causing snoring that can be performed during natural sleep. The objective of this work is development and evaluation of classification schemes for snoring sounds that can provide meaningful diagnostic support.
Materials and methods. Based on two annotated snoring noise databases with different classifications (s-VOTE with four classes versus ACLTE with five classes), identically structured machine classification systems were trained. The feature extractor openSMILE was used in combination with a linear support vector machine for classification.
Results. With an unweighted average recall (UAR) of 55.4% for the s-VOTE model and 49.1% for the ACLTE, the results are at a similar level. In both models, the best differentiation is achieved for epiglottic snoring, while velar and oropharyngeal snoring are more often confused.
Conclusion. Automated acoustic methods can help diagnose sleep-disordered breathing. A reason for the restricted recognition performance is the limited size of the training datasets.

Keywords

Respiratory signs and symptoms · Intrinsic sleep disorders · Obstructive sleep apnea · Data analysis · Drug induced sleep endoscopy

Größe und Charakteristika der beiden resultierenden Datenbanken sind in Tab. 4 aufgeführt. Für die ACLTE-Datenbank wurden Teile der in der s-VOTE-Datenbank enthaltenen Schnachereignisse wiederverwendet und neu annotiert. Die s-VOTE-Datenbank wurde erstmalig unter der Bezeichnung „Munich-Passau Snore Sound Corpus" (MPSSC) publiziert [12, 13] und im Rahmen der Interspeech Computational Paralinguistics Challenge (ComParE) 2017 in einem Wettbewerb zum maschinellen Lernen eingesetzt. Die ACLTE-Datenbank sowie der Vergleich der Klassifikationsergebnisse beider Datenbanken wird in diesem Artikel zum ersten Mal veröffentlicht.

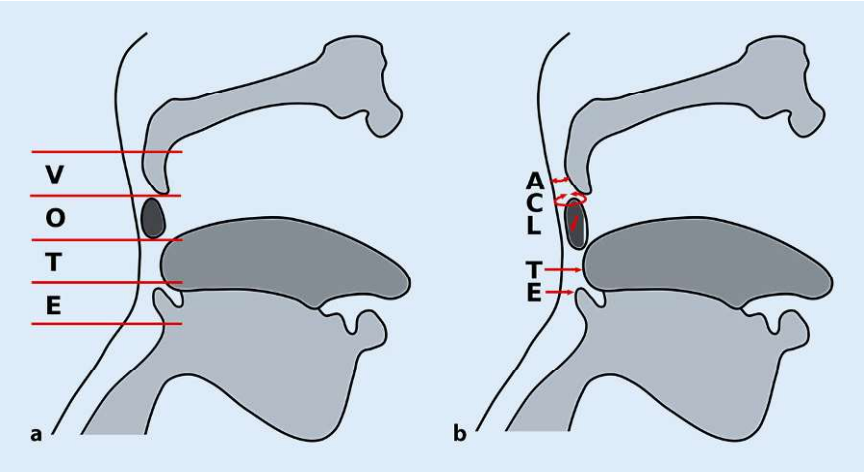


Abb. 1 ▲ **a** Vibrationsebenen, Sagittalschnitt durch den Kopf; s-VOTE-Schema (*V* Velum, *O* Oropharynx, *T* Zungengrund und *E* Epiglottis). **b** Vibrationsbereiche und Vibrationsorientierungen, Sagittalschnitt durch den Kopf; ACLTE-Schema (*A* anterior-posteriore Vibration des weichen Gaumens und/oder der Uvula, *C* konzentrische Vibration auf Ebene des weichen Gaumens, der Uvula oder im Bereich des Oropharynx, *L* laterale Vibrationen auf Ebene des Oropharynx, *T* Vibration auf Höhe des Zungengrundes, *E* Vibration im Bereich der Epiglottis)

Merkmalsextraktion und Partitionierung

Die **Abb. 2** zeigt den Aufbau des verwendeten maschinellen Lernsystems. Die Schnarchereignisse werden einem Merkmalsextrator zugeführt, der für jedes Schnarchereignis eine Reihe von Merkmalen erzeugt, die bestimme Eigenschaften des zugrunde liegenden Audiosignals in numerischer Form beschreiben. Beispiele solcher Merkmale sind die Grundfrequenz, das Verhältnis der Energieanteile hoher und niedriger Frequenzen, das Verhältnis von tonhaltigen Signalanteilen und Rauschen sowie aus der Stimmanalyse bekannte mikroprosodische Merkmale. Diese Merkmale werden auch als Low Level Descriptors (LLD) bezeichnet. In **Tab. 5** sind alle verwendeten Gruppen der akustischen Merkmale aufgeführt. Der Vektor enthält Informationen über den zeitlichen Verlauf der LLD innerhalb des jeweiligen Audioereignisses. Diese Eigenschaften werden über die jeweils erste Ableitung sowie mehrere statistische Funktionen (z. B. Maximum, Minimum, Median, Mittelwert, Standardabweichung) pro LLD codiert.

Als Merkmalsextraktor wurde das Open-Source-System openSMILE (Speech & Music Interpretation by Large-Space Extraction) [14, 15] eingesetzt. Der resultierende Vektor enthält insgesamt 6373 Merkmale. Dieser Merkmalssatz ist in mehreren Wettbewerben zur maschinellen Klassifikation unterschiedlicher Audiodaten erfolgreich eingesetzt worden [12, 16–18] und wurde daher auch für die vorliegenden Untersuchungen als Basis verwendet.

Der Merkmalsvektor wird gemeinsam mit dem Label einem maschinellen Klassifikator zugeführt. In den akustischen Merkmalen erkennt der Klassifikator Muster, die für die Schnarchgeräusche charakteristisch sind, und kann diese anhand des Labels der jeweiligen Klasse zuordnen. Als Ergebnis des Trainingsvorgangs entsteht ein Modell, mit dessen Hilfe basierend auf den akustischen Merkmalen die jeweilige Klasse erkannt werden kann.

Die beiden Datenbanken wurden jeweils in 3 probandendisjunktive und nach Schnarchklasse, Zentrum, Geschlecht und Alter balancierte Partitionen „training“, „develop“ und „test“ aufgeteilt. Eine detaillierte Beschreibung dieses Arbeitsschritts findet sich in einer Publikation von Janott et al. [13].

Klassifikation

Als maschineller Klassifikator wurde eine Supportvektor-Maschine (SVM) eingesetzt (Liblinear [19]). Eine SVM ermittelt in einer Menge von Elementen unterschiedlicher Klassen eine Trennlinie, die diese Klassen bestmöglich unterteilt. Dabei soll die Trennlinie so angeordnet sein, dass um die Klassengrenzen herum ein möglichst breiter freier Bereich bleibt. Diese Trennlinie wird als Supportvektor bezeichnet, **Abb. 3** zeigt ein zweidimensionales Beispiel einer linearen Unterteilung zweier Klassen durch einen Supportvektor. Die Dimension des Raumes, in der die SVM den Supportvektor errechnet, entspricht in unserem Fall der Größe des Merkmalsvektors, also einem 6373-dimensionalen Raum. Eine ausführlichere Beschreibung der prinzipiellen Funktionsweise einer SVM findet sich z. B. in einer Arbeit von Fan et al. [19].

Es wurde ein linearer Kernel verwendet (Kerneltyp: „L2-regularised L2-loss support, vector classification, dual“). Lineare SVM erzielen gute Ergebnisse gerade bei kleineren Datensätzen und einer vergleichsweise großen Anzahl von Merkmalen, wie es in den vorliegenden Experimenten der Fall ist. Darüber hinaus sind sie durch die Variation des Komplexitätsparameters C gut in ihrem Generalisierungsverhalten steuerbar, indem ein bestimmter Anteil von Fehlern beim Training erlaubt und eine Überanpassung auf die Trainingsdaten vermieden wird. Eine ausführliche Übersicht über unterschiedliche Arten von Klassifikatoren und ihre Leistungsfähigkeit bei der Unterscheidung von Schnarchgeräuschen findet sich in einer Arbeit von Qian et al. [20].

Zunächst wurden auf der „Training-Partition“ Modelle mit unterschiedlichen Komplexitätsparametern ($C = 2^{-30}, 2^{-29}, 2^{-28}, \ldots, 2^{0}$) trainiert. Danach wurden diese Modelle auf der „Develop-Partition“ getestet und der Komplexitätsparameter des Modells mit den besten Ergebnissen ausgewählt. Mit diesem Parameter wurde schließlich ein Modell auf der Kombination aus „Training-“ und „Develop-Partition“ trainiert und auf der „Test-Partition“ getestet.

Um mögliche Unterschiede in den Partitionen auszugleichen, wurden alle Experimente 6-Mal in allen möglichen Permutationen der 3 Partitionen durch-

Tab. 4 Eigenschaften der Datensätze

Klassifikation	Anzahl	Anteil (%)
ACLTE		
Probanden insgesamt	343	100
Alter (Jahre, Spannbreite)	48,6	(20–74)
Männlich	306	89
Weiblich	37	11
Probanden, Datenherkunft		
Essen	278	81
München	24	7
Halle	22	6
Valencia	19	6
Schnarchereignisse		
Gesamt	1115	100
A	521	47
C	172	15
L	263	24
T	37	3
E	122	11
Ereignisse pro Partition		
Training	373	33,5
Development	369	33,1
Test	373	33,5
s-VOTE		
Probanden insgesamt	219	100
Alter (Jahre, Spannbreite)	49,8	(24–78)
Männlich	205	94
Weiblich	14	6
Probanden, Datenherkunft		
Essen	164	75
München	25	11
Halle	30	14
Schnarchereignisse		
Gesamt	828	100
V	484	58
O	216	26
T	39	5
E	89	11
Ereignisse pro Partition		
Training	282	34,1
Development	283	34,2
Test	263	31,8

ACLTE 5 Klassen, Kombinationen aus Vibrationsort und -orientierung, VOTE: *V* Velum, *O* Oropharynx, *T* Zungengrund, *E* Epiglottis

Tab. 5 Verwendete Gruppen von Low Level Descriptors (LLD)

Nr.	Anzahl LLD	Beschreibung	Kurzbezeichnung
1	1	Effektivwert der Energie	„RMS energy“
2	1	Nulldurchgangsrate	ZCR
3	1	Wahrscheinlichkeit der Stimmhaftigkeit	„Voicing probability“
4	1	Geglättete Grundfrequenzkontur (Pitch)	Pitch
5	1	Logarithmisches Verhältnis der Energie der harmonischen Signalanteile zu Rauschsignalanteilen („signal-to-noise ratio“)	Log HNR
6	1	Mikrovariation benachbarter Periodenlängen der Grundfrequenz (Jitter)	Jitter
7	1	Erste Ableitung von Nr. 6	Jitter DDP
8	1	Mikrovariation der Amplitude aufeinanderfolgender stimmhafter Signalperioden (Shimmer)	Shimmer
9	15	Amplitudenanteil der Fast-Fourier-Transformationskoeffizienten (FFT)	FFT mag
10	14	Mel-Frequenz-basierte Cepstralkoeffizienten (MFCC)	MFCC
11	26	Aus dem Mel-Frequenzspektrum abgeleitete wahrnehmungsbasierte lineare Prädiktion (PLP) der Cepstralkoeffizienten	Audspec PLP
12	1	Summe der Koeffizienten von Nr. 11	Audspec PLP total
13	1	Auf relativer spektraler Transformation (RST) basierende Filterung des auditiven Spektrums	Audspec RASTA

Tab. 6 Klassifikationsergebnisse, basierend auf dem ComParE-Merkmalssatz, gemittelt über alle Permutationen

Klassifikationsschema	s-VOTE	ACLTE
UAR	55,4 %	49,1 %

ComParE „Interspeech Computational Paralinguistics Challenge“, *UAR* ungewichteter Average Recall

geführt. Mit anderen Worten wurden die 3 Partitionen wie folgt durchgetauscht:

1. Permutation: Partition 1 = „training“; Partition 2 = „develop“; Partition 3 = „test“
2. Permutation: Partition 1 = „training“; Partition 3 = „develop“; Partition 2 = „test“
3. Permutation: Partition 2 = „training“; Partition 1 = „develop“; Partition 3 = „test“
4. Permutation: Partition 2 = „training“; Partition 3 = „develop“; Partition 1 = „test“
5. Permutation: Partition 3 = „training“; Partition 1 = „develop“; Partition 2 = „test“
6. Permutation: Partition 3 = „training“; Partition 2 = „develop“; Partition 1 = „test“

Ergebnisse

Die **Abb. 4** zeigt den Aufbau des Systems für die Klassifikation der von den Trainingsdaten unabhängigen Testdaten. Der aus den Audiosignalen erzeugte Merkmalsvektor wird dem vorher trainierten Modell zugeführt, welches die wahrscheinlichste Klasse ausgibt, zu der das Audioereignis gehört.

Als Maß für die Erkennungsgenauigkeit dient in den vorliegenden Experimenten der Average Recall (AR), definiert als über alle Klassen gemittelter Anteil der richtig zugeordneten Ereignisse. Ein ideales Modell, welches alle Ereignisse richtig erkennt, erreicht einen AR von 100 %. Der ungewichtete Average Recall (UAR) berechnet sich als ungewichteter Mittelwert der klassenspezifischen Recalls über alle Klassen:

$$\mathrm{UAR} = \sum(\text{klassenspezifischer Recall}) * 100\,\%/\,\text{Klassenanzahl}.$$

Ein Wert von 100 %/Klassenanzahl entspricht der Zufallswahrscheinlichkeit.

Die **Tab. 6** zeigt die Klassifikationsergebnisse für beide Klassifikationsschemata. Das 5-Klassen-Modell des ACLTE-

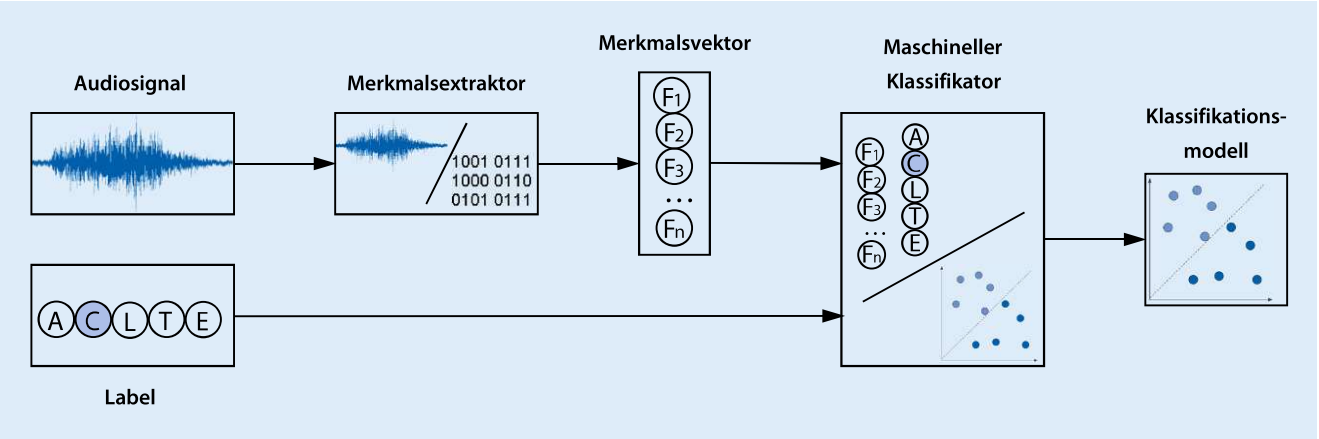


Abb. 2 ▲ Schematischer Aufbau des verwendeten maschinellen Lernsystems (Trainingsphase). *ACLTE* Klassifikation mit 5 Klassen, Kombinationen aus Vibrationsort und -orientierung

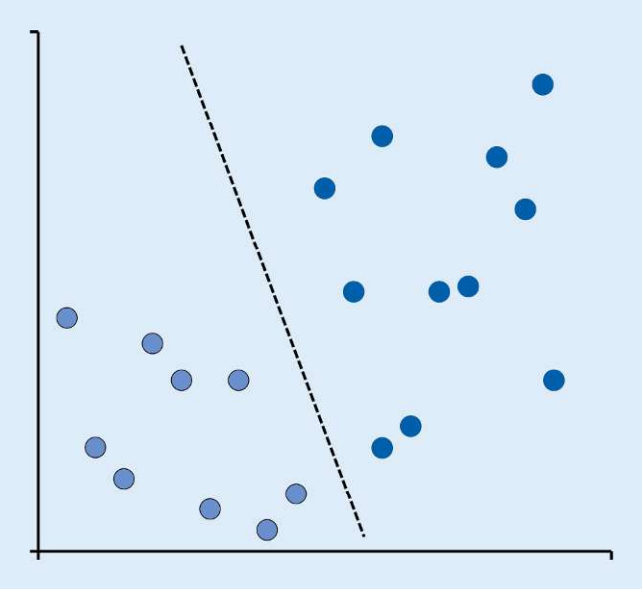


Abb. 3 ◀ Zweidimensionales Beispiel einer linearen Trennung von Elementen zweier Klassen durch einen Supportvektor (*gestrichelte Linie*)

Schemas erreicht einen etwas niedrigeren UAR als das 4-Klassen-Modell des s-VOTE-Schemas, was aufgrund der höheren Anzahl an Klassen plausibel ist. Insgesamt liegen die Ergebnisse aber etwa auf gleichem Niveau.

In **Abb. 5** ist der klassenspezifischen Recall beider Schemata dargestellt. Bemerkenswert ist, dass in beiden Schemata, die auf Epiglottisebene entstehenden Schnarchgeräusche am besten differenziert werden. Außerdem fällt eine ebenfalls gute Erkennungsrate der velaren Schnarchgeräusche auf. In der V-Klasse des s-VOTE-Schemas sind sowohl zirkuläre als auch anterior-posteriore Vibrationsmuster auf Velumebene enthalten, sodass sich aus den Ergebnissen im Vergleich schließen lässt, dass insbesondere Vibrationen in a-p-Richtung gut differenziert werden können. In beiden Schemata fällt die Erkennungsleistung von Zungengrundschnarchen am schlechtesten aus und liegt beim s-VOTE-Schema in etwa auf Zufallsniveau.

Die **Abb. 6 und 7** zeigen die Konfusionsmatrizen beider Schemata. Dabei werden die Erkennungsergebnisse des trainierten Modells auf den Test-Partitionen für die einzelnen Klassen im Detail dargestellt. Pro Zeile zeigt die Darstellung die prozentuale Häufigkeit, mit welcher Klasse die jeweiligen Ereignisse vom Klassifikationsmodell verwechselt wurden. Die Diagonale (grün hinterlegte Felder) enthält den richtig erkannten Anteil der jeweiligen Klasse und damit den klassenspezifischen Recall. Konfusionswerte ≥20 % sind markiert.

Im s-VOTE-Schema tritt die häufigste Konfusion zwischen velaren und oropharyngealen Schnarchereignissen auf, während Zungengrundschnarchen besonders häufig als oropharyngeales Schnarchen, aber auch als Epiglottisschnarchen fehlerkannt wird. Im ACLTE-Schema zeigt sich eine häufige Verwechslung von zirkulären und lateralen Vibrationen auf Velum-/Oropharynxebene. Zudem werden zirkuläre Vibrationen ebenfalls häufig als anterior-posteriores Velumschnarchen fehlerkannt. Auch in diesem Schema wird Zungengrundschnarchen mit (lateralem) oropharyngealem Schnarchen und mit Epiglottisschnarchen verwechselt.

Diskussion

Maschinelle Klassifikatoren sind „datenhungrig“: Je größer die verfügbaren Trainingsdatensätze, desto besser gelingt die Generalisierung der in den Daten vorhandenen Muster und desto genauer werden potenziell die Klassifikationsergebnisse. Andererseits ist die Menge der verfügbaren Trainingsdaten in dem vorliegenden Anwendungsfall knapp: Audio- und Videoaufzeichnungen von MISE-Untersuchungen sind nur in begrenzter Zahl verfügbar, und der Aufwand für die Auswahl und Annotation der Daten ist erheblich. Um trotz begrenzter Datenmenge möglichst

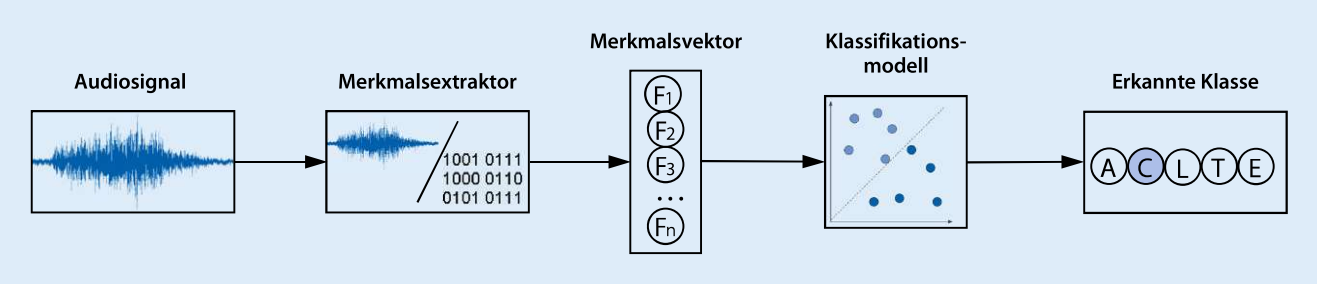


Abb. 4 ▲ Schematischer Aufbau des Klassifikators (Testphase). *ACLTE* Klassifikation mit 5 Klassen, Kombinationen aus Vibrationsort und -orientierung

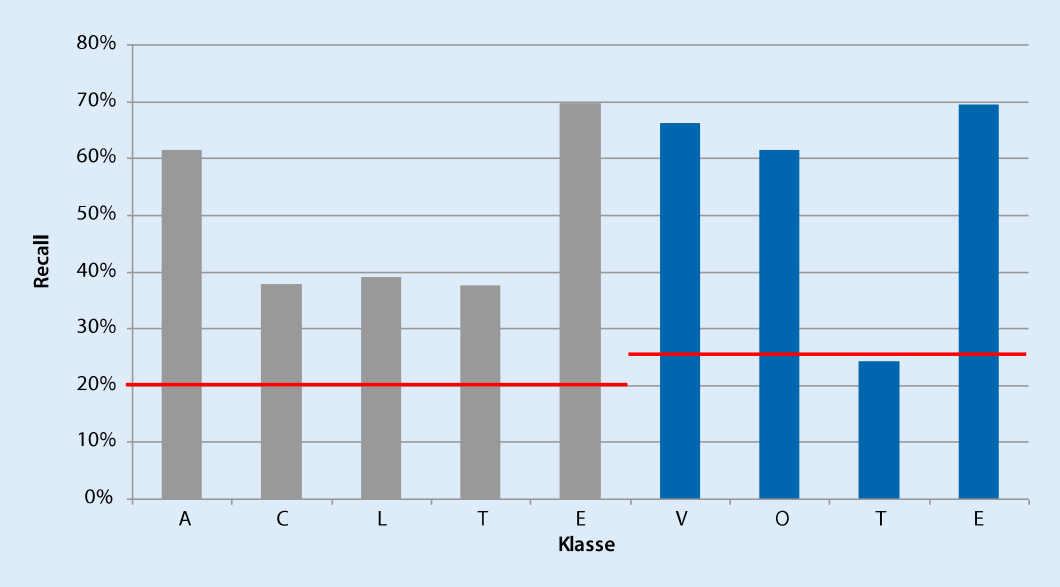


Abb. 5 ◀ Klassenspezifischer Recall ACLTE versus s-VOTE, gemittelt über alle Permutationen. *Rote waagerechte Linie* Zufallswahrscheinlichkeit

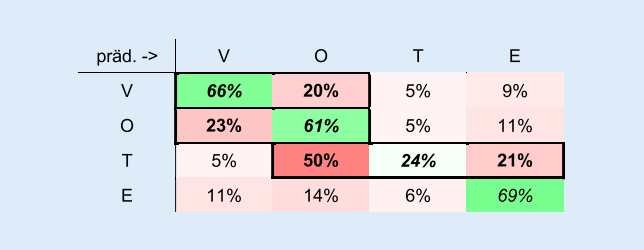
gute Ergebnisse zu erzielen, wurden 2 Klassifizierungsschemata verglichen, die relevante anatomische Informationen in einer möglichst geringen Anzahl unterschiedlicher Klassen codieren.

Neben der Ebene innerhalb der oberen Atemwege, auf denen das Schnarchgeräusch entsteht, enthält auch die Vibrationsrichtung wichtige und therapierelevante Informationen. So ist auf Ebene des Velo- und Oropharynx bei anterior-posteriorer Schwingungsrichtung eine isolierte Vibration des weichen Gaumens bzw. der Uvula zu vermuten. Zirkuläre Vibrationen weisen auf eine Beteiligung der Pharynxwände hin, während bei lateraler Schwingung mit hoher Wahrscheinlichkeit die Gaumenmandeln wesentlich zur Vibration beitragen. Diese differenziertere Aussage des ACLTE-Schemas gegenüber dem s-VOTE-Schema wird allerdings durch eine zusätzliche Klasse erkauft, was höhere Ansprüche an das maschinelle Lernsystem stellt.

Zusammengefasst gelingt auf Basis der vorliegenden Ergebnisse eine gute Differenzierung von Epiglottisschnarchen, außerdem wird isoliertes velopharyngeales Schnarchen in anterior-posteriorer Orientierung (also typisches Weichgaumenschnarchen) gut erkannt. Für die Klinik ist diese Information wertvoll, da diese Schnarcharten unterschiedlich therapiert werden müssen. Die Differenzierung unterschiedlicher Vibrationsmuster insbesondere auf Höhe des Oropharynx gelingt nur mäßig. Hier besteht weiterer Forschungsbedarf, da gerade die Form der Atemwegsverengung (zirkulär oder lateral) eine für die Therapie wichtige Information darstellt. Überraschenderweise tritt isoliertes Zungengrundschnarchen bei den untersuchten Daten sehr selten auf und wird daher auch nicht gut erkannt. Diese Diskrepanz zu wesentlich häufiger diagnostizierten Obstruktionen auf Zungengrundebene mag darin begründet sein, dass Letztere überwiegend unter Beteiligung velo- und oropharyngealer Strukturen als Multilevel-Obstruktionen auftreten. Multilevel-Schnarchereignisse wurden jedoch bei der Erstellung der Datenbank bewusst ausgeschlossen.

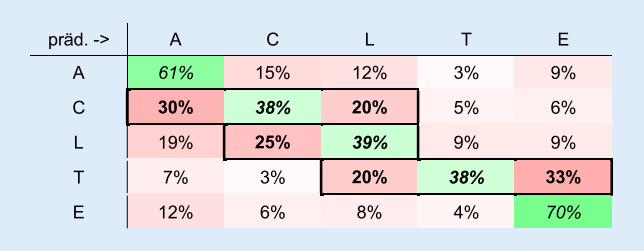
Einen Erklärungsansatz für die unterschiedlichen klassenspezifischen Recalls liefern die akustischen Eigenschaften der oberen Atemwege. Die Strecke vom Vibrationsort bis zur Schallabgabe an die Umgebung an Nasenlöchern und Lippen wirkt als Ansatzrohr, dass ein akustisch wirksames Filter darstellt, dessen Frequenzgang von seiner Länge und seinem Querschnittsverlauf abhängt [21].

Die gute Differenzierung von Epiglottisschnarchen kann entsprechend darin begründet liegen, dass das Ansatzrohr bei Vibrationen auf Epiglottisebene eine



präd. ->	V	O	T	E
V	**66%**	**20%**	5%	9%
O	**23%**	**61%**	5%	11%
T	5%	**50%**	**24%**	**21%**
E	11%	14%	6%	**69%**

Abb. 6 ◀ Über alle Permutationen gemittelte Konfusionsmatrix, s-VOTE. VOTE: *V* Velum, *O* Oropharynx, *T* Zungengrund, *E* Epiglottis, *präd.* prädiziert



präd. ->	A	C	L	T	E
A	**61%**	15%	12%	3%	9%
C	**30%**	**38%**	**20%**	5%	6%
L	19%	**25%**	**39%**	9%	9%
T	7%	3%	**20%**	**38%**	**33%**
E	12%	6%	8%	4%	**70%**

Abb. 7 ▲ Über alle Permutationen gemittelte Konfusionsmatrix, ACLTE. *ACLTE* Klassifikation mit 5 Klassen, Kombinationen aus Vibrationsort und -orientierung; *präd.* prädiziert

im Vergleich zu velaren oder oropharyngealen Vibrationen größere Gesamtlänge hat. Zudem besteht eine Engstelle im Querschnittsverlauf auf Höhe des Zungengrundes mit charakteristischem Einfluss auf die Übertragungsfunktion, die bei allen anderen Klassen nicht vorhanden ist. Analog lässt sich die häufigere Verwechslung der Klassen A, C und L bzw. V und O dadurch erklären, dass die Vibrationsorte nah beieinander liegen. Für diesen Erklärungsansatz spricht zudem, dass die am besten differenzierenden Merkmalsuntergruppen die spektralen Signaleigenschaften und damit die Filtereigenschaften des Ansatzrohres beschreiben [9, 13]. Im Vergleich dazu differenzieren die Merkmale, welche die Eigenschaften des anregenden Signals beschreiben, also die Charakteristik der Schallquelle anstelle von Länge und Form der oberen Atemwege, weniger gut.

Für die Datenbanken wurden nur solche Schallereignisse verwendet, die von 2 Annotatoren eindeutig einer Klasse zugeordnet werden konnten. Dennoch ist der Übergang zwischen unterschiedlichen Vibrationsorten und -orientierungen, gerade im Bereich des Oropharynx, in der Realität fließend. Dies kann eine weitere Erklärung für die Verwechslung insbesondere der Klassen C und L sein. Möglicherweise entscheidet das trainierte Modell anhand der akustischen Eigenschaften in Grenzfällen anders als ein menschlicher Bewerter auf Basis der Videobilder.

Die schlechte Erkennungsrate des Zungengrundschnarchens hingegen lässt sich durch die geringe Anzahl an Schallereignissen in dieser Klasse erklären. Mit nur 3 % (ACLTE) beziehungsweise 5 % (s-VOTE) aller Ereignisse ist die T-Klasse deutlich unterrepräsentiert. Das ist nicht erstaunlich, es ist bekannt, dass isolierte Vibrationen und Obstruktionen auf Zungengrundebene vergleichsweise selten auftreten [22]. Die Menge der Trainingsdaten ist entsprechend sehr gering für eine Maschinenlernaufgabe, die Ergebnisse können daher stark zufallsbedingt sein. Dafür spricht ebenfalls die Tatsache, dass die Abweichungen der klassenspezifischen Recalls der T-Klasse in den einzelnen Permutationen sehr groß sind [13].

Generell ist als methodische Einschränkung der vorliegenden Daten anzumerken, dass die Klassifikation der einzelnen Schnarchepisoden endoskopisch visuell erfolgt ist. Hierbei ist relevant, dass nur die oberste Obstruktionsebene direkt einsehbar ist. Potenzielle Obstruktionsebenen kaudal der einsehbaren könnten übersehen werden. Es ist jedoch möglich, das Endoskop über die Obstruktionsebene zu schieben und zu beobachten, was kaudal davon geschieht. Bei Vorschieben des Endoskops über die kraniale Obstruktionsebene wird diese meist nur wenig beeinträchtigt, und eine zweite kaudale Ebene kann detektiert werden. Bei der Auswahl der Schnarchepisoden wurde darauf geachtet, dass Vibrationen in nur einer Ebene auftraten. Multilokuläre Obstruktionen wurden bewusst ausgeschlossen, um eine möglichst klare Klassifikation für die folgenden akustischen Auswertungen zu erhalten. Gleiches gilt für Aufnahmen, bei denen eine akustische Beeinträchtigung durch übermäßige Schleim- und Speichelansammlungen festgestellt wurde.

Es kann vermutet werden, dass multilokuläre Vibrationsmuster komplexere akustische Charakteristika aufweisen als unilokuläre und auch übermäßige Speichelbildung zu einer akustischen Alteration der Geräusche führt. Hierzu besteht Forschungsbedarf in zukünftigen Projekten.

Für eine breite klinische Routineanwendung der akustischen Analyse wird es differenzierter Auswertungsalgorithmen und im Idealfall einer Korrelation mit polysomnographischen Parametern bedürfen, um das Phänomen des Schnarchens in all seiner Komplexität zu erfassen und akustische Charakteristika als Diagnostikkriterium nutzen zu können.

Fazit für die Praxis

- **Eine automatische Unterscheidung verschiedener Entstehungsorte und Vibrationsformen von Schnarchgeräuschen mittels maschineller Klassifikation ist prinzipiell möglich.**
- **Einschränkungen in der Erkennungsgenauigkeit sind teilweise durch die anatomischen Eigenschaften der oberen Atemwege bedingt, haben aber auch ihre Ursache in der geringen Größe der derzeit zur Verfügung stehenden Datensätze.**
- **Es ist zu erwarten, dass sich die Ergebnisse mit größeren Trainingsdatensätzen verbessern lassen.**
- **Auch wenn eine Diagnostik mittels künstlicher Intelligenz die bestehende diagnostische Praxis in absehbarer Zeit nicht ersetzen wird, so können automatisierte Verfahren in der Zukunft dennoch eine wertvolle**

Unterstützung bei der schonenden Untersuchung der Ursachen schlafbezogener Atmungsstörungen bieten.

Korrespondenzadresse

Dipl.-Ing. C. Janott
Munich School of BioEngineering, Technische Universität München
Boltzmannstraße 11, 85748 Garching, Deutschland
c.janott@gmx.net

Einhaltung ethischer Richtlinien

Interessenkonflikt. C. Janott ist der Erfinder eines patentierten Verfahrens und Systems zur Ermittlung anatomischer Ursachen für die Entstehung von Schnarchgeräuschen (DE102012219128B4). M. Schmitt, C. Heiser, W. Hohenhorst, M. Herzog, M. Carrasco Llatas, W. Hemmert und B. Schuller geben an, dass kein Interessenkonflikt besteht.

Für diesen Beitrag wurden von den Autoren keine Studien an Menschen oder Tieren durchgeführt. Für die aufgeführten Studien gelten die jeweils dort angegebenen ethischen Richtlinien.

Literatur

1. Croft C, Pringle M (1991) Sleep nasendoscopy: A technique of assessment in snoring and obstructive sleep apnoea. Clin Otolaryngol Allied Sci 16(5):504–509
2. Kent D, Rogers R, Soose R (2015) Drug-induced sedation endoscopy in the evaluation of OSA patients with incomplete oral appliance therapy response. Otolaryngol Head Neck Surg 153(2):302–307
3. De Vito A, Carrasco Llatas M, Ravesloot MJ, Kotecha B, De Vries N, Hamans E et al (2018) European position paper on drug-induced sleep endoscopy: 2017 Update. Clin Otolaryngol 43(6):1541–1552
4. Heiser C, Fthenakis P, Hapfelmeier A, Berger S, Hofauer B, Hohenhorst W et al (2017) Drug-induced sleep endoscopy with target-controlled infusion using propofol and monitored depth of sedation to determine treatment strategies in obstructive sleep apnea. Sleep Breath 21(3):737–744
5. Janott C, Pirsig W, Heiser C (2014) Akustische Analyse von Schnarchgeräuschen. Somnologie 18(2):87–95
6. Schmitt M, Janott C, Pandit V, Qian K, Heiser C, Hemmert W et al (2016) A bag-of-audiowords approach for snore sounds' excitation localisation. Proceedings 14. ITG Conference on Speech Communication, IEEE, S 1–5
7. Qian K, Janott C, Zhang Z, Heiser C, Schuller B (2016) Wavelet features for classification of VOTE snore sounds. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016, S 221–225
8. Janott C, Schuller B, Heiser C (2017) Acoustic information in snoring noises. HNO 65:107–116
9. Qian K, Janott C, Pandit V, Zhang Z, Heiser C, Hohenhorst W et al (2016) Classification of the Excitation Location of Snore Sounds in the Upper Airway by Acoustic Multi-Feature Analysis. Ieee Trans Biomed Eng. https://doi.org/10.1109/TBME.2016.2619675
10. Qian K, Janott C, Deng J, Heiser C, Hohenhorst W, Herzog M et al (2017) Snore sound recognition: On wavelets and classifiers from deep nets to kernels. In 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)
11. Kezirian EJ, Hohenhorst W, Vries N (2011) Drug-induced sleep endoscopy: the VOTE classification. Eur Arch Otorhinolaryngol 268:1233–1236
12. Schuller B, Steidl S, Batliner A, Bergelson E, Krajewski J, Janott C et al (2017) The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, cold & snoring. Proceedings of INTERSPEECH; 2017; Stockholm, Sweden, S 20–24
13. Janott C, Schmitt M, Zhang Y, Qian K, Pandit V, Zhang Z et al (2018) Snoring classified: The Munich-Passau snore sound corpus. Comput Biol Med 94:106–118
14. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, S 1459–1462
15. Eyben F, Weninger F, Groß F, Schuller B (2013) Recent developments in opensmile, the munich open-source multimedia feature extractor. Proceedings of the 21st ACM International Conference on Multimedia, S 835–838
16. Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F et al (2013) The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism. IProceedings of INTERSPEECH; 2013; Lyon, S 148–152
17. Schuller B, Steidl S, Batliner A, Epps J, Eyben F, Ringeval F et al (2014) The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & physical load. Proceedings of INTERSPEECH; 2014; Singapore, S 427–431
18. Schuller B, Steidl S, Batliner A, Hantke S, Hönig F, Orozco-Arroyave JR et al (2015) The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & eating condition. IProceedings of INTERSPEECH; 2015; Dresden, Germany, S 478–482
19. Fan R, Chang K, Hsieh C, Wang X, Lin C (2008) LIBLINEAR: A library for large linear classification. J Mach Learn Res 9:1871–1874
20. Qian K, Janott C, Zhang Z, Deng J, Baird A, Heiser C et al (2018) Teaching machines on snoring: A benchmark on computer audition for snore sound excitation localisation. Arch Acoust 43(3):465–475
21. Fant G (1970) Acoustic theory of speech production
22. Hessel NS, Vries N (2003) Diagnostic work-up of socially unacceptable snoring. II. Sleep endoscopy. Eur Arch Otorhinolaryngol 259:158–161