

**Kommunikationsstörungen**

**Berichte aus Phoniatrie und Pädaudiologie**

Herausgeber : M. Döllinger

Begründet 1996 von U. Eysholdt

**Pablo Gómez**

**Deep Learning Methods  
for Processing Endoscopic  
High-Speed Video and Laryngeal  
Parameter Estimation**

**SHAKER  
VERLAG**

# Deep Learning Methods for Processing Endoscopic High-Speed Video and Laryngeal Parameter Estimation

Deep-Learning-Methoden für die Bildverarbeitung  
endoskopischer Hochgeschwindigkeitsaufnahmen und  
Schätzung laryngealer Parameter

Der Technischen Fakultät  
der Friedrich-Alexander-Universität  
Erlangen-Nürnberg

zur  
Erlangung des Doktorgrades  
**DOKTOR-INGENIEUR**

vorgelegt von  
**Pablo Gómez**  
aus Erlangen

Als Dissertation genehmigt  
von der Technischen Fakultät  
der Friedrich-Alexander-Universität  
Erlangen-Nürnberg

Tag der mündlichen Prüfung: 4. Juli 2019  
Vorsitzender des Promotionsorgans: Prof. Dr.-Ing. Reinhard Lerch  
Gutachter: Prof. Dr.-Ing. Michael Döllinger  
Gutachter: PD Dr.-Ing. Thomas Wittenberg

Kommunikationsstörungen - Berichte aus Phoniatrie und  
Pädaudiologie

Band 27

**Pablo Gómez**

**Deep Learning Methods for Processing  
Endoscopic High-Speed Video and  
Laryngeal Parameter Estimation**

D 29 (Diss. Universität Erlangen-Nürnberg)

Shaker Verlag  
Düren 2019

**Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Zugl.: Erlangen-Nürnberg, Univ., Diss., 2019

Copyright Shaker Verlag 2019

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-6845-0

ISSN 1436-1175

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren

Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9

Internet: [www.shaker.de](http://www.shaker.de) • e-mail: [info@shaker.de](mailto:info@shaker.de)

# Danksagung

Mit der Promotion endet für mich vorerst ein über 20 Jahre andauernder Bildungsweg. Die letzten drei Jahre dieses Weges habe ich in der Phoniatrie am Universitätsklinikum Erlangen verbracht. An dieser Stelle möchte ich die Gelegenheit ergreifen, um mich zu bedanken bei den vielen Menschen, die mich auf diesem Weg unterstützt und begleitet haben.

In vielerlei Hinsicht war der Beginn meiner Promotion am Universitätsklinikum für mich ein Neuanfang. Informatiker verschlägt es relativ selten in die Forschung an einem Klinikum. Trotzdem habe ich in dem in jeder Hinsicht diversen Team in der Phoniatrie mich schnell wohl gefühlt. Das Teamwork und der Zusammenhalt der Kolleginnen und Kollegen war ein entscheidender Faktor dafür. Der Umgang war stets von einem direkten und ehrlichen Austausch geprägt und ich fand dort immer ein offenes Ohr sowie fachlichen und persönlichen Rat. Persönlich fand ich auch den interdisziplinären Austausch, der von der Physik und Mathematik über die technischen Fächer bis zur Medizin reichte, extrem wertvoll. Dafür möchte ich mich bei allen Kolleginnen und Kollegen aus der Phoniatrie bedanken.

Besonderer Dank gilt an dieser Stelle meinem Doktorvater und Betreuer Prof. Michael Döllinger. Er stand mir immer geduldig zur Seite und der Austausch mit ihm war für mich ein fachlicher und persönlicher Zugewinn. Zu jedem Zeitpunkt nahm er sich die nötige Zeit, um den erfolgreichen Ablauf meiner Promotion, aber auch meiner persönlichen und wissenschaftlichen Karriere zu unterstützen. Gleichzeitig gab er mir auch große fachliche und wissenschaftliche Entfaltungsfreiheit, was ich sehr zu schätzen weiß. Weiterhin möchte ich mich bei den wissenschaftlichen Kolleginnen und Kollegen bedanken, die immer offen waren für Austausch und Zusammenarbeit. Von ihnen konnte ich nicht nur fachlich lernen, sondern auch persönlich und im Bezug auf das wissenschaftliche Arbeiten. Besonderen Dank möchte ich an dieser Stelle aussprechen an Marion und Stefan, die als *alte Hasen* der Phoniatrie mir besonders bei fachlichen Fragen stets halfen. Weiterhin möchte ich mich bei Denis bedanken, dessen Hinweise oft essentiell waren. Meinem Büronachbar Patrick möchte ich für den regen fachlichen Austausch, erfolgreiche Zusammenarbeit und die angenehme, gelassene Büroatmosphäre danken. Ich möchte mich bei Sebastian danken, dessen gute Laune immer ansteckend war und der immer für einen Kaffee bereit war. Besonders bedanken möchte ich mich außerdem bei Andreas für den spannenden Austausch in allen Belangen, die erfolgreiche Zusammenarbeit und das Korrekturlesen dieser Dissertation. Auch dem medizinischen Personal, das immer offen für Fragen war und einen unverzichtbaren Einblick in den klinischen Alltag gewährte, gilt mein Dank. Schlussendlich gilt auch denen, die hier

nicht persönlich erwähnt sind, mein Dank. Der Austausch am Arbeitsplatz war stets freundlich und von Hilfsbereitschaft geprägt.

Auch außerhalb des Arbeitsumfeldes habe ich rege Unterstützung und Zusprache erhalten, für die ich mich an dieser Stelle bei meinen Freundinnen und Freunden sowie meiner Familie bedanken möchte. Insbesondere möchte ich mich bei Maximilian Seitzer, dessen fachliche Expertise und Unterstützung bei mehreren Arbeiten und auch beim Korrekturlesen dieser Dissertation unentbehrlich war, bedanken. Schließlich gilt mein Dank auch meiner Mutter, die mich auf dem über 20 Jahre langen Bildungsweg stets unterstützt hat und auf die ich mich immer verlassen konnte.

Erlangen, den 13. März 2019





# Abstract

The human voice plays a central role in daily life. It is indispensable for many professions and it is often a critical aspect of human interaction. Therefore, any impairments of the voice have severe consequences for the individual and affect society in general. Yet, as the physical processes behind the voice are quite complex, it is paramount to develop appropriate techniques to gain additional insight on the underlying principles behind voice production. This will help to objectively assist diagnostics and treatment.

The inaccessibility of the human vocal folds and their rapid, small-scale oscillation require sophisticated measurement techniques. One of the state-of-the-art solutions is high-speed videoendoscopy (HSV). However, HSV data is regularly affected by insufficient lighting and the usually performed segmentation is often still a semi-automatic process. This thesis presents deep learning-based approaches to enhance and recover even severely underlit HSV videos and for a fully automated segmentation of HSV recordings. Both were tested on about 50 HSV recordings each. The main goal, however, lies in estimating physical parameters of the vocal folds from the obtained segmentations. Building on previous works, an improved numerical model is developed and employed to estimate the subglottal pressure in 288 porcine *ex vivo* HSV recordings. This is the first demonstration that this approach is applicable to *ex vivo* animal data. After that, to speed up the estimation and thus allow more sophisticated models to be employed, a recurrent neural network is used for the estimation on the same data. An innovative method to estimate the pressure by training the network on solely synthetic data from a numerical model is presented. Thus, this thesis improves the complete parameter estimation pipeline with innovative applications of deep learning, optimization and numerical methods.

Both, the enhancement and segmentation methods, provide robust results requiring only consumer-grade hardware. They can process more than 50 images per second on a consumer-grade graphics card. Segmentation accuracy is comparable to prior work at vastly reduced computation time. The low-light enhancement is shown to statistically significantly outperform four other state-of-the-art methods on three image quality metrics. The improved numerical model used for parameter estimation is shown to be able to reproduce the porcine vocal folds oscillation accurately. The error of the fundamental frequency was on average 0.02 Hz, mean amplitude errors were 0.008 cm and the subglottal pressure estimation can clearly capture trends in the data at an error of 2.9 cmH<sub>2</sub>O or 27.5%. Individual accuracy using the optimization is, however, limited in some cases. The developed deep learning approach is shown to vastly decrease the computational burden of the parameter estimation while maintaining accuracy. It

requires merely a fraction of the model evaluations and computation time compared to the optimization. The evaluation per recording was on average 56 796 times faster.

In summary, this thesis clearly demonstrates the potential of parameter estimation approaches and indicates a clinical applicability. Remaining manual steps were removed from the processing pipeline and streamlined to require only clinically feasible amounts of time and remove subjective influences. The accuracy of the estimation was demonstrated on an experimental baseline. The need for a way to employ more sophisticated models was identified and met. The developed neural network estimation approach allows employing models that were previously not considered viable for this task. The approach thereby starts to bridge the gap between approaches that require supercomputers and ones that are clinically feasible. The estimation of physical parameters bears the promise of a more objective diagnosis and treatment selection.

# Zusammenfassung

Die menschliche Stimme hat eine zentrale Rolle im Alltag. Sie ist unverzichtbar für viele Berufe und oftmals der entscheidende Teil menschlicher Interaktion. Jede Beeinträchtigung der Stimme hat daher dramatische Auswirkungen für das Individuum, aber auch die Gesellschaft im Allgemeinen. Da die zugrundeliegenden physikalischen Prozesse der Stimmgebung ausgesprochen komplex sind, ist es unabdingbar, angemessene Methoden zu entwickeln, um einen tieferen Einblick zu erhalten. Dies wird dazu beitragen, Diagnose und Therapie auf objektive Art zu unterstützen.

Die Unzugänglichkeit der menschlichen Stimmlippen sowie ihre schnelle Oszillation auf kleinem Raum erfordern spezielle Mess- und Aufnahmetechniken. Eine der gängigen Lösungen dafür ist die Hochgeschwindigkeits-Videoendoskopie (HSV). HSV Aufnahmen sind oftmals leider nicht ausreichend belichtet und die häufig stattfindende Segmentierung der Aufnahmen ist in der Regel ein semi-automatischer Prozess. Diese Dissertation präsentiert Deep Learning Ansätze für beide Probleme. Damit können einerseits auch deutlich zu dunkle HSV Aufnahmen verbessert und nutzbar gemacht werden und andererseits Videos komplett vollautomatisch segmentiert werden. Die Ansätze werden an jeweils etwa 50 Videos getestet. Das primäre Ziel dieser Arbeit ist jedoch die Schätzung physikalischer Stimmparameter aus den Segmentierungen. Aufbauend auf vorangegangenen Arbeiten wird ein verbessertes numerisches Modell entwickelt, das dann eingesetzt wird, um den subglottalen Druck in 288 HSV Aufnahmen von *ex vivo* Schweinekehlköpfen zu schätzen. Im Anschluss wird ein Ansatz, der ein rekurrentes neuronales Netzwerk verwendet, auf dieselben Daten angewandt, um eine schnellere Schätzung und damit den Einsatz komplexerer Modelle zu ermöglichen. Eine neuartige Methode, um das neuronale Netz ausschließlich mit modell-generierten Daten zu trainieren, wird vorgestellt. In diesem Sinne stellt diese Arbeit Verbesserungen am gesamten Prozess zur Parameterschätzung vor, die auf den innovativen Einsatz von Deep Learning sowie Optimierungsmethoden und numerischen Verfahren setzen.

Sowohl das Verfahren zur Aufhellung also auch das zur Segmentierung der HSV Aufnahmen liefert robuste Resultate ohne spezielle Hardware, so wie es in einem klinischen Umfeld zu erwarten ist. Die Methoden können über 50 Bilder pro Sekunde auf typischer Endanwenderhardware verarbeiten. Die Genauigkeit der Segmentierung ist vergleichbar mit vorangegangenen Arbeiten, jedoch deutlich schneller. Die Aufhellungsmethode liefert statistisch signifikant bessere Ergebnisse im Vergleich mit vier *state-of-the-art* Verfahren auf drei Metriken für Bildqualität. Das verbesserte numerische Modell, das für die Parameterschätzung eingesetzt wird, ist gut geeignet, die Oszillation der Schweinestimm lippen abzubilden. Der Fehler der Fundamentalfre-

quenz liegt im Mittel bei 0.02 Hz, der mittlere Amplitudenfehler bei 0.008 cm. Die Schätzung des Drucks kann mit einem Fehler von 2.9 cmH<sub>2</sub>O oder 27.5% klar Trends in den Daten abbilden. Die individuelle Genauigkeit kann noch nicht garantiert werden. Der entwickelte Deep Learning Ansatz erlaubt eine Schätzung mit vergleichbarer Genauigkeit mit drastisch reduziertem Rechenaufwand. Er benötigt einen Bruchteil der Modellauswertungen und Rechenzeit. Die Schätzung pro Aufnahme ist dadurch 56 796 mal schneller.

Insgesamt zeigt diese Arbeit klar das Potential der Ansätze zur Parameterschätzung auf und impliziert eine klinische Anwendbarkeit. Verbliebene manuelle Schritte wurden automatisiert und in Hinsicht auf Rechenaufwand optimiert, um einen klinischen Einsatz zu ermöglichen und subjektive Einflüsse zu entfernen. Die Genauigkeit der Parameterschätzung wurde auf experimentellen Daten gezeigt. Weiterhin wurde die Notwendigkeit festgestellt, einen Weg zu finden, um komplexere numerische Modelle für die Schätzung verwenden zu können. Der entwickelte Deep Learning Ansatz ist so ein Weg und kann den Einsatz von Modellen ermöglichen, die bisher als zu rechenaufwändig betrachtet wurden. Der vorgestellte Ansatz schlägt damit eine Brücke zwischen Verfahren, die Supercomputer benötigen und denen, die im klinischen Alltag einsetzbar sind. Die Schätzung der Stimmparameter bietet somit die Chance auf eine objektivere Diagnose und Auswahl der Behandlungsmethode.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>High-speed Videoendoscopy</b>	<b>11</b>
2.1	Laryngoscopic Imaging . . . . .	11
2.2	Practical Challenges . . . . .	12
<b>3</b>	<b>Image Processing</b>	<b>15</b>
3.1	Deep Learning . . . . .	15
3.2	Low-Light Image Enhancement of High-Speed Video . . . . .	19
3.2.1	Low-Light Enhancement . . . . .	19
3.2.2	Metrics . . . . .	21
3.2.3	Compared Enhancement Methods . . . . .	23
3.2.4	Neural Network Architecture . . . . .	25
3.2.5	Training Data . . . . .	28
3.3	Results . . . . .	32
3.3.1	Setup . . . . .	32
3.3.2	Validation Data . . . . .	33
3.3.3	Test Data . . . . .	34
3.4	Discussion . . . . .	35
<b>4</b>	<b>Automatic Segmentation</b>	<b>39</b>
4.1	Segmentation of Endoscopic High-speed Video . . . . .	39
4.2	Semantic Segmentation using Deep Learning . . . . .	42
<b>5</b>	<b>Vocal Fold Models and Voice Parameter Estimation</b>	<b>51</b>
5.1	Models of the Vocal Folds . . . . .	51
5.1.1	The Improved Two-Mass-Model . . . . .	53
5.2	Estimation of Physical Voice Parameters . . . . .	58
5.2.1	Degrees of Freedom . . . . .	60
5.2.2	Optimization . . . . .	63

5.3	Estimation of Subglottal Pressure in ex vivo High-Speed Videos . . . . .	66
5.3.1	Setup . . . . .	68
5.3.2	Optimization Results . . . . .	70
5.3.3	Subglottal Pressure Results . . . . .	71
5.3.4	Discussion . . . . .	73
5.4	Shortcomings and Limitations . . . . .	77
<b>6</b>	<b>Voice Parameter Estimation with a Recurrent Neural Network</b>	<b>81</b>
6.1	Deep Learning in Inverse Problems . . . . .	81
6.2	Estimating Subglottal Pressure with an LSTM . . . . .	83
6.3	Results . . . . .	89
6.3.1	Setup . . . . .	89
6.3.2	Experimental Results . . . . .	90
6.4	Discussion . . . . .	93
6.4.1	Subglottal Pressure Estimation . . . . .	93
6.4.2	Runtime . . . . .	94
6.4.3	Parameter Estimation and Inverse Problems . . . . .	95
6.4.4	Benefit and Applicability of the Approach . . . . .	96
<b>7</b>	<b>Conclusion</b>	<b>99</b>
7.1	Project Status . . . . .	99
7.2	Project Outlook . . . . .	101
7.2.1	Technical Innovation . . . . .	102
7.2.2	Experimental Validation and Extension to Different Parameters	102
7.2.3	Clinical Use Cases . . . . .	103
7.3	Project Impact . . . . .	103
	<b>Bibliography</b>	<b>105</b>
	<b>List of Figures</b>	<b>132</b>
	<b>List of Tables</b>	<b>137</b>