Contributions to Statistical Modeling for Minimum Mean Square Error Estimation in Speech Enhancement

Von der Fakultät für Elektrotechnik, Informationstechnik, Physik
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung der Würde
eines Doktor-Ingenieurs (Dr.-Ing.)
genehmigte

Dissertation

von

Balázs Fodor
aus Esztergom, Ungarn

eingereicht am:	17. Juni 2014
mündliche Prüfung am:	24. Oktober 2014

Erstgutachter:	Prof. Dr.-Ing. Tim Fingscheidt Technische Universität Carolo-Wilhelmina zu Braunschweig
Zweitgutachter:	Prof. Dr.-Ing. Gerhard Schmidt Christian-Albrechts-Universität zu Kiel
Prüfungsvorsitzender:	Prof. Dr.-Ing. Thomas Kürner Technische Universität Carolo-Wilhelmina zu Braunschweig

2015

Mitteilungen aus dem Institut für Nachrichtentechnik der
Technischen Universität Braunschweig

Band 39

Balázs Fodor

Contributions to Statistical Modeling for Minimum Mean Square Error Estimation in Speech Enhancement

Shaker Verlag
Aachen 2015

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.: Braunschweig, Techn. Univ., Diss., 2014

Editor of this volume:

Prof. Dr.-Ing. Tim Fingscheidt
Institute for Communications Technology
Technische Universität Braunschweig
Schleinitzstrasse 22
38106 Braunschweig
Germany
e-mail: fingscheidt@ifn.ing.tu-bs.de
phone: +49-531-391-2485
fax: +49-531-391-8218

Printed in Germany.

ISBN 978-3-8440-3571-1
ISSN 1865-2484

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Preface

This dissertation was written during my research activity at the Institute for Communications Technology (in German: Institut für Nachrichtentechnik, IfN) of Technische Universität Carolo-Wilhelmina zu Braunschweig. In my opinion, a doctoral thesis is inconceivable without precious discussions with colleagues and other researchers for instance on conferences as well as without a stable familial background. It is my pleasure to thank all the people who contributed to the success of this work.

I would like to express my gratitude first to my adviser Prof. Dr.-Ing. Tim Fingscheidt. Thank you for your continuous support and competent supervision. Your suggestions, valuable feedback, and our productive discussions during my time at IfN have left a significant imprint on this dissertation.

Next, I would like to thank Prof. Dr.-Ing. Gerhard Schmidt for being a member of the examination board as well as for his interest in my research work and Prof. Dr.-Ing. Thomas Kürner for being the chair of the examination board.

As part of a fruitful research cooperation on the topic of Chapter 5, I worked together with Prof. Dr.-Ing. Timo Gerkmann. Timo, thank you for all your support and competent feedback. I always enjoyed our friendly and rewarding discussions.

I would like to thank my colleagues at IfN for their open-mindedness, helpfulness, and for creating a friendly work atmosphere. Particularly, I am grateful to Dipl.-Ing. Patrick Bauer, Marc-André Jung M.Sc., Dr.-Ing. Florian Pflug, Dipl.-Ing. Simon Receveur, Dipl.-Ing. David Scheler, Dr.-Ing. Suhadi, and Dr.-Ing. Huajun Yu for having a lot of valuable discussions which this dissertation has definitely benefited from.

Furthermore, I am thankful to Dipl.-Ing. Patrick Bauer, Samy Elshamy M.Sc., Marc-André Jung M.Sc., Dr.-Ing. Florian Pflug, Dr.-Ing. Suhadi, Peter Transfeld M.Sc., and Dr.-Ing. Huajun Yu for proofreading chapters of this thesis and providing constructive feedback regarding this dissertation.

I would like to thank Prof. István Kollár for being my adviser at Budapest University of Technology and Economics prior to my time at IfN. István, I liked your lectures and the research work with you very much. You definitely contributed the most to the decision that I wanted to keep on doing research and work towards a doctoral degree.

I am grateful to my parents for all their support and for laying the groundwork for my achievements such as this thesis. I would like to extend my deepest gratitude to my wife Kristina for her unconditional love, patience, and tireless support, especially during the entire developing process of this dissertation.

Braunschweig, December 2014

Balázs Fodor

Abstract

This thesis deals with minimum mean square error (MMSE) speech enhancement schemes in the short-time Fourier transform (STFT) domain with a focus on statistical models for speech and corresponding estimators.

MMSE speech enhancement approaches taking speech presence uncertainty (SPU) into account usually consist of a common MMSE estimator for speech and an *a posteriori* speech presence probability (SPP) estimator. It is shown that both estimators should be based on the same statistical speech model, as they are in the same estimation framework and assume the same *a priori* knowledge. In order to give a synopsis of consistent MMSE estimation under SPU, typical common MMSE estimators and *a posteriori* SPP estimators are recapitulated. Furthermore, a new specific *a posteriori* SPP estimator is derived based on a novel statistical model for speech. Then, a synopsis of approaches to consistent MMSE estimation under SPU is given.

In the context of statistical modeling, we enhance a modern *a posteriori* SPP estimation approach based on fixed parameters. More precisely, the conservative speech model of this reference approach is replaced by an improved one. Then, a new *a posteriori* SPP estimator is derived and its fixed parameters are trained. The resulting proposed approach unifies the advantages of fixed parameters and a novel statistical speech model.

Although both speech enhancement and error concealment deal with distorted (speech) signals, there has not yet been an attempt to relate both fields to each other. However, since there are many commonalities between these disciplines, many interesting links between them are discussed based on recursive MMSE estimation. Furthermore, besides these commonalities, also interesting differences are analyzed and a general advantage of error concealment is identified. Based on this finding, research perspectives for the field of speech enhancement are sketched, inspired by error concealment.

This thesis provides a new statistical framework for *recursive* MMSE speech enhancement. This advantageously allows for applying the improved statistical models from classical, non-recursive speech enhancement to the recursive case. As a specific enhancement scheme, we extend recursive MMSE estimation by taking SPU into account.

Finally, a new reference-free signal-to-noise ratio (SNR) measurement approach is proposed in this thesis. This approach aims at estimating the SNR of a speech signal distorted by car noise as close as possible to reference-based measurement approach according to ITU-T Recommendation P.56, but in a reference-free fashion. The proposed approach achieves small estimation errors and shows high correlation with the ITU-T P.56 measurement within a typical SNR range. Furthermore, it provides relaxed computational complexity and can be applied to narrowband and wideband signals. Within ITU-T Study Group 12, the Focus Group on Car Communication (FG CarCOM) has decided to adopt the new reference-free SNR measurement approach for the draft of a recommendation proposal.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit Störgeräuschreduktion (engl. speech enhancement) für Sprachsignale mittels frequenzbereichsbasierter MMSE-Schätzer (minimum mean square error, engl. für kleinster mittlerer quadratischer Fehler). Hierbei liegt ein besonderer Fokus auf den statistischen Sprachmodellen und den resultierenden Schätzregeln.

Spezifische MMSE-Verfahren, die die Unsicherheit der Sprachpräsenz (engl. speech presence uncertainty, SPU) berücksichtigen, bestehen aus einem allgemeinen MMSE-Sprachschätzer und einem Schätzer der Wahrscheinlichkeit von Sprachanwesenheit (engl. speech presence probability, SPP). Es wird gezeigt, dass beide Schätzer auf demselben statistischen Sprachmodell basieren, zudem werden üblicherweise verwendete allgemeine MMSE- und SPP-Schätzer rekapituliert. Darüber hinaus wird ein neuer SPP-Schätzer hergeleitet, der auf einem verbesserten statistischen Sprachmodell basiert. Danach wird ein Überblick über konsistente MMSE-Schätzverfahren mit SPU gegeben.

Im Kontext der statistischen Sprachmodellierung wird auch ein spezifisches, auf festen Parametern basierendes SPU-Verfahren weiterentwickelt. Das konservative Sprachmodell dieses SPU-Verfahrens wird durch ein verbessertes ersetzt und ein weiterer neuer SPP-Schätzer wird hergeleitet. Anschließend werden die festen Parameter der resultierenden Schätzregel trainiert. Dieses weiterentwickelte Verfahren vereint die Vorteile der festen Parameter und des verbesserten Sprachmodells.

Obwohl sich Störgeräuschreduktion und Fehlerverdeckung (engl. error concealment) mit der Aufgabe beschäftigen, gestörte (Sprach-)Signale zu verbessern, werden diese Verfahren typischerweise nicht in Beziehung zueinander gesehen. Da es jedoch viele Gemeinsamkeiten zwischen beiden Disziplinen gibt, werden bisher unbekannte Bezüge diskutiert. Darüber hinaus werden auch Unterschiede behandelt und ein grundsätzlicher Vorteil von Fehlerverdeckungsverfahren identifiziert. Motiviert durch diese Erkenntnis werden Forschungsperspektiven für das Themenfeld der Störgeräuschreduktion aufgezeigt.

Ferner wird eine neue, statistische Darstellung von rekursiver MMSE-Schätzung der Sprache präsentiert. Diese ermöglich es, die modernen statistischen Modelle der klassischen, nicht-rekursiven Verfahren auf den rekursiven Fall anzuwenden. In diesem Kontext wird die rekursive MMSE-Schätzung mit einem SPU-Verfahren erweitert.

Schließlich wird ein neues, referenzfreies Messverfahren für das Signal-Rausch-Verhältnis (SNR) vorgestellt. Das Ziel des Verfahrens ist, das SNR eines von Fahrzeuggeräuschen gestörten Sprachsignals referenzfrei zu schätzen. Das Schätzergebnis soll so nah wie möglich am referenzbasierten Messverfahren nach ITU-T Recommendation P.56 liegen. Das neue Verfahren zeichnet sich durch kleine Messfehler und eine hohe Korrelation der Messwerte zum Referenzverfahren aus und kann mit Schmalband- sowie Breitbandsignalen verwendet werden. Die *Focus Group on Car Communication* (FG CarCOM) der ITU-T Study Group 12 hat beschlossen, das Verfahren in den Entwurf eines zukünftigen Standards aufzunehmen.

Contents