DISSERTATION

Ziyue Zhao

Contributions to Neural Network-Based Speech Processing: Nonlinear Speech Prediction, Decoder Postprocessing, and Perceptual Loss Functions

Technische Universität Braunschweig

Institut für Nachrichtentechnik

Contributions to Neural Network-Based Speech Processing: Nonlinear Speech Prediction, Decoder Postprocessing, and Perceptual Loss Functions

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

zur Erlangung des Grades eines Doktors
der Ingenieurwissenschaften (Dr.-Ing.)
genehmigte

Dissertation

von

Ziyue Zhao, M.Eng.
aus Xi'an, Shaanxi, China

1. Referent:	Prof. Dr.-Ing. Tim Fingscheidt Technische Universität Braunschweig
2. Referent:	Prof. Dr.-Ing. Gerald Schuller Technische Universität Ilmenau
Vorsitzender:	Prof. Dr.-Ing. Eduard A. Jorswieck Technische Universität Braunschweig

Eingereicht am:	27. 09. 2021
Mündliche Prüfung am:	12. 05. 2022

Druckjahr: 2022

Dissertation an der Technischen Universität Braunschweig,
Fakultät für Elektrotechnik, Informationstechnik, Physik

Mitteilungen aus dem Institut für Nachrichtentechnik der
Technischen Universität Braunschweig

Band 70

Ziyue Zhao

Contributions to Neural Network-Based Speech Processing: Nonlinear Speech Prediction, Decoder Postprocessing, and Perceptual Loss Functions

Shaker Verlag
Düren 2022

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., 2022

Editor of this volume:

Prof. Dr.-Ing. Tim Fingscheidt
Institute of Communications Technology
Technische Universität Braunschweig
Schleinitzstraße 22
38106 Braunschweig
Germany
e-mail: fingscheidt@ifn.ing.tu-bs.de
phone: +49 (0)531 391-2485
fax: +49 (0)531 391-8218

Printed in Germany.

ISBN 978-3-8440-8779-6
ISSN 1865-2484

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren
Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9
Internet: www.shaker.de • e-mail: info@shaker.de

Acknowledgments

Most parts of this thesis were written during my doctorate study in the Institute for Communications Technology (in German: Institut für Nachrichtentechnik, IfN), Technische Universität Braunschweig. It was then finalized during my stay in the Netherlands. In my opinion, a thesis is a complete presentation of the research work from one's doctorate study, and I can never finish it without the support and help from many people. I am pleased to express thanks to all the people who supported me during my doctorate study.

First of all, I would like to heartily appreciate my doctorate supervisor, Prof. Dr.-Ing. Tim Fingscheidt. His enlightening guidance and continuous support are essential to my entire research work and also my thesis. I always enjoy the fruitful discussions between us and appreciate his constructive feedback every time after the discussion.

Then, I would like to thank Prof. Dr.-Ing. Gerald Schuller for being the co-examiner of this thesis as well as for his interest in this research work. Also, I would like to thank Prof. Dr.-Ing. Eduard A. Jorswieck for being the chair of the examination board.

Next, I would like to thank my dear colleagues in the institute. Although I began my doctorate study in a brand new environment compared to my previous experiences, I always enjoyed the friendly work atmosphere here. I can still remember many delightful and warm moments with our colleagues, making me feel the institute was just like my hometown in Germany. Particularly, I would like to thank Andreas Bär M.Sc., Jan Franzen M.Sc., Jonas Löhdefink M.Sc., and Ziyi Xu M.Sc. for proofreading the chapters of this thesis. I would also like to thank Dr.-Ing. Johannes Abel, Jan Baumann M.Sc., Jasmin Breitenstein M.Sc., Dr.-Ing. Samy Elshamy, Dr.-Ing. Sai Han, Marvin Klingner M.Sc., Timo Lohrenz M.Sc., Patrick Meyer M.Sc., Dr.-Ing. Simon Receveur, Renzheng Shi M.Sc., Maximilian Strake M.Sc., and Jan-Aike Termöhlen for all the interesting discussions and supportive help during my time in the institute. Besides, I would like to thank Ms. Eike-Asslo Erichsen-Rua, who gave me much help in the institute. Mr. Rudolf Görke and Ms. Ingrid Kretzschmann are my friends at the institute, and I appreciate the happy moments we shared.

Finally, I am grateful to my parents, Hongjun Zhao and Hui Wang, for their continuous support and understanding during my stay in Germany. Last but not least, I would like to give my deepest gratitude to my wife Xiao Tai for her unreserved love, support, and patience during my entire doctorate study.

Eindhoven, September 2022
Ziyue Zhao

Abstract

Speech processing technologies are omnipresent in our daily communication products and services. Neural networks, as powerful data-driven models, have shown promising performance in various research fields, including speech processing. This thesis focuses on neural network-based speech processing, and it can be divided into three parts as follows.

In the field of speech prediction, a nonlinear speech predictor using the echo state network (ESN) is proposed as a novel adaptive prediction approach. The ESN is a special type of recurrent neural network (RNN) requiring no training beforehand, instead it adaptively updates the weights in its output layer. This proposed nonlinear predictor shows better prediction performance than all baseline prediction methods in the simulations, including a predictor based on a long short-term memory (LSTM) structure. Second, the field of neural network-based speech enhancement puts focus on loss functions. A novel perceptual weighting filter (PWF) loss function motivated by the weighting filter from code-excited linear prediction (CELP) speech coding is proposed. Through this proposed design, the masking property of the human ear is exploited in neural network-based speech enhancement. As the loss function is only required during the training stage, this proposed loss function can be advantageously applied to an existing neural network-based speech enhancement implementation, without altering the network structure. In the experimental part, the proposed loss function is applied to estimate spectral masks either for amplitudes or for complex values. A fully connected neural network (FCNN) and a convolutional neural network (CNN) are both used to evaluate the proposed loss functions, and the simulation results show their superior performance compared to baselines, especially in terms of speech quality and noise attenuation. Finally, neural network-based postprocessing for the enhancement of coded speech is studied. CNN-based postprocessors are proposed either to directly enhance the raw waveform in an end-to-end fashion, or to enhance the cepstral domain features using analysis synthesis. Furthermore, an advanced network structure, the fully convolutional recurrent network (FCRN), is utilized to enhance coded speech in the frequency domain, with the PWF loss function advantageously applied. The proposed postprocessors are comprehensively evaluated for various narrowband and wideband speech codecs under the conditions of clean, codec tandeming, error-prone transmission (i.e., packet loss), and noisy background. The experimental results confirm the effectiveness of the proposed postprocessors with improved speech quality.

Zusammenfassung

Sprachverarbeitung ist in unseren täglichen Kommunikationsmitteln und -diensten allgegenwärtig. Neuronale Netze haben als leistungsfähige datengesteuerte Modelle vielversprechende Resultate in verschiedenen Forschungsbereichen gezeigt, einschließlich der Sprachverarbeitung. Diese Arbeit beschäftigt sich mit Themen der neuronalen netzwerkbasierten Sprachverarbeitung und kann wie folgt in drei Teile gegliedert werden.

Im Themengebiet der Sprachprädiktion wird ein nichtlinearer Sprachprädiktor unter Verwendung des Echo State Network (ESN) als neuartiger adaptiver Prädiktionsansatz vorgeschlagen. Das ESN ist ein spezieller Typ eines rekurrenten neuronalen Netzes (RNN), das kein vorheriges Training erfordert und die Gewichte in seiner Ausgangsschicht adaptiv aktualisiert. Dieser nichtlineare Prädiktor zeigt in Simulationen eine bessere Vorhersageleistung als andere Prädiktionsmethoden im Vergleich, einschließlich eines Prädiktors, der auf einem langen Kurzzeitgedächtnis (LSTM) basiert. Im zweiten Themengebiet der neuronalen netzbasierten Sprachverbesserung stehen vor allem Kostenfunktionen im Vordergrund. Es wird eine neuartige Kostenfunktion des Perzeptivengewichtungsfilters (PWF) vorgeschlagen, die durch das Gewichtungsfilter aus der code-angeregten linearen Vorhersage (CELP) motiviert ist. Durch dieses Verfahren wird die menschliche Wahrnehmung bei der neuronalen netzwerkbasierten Spachverbesserung ausgenutzt. Da die Kostenfunktion nur während der Trainingsphase benötigt wird, kann die vorgeschlagene Kostenfunktion vorteilhaft auf ein bestehendes, netzwerkbasiertes Sprachverbesserungssystem angewendet werden, ohne die Netzwerkstruktur zu verändern. Im experimentellen Teil wird die vorgeschlagene Kostenfunktion angewendet, um spektrale Masken entweder für die Amplituden oder für die komplexen Werte zu schätzen. Ein vollständig verbundenes neuronales Netzwerk (FCNN) und ein Faltungsnetzwerk (CNN) werden verwendet, um die vorgeschlagenen Kostenfunktionen zu evaluieren. Die Simulationsergebnisse zeigen verbesserte Resultate im Vergleich zu Referenzkostenfunktionen, insbesondere in Bezug auf die Sprachqualität und die Rauschunterdrückung. Schließlich wird eine auf neuronalen Netzwerken basierende Nachbearbeitung für die Verbesserung codierter Sprache untersucht. CNN-basierte Postprozessoren werden vorgeschlagen, um entweder direkt die Rohsignalform nach einer Art End-to-End-Verfahren zu verbessern oder die cepstralen Merkmale mittels Analyse Synthese zu verbessern. Darüber hinaus wird eine moderne Netzwerkstruktur, das Fully Convolutional Recurrent Network (FCRN), verwendet, um codierte Sprache im Frequenzbereich zu verbessern, wobei die PWF-Kostenfunktion vorteilhaft eingesetzt wird. Die vorgeschlagenen Postprozessoren werden umfassend für verschiedene Schmalband- und Breitband-Sprachcodecs unter den Bedingungen von ungestörter Sprache, Tandem, fehleranfälliger Übertragung (d.h. Paketverlust), und Hintergrundrauschen evaluiert. Die experimentellen Ergebnisse bestätigen die Wirksamkeit der vorgeschlagenen Postprozessoren mit verbesserter Sprachqualität.

Contents