

Arbeiten über Digitale Signalverarbeitung

Band 38

Christoph Robert Norrenbrock

**Instrumental Quality Estimation
for Synthesized Speech Signals**

Shaker Verlag
Aachen 2014

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.: Kiel, Univ., Diss., 2014

Copyright Shaker Verlag 2014

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 978-3-8440-2669-6

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

Titel: „Instrumental Quality Estimation for Synthesized Speech Signals“
Autor: Christoph Robert Norrenbrock

Zusammenfassung

Im Bereich der digitalen Sprachsignalverarbeitung stellt die Qualitätsbewertung synthetisierter „Text-zu-Sprache“ (TTS)-Signale eine besondere Herausforderung dar. Ein wesentlicher Grund dafür liegt in dem fehlenden Verständnis für den Zusammenhang zwischen den physikalischen Eigenschaften synthetischer Sprache und der resultierenden Qualität. Da TTS-Systeme allerdings verstärkt in praktischen Anwendungen, wie z.B. Sprachdialogsystemen, eingesetzt werden bzw. eingesetzt werden sollen, sind verlässliche und effiziente Mittel der Qualitätsbewertung von TTS-Signalen zur professionellen Auslegung solcher Anwendungen unverzichtbar geworden.

Die vorliegende Arbeit fasst Forschungsbeiträge zur Objektivierung der Wahrnehmung synthetisierter Sprache zusammen. Auf Basis von auditorisch ermittelten Qualitätsdimensionen wird gezeigt, wie sich subjektive Qualitätsurteile von TTS-Signalen instrumentell schätzen lassen. Dabei hat sich herausgestellt, dass die perzeptiven Besonderheiten synthetischer Sprache, wie zum Beispiel kognitive Nichtlinearitäten, ganzheitliche Ansätze zur Herleitung der Schätzmodelle erfordern. Zu diesem Zweck wird untersucht, mit Hilfe welcher Sprachsignaleigenschaften sich die Sprachqualität am besten erfassen lässt und in welcher Form die abgeleiteten Kennwerte vorteilhaft in ein nicht-intrusives Schätzmodell eingebunden werden können.

Hierzu wird der Ansatz der Regulären Perzeption vorgestellt. Dabei wird der Messbereich der qualitätsrelevanten Signaleigenschaften so kodiert, dass die zur Schätzung verwendeten Messparameter bzw. Qualitätselemente einen positiven linearen Zusammenhang zur subjektiven Qualität aufweisen. Die mittels Perzeptiver Regularisierung hergeleiteten Messparameter stellen eine Brücke zwischen physikalischen und perzeptiven Eigenschaften her, indem der objektive Erwartungshorizont einer qualitätsrelevanten Signaleigenschaft explizit angegeben werden kann. Der Ansatz wird mit regularisierten Regressionsmodellen unterschiedlicher Komplexität kombiniert und unter Einsatz von Kreuzvalidierungsmethoden bewertet. Insgesamt stellen sich die günstigsten Ergebnisse bei Verwendung eines Support-Vector-Regression-Modells ein, wobei eine Kombination aus prosodischen sowie MFCC-basierten Messparametern empfohlen wird. Die Übereinstimmung der geschätzten und tatsächlichen Qualitätsurteile liegt bei einer Korrelation von bis zu 0.96 und einem Quadratwurzelfehler von bis zu 0.20 auf einer Skala von 1 bis 5. Alternativ kann bei Verwendung des erwähnten Ansatzes ein vereinfachtes Modell eingesetzt werden, welches auf dem Mittelwert der Qualitätselemente aufbaut.

Abstract

In the field of digital speech-signal processing, the quality assessment of synthesized “text-to-speech” (TTS) signals is recognized as a challenging subject. One main reason is the lacking knowledge of the correspondence between the physical characteristics of synthesized speech and the resulting quality. However, since TTS systems are increasingly used or to be used in practical applications, e.g., in speech-dialog systems, reliable and efficient means of quality assessment have become indispensable for a professional lay-out design of these applications.

The present thesis summarizes research contributions on objectifying synthesized-speech perception. Based on auditorily derived quality dimensions, it is shown how subjective quality ratings of synthetic speech can be estimated instrumentally. It turns out that the perceptual peculiarities of synthesized speech, e.g., cognitive nonlinearities, require holistic approaches for deriving the prediction models. Therefore, it is investigated which speech-signal properties best capture the speech quality, and in which way the derived parameters can be beneficially integrated into a non-intrusive prediction model.

To this end, the approach of regular perception is introduced. The measurement range of quality-relevant signal properties is coded such that the measurement parameters or quality elements exhibit a positive correlation with the subjective quality. The measurement parameters, which are derived through perceptual regularization, set up a bridge between physical and perceptual properties in that the objective expectation horizon of a quality-relevant property is explicitly given. The approach is combined with regularized regression-models of different complexity and assessed by using cross-validation methods. Overall, the best results are obtained with a support-vector-regression model, whereby a combination of prosodic and MFCC-based measurement parameters is recommended. The accordance between predicted and true quality ratings is given by a correlation of up to 0.96 and a root-mean-square error of up to 0.20 on a scale from 1 to 5. Alternatively, a simplified model can be used which is based on the mean of the quality elements.