KEYNOTE 2:

INSTRUMENTAL EVALUATION OF SYNTHESIZED SPEECH QUALITY

Sebastian Möller and Florian Hinterleitner

Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin
sebastian.moeller@telekom.de

Extended Abstract

Whereas methods for synthesizing speech signals from written text have made considerable advances in the past decade, methods for assessing the performance of speech synthesizers and evaluating their fitness for particular applications are still cumbersome. A reason is that quality assessment and evaluation require perception and judgment processes to take place, which ultimately happen only inside a human assessor. Thus, auditory test methods are currently the only way to validly and reliably assess and evaluate synthesized speech quality. Still, advances in speech transmission quality prediction show up ways to model perception and judgment processes to a limited extent, and thus to predict speech transmission quality on the basis of instrumental measurements only. We are thus interested in the question whether such an approach is also feasible with synthesized speech signals and their corresponding degradations originating from the synthesis process.

In this talk, we will discuss several such ways and analyze their (expected) performance on different synthesized speech databases. First, we will briefly review approaches which rely on the availability of a natural reference speech signal, and compare synthesized speech signals to such natural references [1]. This approach is limited by the (non-) availability of natural speech data usually required from the same speaker the synthesis inventory has been built from. Second, we will address approaches which rely on a model of natural speech, and derive quality predictions on the basis of the similarity of the synthesized speech signal to this model [2, 3, 4]. Third, we will review approaches extracting parameters from the synthesized speech signals which coincide with particular types of degradations [5]. Such approaches have been successful with transmitted speech signals, but the parameters heavily depend of the speech databases and speaker gender. Finally, we will show that improvements can be reached by combining different types of approaches [6]. We will justify our claims on the basis of empirical data from typical German synthesizers as well as from the Blizzard Challenges organized as a controlled comparative assessment of synthesized speech. We will address the deficiencies of the current approaches by showing that their performance heavily depends on the used databases, and will identify research which has to be carried out jointly by the synthesis and evaluation communities to overcome the current limitations.

Literatur

[1] Cernak, M., Rusko, M.: An Evaluation of Synthetic Speech Using the PESQ Measure. In: Proc. European Congress on Acoustics, Budapest, 2005, pp. 2725-2728.

[2] Mariniak, A.: A Global Framework for the Assessment of Synthetic Speech Without Subjects. In: Proc. 3rd Europ. Conf. on Speech Process. and Technology (Eurospeech'93), Berlin, 1993, pp. 1683-1686.

[3] Möller, S., Kim, D.-S., Malfait, L.: Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models. Acta Acustica united with Acustica 94, 2008, pp. 21-31.

[4] Falk, T. H., Möller, S.: Towards Signal-Based Instrumental Quality Diagnosis for Text-

to-Speech Systems. IEEE Signal Processing Letters 15, 2008, pp. 781-784.

[5] Falk, T. H., Möller, S., Karaiskos, V., King, S.: Improving Instrumental Quality Prediction Performance for the Blizzard Challenge. In: Proc. Blizzard Challenge Workshop, Brisbane, 6 pages.

[6] Möller, S., Hinterleitner, F., Falk, T. H., Polzehl, T.: Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems. Accepted for: Proc. 11th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech 2010), 26-30 Sept., Makuhari, 2010.

On the presenter

Sebastian Möller was born in 1968 and studied electrical engineering at the universities of Bochum (Germany), Orléans (France) and Bologna (Italy). From 1994 to 2005, he held the position of a scientific researcher at the Institute of Communication Acoustics (IKA), Ruhr-University Bochum, and worked on speech signal processing, speech technology, communication acoustics, as well as on speech communication quality aspects. Since June 2005, he works at Deutsche Telekom Laboratories, TU Berlin. He was appointed Professor at TU Berlin for the subject "Quality and Usability" in April 2007, and heads the "Quality and Usability Lab" at Deutsche Telekom Laboratories.

He received a Doctor-of-Engineering degree at Ruhr-University Bochum in 1999 for his work on the assessment and prediction of speech quality in telecommunications. In 2000, he was a guest scientist at the Institut dalle Molle d'Intélligence Artificielle Perceptive (IDIAP) in Martigny (Switzerland) where he worked on the quality of speech recognition systems. He gained the qualification needed to be a professor (venia legendi) at the Faculty of Electrical Engineering and Information Technology at Ruhr-University Bochum in 2004, with a book on the quality of telephone-based spoken dialogue systems. In September 2008, we worked as a Visiting Fellow at MARCS Auditory Laboratories, University of Western Sydney (Australia) on the evaluation of avatars.

Sebastian Möller was awarded the GEERS prize in 1998 for his interdisciplinary work on the analysis of infant cries for early hearing-impairment detection, the ITG prize of the German Association for Electrical, Electronic & Information Technologies (VDE) in 2001, the Lothar-Cremer prize of the German Acoustical Association (DEGA) in 2003, a Heisenberg fellowship of the German Research Foundation (DFG) in 2005, and the Johann Philipp Reis prize in 2009. Since 1997, he has taken part in the standardisation activities of the International Telecommunication Union (ITU-T) on transmission performance of telephone networks and terminals. He is currently acting as a Rapporteur for question Q.8/12.