VDE ITG **Voice Message** **OKT 2026**

Newsletter des ITG-Fachausschusses Sprachkommunikation (AT 3), 31. Ausgabe

Editorial

Liebe Leserinnen und Leser,

mit dem Start des Herbstes liegen die wichtigsten Konferenzen des Jahres nun hinter uns, und ich hoffe, Sie konnten einige davon besuchen, sind mit spannenden Inspirationen zurückgekehrt, und freuen sich trotz des Wetters auf die etwas ungemütlichere Jahreszeit, um einige der Ideen umzusetzen!

Hier im Newsletter gibt es weitere Ideen, und gute und interessante Neuigkeiten aus unserer Runde - eine ganze Reihe von Journal-Artikeln, Promotionen, eine neue ITU-Recommendation zur Bewertung von Sprachqualität und einen Preis. Die neuen Tools und Ressourcen in unserem Themenbereich möchte ich besonders erwähnen – schauen Sie dazu gern unten in die Kategorie Open Science.

Ganz unten finden Sie wie immer die Liste der wichtigsten Konferenzen in unserem Gebiet, und auch eine Einladung zu einer strukturierten Sitzung, die sich besonders der Frage der Qualität KI-generierter und multimodaler Sprache widmet – mit der Deadline für kurze Abstracts bereits am 1.11.

Mit freundlichen Grüßen
Dorothea Kolossa

Sie wünschen ein Abo oder haben einen Beitrag? Sehr gerne! Bitte melden Sie sich einfach per Email *unter Hinweis darauf, ob Sie nur Abonnent, oder Abonnent und auch möglicher Autor sein möchten! Wir weisen aus datenschutzrechtlichen Gründen darauf hin, dass Sie unter gleicher Emailadresse jederzeit Auskunft über Ihre gespeicherten Daten erfragen können, sowie die Löschung Ihrer Kontaktdaten erwirken können.*

Latest News, Awards

● ITU-T Recommendation P.566: Machine-Learning-Based Multidimensional Speech Quality Analysis: In July 2026, ITU-T Recommendation P.566, Single-ended machine-learning-based models for multi-dimensional speech quality analysis, was approved and is now in force. The Recommendation provides a standardized approach for estimating speech quality directly from a degraded speech signal, without requiring a clean reference signal. In addition to overall speech quality, the model predicts four perceptual quality dimensions: noisiness, discontinuity, coloration, and loudness. The reference implementation, SQ-AST, is based on the Audio Spectrogram Transformer and was developed at the Quality and Usability Lab at TU Berlin. It was trained and evaluated across a large collection of speech-quality datasets and successfully evaluated on previously unseen conditions during the ITU-T standardization process.

● Matias Pizarro erhielt für sein Long-Paper „Lightweight Detection and Model Attribution of Synthetic Speech via Residual Statistical Fingerprints“ gemeinsam mit Mike Laszkiewicz, Dorothea Kolossa und Asja Fischer den Best Student Paper Award der Interspeech in Sydney.



Da KI-generierte Sprache immer realistischer wird, reicht es nicht mehr aus, Deepfakes nur zu erkennen – wir müssen auch identifizieren, welches Modell sie erzeugt hat. Die Arbeit stellt einen effizienten, ressourcen-schonenden Ansatz vor, der auf residualen statistischen Fingerabdrücken basiert. Damit lässt sich nicht nur zwischen echter und synthetischer Sprache unterscheiden, sondern auch ein gefälschtes Audiosignal dem jeweiligen Erzeugermodell zuordnen – über verschiedene Synthesesysteme und Sprachen hinweg.
Dieser Ansatz kann dazu beitragen, realen Bedrohungen wie Identitätsdiebstahl per Stimme, Finanzbetrug oder biometrischen Täuschungs-versuchen zu begegnen. Die Möglichkeit, nicht nur zu erkennen, ob eine Stimme gefälscht ist, sondern auch, welches Modell sie generiert hat, liefert dabei wertvolle Erkenntnisse.

Dissertationen

● Ulrik Kowalk: Sound Source Localization Exploiting Signal and Geometry Conditioning Mechanisms, Universität Oldenburg (Betreuer: Prof. Joerg Bitzer, Prof. Simon Doclo). This thesis investigates DNN-based direction-of-arrival (DOA) estimation, focusing on desired-source localization in the presence of interfering

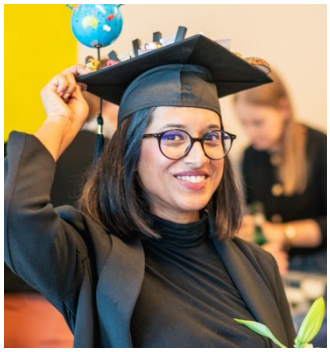


sources and generalization across microphone array geometries. First, a signal-conditioned approach is proposed that exploits an auxiliary microphone near the desired source to emphasize source-specific information in GCC-PHAT features. Second, a geometry-conditioned DNN is developed that incorporates microphone coordinates into the network using feature-wise linear modulation, allowing DOA estimation for varying array geometries without retraining. Experimental results demonstrate improved localization accuracy and robustness compared to conventional approaches, including SRP-PHAT and MUSIC.

● Mattes Ohlenbusch: DNN-based own voice reconstruction for hearables with an in-ear microphone, Universität Oldenburg (Betreuer: Prof. Simon Doclo). This thesis investigates DNN-based own voice reconstruction for hearables, combining noisy outer microphone signals with noise-robust but bandlimited in-ear microphone signals. First, a phoneme-dependent model of the time-varying own voice transfer characteristics is proposed, enabling realistic data augmentation with only limited recorded training data. Second, low-complexity DNN-based reconstruction approaches are developed to achieve high speech quality under practical computational constraints. Third, personalization approaches are investigated to further improve reconstruction quality for individual talkers. Experimental results, including subjective listening tests, demonstrate substantial improvements in speech quality, particularly in challenging noisy environments.



● Wafaa Wardah: "Transformer-Based Modeling for Speech Quality Prediction in Telecommunications", Technische Universität Berlin (Supervisor: Prof. Dr.-Ing. Sebastian Möller). The dissertation was successfully defended on 31 August 2026. It investigates transformer-based foundation models for single-ended speech quality prediction, comparing AST, wav2vec 2.0, HuBERT, WavLM, and Whisper for predicting overall quality and the perceptual dimensions noisiness, discontinuity, coloration, and loudness. The work examines single- versus multi-task learning, cross-dataset generalization, robustness to unseen degradation conditions, and transfer to synthetic speech. Among the investigated models, AST showed particularly strong transferability and was developed into SQ-AST, whose generalization was demonstrated in an independent ITU-T blind evaluation. The work ultimately contributed to ITU-T Recommendation P.566 for single-ended machine-learning-based multidimensional speech quality analysis.



Journalartikel und Special Issues

● A. Nagathil, B. Lentz, G. Mojica, I. Bruce, C. Völter, R. Martin (2026). A computationally efficient music processing algorithm for cochlear implants based on an auditory adaptation model. Listening to music is a major challenge for most cochlear implant (CI) users because of device limitations and physiological constraints that degrade musical cues and lead to increased listening effort. Music pre-processing can enhance appreciation, but existing algorithms are computationally too demanding for real-time implementation in current CIs. This work proposes a lightweight algorithm based on an auditory adaptation model that attenuates slowly varying components while preserving salient elements such as vocals and drums. In listening experiments with ten CI users, the method was significantly preferred over unprocessed music and was on par with a more computationally demanding reference approach.

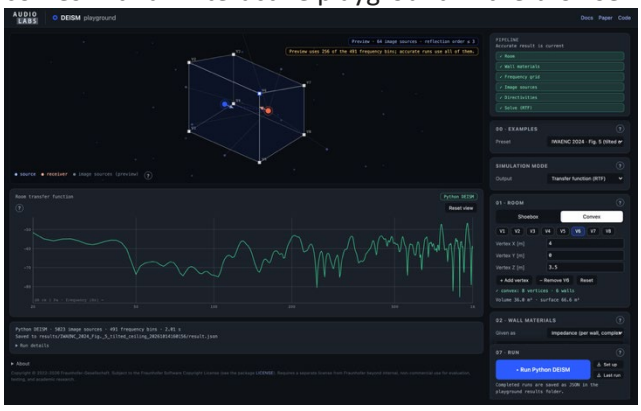
● K. Brümann, S. Doclo (2026), Multi-Source Position and Direction-of-Arrival Estimation Based on Euclidean Distance Matrices. Conventional steered-response power (SRP) methods for source localization require computationally expensive multidimensional continuous optimization, particularly when high localization accuracy is desired. In this paper, we propose novel methods for multi-source position and DOA estimation based on Euclidean distance matrices and their associated Gram matrices, using estimated time differences of arrival between microphones. The proposed position estimation method reduces the optimization to a single continuous variable per source, while the DOA estimation method entirely avoids continuous optimization. Experiments with two sources in noisy and reverberant environments show that the proposed methods outperform SRP-

based approaches in terms of localization accuracy, while also achieving lower run times.

● P. Didier, T. van Waterschoot, S. Doclo, J. Bitzer, P. Behmandpoor, H. Gode, M. Moonen (2026), Distributed Multichannel Wiener Filtering for Wireless Acoustic Sensor Networks. Existing distributed approaches for speech enhancement in acoustic sensor networks, such as the distributed adaptive node-specific signal estimation (DANSE) algorithm, achieve centralized multichannel Wiener filter (MWF) performance through iterative updates. In this paper, we propose a novel distributed multichannel Wiener filter (dMWF) that achieves optimal performance without iterations. The key principle is the exchange of low-dimensional estimates of source contributions jointly observed by pairs of nodes. The dMWF achieves centralized MWF performance using only a single set of estimated second-order statistics, without requiring prior knowledge of which sources are observed by each node. Simulations show that dMWF outperforms DANSE after short operation times, highlighting the benefits of its non-iterative design.

Open Science

● Have you ever wanted to see how a room response changes when you move a loudspeaker, tilt the ceiling, or swap a directivity interactively? The **Diffraction-Enhanced Image Source Method (DEISM)** developed at the International Audio Laboratories Erlangen now comes with an interactive playground in the browser.



After `pip install deism`, run `deism-playground` to see a 3D view of the room next to the transfer function and impulse response. Sources and receivers can be moved, vertices of convex rooms added and dragged, and wall materials and sampled loudspeaker directivities changed, with a live preview after every edit. One click runs the full accurate simulation in Python, and the current setup can be exported as an editable Python script. Presets reproduce figures from our **JASA** and **IWAENC** papers, and further examples cover subtle acoustic path-length variations induced by local and global temperature fluctuations within the room, as described in a recent **JASA Express Letters** paper by Jeremy Lawrence et al. A faster image-source search speeds up the full runs as well: the IWAENC room at reflection order 15 now takes about 5 s rather than 85 s.

● GADVOX: An Open Research Resource for Voice-Based Mental Health Research: The German Anxiety and Depression Voice Examination Dataset (GADVOX) was presented at Interspeech 2026. GADVOX is the first openly accessible German-language speech dataset designed for automated detection and severity estimation of depression and anxiety. It comprises data from 1,004 participants, combining responses to ten structured free-speech prompts with validated PHQ-9 and GAD-7 scores for depression and anxiety severity. This enables regression and multi-target modelling rather than limiting research to binary classification. The dataset underwent a multi-stage quality-assurance process, and its metadata additionally include speech-quality estimates obtained using ITU-T Rec. P.566. In line with our Open Science efforts, the dataset and its metadata will be made available to researchers through a controlled access procedure due to the sensitive nature of the data. A dedicated platform has been established at http://gadvox.de, where researchers can already register their interest. Currently, an approval process is still ongoing. If you are interested in using GADVOX for non-commercial scientific research, you can register your interest with the platform, and we will inform you by email once the dataset is released.

Tagungen und Special Sessions

● Call for Contributions: Structured Session at DAGA 2027. A structured session entitled "Quality Assessment of AI-Generated and AI-Enhanced Audiovisual Communication" will be organized at DAGA 2027, taking place in Darmstadt from 8-11 March 2027. The session is organized by Wafaa Wardah, Robert P. Spang, and Sebastian Möller and will bring together research on subjective and instrumental assessment of AI-generated and AI-enhanced speech in audiovisual communication.

Topics include perceptual quality dimensions, audiovisual interactions and coherence, speech and audio enhancement, generative speech, and machine-learning-based quality prediction. The session will be held in English, with approximately 5-8 invited contributions. Interested researchers are warmly invited to submit an abstract by 1 November 2026.

● Im Folgenden sehen Sie wieder die Tagungen in unserem Themenfeld im Überblick, mit vielen Möglichkeiten, sich hier in Deutschland zu vernetzen, und mit wichtigen Tagungen in Europa, Nord- und Südamerika:

SLT	13.-16.12.2026, Palermo, Italien [Keine Anmeldung von Beiträgen mehr]
DAGA	08.-11.03.2027, Darmstadt, Einreichungen bis 01.11.26
ICASSP	16. - 21.05.2027, Toronto, Kanada [Keine Anmeldung von Beiträgen mehr]
Interspeech	29.08. - 02.09.2027, São Paulo, Brasilien Einreichungen bis 09.02.27
EUSIPCO	30.08. - 03.09.2027, Darmstadt, Einreichungen bis 06.02.27
ITG Speech Communication	15.09. - 17.09.2027, Hamburg, Einreichungen bis 09.06.27
WASPAA	19.09. - 22.09.2027, Levi, Finnland Einreichungen bis 21.04.27