VDE ITG Voice Message JUN 2026
Newsletter des ITG-Fachausschusses Sprachkommunikation (AT 3), 30. Ausgabe

Editorial

Liebe Leserinnen und Leser,

nun ist es tatsächlich schon wieder Sommer – sowohl kalendarisch als auch meteorologisch. Entsprechend beginnt auch bald die Zeit der Konferenzbesuche und, für einige von uns, die vorlesungsfreie Zeit. Für viele sind das die besten Monate, um nicht nur in Urlauben und auf Konferenzen neue Inspirationen zu finden, sondern auch tatkräftig an den verschiedenen Projekten zu arbeiten, die in unseren Gruppen aktuell verfolgt werden. Sehr viele davon drehen sich um die Frage, wie ein analytisches Verständnis der Signaltheorie, Statistik, Physik und Physiologie, die unseren Arbeiten zugrunde liegen, bestmöglich mit dem immensen Potential der künstlichen Intelligenz zusammenwirken kann, um nicht nur sehr gute, sondern auch sehr zuverlässige und reproduzierbare Ergebnisse zu erreichen.

In der neuen Voice Message finden Sie in diesem spannenden Spannungsfeld wieder aktuelle Ideen und Ergebnisse aus unserer Community, zusammen mit interessanten Stellenangeboten (mit kurzen, verbleibenden Fristen) und dem üblichen Überblick über unsere Konferenzen und Meetings, ganz am Ende des Newsletters. Ich hoffe, dass Sie hier auch interessante Informationen und Ideen finden, und wünsche darüber hinaus einen guten, angenehmen und nicht zu heißen Sommer, mit Zeit und Raum für Inspirationen.

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

● Das DFG-ANR Projekt „Akustik-bewusstes tiefes Lernen für die Sprachverarbeitung mithilfe verteilter Mikrofonarrays“ ist als Kooperation zwischen Prof. Simon Doclo (Universität Oldenburg), Prof. Romain Serizel (Université de Lorraine) und Dr. Antoine Deleforge (Universität Straßburg) gestartet. Ziel des Projekts ist die Entwicklung eines hybriden Frameworks, das datengetriebene und physikbasierte Ansätze zur Sprachverbesserung in akustischen Sensornetzwerken kombiniert. Eine zentrale Forschungsfrage ist dabei, welcher Grad an Realitätsnähe bei der akustischen Modellierung und Simulation erforderlich ist. Darüber hinaus wird untersucht, wie partiell verfügbare Informationen über die akustische Szene in Deep-Learning-basierte Algorithmen integriert werden können.

● The assistive listening team in the mtec group at TU Berlin – Prachi Sharma, Ferdinand Campe and Dorothea Kolossa – is happy to have won the Aria Smart Glasses track in Task 2 of the ECHI Chime challenge, which is concerned with target speech extraction for hearing-impaired listeners. The submitted systems were compared regarding their performance on microphone array data recorded on Aria smart glasses, with perceptual quality as the core criterion. The challenge only allowed for models that can operate under strict real-time constraints, allowing at most 20 ms of latency at 16 kHz sampling rate. The winning approach enhanced speaker selectivity of a bounded masking network through cross-attention-based feature-wise linear modulation (FiLM). The results clearly show a trade-off between noise suppression and speech naturalness, and support the idea of using perceptually motivated designs for assistive listening in real-world scenarios.

Journalartikel

● S. Welker, B. Lay, M. Hillemann, T. Peer, T. Gerkmann (2026), Real-Time Streamable Generative Speech Restoration with Flow Matching. Diffusion-

based generative models have recently reshaped speech processing, delivering highly natural outputs and opening new research directions. Yet their use in real-time communication has been limited by heavy computational demands from requiring many network evaluations. Here we introduce Stream.FM, a frame-causal flow-based generative model that can solve six different speech restoration tasks in real-time, with 48 ms total latency on a consumer GPU: speech enhancement, dereverberation, bandwidth

extension, codec artifact removal, STFT phase retrieval, and Mel vocoding. To achieve this, we propose a buffered streaming inference, an optimized DNN architecture, and learned few-step numerical solvers that improve quality within a fixed compute budget. Extensive evaluations, including MUSHRA tests, show that Stream.FM sets a new state of the art for generative streaming speech restoration. A demo video, audio examples, model code, and checkpoints are available here.

● J. Steiner, F. Hilgemann, P. Jax (2026), Investigations on Drone-mounted Active Noise Control. Although drones have many potential applications including parcel delivery and surveillance, their deployment in urban areas is limited by their noise levels. As a mitigation approach, this work considers active noise control (ANC) with drone-mounted secondary sound sources and reference sensors. The feasibility of this approach is investigated in a controlled acoustic environment. Because commercially available drones lack the hardware required for ANC, two drones were equipped with loudspeakers and microphones for measurement purposes. ANC filter design requires information about the electro-acoustic paths, which are investigated based on measurements in a hemi-anechoic chamber. This facilitates the development of a multichannel feedforward control structure, and its evaluation through simulations. The investigation indicates the feasibility of such systems, as spatially robust median attenuation of 9.4 dB and a peak attenuation of over 15 dB is achieved in simulations. Although further investigation is needed to verify the results, this work documents a principally viable system and identifies its fundamental challenges.

● H. Gode, S. Doclo (2026), Closed-Form Successive Relative Transfer Function Vector Estimation based on Blind Oblique Projection Incorporating Noise Whitening. This paper addresses online estimation of relative transfer function (RTF) vectors for multiple sound sources in noisy and reverberant environments. We propose a generalization of the blind oblique projection (BOP) method, replacing iterative optimization with a closed-form solution and incorporating noise whitening and subtraction to improve robustness at low SNRs. Simulations in highly dynamic scenarios with random source activations and deactivations show that the proposed BOP method outperforms conventional BOP and other baseline methods in terms of computational efficiency and signal-to-interferer-and-noise-ratio improvement when using the estimated RTF vectors in an LCMV beamformer.

Dissertationen

● Tobias Kabzinski: Signal Processing Algorithms for Adaptive Loudspeaker-Based Binaural Audio Reproduction, RWTH Aachen (Betreuer: Prof. Peter Jax).

This thesis addresses the limitations of conventional loudspeaker-based binaural reproduction, which suffers from acoustic model errors. An adaptive system with microphones at the listener's ears is proposed, combining crosstalk cancellation (CTC) with continuous system identification to track time-varying acoustic paths. CTC algorithms are developed from control-theory and FIR perspectives, and system identification methods are formulated using a state-space approach for both online and offline use. The methodology is validated through simulations and acoustic measurements, showing effective adaptation to listeners and acoustic environments.

● Navin Raj Prabhu: "Probabilistic and Generative Modeling for Emotion Recognition and Synthesis", Universität Hamburg (Betreuer: Prof. Timo Gerkmann, Prof. Nale Lehmann-Willenbrock).
Diese Dissertation widmet sich dem Social Signal Processing mit Fokus auf Affekt. Sie leistet Beiträge in der Erkennung und der Synthese affektiver Ausdrücke. Auf der Erkennungsseite werden probabilistische Methoden entwickelt, die Annotationsunsicherheit mittels Bayes'scher neuronaler Netze und Label Distribution Learning modellieren. Darüber hinaus werden temporale Annotationen von Gruppenaffekt in Gruppeninteraktionen erhoben und mittels neuronaler Graph-Netze modelliert. Auf der Syntheseseite werden generative Modelle für die Sprachkonvertierung auf In-the-wild-Daten ohne Parallelkorpora eingesetzt. Die Arbeit legt damit eine Grundlage für sozial intelligente KI-Systeme, die Affekt in realen Gruppeninteraktionen sowohl wahrnehmen als auch angemessen ausdrücken können.

● Marvin Tammen: Combining model-based and learning-based approaches for speech enhancement, Universität Oldenburg (Betreuer: Prof. Simon Doclo)

This thesis investigates hybrid speech enhancement algorithms in which quantities required by a model-based enhancement stage are estimated by a learning-based stage. First, we propose a hybrid single-microphone speech enhancement approach that embeds the multi-frame minimum variance distortionless response filter in a deep learning framework and imposes structure on temporal covariance matrices. Second, we extend this approach to multi-microphone speech enhancement for binaural hearing devices, imposing structure on spatio-temporal correlation vectors. Third, we introduce a spatial regularization procedure that improves interpretability of the estimated quantities by enforcing spatial structure.

● Ragini Sinha: Deep neural network-based approaches for single-channel speaker-conditioned target speaker extraction, Universität Oldenburg (Betreuer: Prof. Simon Doclo).

This thesis develops and evaluates novel DNN-based architectures for single-channel speaker-conditioned target speaker extraction (SC-TSE) using reference speech as auxiliary information. First, we propose three customized long short-term memory cell variants for time-frequency-domain target speaker extraction, designed to optimize how target speaker information is retained and updated. Second, we propose two conformer-based architectures for time-domain target speaker extraction and improve the real-time suitability by replacing conventional multi-head self-attention (MHSA) with linear MHSA. Third, we subjectively evaluate two SC-TSE algorithms through listening tests with normal-hearing and hearing-impaired listeners.

Stellenangebote

● Die TU Graz hat zwei Tenure-Track Professuren für Kommunikationsakustik mit den Schwerpunkten „Human and Machine Listening“ ausgeschrieben, siehe https://jobs.tugraz.at/en/jobs/1f6ebe71-e41d-48ac-a06d-4292280ab412. Mindestens eine der beiden Stellen wird mit einer Frau besetzt. Bewerbungsschluss ist bereits der 8. Juli 2026. Nähere Auskünfte bei Franz Pernkopf und Gernot Kubin.

Tagungen

Hier finden Sie die Tagungen des Jahres im Überblick:

IJCAI-ECAI	15. - 21.08.2026, Bremen, Deutschland [Keine Anmeldung von Beiträgen mehr]
EUSIPCO	31.08. - 04.09.2026, Brügge, Belgien [Keine Anmeldung von Beiträgen mehr]
IWAENC	07.09. - 10.09.2026, Cremona, Italien [Keine Anmeldung von Beiträgen mehr]
FA	08.09.2026 - 12.09.2026, Graz, Österreich [Keine Anmeldung von Beiträgen mehr]
Interspeech	27.09. - 01.10.2026, Sydney, Australien [Keine Anmeldung von Beiträgen mehr]
SLT	13.-16.12.2026, Palermo, Italien [Keine Anmeldung von Beiträgen mehr]
DAGA	08.-11.03.2027, Darmstadt, Deutschland Einreichungen bis 01.11.26