Final Report Ha3455/19-1 & SCHL2043/2-1:
DE: Automatische Transkription von Gesprächssituationen
EN: Automatic Transcription of Conversations

1 General Information

DFG reference numbers

Ha3455/19-1 and SCHL2043/2-1

Project number

448568305

Project Title

DE: Automatische Transkription von Gesprächssituationen
EN: Automatic Transcription of Conversations

Names of the applicants and official addresses

Prof. Dr.-Ing. Reinhold Häb-Umbach	PD Dr. Ralf Schlüter
Universität Paderborn	RWTH Aachen University
Fachgebiet Nachrichtentechnik	Machine Learning and Human Language Technology (MLHLT) Group, Lehrstuhl Informatik 6
Warburger Straße 100	Theaterstraße 35–39
33098 Paderborn	52062 Aachen
Email: haeb@nt.uni-paderborn.de	Email: schlueter@ml.rwth-aachen.de
Tel. (05251) 60 3625	Tel. (0241) 80 21612

Names of co-applicants

Names of cooperation partners

Reporting period (entire funding period)

1 July 2021 – 31 December 2024 (incl. cost-neutral project extension)

2 Summary

Multi-talker conversational speech recognition is concerned with transcribing meetings recorded with distant microphones. The difficulty of the task can be attributed to three factors. First, the recording conditions are challenging: The speech signal captured by microphones from a distance is noisy and reverberated and often contains nonstationary acoustic distortions, which makes it hard to decode. Second, there is a significant percentage of time with overlapped speech, where multiple speakers talk at the same time. Finally, the interaction dynamics of the scenario are challenging because speakers articulate themselves in an intermittent manner with alternating segments of speech inactivity, single-, and multi-talker speech.

This project was concerned with developing a transcription system that can operate on arbitrarily long input, correctly handles segments of overlapped as well as non-overlapped speech, and transcribes the speech of different speakers consistently into separate output streams. Such a multi-talker Automatic Speech Recognition (ASR) system typically consists of the following three components: a source separation and enhancement block, a diarization stage, that attributes segments of input speech to speakers, and an ASR stage, whereby different orders of processing have been proposed. Those orders differ in when to do diarization.

While existing approaches employed separately trained subsystems for diarization, separation, and recognition, our research hypothesis was that a joint approach, which is optimized under a single training objective, should lead to superior solutions compared to the separate optimization of individual components. Such a coherent formulation, however, would not necessarily mean that the three aforementioned tasks had to be carried out in a single, monolithic (probably neural) integrated system.

Indeed, the research carried out showed that it is beneficial to have separate subsystems, however, with a tight coupling between them. Examples of such systems we developed are

- TS-SEP [1], which carries out diarization and separation/enhancement, with a tight coupling in-between.
- Mixture encoder [2], which leverages explicit speech separation, but also forwards the not yet separated speech to the ASR module to mitigate error propagation from the separator to the recognizer.
- Joint diarization and separation [3], realized by a statistical mixture model, which integrates a mixture model for diarization and one for separation, that share a common hidden state variable.
- Transcription-supported diarization [4], which uses sentence- and word-level boundaries of the ASR module to support speaker turn detection.

Furthermore, we developed new approaches to the individual subsystems and shared several tools and data sets with the research community.

3 Progress Report

Meeting transcription consists of three tasks

- Source separation and enhancement: the separation of partially overlapping speech, and the suppression of additive noise, reverberation or any other distortions
- Diarization: The segmentation of an arbitrarily long input into segments of constant speaker activity, and the assignment of speakers to those segments
- Recognition: Transcription of the speech signal.

While prior art in meeting transcription was a modular approach, where source separation and enhancement, diarization, and ASR were treated in isolation, the objective of this project was to explore the joint consideration of the tasks under a common training objective. Our hypothesis was that this should lead to improved performance. Being aware of the fact that the modular approach also bears advantages, such as flexibility by being able to plug together different solutions for the individual tasks, and better assessment of the performance of each task, we were not aiming at a monolithic all-neural approach. Instead, we wanted to combine the advantages of the two paradigms, by keeping some modularity, while at the same time jointly optimizing them.

This led to systems that jointly treat diarization and separation, separation and recognition, as well as recognition and diarization. Note that, as a deviation of the original plan, we did not optimize all three components under a single training loss. One reason for this was that the state-of-the-art solution for source separation that became available during the course of the project [O1] was so demanding w.r.t. compute power and memory that a joint optimization of all components was no longer feasible.

The progress report is roughly organized along the original proposal, starting with the description of the data sets we published (work package (WP) 1 of the proposal) and the research on evaluation metrics (WP2). We then report on source separation and diarization (WP3 and WP4), on ASR, including ASR and diarization (WP5), as well as on training paradigms and joint separation and ASR approaches (WP6 and WP7).

3.1 Data Sets and Evaluation Metrics (WP1 and WP2)

Research on meeting transcription requires dedicated data sets and evaluation metrics.

At the time of the submission of the proposal for this project, Microsoft published a data set for the evaluation of meetings. LibriCSS [O2] consists of 10 hours of meeting data that is simulated based on the LibriSpeech test-clean set [O3]. Eight loudspeakers were placed around a conference table and played back sentences of the LibriSpeech corpus with varying degrees of speaker overlap, ranging from 0% to 40%. The recording device was a seven-channel circular microphone array.

While this was an excellent test data set, a corresponding training data set was missing. In a parallel project we thus developed a “meeting generator” [O4]. This is a software tool that takes utterances from a source data set (e.g., LibriSpeech or Wallstreet Journal), reverberates it with artificially generated room impulse responses simulating a playback from a configurable number of loudspeakers and a configurable number and placement of microphones for signal capture. Also, the percentage of overlapped speech can be chosen. With this tool, appropriate data for training a meeting transcription system can be generated artificially.

Another contribution from a parallel project is the LibriWASN [O5] data set. This is an additional re-recording of the same synthetic meetings as LibriCSS, however, in a distributed setup with multiple recording devices, including smartphones and microphone arrays. Re-recordings were done in two different rooms, exhibiting a room reverberation time (T_{60}) of 200 ms (LibriWASN$_{200}$) and 800 ms (LibriWASN$_{800}$). This data set can be used for developing sampling clock synchronization algorithms (not part of this project), and for testing meeting transcription in a distributed microphones scenario.

Concerning evaluation metrics, the classical word error rate (WER), which counts the number of substitutions, deletions and insertions in the ASR output, divided by the number of words in the ground truth transcript, reaches its limits when evaluating meeting transcription systems, which encompass multiple speakers and multiple ASR output streams. A sensible metric must assign recognition outputs to speakers, be able to measure performance including and excluding speaker assignment errors, and be able to handle transcripts of arbitrary length. We carried out an in-depth investigation of existing WER metrics for meeting transcription to find out how well they aligned with (our) human intuition and proposed several extensions, which included, among others

- The introduction of time constraints to avoid the matching of hypothesized and ground truth transcriptions, if the two are temporally too far apart. This not only improved the plausibility of the alignment, but also the computational speed.
- A diarization-aware and a diarization invariant concatenated minimum permutation word error rate (cpWER), which allowed to quantify the contribution of speaker attribution errors to the overall error rate.

We published a toolkit, *MeetEval*, with efficient implementations of a range of WER definitions, and which also includes a visualization tool, to help system developers spot potential issues of the transcription system. This tool has been presented to the public in a Show & Tell session of ICASSP 2024. Furthermore, its use has been recommended in the official GitHub Repository of the CHiME challenge, a famous community effort to foster progress and comparability in meeting transcription research.[1]

In addition to conference publications [5], [16], a journal article on this topic is currently under review [6].

3.2 Diarization and Source Separation (WP3 and WP4)

We first investigated source separation and diarization separately and then proposed joint systems.

Concerning source separation, we looked into the then (and still) state-of-the-art tf-GridNet [O1] architecture and made several modifications such that it achieved the same separation performance, however at only one sixth of the computational effort. This was mainly achieved by replacing the BLSTM layers by a simpler architecture called TAC (transform-average-concatenate) [O6] and by adding dynamic range compression, as well as switching to a source-aggregated logarithmic Mean Absolute Error loss function. We then were able to use this separation architecture as a front-end for the transcription of meeting data [7], by integrating the modified tf-GridNet into a continuous source separation (CSS) processing stage: The separator is applied on a sliding window of the observation with a certain overlap between adjacent windows. With this we achieved a new single-microphone state-of-the-art WER on the LibriCSS corpus.

If input from a microphone array is available, a spatial mixture model (SMM) is an attractive alternative to a neural separation. We developed an initialization scheme for SMMs that takes advantage of speaker activity patterns that are typical of meetings, where there are regularly segments of speech where only a single speaker is active. Those segments serve to find initial values of the SMM parameters, which are then refined on all data. The proposed system carries out diarization, while at the same time delivering time-frequency masks used to estimate acoustic beamformer coefficients of signal enhancement and separation [8]. Recognition performance was close to the best reported results on the LibriCSS data set.

In an attempt to leverage both the power of neural networks to capture the spectral profile of speakers and the capability of SMMs to represent spatial information we devised an integrated mixture model which consists of a spectral mixture model (with von-Mises-Fisher distributions) of speaker embeddings and a SMM (with complex Angular Central Gaussian distributions) of multichannel microphone signals in the STFT domain, the two sharing a common hidden state variable. This model elegantly integrates spectral and spatial information to compute time-frequency masks for a downstream acoustic beamformer, again jointly carrying out diarization and enhancement [3].

We also developed an all-neural approach to joint diarization and separation. It is an extension of the well-known target-speaker voice activity detection (TS-VAD) [O7], which is known for its

[1] `https://github.com/chimechallenge`

excellent diarization performance on the CHiME-5 challenge data. The output of the last layer of the network, which contains the diarization information, is broadcast along the now introduced frequency axis and serves as the initialization of a neural-network for time-frequency source activity estimation. Its output, which indicates, for each source, at which time-frequency bin it is dominant, can then be used for source separation/extraction, either by mask multiplication or by beamforming, depending on whether single-channel or multi-channel input is available. In both cases, the system achieved a new state-of-the-art WER on the LibriCSS meeting data recognition task [1].

3.3 ASR for Meeting Transcription (WP5)

Today, many ASR architectures compete. To provide an overview of current end-to-end (E2E) ASR architectures in relation to classical hybrid neural network hidden Markov models (HMM), the MLHLT PI co-authored a widely received survey [O8]. The survey provides a taxonomy of E2E ASR models and corresponding improvements, and discusses their properties and their relationship to classical hidden Markov model (HMM) based ASR architectures, also covering the different approaches to time alignment of the different ASR architectures, providing an important connection between ASR model training, speech segmentation and speaker diarization. Encoder structures may be shared upon ASR architectures, and in [O9], a conformer encoder was established for the hybrid HMM architectures, which formed the basis of the jointly trained mixture encoder approach introduced in this project. Further work investigated the alignment behaviour of specific ASR architectures: alignment ambiguity and potential remedies for the attention-encoder-decoder architecture [O10], and a comparative study on training and alignment of hybrid HMM and CTC models [9], with subsequent work focusing on alignment quality [O11]. [9] provides an in-depth investigation, enabling end-to-end from scratch training for the hybrid HMM architecture based on factored modeling and comparing it to the CTC architecture. Factored hybrid HMM modeling avoids the use of secondary knowledge sources, i.e. phonetic clustering and prior alignments, and thus enables end-to-end from-scratch training. By counteracting convergence issues, from scratch training is improved, leading to better alignment behaviour and word error rate improvements. Also, both for hybrid HMM and CTC, improvements in word error rate are shown not to be connected to improved alignment, necessarily.

To perform efficient recognition, the open source C++ RWTH ASR Toolkit (RASR) [O12] [O13] developed by the MLHLT group has been used. This toolkit allows state-of-the-art large-vocabulary speech recognition search with various output label units, pruning options, language model integration and lookahead techniques. It was originally designed with classical ASR architectures such as Gaussian mixture or hybrid HMM in mind but has recently been extended to cover a wider range of ASR modeling choices such as CTC, Neural Network Transducers or Attention-Encoder-Decoder models [O14]. Using this toolkit for a comparable search procedure, different types of ASR models were internally evaluated on separated meeting data signals. This comparison leads to the conclusion that the CTC and Neural Network Transducer models are outperformed by the classical hybrid HMM in this scenario, although they usually tend to be on par or better on single-speaker tasks such as LibriSpeech [O8], which suggests that the hybrid HMM approach might be more robust to artifacts produced by the speech separation front-end. These insights motivated the use of hybrid HMM ASR architectures in further publications [2] [7] [17].

While the developed systems described in Section 3.2 performed diarization *before* ASR, there are also good arguments to do diarization *after* ASR: The syntactic information about word and sentence boundaries provided by ASR can benefit the segmentation. However, this requires the front-end to be able to process input of arbitrary length. The concept of CSS mentioned in Sec. 3.2 allows for that. The system we proposed in [4] extracts so-called d-vectors from the segments after the CSS module to assign speakers to segments of speech, while the segmentation itself is refined after the ASR module by using its word and sentence boundary information. This resulted in state-of-the-art cpWER for the full meeting recognition pipeline.

3.4 Training Paradigms and Joint Training of Cascaded Systems (WP6 and WP7)

Cascaded modular systems, which first perform source separation on overlapped multi-speaker audio and then afterwards run conventional single-speaker ASR on the separated signals have been commonplace in the community for a while [O15]. This way, the source separation front-end can be trained and evaluated independently and any pre-trained single-speaker model can be used as the ASR back-end. However, this pipeline enforces intermediate hard decisions for the separation which can include errors that are propagated and can hardly be corrected by the ASR back-end. Existing mitigation approaches include fine-tuning of the ASR module on separated audio signals or even joint training of separation front-end and ASR back-end [O16]. We introduced another approach called the mixture encoder which allows the ASR module to process the original overlapped signal in addition to the separated signals [2]. The goal of this approach was to allow the ASR module to not fully rely on the hard separation decisions and potentially correct errors made by the front-end. In this model, there are two disjoint encoder blocks with one separation encoder processing the separated audio signal and one mixture encoder processing the original overlapped audio. The outputs of these disjoint encoder blocks are concatenated and further processed by a mixture-aware speaker (MAS) encoder. A further extension of the model architecture uses a combination-layer which jointly encodes the outputs of the MAS encoders based on the primary- and cross-signals produced by the speech separator. The training of the models was done in three phases: 1) pre-train the speech separation front-end separately using a signal-level criterion, 2) freeze the separator and pre-train the ASR back-end with mixture encoder on the separated signals together with the overlapped audio using an ASR criterion, 3) unfreeze the separator and jointly optimize both modules using the same ASR criterion. We compared the model with mixture encoder and combination layer to a cascaded baseline model without additional encoders that has been trained in the same way. The comparison was carried out on the SMS-WSJ dataset [O17] which contains speech mixtures with source data taken from the Wallstreet Journal dataset. Overall, we observed around a 7% relative reduction in WER when using the mixture encoder model compared to the baseline, proving that the ASR model is able to catch and correct some separation artifacts [2].

In [7], the mixture encoder was applied to a scenario with an arbitrary number of speakers and dynamic overlap ratios and was therefore tested on LibriCSS. Furthermore, it was combined with a stronger separation front-end compared to [2], namely tf-GridNet [O1]. The training procedure was similar to [2] but without joint training of front- and back-end and instead with an additional phase of sequence discriminative training of the back-end. While the obtained results constituted a new state-of-the-art on LibriCSS, we also observed that tf-GridNet advances the separation quality to a level where the effect of additional mixture information vanishes.

Given the state-of-the-art results on LibriCSS, we further investigated the remaining errors by measuring the performance using oracle information and by studying the impact of leakage which is a typical separation artifact [7]. We presented results using oracle segmentation from the LibriCSS annotation and additionally oracle signals from LibriSpeech where no reverberation or overlap occurs. The relative improvements over our best system were 27% with oracle segmentation and 63% with oracle signals and segmentation. This shows that there is still a large gap between meeting transcription and clean single-speaker transcription.

In [7], we proposed a framework to measure the effect of leakage in order to get a deeper understanding of the influence of separation errors on the recognition performance. Unlike existing metrics, it handles silence appropriately and guarantees the temporal locality of detected errors by using Hamming distances of frame-wise word-level alignments. We showed that the cross-talker is well suppressed and there are hardly any leaked words from the cross channel that play a role in the search space for the primary channel [7]. In a further submission that we are currently preparing, we were able to show that there is significant leakage from primary channel into the cross channel in areas where only the primary speaker is active [17]. Motivated by the large performance gap to the oracle segmentation in [7] and the occurrence of leaked words in the cross channel, we studied the errors that occur during segmentation [17]. We used the MeetEval visualization toolkit to find typical problems and measured the impact of each type by eliminating the errors from our Voice Activity Detection (VAD) segmentation using heuristics that exploit oracle information. Using these eliminations, mainly by recovering parts of segments that were dropped by VAD, we were able to explain more than half of the performance gap to the oracle segmentation.

3.5 Deviations from Original Concept

There was no significant deviation from the original concept we outlined in the project proposal. However, there were some minor deviations from the plan as follows

- Task 3.2 Research on reverberation robustness was reduced in favor of increasing research on source separation, as this turned out to open options for joint optimization with the ASR back-end, see Section 3.4.
- For the same reason (better compatibility with joint optimization with back-end) we more focused on CSS than on the originally planned recursive speaker separation (Task T4.1)
- End-to-end model training (Task T6.2) was done only with the separator/enhancement and the ASR subsystem, excluding diarization, because of complexity reasons.
- Decoding aware source separation (Task T7.3) was abandoned, because the self-supervised model WavLM [O18] published by Microsoft and used for several of our experiments essentially solves this task: this data-driven feature extraction is learned from partially overlapped speech and given the target speech signal as guide, thus effectively incorporating back-end information into the front-end.

3.6 Contributors

The many publications with reference to this DFG project were only possible with the support of other team members who worked on related topics and with whom we collaborated.

On the Paderborn side, we wish to mention Tobias Cord-Landwehr, Thilo von Neumann, Jörg Schmalenströer and Tobias Gburrek. The contribution of the latter two, the LibriWASN data sets mentioned in Sec. 3.1, was done with support from the DFG Research Unit FOR2457 “Acoustic Sensor Networks” (project number 282835863).

On the Aachen side, we have to mention Christoph Lüscher, Tina Raissi, Robin Schmitt, Jingjing Xu, Mohammad Zeineldeen, Albert Zeyer and Wei Zhou. Lüscher, Schmitt, Xu, Zeineldeen and Zeyer contributed to the conformer encoder used in the ASR systems, Zhou contributed with a generic search implementation covering a variety of ASR architectures, Raissi and Zhou investigated and compared the alignment behavior of HMM- and CTC-based ASR models.

3.7 Quality Enhancing Measures

We applied rather conventional but well-established measures to maintain high quality of the results produced. These are

- Regular code review by peers

- Publication of software as open source for the community, both to provide a service to the community and to gather feedback
- Development and use of analysis and visualization tools as mentioned above. Those tools allow inspection into the inner workings of the developed algorithms
- Use of widely used data sets to ease comparison and reproducibility by other research groups
- All publications have undergone a peer review process

3.8 Dissemination of Research Data, Software

We published code generated within this project on github, see Sec. 4.2 for a list.

3.9 Scientific Events, Science Communication

PI Haeb-Umbach gave an interview on this project ("Spracherkennungssysteme für schwierige Situationen") in the feature "Computer und Kommunikation" of Deutschlandfunk, which was broadcast on March 8, 2025.[2]

3.10 Other References

[O1] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B. Kim, S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," *Proc. IEEE ICASSP*, 2022.

[O2] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE ICASSP*, 2020, pp. 7284–7288.

[O3] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[O4] T. Cord-Landwehr, T. Von Neumann, C. Boeddeker, R. Haeb-Umbach, "MMS-MSG: A multi-purpose multi-speaker mixture signal generator," in *Proc. IEEE IWAENC*, 2022, pp. 1–5.

[O5] J. Schmalenstroeer, T. Gburrek, R. Haeb-Umbach, "LibriWASN: a data set for meeting separation, diarization, and recognition with asynchronous recording devices," in *Speech Communication; 15th ITG Conference*, VDE, 2023, pp. 86–90.

[O6] Y. Luo, Z. Chen, N. Mesgarani, T. Yoshioka, *End-to-end microphone permutation and number invariant multi-channel speech separation*, 2020.

[O7] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. ISCA Interspeech*, 2020, pp. 274–278. DOI: `10.21437/Interspeech.2020-1602`.

[O8] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM TASLP*, vol. 32, pp. 325–351, 2023.

[O9] M. Zeineldeen, J. Xu, C. Lüscher, R. Schlüter, H. Ney, "Improving the training recipe for a robust conformer-based hybrid model," in *Proc. ISCA Interspeech*, 2022.

[O10] R. Schmitt, A. Zeyer, M. Zeineldeen, R. Schlűter, H. Ney, "The conformer encoder may reverse the time dimension," in *Proc. IEEE ICASSP*, 2025, pp. 1–5. DOI: `10.1109/ICASSP49660.2025.10889452`.

[O11] T. Raissi, C. Lüscher, S. Berger, R. Schlüter, H. Ney, "Investigating the Effect of Label Topology and Training Criterion on ASR Performance and Alignment Quality," in *Proc. ISCA Interspeech*, 2024. DOI: `10.21437/Interspeech.2024-830`.

[O12] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, H. Ney, "RASR - the rwth aachen university open source speech recognition toolkit," Dec. 2011.

[O13] S. Wiesler, A. Richard, P. Golik, R. Schlüter, H. Ney, "Rasr/nn: The rwth neural network toolkit for speech recognition," in *Proc. IEEE ICASSP*, 2014, pp. 3281–3285. DOI: `10.1109/ICASSP.2014.6854207`.

[2] `https://www.deutschlandfunk.de/spracherkennungssysteme-fuer-schwierige-situationen-int-reinhold-haeb-umbach-100.html`

[O14] W. Zhou, E. Beck, S. Berger, R. Schlüter, H. Ney, "RASR2: The RWTH ASR Toolkit for Generic Sequence-to-Sequence Speech Recognition," in *Proc. ISCA Interspeech*, 2023, pp. 4094–4098. DOI: 10.21437/Interspeech.2023-1062.

[O15] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, J. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, 2016, pp. 545–549. DOI: 10.21437/Interspeech.2016-1176.

[O16] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in *Proc. IEEE ICASSP*, 2020, pp. 7004–7008. DOI: 10.1109/ICASSP40776.2020.9053461.

[O17] L. Drude, J. Heitkaemper, C. Boeddeker, R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[O18] X. Chang, T. Maekaku, Y. Fujita, S. Watanabe, "End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation," in *Proc. ISCA Interspeech*, 2022, pp. 3819–3823. DOI: 10.21437/Interspeech.2022-10839.

4 Published Project Results

4.1 Publications with scientific quality assurance

[1] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, J. Le Roux, "TS-SEP: Joint Diarization and Separation Conditioned on Estimated Speaker Embeddings," *IEEE/ACM TASLP*, vol. 32, pp. 1185–1197, 2024, ISSN: 2329-9290. DOI: 10.1109/taslp.2024.3350887.

[2] S. Berger, P. Vieting, C. Boeddeker, R. Schlüter, R. Haeb-Umbach, "Mixture encoder for joint speech separation and recognition," in *Proc. ISCA Interspeech*, 2023. DOI: 10.21437/interspeech.2023-1815.

[3] T. Cord-Landwehr, C. Boeddeker, R. Haeb-Umbach, "Simultaneous Diarization and Separation of Meetings through the Integration of Statistical Mixture Models," in *Proc. IEEE ICASSP*, (Accepted for publication), IEEE, 2025. DOI: 10.48550/arXiv.2410.21455.

[4] T. von Neumann, C. Boeddeker, T. Cord-Landwehr, M. Delcroix, R. Haeb-Umbach, "Meeting recognition with continuous speech separation and transcription-supported diarization," in *Proc. IEEE ICASSP Workshops*, 2024. DOI: 10.1109/icasspw62465.2024.10625894.

[5] T. von Neumann, C. Boeddeker, K. Kinoshita, M. Delcroix, R. Haeb-Umbach, "On word error rate definitions and their efficient computation for multi-speaker speech recognition systems," in *Proc. IEEE ICASSP*, IEEE, 2023. DOI: 10.1109/icassp49357.2023.10094784.

[6] T. von Neumann, C. Boeddeker, M. Delcroix, R. Haeb-Umbach, "Word error rate definitions and algorithms for long-form multi-talker speech recognition," *IEEE/ACM TASLP*, 2025, (under review).

[7] P. Vieting, S. Berger, T. von Neumann, C. Boeddeker, R. Schlüter, R. Haeb-Umbach, "Combining TF-GridNet and Mixture Encoder for Continuous Speech Separation for Meeting Transcription," in *Proc. IEEE SLT*, 2024. DOI: 10.1109/SLT61566.2024.10832307.

[8] C. Boeddeker, T. Cord-Landwehr, T. von Neumann, R. Haeb-Umbach, "An initialization scheme for meeting separation with spatial mixture models," in *Proc. ISCA Interspeech*, 2022. DOI: 10.21437/interspeech.2022-10929.

[9] T. Raissi, W. Zhou, S. Berger, R. Schlüter, H. Ney, "HMM vs. CTC for Automatic Speech Recognition: Comparison Based on Full-Sum Training from Scratch," in *Proc. IEEE SLT*, 2022. DOI: 10.1109/SLT54892.2023.10022967.

[10] T. Cord-Landwehr, C. Boeddeker, C. Zorilă, R. Doddipatla, R. Haeb-Umbach, "Geodesic Interpolation of Frame-Wise Speaker Embeddings for the Diarization of Meeting Scenarios," in *Proc. IEEE ICASSP*, 2024. DOI: 10.1109/icassp48485.2024.10445911.

[11] C. Boeddeker, T. Cord-Landwehr, R. Haeb-Umbach, "Once more Diarization: Improving meeting transcription systems through segment-level speaker reassignment," in *Proc. ISCA Interspeech*, 2024. DOI: 10.21437/interspeech.2024-1286.

[12] T. Cord-Landwehr, C. Boeddeker, C. Zorilă, R. Doddipatla, R. Haeb-Umbach, "Frame-wise and overlap-robust speaker embeddings for meeting diarization," in *Proc. IEEE ICASSP*, 2023. DOI: 10.1109/icassp49357.2023.10095370.

[13] T. Cord-Landwehr, C. Boeddeker, C. Zorilă, R. Doddipatla, R. Haeb-Umbach, "A teacher-student approach for extracting informative speaker embeddings from speech mixtures," in *Proc. ISCA Interspeech*, 2023. DOI: 10.21437/interspeech.2023-1379.

[14] T. Cord-Landwehr, C. Boeddeker, T. von Neumann, C. Zorila, R. Doddipatla, R. Haeb-Umbach, "Monaural source separation: From anechoic to reverberant environments," in *Proc. IEEE IWAENC*, 2022. DOI: 10.1109/IWAENC53105.2022.9914794.

[15] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, R. Haeb-Umbach, "SA-SDR: A novel loss function for separation of meeting style data," in *Proc. IEEE ICASSP*, 2022. DOI: 10.1109/icassp43922.2022.9746757.

4.2 Other publications and published results

[16] T. von Neumann, C. Boeddeker, M. Delcroix, R. Haeb-Umbach, "MeetEval: A toolkit for computation of word error rates for meeting transcription systems," in *Proc. CHiME 2023 Workshop*, 2023. DOI: 10.21437/CHiME.2023-6.

[17] P. Vieting, S. Berger, T. von Neumann, C. Boeddeker, R. Schlüter, R. Haeb-Umbach, "Error Analysis for TF-GridNet-based Meeting Transcription," in *to be submitted to ITG 2025*, 2025.

[18] C. Boeddeker, T. Cord-Landwehr, T. von Neumann, R. Haeb-Umbach, "Multi-stage diarization refinement for the CHiME-7 DASR scenario," in *Proc. CHiME 2023 Workshop*, 2023. DOI: 10.21437/chime.2023-10.

[19] T. Gburrek, C. Boeddeker, T. von Neumann, T. Cord-Landwehr, J. Schmalenstroeer, R. Haeb-Umbach, *A meeting transcription system for an Ad-Hoc acoustic sensor network*, 2022. DOI: 10.48550/ARXIV.2205.00944.

Software: *MeetEval:* A Toolkit for Computation of Word Error Rates for Meeting Transcription Systems [5], [16], [6] and MeetEval alignment visualization tool; published on GitHub: `https://github.com/fgnt/meeteval`

Software: Once more Diarization: Improving meeting transcription systems through segment-level speaker reassignment [11]; published on GitHub: `https://github.com/fgnt/speaker_reassignment`

Software: TS-SEP: Joint Diarization and Separation Conditioned on Estimated Speaker Embeddings [1]; published on GitHub: `https://github.com/fgnt/tssep_data`

4.3 Patents (applied for and granted)