Fully End-to-End Multi-Channel, Multi-Talker Speech Recognition by Leveraging Foundation Models

Team Leaders	Lukáš Burget[1], Samuele Cornell[2]
Senior Members (confirmed)	Yoshiki Masuyama[3], Shinji Watanabe[1], Matthew Wiesner[4], Matt Maciejewski[4] Paola Garcia[4]
Senior Members (interested)	Jon Barker[4], Michael Mandel[5], Marc Delcroix[7], Naoyuki Kanda[5], Jun Du[8] Reinhold Haeb-Umbach[9]
Affiliations:	[1]Brno University of Technology, Czechia [2]Carnegie Mellon University, USA [3]Mitsubishi Electric Research Laboratories, USA [4]University of Sheffield, UK [5]Meta, USA [6]Johns Hopkins University, USA [7]NTT Corporation, Japan [8]University of Science and Technology of China, China [9]Padeborn University, Germany

Motivation

The field of robust speech processing has recently been revolutionized by the adoption of large-scale data self-supervised learning (SSL) and weakly-supervised learning methods. Yet, due to the training strategy of such foundation models, or the actual data that was used, we argue that there is still untapped potential especially when considering everyday conversational speech. Current speech foundation models, such as e.g. wav2vec 2.0 [1] and HuBERT [2] mostly leverage single-speaker speech data (or a synthetically augmented version of it as in WavLM [3]) and thus, while still effective, their training data is inherently mismatched with conversational speech. Regarding weakly supervised models, Whisper [4] likely used significant "speech-in-the-wild" data from web videos during training. However, due to its design, it fails to fully leverage it as it does not really have multi-speaker or diarization capability.

As such, current top-performing systems for multi-talker, multi-channel ASR (e.g most submissions to CHiME-7 and 8 DASR and NOTSOFAR-1[1] challenges) typically consist of a pipeline of independently trained subsystems, which include speaker diarization and source separation to isolate speakers, and single-channel single-speaker ASR applied to the separated audio. Thus errors made by each component propagate, negatively impacting overall performance.

At the same time, while current pre-trained models such as Whisper can be fine-tuned and adapted for the task at hand when inserted in a pipeline, these models are also fundamentally computational intensive and not really suited for processing multi-channel data. In fact, one common technique for multi-channel extension is running some layers in parallel and adapting the model with transform-average-concatenate (TAC) [5] adapters as done in our recent work [6] with WavLM. However, this is still computational intensive due to the fact that these models employ quadratic self-attention mechanisms and, more crucially, expensive convolutional front-ends. This computational burden also fundamentally limits the context of these models which is limited to few tens of seconds, potentially impacting the achievable upper-limit performance on long-form meeting scenarios. Whisper and OWSM [7] obviate to this issue by using the previous transcription result as a prompt for the current context, but this is known to cause hallucinations especially with multi-talker speech, again, due to error propagation [8]. On the other hand recent works suggest that long-term modeling is useful for ASR [9] and, crucially, diarization [10, 11].

Research Proposal

Our aim is to advance the research towards robust speech processing of conversational scenarios by tackling this important problem from two different but complementary directions which are summarized in Figure 1.

1. We will assemble a multi-channel, multi-talker ASR system using existing pre-trained subsystems, allowing for further fine-tuning on available training/adaptation multi-channel data. The advantage of this approach is that each subsystem can be be pre-trained on large datasets designated to handle simpler, related tasks (e.g. by leveraging Whisper for single-channel, single-speaker ASR). For this approach we plan to start from our submitted system in the recent CHiME-8 NOTSOFAR-1 challenge [12], which won the jury prize, and move towards end-to-end integration of its diarization and target-speaker ASR (TS-ASR) components.

2. We plan to build a "Whisper-style" large-scale pre-trained model that is 1) more computational "friendly", by exploring more efficient modeling mechanisms such as state-space models [13] and sparse attention mechanisms [14] and 2) we want to continue to expand the multi-task framework of Whisper and build a model that is more effective for long-form conversational speech. In detail we want the model to be able to perform diarization jointly with recognition by predicting speaker-id tokens as depicted in Figure 1 left and extend the audio context

[1] https://www.chimechallenge.org/

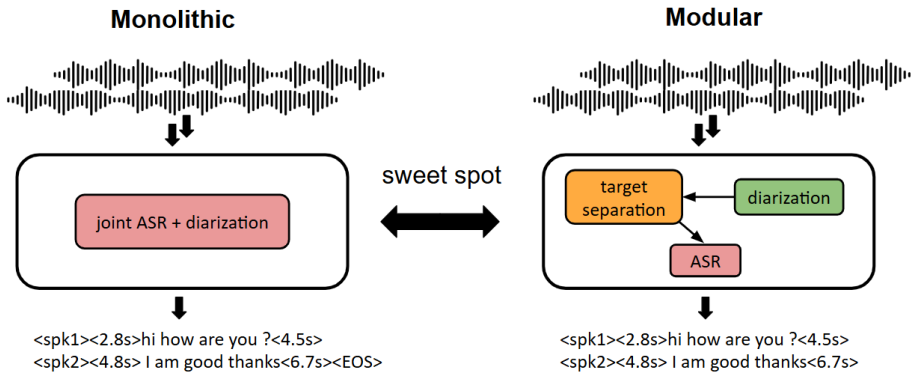

Figure 1: We want to improve conversational speech recognition by exploring two complementary directions: 1) modular end-to-end integration of existing diarization and ASR models and 2) creating a more computational efficient "Whisper-style" model that can perform joint ASR and diarization on long-form data.

to several minutes as in [9]. Also for this direction we can start from recent works [15, 16]. Regarding data, we plan to use the recently proposed YODAS dataset [17] for pre-training such model. We believe that the potential usefulness of such model can also extend beyond mere diarization and ASR. Whisper is often commonly used as a general audio representation extractor, and a model trained to perform also diarization together with ASR could be even more useful in this regard.

Our plan is to work on both directions simultaneously and integrate this newly designed pre-trained model (2) within the modular system (1) by leveraging it directly for multi-talker ASR and compare the performance with our Whisper-based TS-ASR system. As said before, a more efficient large scale pre-trained model is very important as it can drastically make the overall system less computationally cumbersome and more viable in real-world application scenarios (TS-ASR with Whisper requires to perform inference independently for each speaker in the conversation simultaneously).

Our proposal has a strong emphasis on multi-channel and we would like to build a system that is fundamentally array agnostic and can thus generalize to different recording setups. One direction we want to explore, since multi-channel data is scarce, is to use the TAC-based fine-tuning approach from [6]. This is another instance where the two approaches are complimentary since we can design the pre-trained model from the ground-up not only to be more computationally efficient but also to be able to exploit phase information (by e.g. using complex short-time Fourier transform as input feature) thus facilitating a multi-channel extension.

During the JSALT workshop, we hope to address several key questions: what subsystems should constitute the complete model (e.g. diarization + TS-ASR [12] vs. direct multi-talker ASR [15, 18] or continuous speech separation + joint ASR + diarization as in [19]) ? Regarding the direction of building a "Whisper-style" pre-trained model: is a fully end-to-end approach to meeting transcription viable ? Do we still need clustering (e.g. as in [16]) or downstream diarization to ensure speaker-id consistency across the whole meeting ? With what technique/mechanism we can achieve a good trade-off between performance and computational requirements ? How can we integrate different components (especially diarization and ASR) while ensuring the entire pipeline remains differentiable ? How can we deal with multiple channels when we have pre-trained models on single channel data ? Is the approach in [6] the best we can do ?

Finally, regarding evaluation data, this proposal is closely aligned with the objectives of the CHiME challenge, and it has received full support from the CHiME steering committee. As such, we plan to integrate the upcoming CHiME-9 challenges and align our model benchmarking activities accordingly.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[3] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[5] Yi Luo, Zhuo Chen, Nima Mesgarani, and Takuya Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. of ICASSP*. IEEE, 2020, pp. 6394–6398.

[6] Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukas Burget, "Leveraging self-supervised learning for speaker diarization," *Submitted to ICASSP*, 2024.

[7] Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, et al., "Reproducing whisper-style training using an open-source toolkit and publicly available data," in *Proc. of ASRU*. IEEE, 2023, pp. 1–8.

[8] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," *Proc. of Interspeech*, 2023.

[9] William Chen, Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe, "Train long and test long: Leveraging full document contexts in speech processing," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13066–13070.

[10] Tae Jin Park, Kunal Dhawan, Nithin Koluguri, and Jagadeesh Balam, "Enhancing speaker diarization with large language models: A contextual beam search approach," in *Proc. of ICASSP*. IEEE, 2024, pp. 10861–10865.

[11] Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao, "DiarizationLM: Speaker diarization post-processing with large language models," *Proc. of Interspeech*, 2024.

[12] Alexander Polok, Dominik Klement, Matthew Wiesner, Sanjeev Khudanpur, Jan Černockỳ, and Lukáš Burget, "Target speaker asr with whisper," *Submitted to ICASSP 2025*, 2024.

[13] Tri Dao and Albert Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," in *International Conference on Machine Learning*.

[14] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, and Furu Wei, "Longnet: Scaling transformers to 1,000,000,000 tokens," in *International Conference on Learning Representations*, 2023.

[15] Naoyuki Kanda, Xiong Xiao, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr," in *Proc. of ICASSP*. IEEE, 2022.

[16] Samuele Cornell, Jee-weon Jung, Shinji Watanabe, and Stefano Squartini, "One model to rule them all? towards end-to-end joint speaker diarization and speech recognition," *arXiv preprint arXiv:2310.01688*, 2023.

[17] Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe, "Yodas: Youtube-oriented dataset for audio and speech," in *Proc. of ASRU*. IEEE, 2023, pp. 1–8.

[18] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," *Interspeech*, 2020.

[19] Naoyuki Kanda, Jian Wu, Xiaofei Wang, Zhuo Chen, Jinyu Li, and Takuya Yoshioka, "Vararray meets t-sot: Advancing the state of the art of streaming distant conversational speech recognition," in *Proc. of ICASSP*. IEEE, 2023, pp. 1–5.