ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889116

MSA-ASR: Efficient Multilingual Speaker Attribution with frozen ASR Models

Thai-Binh Nguyen
Karlsruhe Institute of Technology
Karlsruhe, Germany
thai-binh.nguyen@kit.edu

Alexander Waibel
Carnegie Mellon University
Pennsylvania, USA
alexander.waibel@cmu.edu

***Abstract*—Speaker-attributed automatic speech recognition (SA-ASR) aims to transcribe speech while assigning transcripts to the corresponding speakers accurately. Existing methods often rely on complex modular systems or require extensive fine-tuning of joint modules, limiting their adaptability and general efficiency. This paper introduces a novel approach, leveraging a frozen multilingual ASR model to incorporate speaker attribution into the transcriptions, using only standard monolingual ASR datasets. Our method involves training a speaker module to predict speaker embeddings based on weak labels without requiring additional ASR model modifications. Despite being trained exclusively with non-overlapping monolingual data, our approach effectively extracts speaker attributes across diverse multilingual datasets, including those with overlapping speech. Experimental results demonstrate competitive performance compared to strong baselines, highlighting the model's robustness and potential for practical applications.**

***Index Terms*—speaker-attributed, asr, multilingual**

I. Introduction

Speaker-attributed automatic speech recognition (SA-ASR) involves transcribing all speech within a multi-speaker recording and accurately attributing each spoken word to the correct speaker. Specifically, suppose a recording X contains speech from K different speakers. In that case, the objective is to generate a set of transcriptions $Y = \{y_1, y_2, ..., y_K\}$, where each y_k corresponds to the sequence of words spoken by the speaker k. Addressing this challenge is crucial for various applications, from meeting transcription [1]–[4] to conversational AI [5], [6], where accurate speaker attribution is essential. Current approaches to this problem typically fall into two categories: modular strategies, which break down the task into separate components, and joint methods, which attempt to solve the problem in a unified framework.

Modular SA-ASR systems [7]–[11] decompose the SA-ASR task into sequential processing stages, typically involving speech separation, speaker diarization, and target speaker voice activity detection. These systems often assign speaker labels to speech segments prior to ASR. While modularity offers flexibility and the potential to leverage advancements in individual components, it can suffer from suboptimal performance due to the independent training of modules. Misalignments between the training objectives of these components can hinder the overall system's efficacy. In contrast, joint SA-ASR systems [8], [12] address these limitations by processing the entire SA-ASR task end-to-end, potentially achieving improved performance and coherence.

Joint SA-ASR systems, also known as E2E SA-ASR models, typically consist of two main components: the ASR module and the Speaker module. The ASR module generally generates a sequence of tokens, while the Speaker module produces speaker embeddings for each token. These embeddings are then used for speaker identification [13], [14]. In studies such as [15]–[17], speaker embeddings have been shown to enhance ASR performance by providing additional features for the ASR decoder layer. In some cases [18], rather than outputting speaker embeddings explicitly, they are utilized to help the ASR directly generate speaker labels. Often [12], [14], [15], [18], the ASR module not only generates a sequence of tokens but also produces special tokens (e.g., $< cc >, < sc >$) to indicate a change in speaker.

While much research on E2E SA-ASR has centered on enhancing ASR performance by incorporating speaker information, our approach offers a different perspective. Traditional joint models often involve fine-tuning the ASR component to add capabilities like generating speaker change tokens or managing overlapping speech, usually relying on limited, language-specific datasets. However, since overlapping speech constitutes roughly 10% of multi-talker data (as illustrated in Table I) and given the success of models like Whisper, which are trained on large and diverse datasets, we question the necessity of extensive ASR fine-tuning. The Whisper study [19] demonstrates that models trained on comprehensive and varied datasets can effectively handle a wide range of speech recognition tasks. In contrast, models tailored to specific datasets may excel within those domains but often lack broader robustness. We suggest that by focusing on the speaker module and utilizing a robust pre-trained ASR model, effective SA-ASR can be achieved without compromising generalizability.

To demonstrate generalizability, we extend SA-ASR to handle multilingual speech. Multilingual SA-ASR studies are rare, with the last in 2002 on speaker ID via multilingual phone strings [20], [21]. Research in zero-shot multilingual transfer learning [22] has shown that a frozen multilingual pre-trained model can be trained for a task in one language and then used to make predictions in another. Additionally, studies [23], [24] have explored training models for multi-speaker speech recognition using standard ASR datasets. Building on this foundation, we leverage the Whisper model's capacity to process multiple languages and follow the approach suggested by [23] to adapt regular English ASR datasets. This allows

Dataset	Overlap (%)	Duration (hours)
ESTER 1&2 [25]	0.67	260
ETAPE [25]	5.29	105
EPAC [25]	1.11	34
REPERE [25]	3.36	58
DIHARD [25]	11.6	34
AMI [25]	13.87	96
FISHER [26]	13.53	984
CHIME 6 dev/eval [27]	21 / 15	4.5 / 5.1

TABLE I
TOTAL DURATION AND PROPORTION OF OVERLAPS DURATION FOR DIFFERENT SPEECH CORPORA.

us to create a new E2E SA-ASR model (MSA-ASR) that can predict speakers across different languages, introducing a novel method for multilingual SA-ASR. Through benchmarking across various datasets, we demonstrate the effectiveness of our proposed approach in handling SA-ASR tasks across different languages and conditions.

II. Approach

A. Modeling

In designing our MSA-ASR model (Figure 1), we sought to blend the advantages of both modular and joint SA-ASR systems. The model consists of two main components: the ASR and Speaker modules. By keeping the frozen ASR module as a modular system, we ensure the speech recognition process remains stable and generalizable across diverse languages and domains. Meanwhile, the Speaker module is fine-tuned to work in harmony with the ASR system as a joint system.

ASR module is a transformer seq2seq model containing an encoder and decoder. The encoder takes the input signal X to produce hidden features $H^{\text{asr}} \in \mathbb{R}^{L \times f^e}$ where f^e and L are the feature dimension and the length of the feature sequence. The ASR decoder then iterative estimates sequence $W = [w_1, ..., w_N]$. At each decoder step, the ASR decoder calculates the output $w_n = ASRDecoder(w_{[0:n-1]}, H^{\text{asr}}) \in \mathcal{V}$ ($\mathcal{V}$ is the ASR vocabulary) given previous token $w_{[0:n-1]}$ and encoder hidden features H^{asr}.

The Speaker module predicts a sequence of speaker embeddings $E = [e_1, e_2, ..., e_N] \in \mathbb{R}^{N \times f^d}$ where f^d denotes the speaker embedding dimension. Each embedding e_n corresponds to a token w_n in the ASR-generated sequence. This module, like the ASR, uses a transformer architecture. Speaker encoder transforms X into hidden features $H^{\text{spk}} \in \mathbb{R}^{L \times f^e}$ (same shape as ASR encoder output). Figure 2 illustrates the architecture of the Speaker decoder. The word and position embeddings are shared between the ASR and Speaker decoder modules to ensure alignment between their outputs. The cross-attention of the first K layers in the Speaker decoder has been customized to take H^{asr} as the key, H^{spk} as the value. In the rest of $(D-K)$ layers, the key has been calculated from H^{spk}.

The Speaker module's training objective is to optimize E to closely match the target speaker embedding sequence $T = [t_1, t_2, ..., t_N] \in \mathbb{R}^{N \times f^d}$ using cosine similarity loss (4). We also want the model to distinguish between speakers inside an utterance. To do this, we first calculate the pairwise cosine similarity between the embeddings within the output sequence

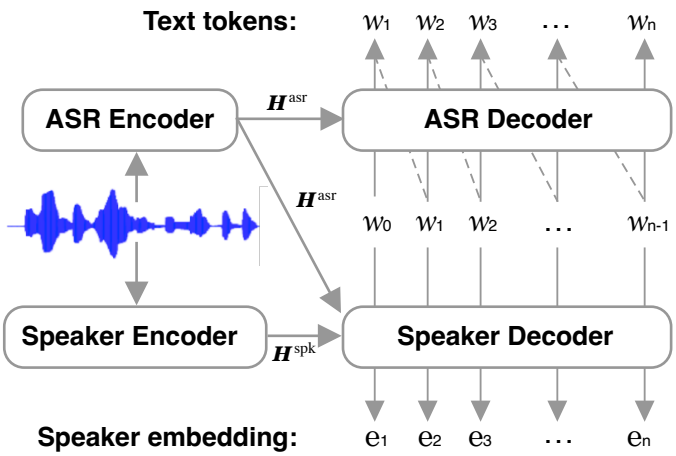

Fig. 1. Overview of MSA-ASR model. ASR decoder processes tokens sequentially during inference, while the Speaker decoder can generate speaker embeddings in parallel.

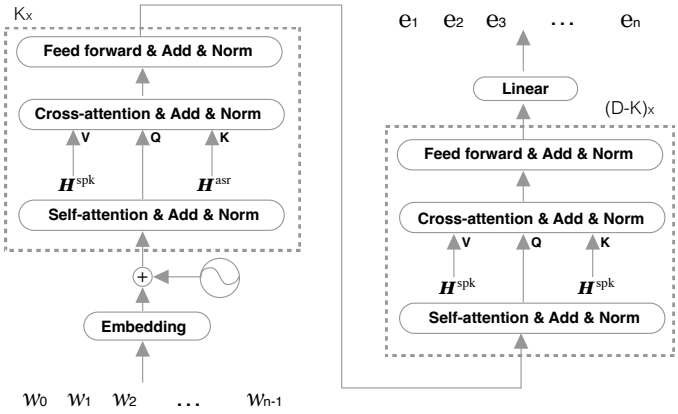

Fig. 2. Speaker decoder architecture. Similar as standard transformer decoder, but cross-attention uses key and value from a different encoders.

E (denoted as C_{ee}) (1), between the embeddings within the target sequence T (denoted as C_{tt}) (3), also the pairwise cosine similarity between E and T (denoted as C_{et}) (2). Then an MSE loss (5, 6) used to make both C_{ee} and C_{et} close to C_{tt}. Our final Embedding Alignment and Discrimination loss (EAD) will be the weighted sum of 3 losses (7).

$$C_{ee} = [cos(e_i, e_j)]_{i,j=1}^{N} \in \mathbb{R}^{N \times N} \quad \text{pairwise similarity within } E \tag{1}$$

$$C_{et} = [cos(e_i, t_j)]_{i,j=1}^{N} \in \mathbb{R}^{N \times N} \quad \text{between } E \text{ and } T \tag{2}$$

$$C_{tt} = [cos(t_i, t_j)]_{i,j=1}^{N} \in \mathbb{R}^{N \times N} \quad \text{within } T \tag{3}$$

$$L_1 = \sum_{i=1}^{N} (1 - cos(t_i, e_i)) \in \mathbb{R} \quad \text{cosine similarity loss with E, T} \tag{4}$$

$$L_2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (C_{ee}[i,j] - C_{tt}[i,j])^2 \in \mathbb{R} \quad \text{MSE loss} \tag{5}$$

$$L_3 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (C_{et}[i,j] - C_{tt}[i,j])^2 \in \mathbb{R} \quad \text{MSE loss} \tag{6}$$

$$L = \alpha L_1 + \beta L_2 + \gamma L_3 \quad \text{EAD Loss} \tag{7}$$

Our EAD loss function is similar to triplet and contrastive

loss (widely used for speaker verification tasks [28]), as all aim to distinguish between different speakers in the embedding space. Like triplet loss, which pulls similar samples together and pushes dissimilar ones apart, and contrastive loss, which minimizes distances within similar pairs, our loss uses cosine similarity (L_1) to align output and target embeddings while maintaining internal pairwise relationships (L_2, L_3). Our EAD loss is chosen because we don't have the ground truth of speaker labels but only the target speaker embedding T, so these target functions help in effective embedding differentiation and alignment without needing explicit labels. A detail of how T has been constructed is presented in the next section.

B. Data Processing

We extend the approach outlined in [23], which processed speaker turns in conventional ASR datasets for multi-talker speech recognition. While the previous work focused on detecting speaker changes by combining random turns, it did not account for the need to extract consistent speaker embeddings, as each speaker might only appear once, limiting the model's ability to differentiate between speakers.

To overcome this, we employed a pre-trained speaker embedding model, TitaNet-L [29], trained on 7,000 hours of diverse speech datasets, to compute an embedding for each speaker's turn. Since clustering turns from the same speaker is inherently complex, we opted not to rely on cluster labels. Instead, as detailed in section II-A, we used the original speaker embeddings as weak labels, avoiding the need for explicit speaker labels. We identified similar turns by selecting those with a cosine similarity score above the threshold θ.

The training samples were created by pairing turns randomly. The target sequence of speaker embeddings T is constructed by aligning each turn's transcript with its corresponding speaker embedding. Each turn was paired with at least one similar counterpart, and we limited each sample to five distinct turn groups with no overlapping similar turns. This process ensured that each sample contained up to 30 seconds of no overlapping speech from a maximum of five speakers. Similar to [23], we also incorporated random noise and reverberation to enhance the data.

III. Experiments

A. Datasets

To benchmark systems, we utilize three types of datasets: multilingual (Voxpopuli [30]), monolingual (AMI-IHM [31], LibriCSS [32]), and mixed-language (in-house data). Since AMI-IHM and LibriCSS are typically employed for benchmarking English SA-ASR systems, we will not delve into their details here. The processing for the other datasets is as follows:

Voxpopuli is a multilingual speech corpus featuring one speaker per sample across 16 languages. To adapt it for the multi-talker benchmark, we used the approach from [23], where test set utterances are randomly concatenated. This produces samples with an average of 2.5 speakers, 20 seconds of audio, and up to 5 non-overlapping turns.

Language	Diarization+ASR	MSA-ASR (Our)	ASR
English	15.52	12.90	12.24
German	26.24	16.54	14.28
French	32.20	16.53	13.95
Spanish	20.75	13.73	11.32
Polish	34.94	16.15	10.31
Italian	33.52	23.71	20.06
Romanian	38.48	23.65	18.05
Hungarian	29.77	28.12	19.82
Czech	35.47	28.60	16.27
Dutch	29.81	18.12	14.68
Finnish	37.30	20.51	15.85
Croatian	37.52	34.52	28.05
Slovak	34.44	27.24	16.07
Slovenian	41.32	30.90	27.33
Estonian	44.65	39.59	37.15
Lithuanian	69.17	40.57	34.04

TABLE II
COMPARISON OF CPWER (%) ACROSS MULTIPLE LANGUAGES FOR NON-OVERLAPPING MULTI-TALKER VOXPOPULI USING DIARIZATION WITH ASR (PYANNOTE + WHISPER LARGE-V2), OUR SA-ASR SYSTEM, AND ASR WITHOUT SPEAKER CONSIDERATION.

In our study, a mixed-language dataset includes samples with multiple languages. Our dataset features English, German, Turkish, and Vietnamese. Each session involves two speakers discussing a scientific paper, one speaking in English and the other in one of the other languages. The dataset totals 45 minutes, with language distribution as follows: English (44%), German (9%), Turkish (12%), and Vietnamese (35%). The overlap rate is approximately 3%.

B. Modeling and Metric

All systems will be evaluated using the concatenated minimum permutation word error rate (cpWER) [33], depending on ASR performance and speaker labels. For the multilingual and mixed-language datasets, the baseline system is Diarization + ASR where diarization is Pyannote 3.1 [34] and ASR is Whisper large-v2. The baseline for the monolingual English dataset will vary between modular and joint systems, as outlined in section I.

Our MSA-ASR model employs Whisper large-v2 as the ASR component. The Speaker model has 12 layers for each encoder and decoder. The first $K = 1$ layers of the Speaker decoder use H^{asr} as the key. $\alpha = \beta = \gamma = 1$ for the EAD loss. $\theta = 0.7$ is used for grouping similar speaker turns. We train this model for 250,000 steps with batch size 80 (equivalent to 40 minutes of audio), using AdamW optimization with a learning rate 1e-4. We utilize spectral clustering [35] for speaker assignment. Our MSA-ASR model is only trained on the data that has been processed as described in section II-B without fine-tuning for the in-domain data (section III-A).

C. Results

Table II shows the benchmark result on the multi-talker Voxpopuli dataset. The first column is the results of using diarization followed by ASR. The second column shows the performance of our system. The third column is the baseline ASR performance, which does not account for speaker labels. This baseline ASR value is the lower bound WER, as the other systems incorporate speaker information into the ASR output. Overall, our system introduces a 29.3% relative error increase

TABLE III
COMPARISON OF CPWER (%) FOR LIBRICSS

System	Overlap ratio in %					
	0S	0L	10	20	30	40
LSTM SOT-SA-ASR [12]	10.3	15.8	13.4	17.1	24.4	28.6
Conformer SOT-SA-ASR [16]	8.6	12.7	11.2	11.3	16.1	17.5
TS-VAD + ASR [7]	9.5	11.0	16.1	23.1	33.8	40.9
NME-SC + SOT-SA-ASR [13]	9.0	12.2	**8.7**	**10.9**	**13.7**	**13.9**
MSA-ASR (our)	**7.5**	**8.1**	11.5	27.9	41.7	46.5

TABLE IV
COMPARISON OF CPWER (%) FOR AMI-IHM. ALL SYSTEMS USE GOLD VAD.

System	Dev	Eval
Transformer SOT-SA-ASR [8]	**14.5**	15.0
NME-SC + SOT-SA-ASR [13]	16.3	15.1
MSA-ASR (our)	15.6	**14.3**
Gold transcript + MSA-ASR (our)	2.7	1.9

System	en_de	en_tr	en_vi	avg
Diarization + ASR	6.81	15.76	18.54	13.70
MSA-ASR (our)	5.71	5.98	11.54	7.74

TABLE V
COMPARISON OF CPWER (%) FOR MIX-LANGUAGES MEETING DATASET.

over the baseline ASR, whereas the Diarization + ASR system increases it by 92%.

Although both systems perform well in English, for languages with large datasets (over 1,000 hours, detail in whisper paper [19]) used to fine-tune the ASR model, such as German, French, Spanish, Polish, Italian, Dutch, and Finnish, our system results in a 26% relative error increase, compared to a 120% increase with Diarization + ASR. For languages with smaller datasets, including Romanian, Hungarian, Czech, Croatian, Slovak, Slovenian, Estonian, and Lithuanian, our system introduces a 35% relative error increase, while Diarization + ASR results in a 76% increase. These results demonstrate that, compared to the language-independent diarization approach, using the same ASR model, our system more effectively leverages the generalizability of the ASR model to handle multilingual scenarios, particularly in languages with larger datasets.

Table III compares different systems on the LibriCSS dataset across various overlapping ratios. All systems, except TS-VAD, utilize a VAD model to segment long audio into smaller chunks. In our setting, we use Silero VAD [36]. The LSTM SOT SA-ASR [12] and Conformer SOT SA-ASR [16] systems are similar to ours but enhance the ASR model by incorporating speaker embeddings. TS-VAD is a target speaker voice activity detection system, which, in [7], is followed by an ASR model to transcribe the detected speaker. NME-SC is a diarization system that integrates Conformer SOT SA-ASR as in [13]. Our MSA-ASR system outperforms the others in scenarios without overlapping speech but shows decreased performance as the overlap ratio increases. This is expected, as our ASR model is frozen, whereas other systems are fine-tuned for this specific data type.

One of the significant advantages of our MSA-ASR model is its ability to run the Speaker model independently of the ASR model. This feature provides greater flexibility and efficiency in processing. Table IV presents benchmark results for various joint systems on the AMI-IHM dataset, using the ideal scenario where gold VAD labels are available. In this scenario, all systems focus solely on ASR and assigning speaker information. Our MSA-ASR model, even without fine-tuning on AMI-IHM, demonstrates competitive performance compared to other state-of-the-art models like the Transformer SOT SA-ASR [8] and NME-SC + SOT SA-ASR [13]. The final row of table IV highlights our model's unique capability to directly accept gold transcripts for assigning speaker embeddings, resulting in exceptional performance.

Table V highlights the performance of our system on real multilingual meeting data from a mixed-language dataset. The long audio recordings were segmented into smaller chunks using Silero VAD and then processed using our MSA-ASR model. Since Whisper large-v2 is a multilingual ASR model, it effectively manages multilingual audio. Compared to the Diarization + ASR system, our MSA-ASR model delivers significantly better results. Error analysis shows that while both ASR and diarization have acceptable error rates, averaging 7.67% and 7.38%, respectively, combining diarization with ASR leads to a higher cpWER than using a joint ASR and speaker model like our MSA-ASR.

Table II and table V highlight the advantages of our joint speaker attribute system, which effectively handles multilingual scenarios despite being fine-tuned only on an English dataset. While other joint systems often need fine-tuning for each specific language (due to data constraints) and are limited to those trained languages, our system demonstrates the ability to generalize and adapt to multiple languages without additional fine-tuning.

IV. CONCLUSION

This study introduces a novel approach for integrating ASR and Speaker models into a unified speaker-attribute speech recognition system, capable of handling multilingual datasets while using only standard monolingual ASR data. By fine-tuning exclusively on the speaker model, we preserve the original performance of the ASR model. Our system, although primarily optimized for non-overlapping data, also demonstrates robust performance across a range of diverse benchmarks. This highlights its capability to manage real-world scenarios with varying complexities effectively. We have made our pre-trained model and dataset publicly available for further research at hf.co/nguyenvulebinh/MSA-ASR.

V. ACKNOWLEDGMENT

The authors gratefully acknowledge support from Carl Zeiss Stiftung under the project Jung bleiben mit Robotern (P2019-01-002). This work was also partially supported by the European Union's Horizon research and innovation programme (grant No. 101135798, project Meetween), the Volkswagen Foundation project "How is AI Changing Science? Research in the Era of Learning Algorithms" (HiAICS), and KIT Campus Transfer GmbH (KCT) staff in accordance to the collaboration with Carnegie-AI.

REFERENCES

[1] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel, "Multimodal people id for a multimedia meeting browser," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, 1999, pp. 159–168.

[2] R. Gross, M. Bett, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Towards a multimodal meeting record," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, vol. 3. IEEE, 2000, pp. 1593–1596.

[3] R. Stiefelhagen, J. Yang, and A. Waibel, "Estimating focus of attention based on gaze and sound," in *Proceedings of the 2001 workshop on Perceptive user interfaces*, 2001, pp. 1–9.

[4] R. S, J. Yang, and A. Waibel, "Estimating focus of attention based on gaze and sound," in *Workshop on Perceptive User Interfaces*. Association for Computing Machinery, 2001.

[5] A. Waibel, H. Steusloff, R. Stiefelhagen *et al.*, "Chil: Computers in the human interaction loop," 2005.

[6] A. Waibel and C. Fuegen, "Simultaneous translation of open domain lectures and speeches," 2012, uS Patent 8,090,570.

[7] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *2021 IEEE spoken language technology workshop (SLT)*. IEEE, 2021, pp. 897–904.

[8] N. Kanda, X. Xiao, J. Wu, T. Zhou, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "A comparative study of modular and joint approaches for speaker-attributed asr on monaural long-form audio," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 296–303.

[9] C. Cui, I. A. Sheikh, M. Sadeghi, and E. Vincent, "Improving speaker assignment in speaker-attributed asr for real meeting applications," in *The Speaker and Language Recognition Workshop*, 2024.

[10] F. Yu, Z. Du, S. Zhang, Y. Lin, and L. Xie, "A Comparative Study on Speaker-attributed Automatic Speech Recognition in Multi-party Meetings," in *Proc. Interspeech 2022*, 2022.

[11] T.-B. Nguyen and A. Waibel, "Convoifilter: A case study of doing cocktail party speech recognition," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 565–569.

[12] N. Kanda, X. Chang, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Investigation of end-to-end speaker-attributed asr for continuous multi-talker recordings," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 809–816.

[13] N. Kanda, X. Xiao, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8082–8086.

[14] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, "Streaming speaker-attributed asr with token-level speaker embeddings," in *Interspeech*, 2022.

[15] Y. Li, F. Yu, Y. Liang, P. Guo, M. Shi, Z. Du, S. Zhang, and L. Xie, "Saparaformer: Non-autoregressive end-to-end speaker-attributed asr," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.

[16] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "End-to-End Speaker-Attributed ASR with Transformer," in *Proc. Interspeech 2021*, 2021, pp. 4413–4417.

[17] M. Shi, Z. Du, Q. Chen, F. Yu, Y. Li, S. Zhang, J. Zhang, and L.-R. Dai, "CASA-ASR: Context-Aware Speaker-Attributed ASR," in *Proc. INTERSPEECH 2023*, 2023, pp. 411–415.

[18] S. Cornell, J.-w. Jung, S. Watanabe, and S. Squartini, "One model to rule them all? towards end-to-end joint speaker diarization and speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 856–11 860.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[20] Q. Jin, T. Schultz, and A. Waibel, "Speaker identification using multilingual phone strings," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2002, pp. I–145.

[21] ——, "Phonetic speaker identification." in *INTERSPEECH*, 2002, pp. 1345–1348.

[22] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4996–5001.

[23] T.-B. Nguyen and A. Waibel, "Synthetic conversations improve multi-talker asr," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 461–10 465.

[24] M. Yang, N. Kanda, X. Wang, J. Wu, S. Sivasankaran, Z. Chen, J. Li, and T. Yoshioka, "Simulating realistic speech overlaps improves multi-talker asr," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[25] M. Lebourdais, M. Tahon, A. Laurent, S. Meignier, and A. Larcher, "Overlaps and gender analysis in the context of broadcast media," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 3264–3270.

[26] Abdullah, "Detecting double-talk (overlapping speech) in conversations using deep learning," 2017. [Online]. Available: https://publica.fraunhofer.de/handle/publica/281843

[27] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and counting overlapping speakers in distant speech scenarios," in *Interspeech*, 2020, pp. 3107–3111.

[28] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "A comparison of metric learning loss functions for end-to-end speaker verification," in *International Conference on Statistical Language and Speech Processing*. Springer, 2020, pp. 137–148.

[29] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8102–8106.

[30] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003.

[31] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[32] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7284–7288.

[33] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *CHiME 2020*, 2020.

[34] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023.

[35] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5239–5243.

[36] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," 2021.