Interspeech 2024
1-5 September 2024, Kos, Greece

Specializing Self-Supervised Speech Representations for Speaker Segmentation

Séverin Baroudi[1], *Thomas Pellegrini*[2], *Hervé Bredin*[2]

[1] Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon, France
[2] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
severin.baroudi@lis-lab.fr, thomas.pellegrini@irit.fr, herve.bredin@irit.fr

Abstract

Self-supervised speech representation learning has been shown to be very effective for a wide range of speech processing downstream tasks. However, most of these models have been pretrained using clean pre-segmented single-speaker utterances, which is not representative of tasks involving realistic multi-speaker conversational speech like speaker diarization. WavLM pre-training mitigates this domain mismatch using artificial mixtures of single-speaker utterances, and outperforms other pretrained models such as wav2vec2 or HuBERT for speaker diarization. We propose to further specialize WavLM for speaker diarization in two ways: pre-training on real-world multi-speaker conversational speech, and crafting targets of pre-training pretext task to benefit the most to speaker diarization. When finetuned with recently proposed powerset multi-class cross entropy loss, we outperform, often by a large margin, the state-of-the-art on most speaker diarization benchmarks.
Index Terms: Speaker diarization, self-supervised representation learning

1. Introduction

Self-supervised learning (SSL) of speech representations has shown great promises throughout recent years. Contrastive models such as wav2vec [1] and wav2vec 2.0 [2] layed the foundations towards improving many speech processing downstream tasks. Drawing inspiration from natural language processing, models such as HuBERT [3] and WavLM [4] employ clustering and masked prediction of discrete hidden units to acquire meaningful audio representations. These models were shown to outperform previous SSL speech models in a number of downstream tasks, in particular in tasks related to speaker identity. They are further introduced in Section 2.

We focus on the interaction between SSL and speaker diarization, the task of partitioning the recording of a multi-speaker conversation into segments based on speaker identity. Conducting an initial experiment on DIHARD 3 [5] in Section 3, we start with a strong baseline system utilizing a trainable SincNet feature extractor [6] which is outperformed by off-the-shelf SSL models such as wav2vec 2.0, HuBERT, or WavLM. Our findings also corroborate prior literature, highlighting WavLM as the most effective SSL model among them. One distinguishing feature of WavLM, in contrast to HuBERT, is its use of the utterance mixing training strategy during pre-training. While this strategy – simulating noisy/overlapped speech – may explain why WavLM excels in modeling speaker characteristics compared to its counterparts, we hypothesize that artificial mixtures still fall short of real-world conversational speech recordings.

Section 4 describes our primary contribution that addresses this domain mismatch by constructing a multi-speaker, multi-domain, and multi-lingual conversational dataset for pre-training a randomly initialized WavLM Base architecture. When evaluated on DIHARD 3 and other major speaker diarization benchmarks, our proposed *conversational WavLM* significantly outperforms the off-the-shelf version.

Section 5 summarizes our second contribution towards optimizing this training strategy through a second pre-training iteration, based on a prior layer-wise performance analysis of WavLM. Focusing on the layer that contributes the most to diarization, we generate targets/pseudo-labels from representations that hold significance for our task, aiming to refine WavLM by emphasizing hidden units more closely related to speaker-related content, as opposed to phonetic content. This iterative approach leads to new state-of-the-art results across most benchmarks, with an averaged diarization error rate of 17.1% – a significant improvement from a previous 20.1% rate [7].

2. Overview of HuBERT and WavLM

The objective of employing a pretext task in pre-training self-supervised learning speech models is to facilitate an understanding of the fundamental structure of speech. An SSL model should demonstrate the capability to: i) discern a list of representative speech events and ii) predict events from masked segments of speech. Based on this ideal goal, both HuBERT and WavLM employ clustering, in order to discover speech events in an unsupervised way. The initial generation of hidden units is obtained using Mel frequency cepstral coefficients (MFCC). Subsequently, speech data is annotated with frame-level pseudo-labels.

Major SSL models consist of a trainable Convolutional Neural Network (CNN), followed by a Transformer network, and a concluding linear layer responsible for predicting pseudo-labels. In downstream applications, including automatic speech recognition (and speaker diarization in our case), the final layer is omitted. The outputs from one or more layers of the Transformer serve as speech representations inputted into a network tailored for the specific application. The Transformer outputs are particularly valuable due to the Transformer's capacity to learn contextualized representations through self-attention. One point of distinction between HuBERT and WavLM is the utilization of a gated relative position bias in WavLM, as opposed to a convolutional relative position embedding in HuBERT.

The optimization objective during the pre-training phase is a mask prediction loss, defined through cross-entropy. Following the initial pre-training step, referred to as the first iteration, HuBERT exhibits the ability to generate features richer than MFCCs within its Transformer layers. Consequently, a second

10.21437/Interspeech.2024-962

generation of targets is created using an intermediate layer of the first iteration of the HuBERT model, incorporating an increased number of clusters.

WavLM capitalizes on the features produced by HuBERT, bypassing the initial MFCC-based pre-training iteration and opting for a single pre-training step. As mentioned in the introduction, it attempts to emphasize multi-speaker content through the implementation of an utterance mixing strategy applied to the input audio. This involves selecting two utterances from a batch and overlapping them using diverse mixing techniques. The model is then trained to predict denoised targets, *i.e.* pseudo-labels generated from the original utterance (the one before mixing).

3. Speaker segmentation with pretrained self-supervised speech representations

The starting point of our study is the recently proposed hybrid approach based on *powerset* speaker segmentation model [7]. The latter ingests 10s audio chunks, performs frame-wise feature extraction with a trainable *SincNet* [6], processes the sequence of frames with four bi-directional LSTM layers (2.1 M parameters), and performs multi-class classification using two hidden linear layers (of hidden size 128), followed by a third one (for a total of 50k parameters) leading into the *powerset* space (one class for non-speech, one class for each single speakers, and one for each pair of overlapping speakers). At inference time, this model slides over the whole audio recording and a subsequent clustering step (based on speaker embedding) stitches the local diarization results of the short windows into a coherent whole [8].

We propose to study the impact of replacing the *SincNet* feature extraction blocks by several pretrained (and frozen) self-supervised speech representation models from the literature, in their Base version (*i.e.,* Transformers with 12 layers). We use a learnable weighted average of the representations extracted by each layer as replacement of the *SincNet* block. We denote α_i the weight assigned to i^{th} transformer layer (with $\sum \alpha_i = 1$). In all cases, we train the speaker segmentation model on 10-s audio chunks, with the reasonable assumption that at most 3 speakers can be active in each chunk.

For this initial study, we only focus on the improvement of the speaker segmentation step, and not the whole speaker diarization pipeline. More precisely, we train the original and modified segmentation models using approximately 77 % of the development set of DIHARD 3 dataset [5] and evaluate their average performance as diarization error rate (DER) on 10s audio chunks sampled from the remaining 23 % only. The reported numbers in Table 1 shall therefore not be confused with standard diarization error rates on DIHARD 3 test set (which will be reported later in the paper). The curious reader can safely ignore the lower part of the table for now (we will also come back to those numbers later).

As expected, replacing *SincNet* by any of the aforementioned pretrained models improves the performance of the speaker segmentation model significantly (almost 10% relative), mostly on the false alarm component of the diarization error rate. This indicates that the segmentation model is getting better at detecting when a speaker is active. WavLM is the clear winner thanks to better missed detection (compared to HuBERT) and false alarm (compared to wav2vec2.0). On the basis of this preliminary study, we focus on pushing WavLM improvement even further throughout the rest of the paper.

Table 1: *Impact of off-the-shelf (upper part) and proposed (lower part) self-supervised speech representation models on speaker segmentation. 10s DER: average diarization error rate on 10s audio chunks sampled from DIHARD 3 development set, FA: false alarm rate, MD: missed detection rate, SC: speaker confusion rate.*

Feature extraction	FA%	MD%	SC%	10s DER%
SincNet [6]	6.1	8.1	4.9	19.3
wav2vec2.0 [2]	4.8	8.1	4.6	17.6
HuBERT [3]	4.3	8.6	4.7	17.7
WavLM [4]	4.4	8.3	4.5	17.3
Conversational WavLM #1	4.1	7.2	2.5	13.9
Conversational WavLM #2	4.4	6.8	2.4	13.7

4. Pre-training with conversational speech

As suggested in [4], the main reason why WavLM gets better results on the speaker diarization downstream task is the use of utterance-level mixing strategy that supposedly familiarizes the model with multi-speaker conversational overlapping speech.

Yet, this augmentation technique remains very artificial and only leads to an imperfect simulation of the conditions of use of the model in realistic multi-speaker conversational speech:

- training utterances contain mostly speech because self-supervised representation techniques were first introduced for automatic speech recognition and assumed that voice activity detection (VAD) was performed before ASR at inference time — this is not the case for end-to-end speaker diarization models that are supposed to take care of VAD;
- before mixing, training utterances usually do not contain speaker changes because they are coming from audiobooks – which does not prepare the model for inference on multi-speaker audio chunks;
- two utterances used for creating training mixtures are sampled from two different audiobooks, and are therefore recorded in two different acoustic environments – this is not so frequent for multi-speaker conversational speech where both speakers are usually in the same room.

Therefore, our second contribution is to pre-train, from scratch, a WavLM model on a large and diverse un-segmented multi-speaker conversational speech dataset.

4.1. Conversational speech dataset

We assembled such a multi-domain multi-lingual dataset by gathering many speaker diarization datasets. Depicted in Figure 1, they cover a large variety of domains, such as daily-life activities [9, 10], meetings [11, 12, 13], or TV broadcasts [14]. The majority of this compound conversational dataset is in English (30%) or Mandarin (35%) but other languages are included such as French [14], German, or Spanish [15] to name a few. In total, the proposed dataset contains 663 hours of audio, 27% (resp. 11%) of which is non-speech (resp. overlapping speech), to be compared to 960 hours of Librispeech [16] (that is made almost entirely of speech and contains no overlapping speech) used for pre-training the off-the-shelf WavLM Base model [4].

4.2. Pre-training

We follow the exact same training recipe as the official off-the-shelf WavLM Base model [4]. As suggested earlier, the only

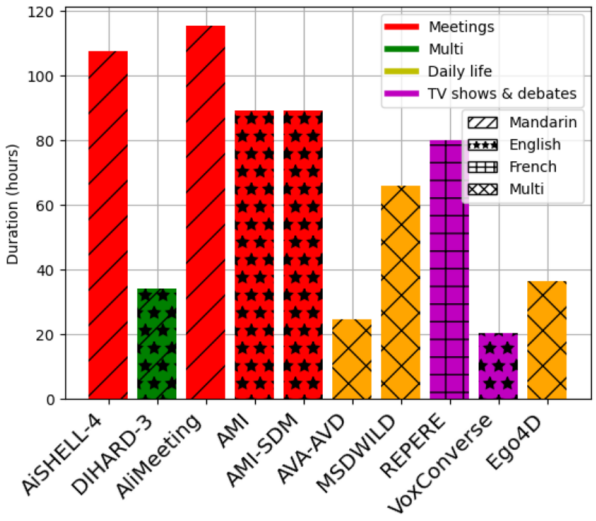

Figure 1: *Conversational speech dataset used for pre-training from scratch the proposed Conversational WavLM. "Multi" refers to datasets that contain more than a single domain or language.*

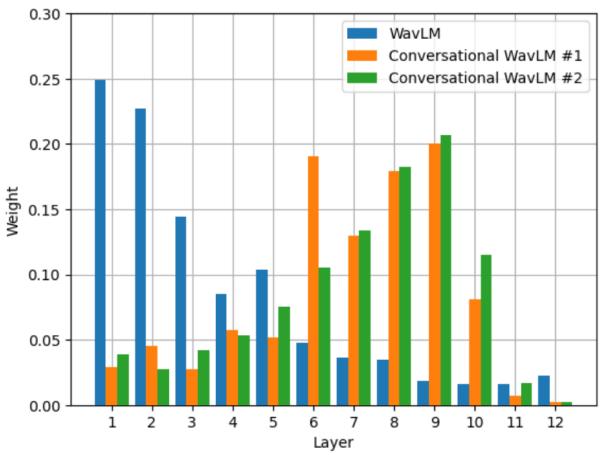

Figure 2: *Layer-wise analysis of the off-the-shelf WavLM and both iterations of our conversational WavLM for Speaker Diarization.*

difference lies in the training set and its preparation. More precisely, we use the aforementioned conversational dataset in place of 960 hours from *Librispeech*, and we split it into 10-s-long training chunks without any kind of pre-segmentation or constraints on the number of speakers nor on the amount of non-speech they contain. Note that we also stick with the exact same utterance mixing augmentation strategy as the one used in the original WavLM Base model, despite the fact that training chunks may already contain multiple, possibly overlapping, speakers.

4.3. Conversational WavLM #1

Like for off-the-shelf Base models in Section 3, we evaluate its performance for the downstream task of speaker diarization of 10-s chunks sampled from DIHARD 3 development set. Table 1 shows that our proposed conversational WavLM #1 significantly outperforms its off-the-shelf counterpart, bringing a 20% relative decrease in diarization error rate.

While decreasing missed detection rate usually mechanically implies increasing speaker confusion rate, conversational WavLM #1 manages the *tour de force* of improving on every front — false alarm, missed detection, and speaker confusion by almost 50% relative, showing that it is much more capable of discriminating speakers.

5. Crafting targets for speaker diarization

Following the original WavLM recipe [4], our proposed conversational WavLM #1 is trained using targets obtained through clustering, k-means with 500 clusters, on the output frames of the fifth Transformer layer of the off-the-shelf HuBERT Base model. Going against the bandwagon of training self-supervised representations that generalize for many downstream tasks, we propose to study whether one can specialize pre-training for the speaker diarization downstream task, by selecting a more suitable Transformer layer for targets.

5.1. Layer-wise analysis

Using Fig. 2, we start by analyzing the distribution of weights α_i, introduced in Section 3, assigned to Transformer layers to evaluate their contribution to our downstream task. On the one hand, the off-the-shelf WavLM (in blue) has its first layers contributing the most to the diarization downstream task. This is consistent with what the WavLM authors [4] observed on the *SUPERB* benchmark [17], according to which downstream tasks related to phonetic content mainly activate higher layers, while speaker-related tasks activate lower layers. On the other hand, the distribution of conversational WavLM #1 layer weights (in orange) is shifted towards higher layers and mostly activates middle-to-higher layers (between 6 and 9), indicating that a a larger part of the model actually contributes to speaker diarization. By switching from audiobook to conversational speech, the model has already been specialized for this task.

5.2. Conversational WavLM #2

Our third contribution involves a further attempt to fine-tune the pre-trained model by generating targets using the output layer 8 of conversational WavLM #1. We then proceed to retrain the entire model from scratch, with the expectation that these targets will prove more beneficial for the diarization downstream task.

Compared to the usual WavLM Base training recipe, we increased the number of training steps to 500k in total, as we witnessed a linear increase in the performance of our WavLM for diarization during the first iteration, and also that training this second model longer would improve the results even further. The learning rate increases for the first 12% steps, before decreasing as is done in the first iteration. We also increased the total amount of seconds that each GPU can hold, from 87.5 to 135 seconds. All our models were trained on 32 GPUs V100-32GB.

The resulting model is called conversational WavLM #2, and its results are reported in Table 1. It offers slightly better performance than WavLM #1. The difference mainly comes from a better balance between false alarm and missed detection rates — and once again, a decrease in both speaker confusion

Table 2: *Evaluation of off-the-shelf (upper part) and proposed (lower part) self-supervised speech representation models on an entire diarization pipeline, using test sets from various diarization corpuses using ResNet-34 pretrained speaker embedding model. "Finetuned Conversational WavLM #2" refers to Conversational WavLM #2 in which ONLY the segmentation part has been finetuned and clustering thresholds have been optimized on each dataset. DER: Diarization Error Rate computed over whole audio files.*

Feature Extraction	AISHELL-4	AliMeeting	AMI	AMI-SDM	AVA-AVD	DIHARD 3	MSDWILD	REPERE	VoxConverse	Average
SincNet [6]	12.3	24.3	19.0	22.2	49.0	21.6	24.5	7.8	11.2	21.3
WavLM [4]	13.4	22.8	16.6	20.2	50.2	19.4	23.7	9.5	11.2	20.8
Conversational WavLM #1	12.8	19.0	17.0	19.9	44.5	17.8	21.6	7.6	10.0	18.9
Conversational WavLM #2	13.5	18.8	15.7	**18.4**	44.9	17.2	20.8	7.4	9.9	18.5
Finetuned Conversational WavLM #2	**10.6**	**18.4**	**14.8**	18.9	**39.9**	**16.7**	**19.6**	**6.8**	8.5	**17.1**
State-of-the-art (Dec. 2023)	13.2 [7]	23.3 [7]	16.9 [18]	19.5 [18]	46.4 [7]	16.8 [18]	27.1 [7]	8.2 [7]	**6.1** [19]	20.1

and missed detection rates. The second iteration is depicted in green in Figure 2: the distribution of layer weights is shifted even more towards higher layers.

6. Benchmark results

To further investigate the performance of our proposed conversational WavLM, we integrate them into a full speaker diarization pipeline, following the paradigm described in [8]. It falls into the category of hybrid systems that integrate local supervised speaker segmentation models with global unsupervised clustering [20, 21, 7]. Long-form conversations are first divided into 10s overlapping audio chunks, each processed by the considered speaker segmentation model independently. Then, ResNet34 speaker embeddings are computed with Wespeaker open source toolkit [22] for each active speaker of each audio chunk based, and clustered to yield the final diarized segments for the entire conversation. For conversational WavLM #1 and #2, the clustering threshold is tuned globally on the development set of the whole conversation dataset (to minimize diarization error rate), and the resulting pipeline is applied directly to each test set. *Finetuned conversational WavLM #2* goes one step further and finetunes the speaker segmentation model using the training set of each benchmark separately, as well as optimizes the clustering threshold using their respective development set.

Table 2 shows that WavLM leads to an overall better average performance than SincNet. Yet, some datasets (such as AVA-AVD, AISHELL-4 or REPERE) still put SincNet on top, highlighting that representations extracted from off-the-shelf WavLM models do not always perform better than the baseline. On the other hand, proposed conversational WavLM #1 outperforms the SincNet baseline by 13% relative on average (with the notable exception of AISHELL-4). When compared to off-the-shelf WavLM, DERs are consistently lower for every dataset (except for AMI), showing that proposed conversational WavLM provides representations more suitable to speaker diarization.

Switching to targets that contribute to diarization the most, conversational WavLM #2 manages to lower DERs significantly on some datasets that previously showed little to no improvement over conversational WavLM #1 (or even SincNet) – mainly AMI and AMI-SDM. DIHARD 3 is one of the datasets that benefit the most from the second iteration, with a relative performance increase of 26% over SincNet and 11% over off-the-shelf WavLM. This demonstrates that our primary goal of mitigating domain mismatch faced by off-the-shelf models has been achieved. Finally, finetuning the segmentation model and adapting the clustering threshold to each dataset, we set a new state-of-the-art on almost every benchmark (AISHELL-4, AliMeeting, AMI, AMI-SDM, AVA-AVD, DIHARD 3, MSDWILD and REPERE) without resorting to a fusion of several systems that is often used in the literature.

7. Conclusions

In this paper, we investigated the use of self-supervised speech representations for the task of speaker diarization. We first conducted a performance study over three prominent SSL models (wav2vec2.0, HuBERT and WavLM) pretrained on librispeech-960h, before questioning their direct contribution towards diarization. While WavLM relies on artificial mixtures to approach real conversational recordings, we demonstrate that directly pre-training the model on an assembled conversational dataset made of real-world mixtures managed to significantly improve diarization performance. Finally, we specialized our conversational WavLM by generating targets derived from layers that are efficient for diarization, based on a layer-wise study. By doing so, we obtained state-of-the-art performance on AiSHELL-4, AliMeeting, AMI, AMI-SDM, AVA-AVD, DIHARD 3, MSDWILD and REPERE. Future research directions include:

- Corpus expansion: while our assembled conversational corpus derives from diarization datasets only, adding more hours of multi-speaker content while still keeping diversity across domains and languages might enhance overall performance further.
- Denoising process understanding: although the utterance mixing strategy was crucial for simulating multi-speaker content in WavLM, it is not clear how the denoising process operates when both multi-speaker content and utterance mixing are applied simultaneously, as it is the case when pre-training our conversational WavLMs.
- Further pre-training iterations: specializing targets in Conversational WavLM #1 for diarization improved the performance of Conversational WavLM #2. Exploring additional pre-training iterations using the identified optimal layer may enhance the contribution of transformer layers, particularly in later stages, potentially improving WavLM's effectiveness for speaker diarization.

8. Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation AD011014274R1 made by GENCI, and supported by the Agence de l'Innovation Défense under the grant number 2022 65 0079.

9. References

[1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," 09 2019, pp. 3465–3469.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1–14, 10 2022.

[5] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third DIHARD diarization challenge," in *Interspeech 2021*. ISCA, pp. 3570–3574. [Online]. Available: https://www.isca-archive.org/interspeech_2021/ryant21_interspeech.html

[6] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[7] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, 2023, pp. 3222–3226.

[8] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023, pp. 1983–1987.

[9] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the World in 3,000 Hours of Egocentric Video," in *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] T. Liu, S. Fan, X. Xiang, H. Song, S. Lin, J. Sun, T. Han, S. Chen, B. Yao, S. Liu, Y. Wu, Y. Qian, and K. Yu, "MSDWild: Multimodal Speaker Diarization Dataset in the Wild," in *Proc. Interspeech 2022*, 2022, pp. 1476–1480.

[11] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario." [Online]. Available: http://arxiv.org/abs/2104.03603

[12] AMI corpus overview. [Online]. Available: https://groups.inf.ed.ac.uk/ami/corpus/overview.shtml

[13] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, "M2met: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," number: arXiv:2110.07393. [Online]. Available: http://arxiv.org/abs/2110.07393

[14] J. Kahn, O. Galibert, L. Quintard, M. Carre, A. Giraudel, and P. Joly, "A presentation of the REPERE challenge," in *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/6269851/

[15] E. Z. Xu, Z. Song, S. Tsutsui, C. Feng, M. Ye, and M. Z. Shou, "AVA-AVD: Audio-visual speaker diarization in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3838–3847. [Online]. Available: http://arxiv.org/abs/2111.14448

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5206–5210. [Online]. Available: http://ieeexplore.ieee.org/document/7178964/

[17] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[18] M.-K. He, J. Du, Q.-F. Liu, and C.-H. Lee, "Ansd-ma-mse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1561–1573, 2023.

[19] F. Landini, M. Diez, T. Stafylakis, and L. Burget, "DiaPer: End-to-end neural diarization with perceiver-based attractors." [Online]. Available: http://arxiv.org/abs/2312.04324

[20] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7198–7202.

[21] Kinoshita, Keisuke and Delcroix, Marc and Tawara, Naohiro, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," 05 2021.

[22] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.