Interspeech 2024
1-5 September 2024, Kos, Greece

On the calibration of powerset speaker diarization models

Alexis Plaquet, Hervé Bredin

IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
firstname.lastname@irit.fr

Abstract

End-to-end neural diarization models have usually relied on a multilabel-classification formulation of the speaker diarization problem. Recently, we proposed a powerset multiclass formulation that has beaten the state-of-the-art on multiple datasets. In this paper, we propose to study the calibration of a powerset speaker diarization model, and explore some of its uses. We study the calibration in-domain, as well as out-of-domain, and explore the data in low-confidence regions. The reliability of model confidence is then tested in practice: we use the confidence of the pretrained model to selectively create training and validation subsets out of unannotated data, and compare this to random selection. We find that top-label confidence can be used to reliably predict high-error regions. Moreover, training on low-confidence regions provides a better calibrated model, and validating on low-confidence regions can be more annotation-efficient than random regions.

Index Terms: speaker diarization, calibration, powerset classification

1. Introduction

The speaker diarization task can be defined as taking an audio excerpt and answering the question "who spoke when?", without concern for the exact identities of the speakers. Solving this task provides the exact beginning and end of each speaker turn, which proves very useful when combined with other tasks output such as transcribed text.

Classifiers usually provide some notion of "confidence" along with the predicted output. In an ideal world, confidence would always be linked to the epistemic or aleatoric uncertainty. However, deep learning classifiers are famously overconfident and predict high probabilities for unknown classes and classes where the model is wrong [1]. Despite these limitations, model output probabilities are still one of the only tools available to estimate uncertainty contained in the predictions of deep learning models. This has led to a number of different usages: out-of-domain detection [2], semi-supervised learning [3], or active learning.

Research on the calibration of End-to-end Neural Diarization (EEND) models and its application is lacking. The goal of this paper is to study the calibration of the powerset speaker diarization model proposed in [4], and to use model confidence to select data of interest for training and validation purposes. We study in detail the calibration of the model on in-domain and out-of-domain data, and observe what kind of data is represented in low confidence predictions. Moreover, we study the selection of low-data validation and training set with confidence-based strategies.

2. Model calibration

In a multiclass setting, a model is deemed "top-label-calibrated" if the maximum of its output scores (the score of the predicted class) is equal to the probability of it being the correct prediction. For example, a top-label-calibrated model outputting probabilities $(c_1, c_2, c_3) = (0.7, 0.1, 0.2)$ has a 70% chance of being correct about c_1 being the correct class. Stricter definitions of calibration exist, such as classwise-calibrated [5] and jointly-calibrated [6], but they are out of the scope of this paper.

There is no easy way to guarantee a degree of model cal-

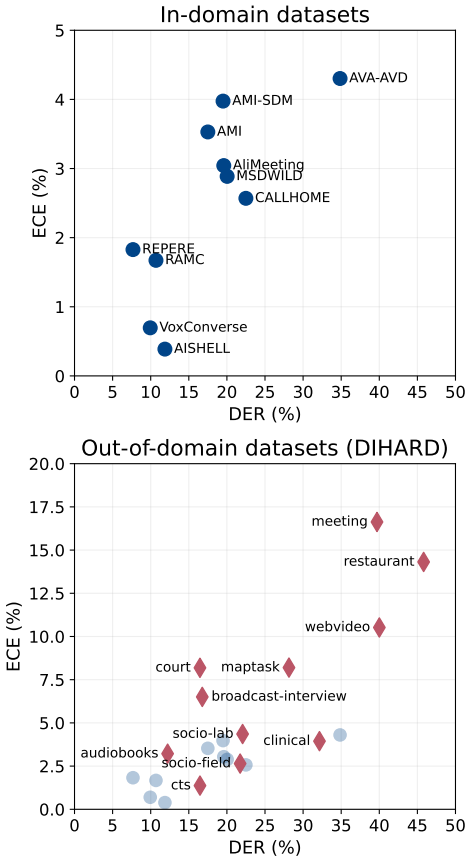

Figure 1: *Calibration error as a function of the DER of the powerset segmentation model. In-domain datasets (top figure) are plotted with blue circles, out-of-domain datasets (bottom figure) are plotted with red diamonds. To give a frame of reference, the blue circles of in-domain datasets are also overlayed in transparency on the bottom figure.*

10.21437/Interspeech.2024-1060

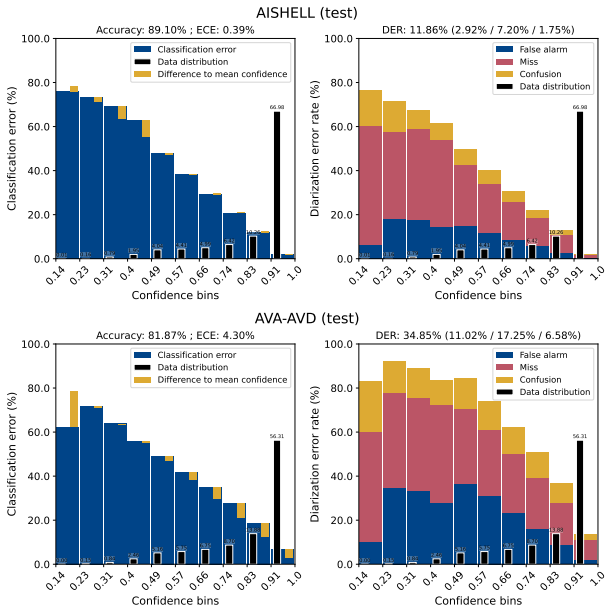

Figure 2: *Best and worst in-domain calibration, measured by ECE. The left column is a classical reliability diagram, a perfect ECE would mean no "difference to mean confidence" in every bin, resulting in a diagonal plot. The right column is the same plot but with DER instead of classification error.*

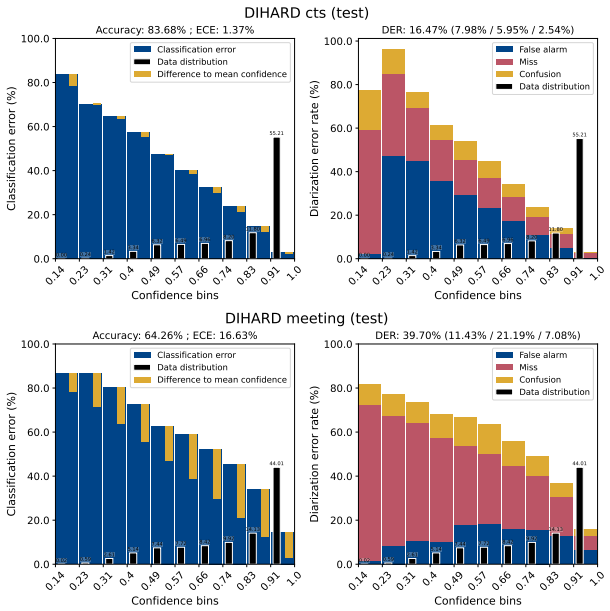

Figure 3: *Reliability diagram and binwise DER distributions for the best and worst calibrated domains in DIHARD.*

ibration in deep learning. Models are not designed to achieve top-label-calibration, and tend to be naturally overconfident [1]. There are two main ways to approach top-label-calibration: encourage better calibration during training, and post-hoc calibration methods [7]. Training-time calibration methods are diverse and encompass regularization methods [8] or loss modifications [9]. Simple post-hoc calibration methods work well [1], but are very sensitive to data shift: calibrating on a domain will only help on data that is part of, or very close to this domain [10]. This means that it is extremely difficult to design a model that is calibrated on any data it encounters: "top-label-calibration" is not free and has huge annotation cost.

We study the "top-label-calibration" of the local speaker diarization model proposed in [4]. To do so, we trained a standard powerset PyanNet model with classes for at most 2 simultaneous speakers, and up to 3 distinct speakers in 5s chunks. The model is trained until convergence after 89 hours of training on a single NVIDIA V100 GPU. We rely on a compound training dataset made of the concatenation of the training subsets of AISHELL-4 [11], AliMeeting [12], AMI [13], AVA-AVD [14], CallHome [15], Displace [16], Ego4D [17], MSDWild [18], MagicData-RAMC [19], REPERE [20], and VoxConverse [21]. Any of the test subsets from the compound dataset is considered as "in-domain". We use the 11 distinct sub-domains of DIHARD 3 [22] dataset as "out-of-domain" datasets (*i.e.* that have never been seen during training).

In the spirit of reproducible research, all relevant model checkpoints, metrics, metadata and code are available at `github.com/FrenchKrab/IS2024-powerset-calibration`.

2.1. Metrics

In-domain performance is assessed with two metrics: diarization error rate (DER) and expected calibration error (ECE).

To compute ECE, model predictions are grouped by their confidence into different bins. As the powerset model outputs softmax probabilities, we define the confidence as the probability of the predicted class. ECE is defined as:

$$\mathrm{ECE} = \sum_{i=0}^{B} \mathrm{prop}(b_i) \cdot |\mathrm{acc}(b_i) - \mathrm{conf}(b_i)|$$

Where $\mathrm{prop}(b_i)$ is the proportion of predictions in bin b_i, and $\mathrm{acc}(b_i)$ and $\mathrm{conf}(b_i)$ the average accuracy and confidence in b_i. Multiple binning schemes and distances $|\mathrm{acc}(b_i) - \mathrm{conf}(b_i)|$ can be used. In our experiments and figures, we use $N = 10$ bins uniformly distributed in $[\frac{1}{\text{class count}}; 1]$, and the L1 distance. We experimented with Adaptive ECE (where all bins contain the same number of samples), and varied the bin size from 10 to 20, but we did not find any meaningful differences and hence we do not report the different variants.

The DER is a standard speaker diarization metric defined as

$$\mathrm{DER} = \frac{\text{False alarm} + \text{Missed detection} + \text{Speaker confusion}}{\text{Total speech}}$$

It is commonly expressed in percentages but can go over 100% as false alarm can exceed the total duration of speech.

Since our focus is the local segmentation model working on 5 seconds chunks only, we disregard the usual subsequent steps of the diarization pipeline (embedding extraction and clustering) [23]. Metrics are directly computed on the (sliding) outputs of the local segmentation model after each window has been aligned with the reference. We call this "local DER" and it should not be compared to the values of DER that are usually reported in the literature.

2.2. In-domain and out-of-domain calibration

The ECE and DER on the test subsets of all datasets are shown in Figure 1. We observe some correlation between DER and ECE: domains with higher DER tend to have higher ECE as

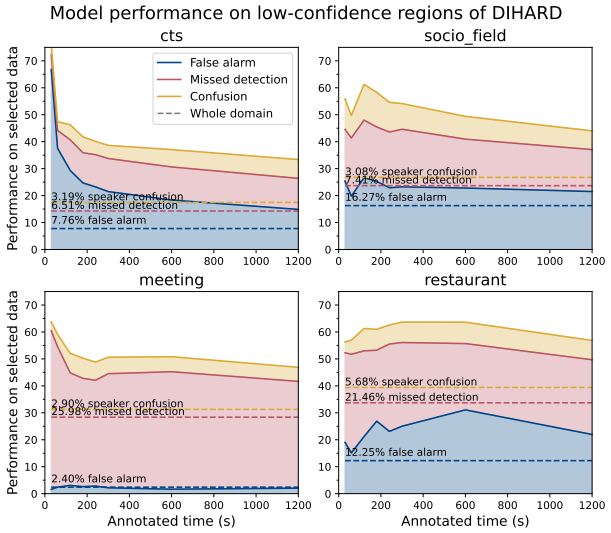

Figure 4: *Composition of the diarization error rate when sampling 5 seconds chunks, lowest confidence chunks are selected first. The dashed lines show the composition of the DER on the whole test set.*

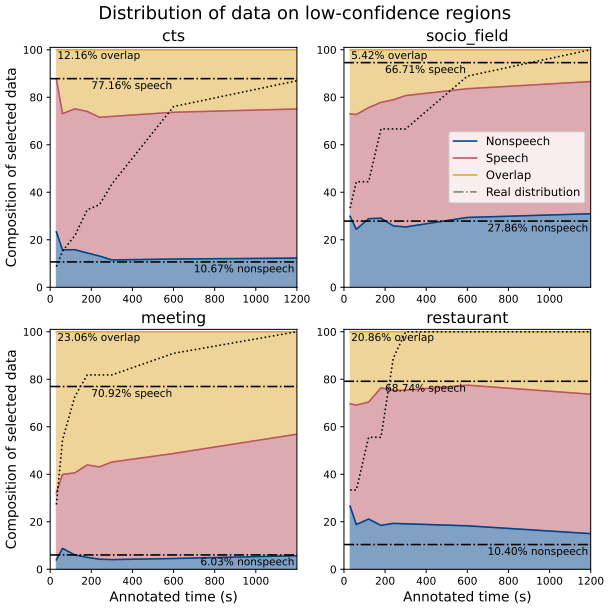

Figure 5: *Composition of the data categorized in nonspeech, speech and overlap, when sampling 5 seconds chunks like in Figure 4. Dashed lines show the average distribution on the whole test set.*

well. On in-domain datasets, the ECE goes up to 4.3% on AVA-AVD with a DER of 35%, the worst performing dataset DER-wise and ECE-wise. However, most out-of-domain datasets are badly calibrated, only 4 out of 11 are under 4.3% ECE. And, unsurprisingly, the DER is also worse on out-of-domain datasets.

The left row of Figure 2 shows reliability diagrams on a couple of in-domain datasets. On AISHELL, the model has nearly perfect calibration (the best out of all tested in-domain datasets): the model confidence matches its average accuracy very well. Even though AVA-AVD is the in-domain dataset where the model has the worst calibration, it is still fairly low with only 4.30%. However, we can see a trend where the mis-calibration comes mainly from model overconfidence: the model makes too many errors in high confidence bins.

Figure 3 shows the same plot for out-of-domain datasets. The "meeting" plot displays how the model is overconfident in all bins, which is usually the expected pattern in deep neural networks. We observe a very high "missed detection" rate, which suggests that the model is not sensitive enough to detect the speech and fails to reflect any uncertainty in its output probabilities. The distribution of errors (false alarm, missed detection, speaker confusion) does not depend on the ECE but rather on the mismatch between the domain and the pretrained model's training data.

2.3. Analysis of low-confidence regions

Previous figures provide insights into *framewise* calibration and confidence. However one of the main challenges of speaker diarization is the temporal aspect of the prediction. While in image classification it makes sense to discard or select individual data samples, in speaker diarization it does not always make sense to only keep or select individual frames of data. This is especially true in use-cases involving human annotators, as they need the preceding audio context to make sense of a specific frame. Consequently, in the following figures all selected data is made of continuous chunks of at least 7.5 seconds. We select in priority regions where the average confidence is the lowest (i.e "low-confidence regions") and study model performance and class distribution on them.

In Figure 4, we look at the two best (CTS, Socio-field) and worst (Meeting, Restaurant) calibrated domains. The figure shows the diarization error rate when selecting x seconds of data. In both CTS and socio-field, the DER in low-confidence regions for $x \in [0, 1200]$ seconds is significantly higher than the DER computed on the whole domain. In Meeting and Restaurant, this difference is less significant. Nonetheless, even in the worst calibrated dataset, DER on low-confidence regions stays significantly higher than the average on the domain. This observation is encouraging, as it appears that even when the model is badly calibrated on a domain, low-confidence still strongly correlate with low performance.

Figure 5 shows input data distribution in the same datasets on the same low-confidence regions. We divide the data into three categories: "nonspeech" when no speaker is active, "speech" when only one speaker is active, and "overlap"' when at least two speakers are active simultaneously. In both well- and badly-calibrated datasets, we observe that low-confidence data contains significantly more overlap than the rest of the classes. This can be expected as overlapped speech is difficult, and remains one of the main sources of error in speaker diarization.

3. Annotation-efficient domain adaptation

Although speaker diarization models are getting better and better, applying a pretrained state-of-the-art speaker diarization model on unseen data might not output what we expect. This could be because the model has not yet seen or generalized to this specific kind of data: acoustic conditions, speech type, language. Or it could simply be because annotation standards vary a lot between datasets [23], and the ones learnt by the model might not fit the needs of the user. In any case, if one wants

optimal performance on a new dataset, one will need to specify what is expected from the model. This usually means annotating (at least some of) the data, which is a very costly process, requiring up to dozens of person-hours to label a few hours of audio.

In this part, we borrow from active learning the idea of focusing the human annotation effort to low-confidence regions of the data. This choice is motivated by the results obtained in subsection 2.3 where we observe a high DER on low-confidence regions.

We treat DIHARD 3 eleven distinct datasets as "unlabeled" for these experiments (artificially withholding annotations). We simulate the human annotation process with an "oracle labeler" which provides the withheld annotations when requested. Our goal is to improve the DER while using as little oracle annotations as possible.

3.1. Finding a minimal training subset

In active learning the learning process is usually iterative, but here we limit the research to a single iteration: we select the relevant data to annotate, label the (withheld) selected regions, retrain the model on this new data, and finally evaluate the model performance. The selected regions are 7.5 seconds long to mirror the process described in subsection 2.3. We test two ways to select the regions to label:

- Random: the data used to train the model is selected at random (our baseline).
- Worst confidence: the regions where the average confidence of the model is the lowest are selected (like in subsection 2.3).

We can make some interesting observations on the results that are summarized in Figure 6. First, for all domains but webvideo, 30 seconds of training data is enough to significantly improve the DER. The improvement can be quite important indomains like socio-lab or court. Webvideo behaves this way probably because it is not a homogeneous domain, but a collection of heterogeneous YouTube videos, hence the need for more data. More importantly, we can observe that the querying strategy does not have a strong impact on the DER either way. At equal annotation budget, we can expect a similar DER. However, the confidence-based selection obtains better ECE on almost all domains.

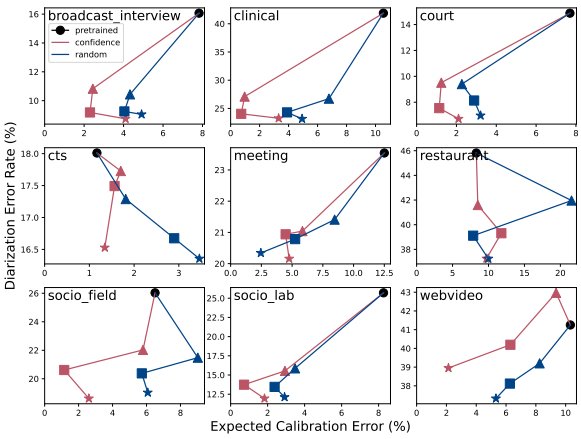

Figure 6: *Evolution of DER and ECE when varying the size of the training dataset. Each marker corresponds to a dataset duration: ▲=30s, ■=2min30s, ★=20min.*

3.2. Finding a minimal validation subset

One limitation of retraining the model for domain adaptation is the need for a validation set. Without a validation set, we might choose a checkpoint where the model has overfitted, which is very likely on low amounts of data.

However, we also need the validation set to be as small as possible if we want the model retraining to effectively be annotation-efficient (as the last subsection used oracle validation). We need to find the smallest validation subset that selects the same best checkpoint as the full validation set.

This task proves to be difficult, even if we allow suboptimal selection of checkpoint (i.e. we allow a small relative difference in DER between the best checkpoint selected from a small validation set, and the real best checkpoint selected from the full validation set).

To evaluate it, we use a fixed set of checkpoints, compute the DER on validation subsets of various lengths, using random and low-confidence selection, and observe how good they approximate the full set. We estimate that between 2 and 5 minutes of data are required to reliably select a checkpoint with less than 10% of relative difference in DER to the best checkpoint. More data is needed if we want to approach more closely the selection of the full validation set. Interestingly, random selection of regions yields a better minimal validation set at low annotation budget (under 5 minutes), but is outclassed by the selection of low-confidence regions when the budget increases.

4. Conclusion

In this paper, we studied the calibration and performance of the powerset speaker diarization model on 12 datasets seen during training, as well as the 11 domains composing DIHARD 3. We found that the model is well calibrated on in-domain datasets, while calibration on out-of-domain datasets is generally worse. Despite this, we observed that diarization error rate on predicted low-confidence regions is always significantly higher than the average on the dataset.

We then simulated the annotation of low-confidence regions on out-of-domain datasets to constitute small training sets. We observed that such sets offer no significant advantage or disadvantage to random region selection in terms of DER, but prove to be better calibrated. Selection of a minimal validation set proves to be a difficult task, but selection of low-confidence regions seems to improve its efficiency given a high enough annotation budget.

These results lead us to believe that top-label confidence can be reliably used to find regions of the data where powerset speaker diarization model performs badly. Uses include out-of-domain detection, semi-supervised learning, and especially active learning. The improvement in calibration after domain adaptation is very encouraging and lead us to believe that active learning with iterative retraining and selection of new low-confidence regions using the better calibrated model might take full advantage of this property.

5. Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation AD011013477R1 made by GENCI. The research described in this paper is partly supported by the Agence de l'Innovation Défense under the grant number 2022 65 0079.

6. References

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 1321–1330.

[2] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[3] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

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

[5] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 694–699. [Online]. Available: https://doi.org/10.1145/775047.775151

[6] A. Perez-Lebel, M. L. Morvan, and G. Varoquaux, "Beyond calibration: estimating the grouping loss of modern neural networks," in *Proceedings of the International Conference on Learning Representations*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2210.16315

[7] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, Sep 2023. [Online]. Available: https://doi.org/10.1007/s10994-023-06336-7

[8] R. Müller, S. Kornblith, and G. Hinton, *When does label smoothing help?* Red Hook, NY, USA: Curran Associates Inc., 2019.

[9] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 23 631–23 644. [Online]. Available: https://proceedings.mlr.press/v162/wei22d.html

[10] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, *Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[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," in *Proc. Interspeech 2021*, 2021.

[12] 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," in *Proc. ICASSP 2022*, 2022.

[13] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal, "The AMI Meetings Corpus," in *Proc. Symposium on Annotating and Measuring Meeting Behavior*, 2005.

[14] 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*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3838–3847. [Online]. Available: https://doi.org/10.1145/3503161.3548027

[15] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-Based Speaker Segmentation using Speaker Factors and Eigenvoices," in *Proc. ICASSP 2008*, 2008.

[16] S. Baghel, S. Ramoji, Sidharth, R. H, P. Singh, S. Jain, P. Roy Chowdhuri, K. Kulkarni, S. Padhi, D. Vijayasenan, and S. Ganapathy, "The DISPLACE Challenge 2023 - DIarization of SPeaker and LAnguage in Conversational Environments," in *Proc. INTERSPEECH 2023*, 2023, pp. 3562–3566.

[17] 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 *Proc. CVPR 2022*, 2022.

[18] 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.

[19] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, L. Xie, and Y. Yan, "Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational(RAMC) Speech Dataset," in *Proc. Interspeech 2022*, 2022, pp. 1736–1740.

[20] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A Presentation of the REPERE Challenge," in *Proc. CBMI 2012*, 2012.

[21] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the Conversation: Speaker Diarisation in the Wild," in *Proc. Interspeech 2020*, 2020.

[22] 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 *Proc. Interspeech 2021*, 2021, pp. 3570–3574.

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