Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

# Count Your Speakers! Multitask Learning for Multimodal Speaker Diarization

*Prabhav Singh*[1], *Jesús Villalba*[1,2], *Najim Dehak*[1,2]

[1]Center for Language & Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[2]HLT Center of Excellence, Johns Hopkins University, Baltimore, MD, USA
{psingh54, jvilla17, ndehak3}@jhu.edu

## Abstract

Recent advances in speaker diarization have explored diverse clustering methods, particularly in multimodal frameworks. However, a critical limitation lies in the clustering stage, where heuristic-based methods often fail to leverage the full potential of multimodal data. For example, threshold-based clustering frequently leads to over-clustering, causing incorrect speaker assignments and elevated DER. To address this, we propose CYS-MSD, a novel framework that fuses audio-visual modalities via a trainable cross-modal attention mechanism. The embeddings are fine-tuned with a multitask objective to jointly predict speaker counts and assign speaker labels, enabling data-driven clustering that adapts to varying speaker scenarios. Additionally, a modality-masking mechanism ensures robustness to missing inputs in real-world conditions. We evaluate CYS-MSD on the AVA-AVD corpus, reporting a 5% reduction in DER over the baseline and an average 2% reduction compared to various SOTA systems.

**Index Terms**: diarization, multitask-learning, multi-modality

## 1. Introduction

Speaker diarization (SD), the task of determining 'who spoke when' in audio or audio-visual streams, is a fundamental problem for conversational analysis. Its applications span a wide range, including meeting transcription, media content analysis, and human-computer interaction [1]. Classical approaches based on clustering speaker embeddings [2, 3] remain prevalent for processing long, multi-speaker recordings due to their robustness with real-world conversational data [4]. However, these probabilistic clustering methods, while improving speaker uncertainty handling [5], still face challenges with over-clustering, often overestimating the true number of speakers in conversations.

Further, multimodal SD has seen significant advancements in recent years, particularly in leveraging audio-visual cues to improve performance [6]. For example, Kang et al. [7] introduced a novel approach that integrates d-vectors with spatial information obtained through acoustic beamforming. Similarly, Ahmad et al. [8] introduced a pre-trained Audio-Visual (AV) synchronization model to enhance clustering, showing promising results on the AMI meeting corpus [9]. However, these approaches often struggle with missing modalities or rely on simplistic fusion strategies that overlook complex intermodal interactions. A recent work by Cheng et al. [10] attempted to address these limitations by proposing a framework that jointly utilizes audio, visual, and semantic cues, formulating multimodal modeling as a constrained optimization problem. While this approach shows promise, it also highlights the ongoing challenge of effectively integrating multiple modalities in spontaneous and unstructured conversations, particularly when dealing with varying quality and availability of different modalities across speakers and time.

Finally, while end-to-end approaches [11, 12] have shown promise in achieving good performance on the CALLHOME dataset [13], clustering-based methods continue to demonstrate superior performance, particularly on datasets that mimic real-world conditions with complex speaker dynamics [14].

Hence, in this work, we focus on three critical challenges that remain unresolved in contemporary diarization research:

1. Systems [6] often report higher accuracy with oracle speaker counts for each utterance, creating a significant gap between evaluation protocols and practical deployment scenarios.
2. Threshold-based clustering approaches are error-prone in real-world settings, especially when speaker overlap is high or the number of speakers varies unpredictably [14, 15].
3. Multimodal approaches, while promising, frequently fail to address missing modalities or rely on simplistic fusion strategies that overlook inter-modal interactions.

To address these challenges, we introduce CYS-MSD (**C**ount **Y**our Speaker - **M**ultitask Learning for Multimodal **SD**), a novel multitask multimodal framework for speaker diarization. Building on recent advances in multimodal architectures, CYS-MSD integrates trainable cross-modal attention with a multitask learning framework to jointly predict speaker counts and utterance speaker labels. Additionally, we propose a robust mechanism for handling missing modalities [16] that ensures reliable performance even when audio or visual information is partially or entirely unavailable—a common challenge in real-world deployments. To our knowledge, this is the first work to explore the benefits of speaker-counting in the process of diarization. The core contributions of CYS-MSD, shown in Figure 1, are as follows:

1. **Learnable Modality Fusion:** We introduce an attention-based mechanism to learn optimal fusion weights for audio and visual embeddings while employing a random masking mechanism to handle missing modalities dynamically.
2. **Multitask Learning:** We design a multitask framework that optimizes embeddings using a combination of Contrastive [17] and BCE loss for segment-level active speaker identification and MSE Loss for speaker count estimation. This auxiliary supervision not only enhances embedding quality but also informs downstream clustering.
3. **Informed Clustering:** Diarization is performed using Agglomerative Hierarchical Clustering (AHC) [18], albeit, guided by the predicted speaker counts for each segment, replacing the need for heuristically tuned thresholds.[1]

[1] Note that this is distinct from tuning a threshold for AHC on val-

10.21437/Interspeech.2025-334

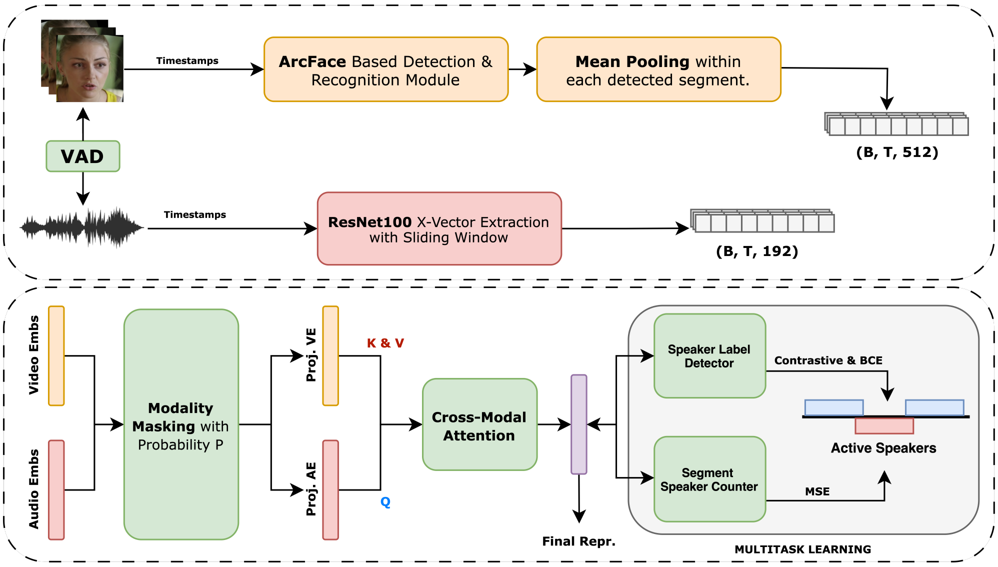

Figure 1: *Architecture: Multitask Learning for Multimodal SD*

We evaluate CYS-MSD on the AVA-AVD corpus [14], a challenging dataset designed to test diarization systems under complex, real-world conditions. Our results demonstrate significant performance gains over SOTA works. Ablation studies further reveal the effectiveness of each component, particularly the random masking mechanism and multitask objective, in enhancing diarization robustness and accuracy. The remainder of this paper is organized as follows: Section 2 presents the proposed method in detail. Section 3 defines the data used for pretraining and evaluation and also identifies the metrics and training setup. Section 4 presents detailed experimental evaluations, comparisons, and ablation studies. Section 5 concludes with a discussion of limitations.

## 2. Proposed Method

In this section, we introduce the core components and methodology of the proposed multimodal speaker diarization system, CYS-MSD.

### 2.1. Normalization & Embedding Extraction

For the audio modality, we utilized a pre-trained X-Vector-based architecture [19] implemented with the Hyperion toolkit[2]. Specifically, we employed a ResNet100 [20] architecture with four stages and residual squeeze-excitation blocks [21], pre-trained on VoxCeleb-2 [22] for the task of speaker recognition. The model incorporated utterance-level mean and standard deviation normalization. For each audio file, we leveraged oracle Voice Activity Detection (VAD) to identify segments containing active speakers. Each segment was processed using a sliding window approach with a window size of 1.5 s and a shift of 0.75 s. From each active segment, 192-dimensional embeddings were extracted after pooling.

For the video modality, we employed the InsightFace library[3] for face recognition and detection. Specifically, we use the BUFFALO_L variant, which incorporates a pre-trained RetinaFace detection module [23] coupled with a ResNet50 recognition backbone fine-tuned on the WebFace600K dataset [24]. The architecture uses ArcFace [25] loss to maximize class separability. For each active segment, all detected faces were processed, and mean pooling was applied to obtain 512-dimensional embeddings for each segment.

### 2.2. Modality Masking & Learnable Attention

To enhance the robustness of our multimodal system and prevent over-reliance on any single modality, we implemented a stochastic modality masking strategy [16] during training. For each batch, modalities were independently masked with probability $p_m$, where audio and visual embeddings were replaced with zero vectors:

$$\mathbf{e}_m = \begin{cases} \mathbf{0}, & \text{with probability } p_m \\ \mathbf{e}_m^{\text{orig}}, & \text{otherwise} \end{cases} \tag{1}$$

where $m \in \{\text{audio}, \text{visual}\}$ denotes the modality, and $\mathbf{e}_m^{\text{orig}}$ represents the original embedding. We note that there theoretically could be cases of both modalities being masked. However, since $p_m$ is kept at 10%, chosen by experimentation, the actual chances of both modalities being zero were low.

Following the masking operation, we employed a multimodal fusion strategy that combines projection layers and multi-head attention. First, we projected the audio and visual embeddings into a common dimensional space:

$$\begin{aligned} \mathbf{E}_a^p &= \text{ReLU}(\mathbf{E}_a \mathbf{W}_a) \\ \mathbf{E}_v^p &= \text{ReLU}(\mathbf{W}_v^2 \, \text{ReLU}(\mathbf{E}_v \mathbf{W}_v^1)) \end{aligned} \tag{2}$$

where $\mathbf{W}_a \in \mathbb{R}^{d_a \times d_f}$ projects audio embeddings to the fusion dimension $d_f$, and the visual embeddings undergo a multi-layer perceptron (MLP) transformation with an intermediate dimension of $d_v/2$ and dropout regularization.

idation data. We propose predicting actual counts of speakers in an utterance and using that to inform AHC with defined number of clusters.

[2] https://github.com/hyperion-ml/hyperion

[3] https://github.com/deepinsight/insightface

We then employed a 4-headed attention mechanism where audio embeddings attended to visual information:

$$\mathbf{A} = \text{Attention}(\mathbf{E}_a^p, \mathbf{E}_v^p, \mathbf{E}_v^p) \tag{3}$$

where $\mathbf{E}_a^p$ serves as the query and $\mathbf{E}_v^p$ provides both keys and values. The final multimodal representation was obtained through a weighted combination:

$$\mathbf{R} = \alpha \mathbf{E}_a^p + (1 - \alpha)\mathbf{A} \tag{4}$$

where $\alpha$ is a pre-defined weighting factor that controls the contribution of the audio modality versus the attention-processed features. This weighted fusion strategy allowed us to maintain strong speaker-discriminative information from the audio stream while incorporating complementary visual cues through attention.

### 2.3. Multitask Training

We employed a multitask learning strategy with two complementary objectives: active speaker identification and speaker counting. This dual-objective approach helps learn more robust multimodal representations while also providing practical utility for downstream diarization tasks.

#### *2.3.1. Speaker Label Detection*

The primary task involves detecting active speakers in each segment using the fused multimodal representations. Given the fused embeddings $\mathbf{R}$, we employed a classification head consisting of two fully connected layers with ReLU activation and dropout. A Sigmoid layer represented the probability of each speaker being active at each time step, for $S$ total speakers. We optimized this objective using binary cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = \text{BCE}(\mathbf{P}, \mathbf{Y}) \tag{5}$$

where $\mathbf{Y}$ represents the ground truth binary speaker labels and $\mathbf{P}$ is the prediction. We note that using a binary label for each speaker helps account for overlapping speech, where multiple speakers can be active simultaneously.

To enhance the discriminative power of the learned representations, we additionally employed a contrastive learning objective using the normalized temperature-scaled cross entropy (NT-Xent) loss [26] within a temporal window. For each segment $i$, we treated segments $j$ within a temporal window $w$ sharing active speakers as positive pairs ($\mathcal{P}_i$). The loss is then computed as:

$$\mathcal{L}_{\text{cont}} = -\log \frac{\sum_{j \in \mathcal{P}_i} \exp(\text{sim}(\mathbf{r}_i, \mathbf{r}_j)/\tau)}{\sum_{k \notin \mathcal{P}_i} \exp(\text{sim}(\mathbf{r}_i, \mathbf{r}_k)/\tau)}, \tag{6}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity and $\tau$ is the temperature.

#### *2.3.2. Speaker Counting*

As a complementary task, we predicted the total number of unique speakers in the utterance. This is accomplished through a separate counting head that processed the mean-pooled fused embeddings:

$$\mathbf{C} = \text{MLP}_{\text{count}}(\text{mean}(\mathbf{R})) \tag{7}$$

where $\mathbf{C} \in \mathbb{R}$ represents the predicted speaker count. This objective is optimized using mean squared error loss and the final training objective combines both tasks with a weighting parameter $\beta$:

$$\begin{aligned} \mathcal{L}_{\text{count}} &= \text{MSE}(\mathbf{C}, \mathbf{C}_{\text{gt}}) \\ \mathcal{L}_{\text{total}} &= \beta(\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cont}}) + (1 - \beta)\mathcal{L}_{\text{count}} \end{aligned} \tag{8}$$

It should be noted that while the primary task of label detection was used to generate final embeddings for clustering, the secondary task was used to predict speaker count for each utterance to allow for informed clustering.

### 2.4. Informed Clustering

For the final speaker diarization output, we employed AHC on the trained representation of embeddings $\mathbf{R}$. The clustering process leveraged both the multimodal representations and the predicted speaker count from our multitask model.

We first normalized all segment-level embeddings using L2 normalization to ensure they lie on a unit hypersphere. Following [27], we computed pairwise cosine distances between all segments and performed average-linkage clustering. The number of clusters was determined by the speaker count prediction from our model, which provided a more informed choice compared to traditional threshold-based approaches [28].

To maintain temporal consistency and reduce fragmentation, we processed segments using overlapping windows during both embedding extraction and clustering. This helped capture speaker transitions more accurately while maintaining speaker homogeneity within segments. The final diarization output was generated such that each segment was assigned a speaker label based on its cluster membership.

## 3. Experiments

In this section, we outline the datasets and metrics. We also describe the hyperparameter selection and training strategy used for our experiments.

### 3.1. Dataset & Metrics

To evaluate CYS-MSD, we chose the AVA-AVD corpus [14] due to its in-the-wild scenarios. The dataset comprises 351 video clips, divided into 243 for training, 54 for validation, and 54 for testing. Each clip is 5 minutes long and may feature up to 24 speakers. Due to the small size of the corpus, we first used the validation data to choose the best hyperparameters, but trained our final model on both the training and validation sets. The test videos were kept unseen for final evaluation.

The primary metric for our evaluation is the Diarization Error Rate (DER)[4]. However, we also report the Speaker Error (SPKE) and Miss Rate (MR). We evaluate our approach first against the baseline, AVR-NET, defined with the release of the AVA-AVD corpus [14]. We also evaluate against SOTA approaches like the one proposed by Cheng et al. [10], AFL-NET [16], and DYVISE [29].

### 3.2. Training Strategy

We trained the model for 25 epochs using the AdamW [30] optimizer with an LR of 1×10$^{-5}$ and a batch size of 16. The fused dimension was set to 256. For multimodal fusion, we employed an audio weighting factor ($\alpha$) of 0.6, found through validation. The contrastive learning temperature ($\tau$) was set to 0.3 with a margin of 1.0. We used equal weighting ($\beta = 0.5$) between the classification and counting objectives during multitask training.

[4] https://github.com/nryant/dscore

# 4. Results

This section presents the experimental results and comparisons on the AVA-AVD corpus. Table 1 demonstrates the effectiveness of CYS-MSD in speaker diarization, achieving lower DER and SPKE compared to prior works[5].

Table 1: *Performance of CYS-MSD*

| Framework | VAD | DER ↓ | MR | SPKE |
|---|---|---|---|---|
| DyViSE | Oracle | 23.46 | **1.98** | 20.86 |
| AFL-NET | Oracle | 22.12 | 2.55 | 21.10 |
| AVA-NET | Oracle | 20.57 | 2.92 | 17.65 |
| Cheng et al. | Oracle | 20.32 | - | 17.40 |
| Ours ($c = 0$) | Oracle | **19.16** | 2.52 | **17.14** |
| Ours ($c = 0.25$) | Oracle | **18.22** | 2.19 | **17.08** |

Our proposed approach achieves a DER reduction of 4.3% and 3% over DyViSE and the baseline AFL-NET, respectively, at a collar of 0 seconds. Compared to the SOTA, we achieve a DER reduction of 1.16% and a SPKE reduction of 0.26%. We note that, at a more common collar of 0.25 seconds, we achieve an additional 1% reduction in DER. DyViSE achieves the lowest MR but has a higher DER and SPKE, suggesting that despite strong speech activity detection, its identity-based clustering remains suboptimal. AFL-NET [16] and AVA-NET [14] incorporate modality masking but still rely on heuristically optimized AHC. For example, AFL-NET uses three modalities (adding lip detection) but is still outperformed by our approach.

## 4.1. Ablation Study

In Table 2, we report the effects of ablation on CYS-MSD. It should be noted that all removals are sequential; that is, each row is obtained by removing the components mentioned in all the above rows. Further, the Dynamic AHC ablation only removes the clustering with predicted speaker counts. The embeddings are still obtained through the multitask learning module.

Table 2: *Ablation Study on CYS-MSD ($c = 0$)*

| Method & Ablation | DER | SPKE |
|---|---|---|
| **CYS-MSD** | **19.16** | **17.14** |
| *w/o Dynamic AHC* | 21.85 | 19.42 |
| *w/o Multitask Learning* | 22.92 | 21.51 |
| *w/o Modality Masking* | 23.08 | 21.76 |
| *w/o Video Modality* | 26.43 | 24.22 |

The baseline CYS-MSD achieves a DER of 19.16% and SPKE of 17.14%. Sequential removal of components reveals their relative importance: removing Dynamic AHC increases DER by 2.69%, while ablating multitask learning further degrades performance to 23.92% DER. This confirms our hypothesis: Clustering with high-quality predictions of the number of speakers in each segment has a direct impact on decreasing the DER. As expected, removing the video modality leads to the highest degradation, with DER increasing to 26.43% (+7.27% absolute from baseline) and SPKE reaching 24.22%, underscoring the vital role of visual information in our multimodal diarization framework.

## 4.2. Underscoring the Importance of Counting

We further perform comparative experiments to evaluate the extent of benefits obtained by using speaker counts for each utterance. Table 3 presents the accuracy in predicting the number of speakers with two different methods. The first uses a search[6] over the thresholds for AHC within the range from 0.1 to 0.9 with a step size of 0.01. The second is the proposed approach with predicted counts for each detected utterance.

Table 3: *CYS-MSD: Heuristic vs Informed Clustering*

| Clustering Approach | Accuracy | DER ↓ |
|---|---|---|
| AHC ($\tau = 0.5$) via Search | 38.19% | 21.85% |
| AHC via Predicted Counts | 78.88% | 19.16% |

It can be seen from the table that using clustering with just a heuristic search leads to low accuracy for speaker counts. This can be attributed to the fact that the dataset has a high variation of unique speakers in each file. However, the multitask module is able to increase the accuracy, which directly translates to a lower DER.

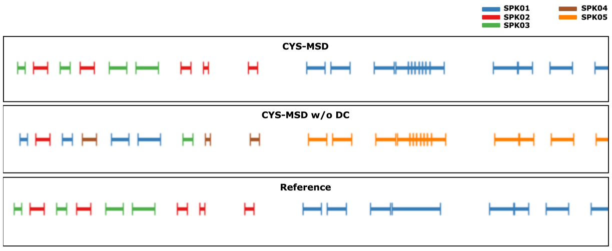

Figure 2: *Sample Diarization Output (Top to Bottom: CYS-MSD, CYS-MSD w/o Informed Clustering, Reference)*

Figure 2 illustrates an example of this scenario. While both systems produce segmentations that are close to the reference, accurately determining the number of speakers significantly improves diarization quality. In the case of CYS-MSD without speaker counting, over-clustering results in the detection of five speakers, whereas the reference indicates only three.

# 5. Conclusion & Limitations

This paper presents CYS-MSD, a novel multitask framework for SD that integrates speaker counting with multimodal fusion. Our approach leverages informed clustering to adaptively refine speaker embeddings, outperforming SOTA methods while demonstrating the superiority of learned speaker grouping over traditional heuristic clustering. By jointly optimizing diarization and speaker counting, we enhance segmentation accuracy and reduce speaker attribution errors, particularly in multi-speaker scenarios. However, a key limitation of our method is its reliance on oracle VAD, which restricts real-world applicability. Future work should focus on integrating similar frameworks with robust VAD models to enable deployable speaker diarization systems.

[5] $c$ represents the collar in this case. Note that the AVA-AVD has overlapping speech and hence we choose to report scores with collar.

[6] We use the validation set to find the best threshold.

# 6. References

[1] V. Mingote, A. Ortega, A. Miguel, and E. Lleida, "Audio-visual speaker diarization: Current databases, approaches and challenges," 2024. [Online]. Available: https://arxiv.org/abs/2409.05659

[2] X. Zhang, W. Wang, and P. Zhang, "Speaker diarization system based on dpca algorithm for fearless steps challenge phase-2," in *Interspeech 2020*, 2020, pp. 2602–2606.

[3] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization," in *Interspeech*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:4547281

[4] K. Kinoshita, M. Delcroix, and N. Tawara, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," in *Interspeech 2021*, 2021, pp. 3565–3569.

[5] P. Singh and S. Ganapathy, "End-to-end supervised hierarchical graph clustering for speaker diarization," 2024. [Online]. Available: https://arxiv.org/abs/2401.12850

[6] Z. Pan, G. Wichern, F. G. Germain, A. S. Subramanian, and J. L. Roux, "Late audio-visual fusion for in-the-wild speaker diarization," *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 174–178, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253255375

[7] W. Kang, B. Roy, and W. Chow, "Multimodal speaker diarization of real-world meetings using d-vectors with spatial features," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6509–6513, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216480639

[8] R. Ahmad, S. Zubair, H. Alquhayz, and A. Ditta, "Multimodal speaker diarization using a pre-trained audio-visual synchronization model," *Sensors*, vol. 19, no. 23, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/23/5163

[9] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. L. Masson, I. McCowan, W. Post, D. Reidsma, and P. D. Wellner, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:6118869

[10] L. Cheng, H. Wang, S. Zheng, Y. Chen, R. Huang, Q. Zhang, Q. Chen, and X. Li, "Integrating audio, visual, and semantic information for enhanced multimodal speaker diarization," *ArXiv*, vol. abs/2408.12102, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271924486

[11] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-End Neural Speaker Diarization with Permutation-free Objectives," in *Interspeech*, 2019, pp. 4300–4304.

[12] C. Wang, J. Li, X. Fang, J. Kang, and Y. Li, "End-to-end neural speaker diarization with absolute speaker loss," in *Interspeech 2023*, 2023, pp. 3577–3581.

[13] A. Canavan, D. Graff, and G. Zipperlen, "Callhome american english speech," Web Download, Philadelphia, 1997, lDC Catalog No.: LDC97S42.

[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] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Černocký, "Bayesian hmm based x-vector clustering for speaker diarization," in *Interspeech 2019*, 2019, pp. 346–350.

[16] Y. Yin, X. Li, Y. Shan, and Y. Zou, "Afl-net: Integrating audio, facial, and lip modalities with a two-step cross-attention for robust speaker diarization in the wild," in *Interspeech 2024*, 2024, pp. 42–46.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20. JMLR.org, 2020.

[18] J. Luque and J. Hernando, "On the use of agglomerative and spectral clustering in speaker diarization of meetings," in *The Speaker and Language Recognition Workshop (Odyssey 2012)*, 2012, pp. 130–137.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[21] N. Torgashov, R. Makarov, I. Yakovlev, P. Malov, A. Balykin, and A. Okhotnikov, "The id r&d voxceleb speaker recognition challenge 2023 system description," 2023. [Online]. Available: https://arxiv.org/abs/2308.08294

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[23] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou, "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 487–10 497.

[25] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, p. 5962–5979, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2021.3087709

[26] W. Ågren, "The nt-xent loss upper bound," 2022. [Online]. Available: https://arxiv.org/abs/2205.03169

[27] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 413–417.

[28] T. Park, N. R. Koluguri, J. Balam, and B. Ginsburg, "Multi-scale speaker diarization with dynamic scale weighting." in *INTERSPEECH*. ISCA, 2022, pp. 5080–5084.

[29] A. Wuerkaixi, K. Yan, Y. Zhang, Z. Duan, and C. Zhang, "Dyvise: Dynamic vision-guided speaker embedding for audio-visual speaker diarization," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, 2022, pp. 1–6.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270