Proceedings Paper:

Clarke, J., Gotoh, Y. orcid.org/0000-0003-1668-0867 and Goetze, S. (2025) Speaker embedding informed audiovisual active speaker detection for egocentric recordings. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Proceedings. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 06-11 Apr 2025, Hyperabad, India. Institute of Electrical and Electronics Engineers (IEEE) ISBN 9798350368758

https://doi.org/10.1109/ICASSP49660.2025.10890414

Speaker Embedding Informed Audiovisual Active Speaker Detection for Egocentric Recordings

Jason Clarke[1], Yoshihiko Gotoh[1], and Stefan Goetze[1,2]
[1]Speech and Hearing (SPandH) group, School of Computer Science, The University of Sheffield, Sheffield, United Kingdom
[2]South Westphalia University of Applied Sciences, Iserlohn, Germany
{jclarke8, y.gotoh, s.goetze}@sheffield.ac.uk, goetze.stefan@fh-swf.de

***Abstract*—Audiovisual active speaker detection (ASD) addresses the task of determining the speech activity of a candidate speaker given acoustic and visual data. Typically, systems model the temporal correspondence of audiovisual cues, such as the synchronisation between speech and lip movement. Recent work has explored extending this paradigm by additionally leveraging speaker embeddings extracted from candidate speaker reference speech. This paper proposes the speaker comparison auxiliary network (SCAN) which uses speaker-specific information from both reference speech and the candidate audio signal to disambiguate challenging scenes when the visual signal is unresolvable. Furthermore, an improved method for enrolling face-speaker libraries is developed, which implements a self-supervised approach to video-based face recognition. Fitting with the recent proliferation of wearable devices, this work focuses on improving speaker-embedding-informed ASD in the context of egocentric recordings, which can be characterised by acoustic noise and highly dynamic scenes. SCAN is implemented with two well-established baselines, namely TalkNet and Light-ASD; yielding a relative improvement in mAP of 14.5% and 10.3% on the Ego4D benchmark, respectively.**

***Index Terms*—Diarization, Audiovisual Active Speaker Detection, Video-based Face Recognition, Speaker Recognition**

I. Introduction

Audiovisual active speaker detection (ASD) revolves around determining the video-framewise speech activity of a candidate speaker. The task is typically formulated as a binary classification problem, where, given a mixed audio signal and sequence of temporally contiguous bounding boxes centered on the candidate speaker's face, a system identifies video frames where the candidate speaker is talking [1]–[6].

Previous ASD research has mainly focused on improving performance for exocentric data (recorded from the third person perspective) [1], [3], [7], [8], where the camera and microphone are typically stationary relative to the scene rendering the recording conditions favourable. With the recent proliferation of wearable devices, however, the flavour of data ASD systems are likely to be deployed upon has shifted to egocentric recordings, where the audiovisual signal is acquired from the first person perspective, and the camera and microphone are dynamic relative to the scene. This change in recording perspective introduces several challenges: (i) low signal-to-noise ratios for speech signals, (ii) highly spontaneous conversations with overlapping speech, (iii) audiovisual distortion caused by the camera wearer's head movements, and (iv) situational obfuscation, where visual cues are occluded [9], [10]. Since the paradigm observed in recent literature involves modelling the correspondence between audiovisual cues indicative of a candidate speaker talking [3], [11]–[13] (like lip movement, cheek posture [4], and audible speech), this paper argues said approaches are not sufficiently robust to handle the aforementioned challenges associated with egocentric recordings. For example, when speech is present in the audio signal but the video signal is heavily corrupted, a typical system based on audiovisual correspondence will only be able to recognise the presence of speech, the system will not be able to attribute it to the candidate speaker [3]. This is illustrated in Figure 1: degraded video frames of the candidate speaker (bottom panel) and active speech in the microphone channel (top panel) that is not spoken by the candidate speaker induces a false activity detection for a speaker-embedding-naive system [13] (blue line in the middle panel of Figure 1). This problem is also demonstrated by the disparity in performance when evaluating ASD systems on exocentric [8] vs egocentric [9], [14], [15] benchmarks, where in the latter, challenging scenes are regularly prevalent.

a)
b)
c)

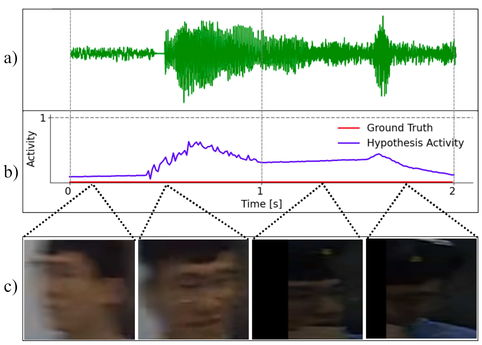

Figure 1. Example of typical false-positive ASD: a) input audio signal; b) ground truth speaker activity of the candidate speaker (inactive throughout) and hypothesised speaker activity by a state-of-the-art speaker-embedding naive ASD system [13]; c) selection of challenging video frames from a typical egocentric video track [10].

Recent work has attempted to mitigate the limitations of established ASD systems [1] by injecting speaker-specific information. Specifically, TS-TalkNet [7] uses a pre-trained speaker recognition model to extract speaker embeddings from reference speech based on the well-known ECAPA-TDNN architecture [16]. These speaker embeddings are then leveraged as an additional source of information. Drawing inspiration from this, this paper proposes the speaker comparison auxiliary network (SCAN), an auxiliary module that can be integrated with various end-to-end ASD systems [1], [13]. Unlike TS-TalkNet [7], SCAN extracts speaker-specific information from two distinct sources: reference speech, i.e. previously diarised speech spoken by the candidate speaker, and the candidate audio signal. This technique has previously leveraged successfully in the domain of personal voice activity detection [17], [18]. The novelty of SCAN lies in its ability to perform framewise comparisons between these two sources via a cross-attention mechanism. This enables SCAN to identify similarities and distinctions between speaker-specific cues in the reference speech and the candidate audio signal with high temporal granularity. By doing so, SCAN provides a

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UKRI [grant number EP/S023062/1]. This work was also funded in part by Meta.

mechanism for effectively disambiguating scenarios with low-quality video signals, resulting in improved ASD reliability and robustness. SCAN constitutes the first contribution of this paper. Additionally, this work demonstrates that existing methods for generating identity-speech libraries, which associate reference speech with the identity of the candidate speaker, are not robust to the challenges of egocentric video. To address this, this paper proposes an improved method for generating identity-speech libraries by extending and finetuning an existing face-recognition model to leverage the temporal context of video data via a self-supervised learning objective. This constitutes the second contribution of this paper.

To summarise, the work outlined in this paper expands upon the concept of target-speaker ASD, tailoring it to address the particular challenges posed by egocentric recordings with the following main contributions:

1) The auxiliary module SCAN, which leverages speaker-specific information to help disambiguate challenging scenes for ASD.
2) A self-supervised method to finetune a pre-trained face recognition model on video data to enroll identity-speech libraries more robust to egocentric recordings.
3) Performance analysis on the egocentric Ego4D-AVD dataset [10] and exocentric AVA-ActiveSpeaker [8] using three existing systems as baselines: two speaker-embedding-naive systems - TalkNet [1] and Light-ASD [13]; and the current state-of-the-art speaker-embedding-informed system TS-TalkNet [7].

II. Speaker Embedding Informed Audiovisual Active Speaker Detection

This section presents an overview of a typical ASD architecture, introduces the proposed SCAN module, and outlines the training protocol for fine-tuning an existing framewise face recognition model on egocentric video to construct more robust identity-speech (face-speaker [7]) libraries.

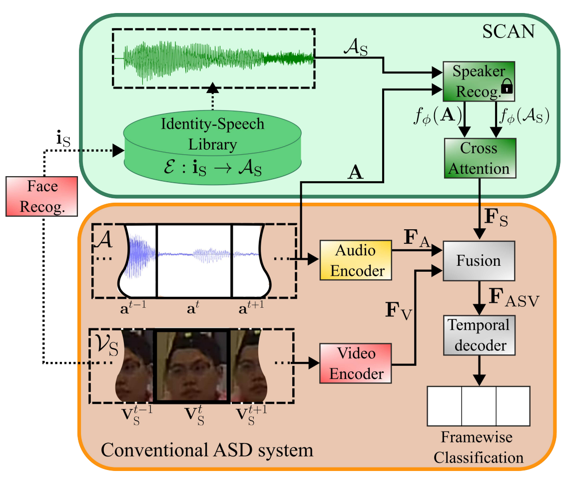

Figure 2. SCAN is shown in the top box which leverages speaker-specific information for framewise comparison of reference speech and input audio signal via cross-attention. The bottom box shows a typical ASD architectural design (baseline). Dotted connections represent non-end-to-end passages in the framework.

A. Baseline Architecture Overview

As shown in the lower part of Figure 2, conventional ASD systems, such as the speaker-embedding-naive systems used as baselines in this paper (i.e. TalkNet [1] and Light-ASD [13]), follow the current paradigm of comprising an audio encoder, a video encoder, a modality fusion mechanism, and a temporal decoder [1], [7], [12], [13]. These systems operate on the video signal $\mathcal{V}_\mathrm{S}$ and the audio signal $\mathcal{A}$. The video signal is a set $\mathcal{V}_\mathrm{S} = \{\mathbf{V}_{\mathrm{S},1}, ..., \mathbf{V}_{\mathrm{S},T}\}$ of temporally contiguous video frames, centred on a single candidate speaker S. Each grayscale image in the set is denoted $\mathbf{V}_{\mathrm{S},t} \in \mathbb{R}^{H\times W}$ with time index $t \in \{1, ..., T\}$, height dimension H, and width dimension W. $\mathcal{A} = \{\mathbf{a}_1, ..., \mathbf{a}_{T_\mathrm{A}}\}$ denotes the mixed audio signal temporally correspondent to $\mathcal{V}_\mathrm{S}$ with time index $t_\mathrm{A} \in \{1, ..., T_\mathrm{A}\}$. The distinction between T and T_A compensates for the discrepancy in modality sampling rates.

Video and audio encoders extract pertinent features from their respective modality inputs, and perform short-term temporal modelling to encapsulate the local inter-frame relationships. These encoders typically resemble 3D-ResNets [2], visual temporal convolutional networks (V-TCNs) [1], or depth-wise separable convolutions [13]. The audio and video encoders embed their respective inputs, resulting in two matrices $\mathbf{F}_\mathrm{A} \in \mathbb{R}^{T\times d}$ and $\mathbf{F}_\mathrm{V} \in \mathbb{R}^{T\times d}$, where d is the embedding dimension of both encoders. A fusion mechanism is then applied to combine these two embeddings, generating a single 2-dimensional output $\mathbf{F}_\mathrm{ASV}$. This fusion is often a simple channel-wise concatenation, summation [13], or an attention-based approach [1], [14]. The temporal decoder performs two tasks: long-term temporal modelling on the mixed-modality embedding to capture the sequential nature of speech and framewise classification to predict the speaker activity of the candidate speaker.

B. Speaker Comparison Auxiliary Network (SCAN)

The TS-TalkNet system, introduced in [7], builds upon the paradigm outlined in the bottom part of Figure 2. It does so by injecting speaker-specific information from pre-diarised reference speech of the candidate speaker $\mathcal{A}_\mathrm{S}$ into the model. A speaker embedding $f_\phi(\mathcal{A}_\mathrm{S})$ is extracted from $\mathcal{A}_\mathrm{S}$ via ECAPA-TDNN [16] which is then fused with $\mathbf{F}_\mathrm{A}$ and $\mathbf{F}_\mathrm{V}$ prior to temporal decoding. This injection of speaker-specific information relating to the candidate speaker results in a significant performance improvement over its respective baseline [1]. This extension is motivated by the need to resolve ambiguous scenarios where visual cues indicative of speech activity are occluded, rendering the scenario intractable via audiovisual correspondence alone.

Inspired by this approach, SCAN employs a distinct modification by extracting speaker embeddings from not only pre-diarised reference speech $\mathcal{A}_\mathrm{S}$, but also from the candidate audio signal $\mathcal{A}$. This enables an explicit comparison between speaker characteristics of the reference speech and the candidate audio signal via the cross-attention mechanism shown in Figure 2. As a result, it will be easier for the network to learn to identify similarities and distinctions between $\mathcal{A}$ and $\mathcal{A}_\mathrm{S}$.

First, overlapping windows temporally centred around each video frame are extracted from the raw waveform input of $\mathcal{A}$, transforming $\mathcal{A}$ to a matrix $\mathbf{A}$. This transformation is performed such that a sufficient duration of audio exists at each time point for meaningful speaker embeddings to be extracted. $\mathcal{A}_\mathrm{S}$ and $\mathbf{A}$ are then embedded by a pre-trained speaker recognition model f_ϕ and a cross-attention mechanism is employed along the temporal dimension of $\mathbf{A}$:

$$\mathbf{F}_\mathrm{S} = \sigma\left(\frac{\mathrm{f}_\phi(\mathbf{A})\mathrm{f}_\phi(\mathcal{A}_\mathrm{S})^\top}{\sqrt{d_\phi}}\right)\mathrm{f}_\phi(\mathcal{A}_\mathrm{S}) \tag{1}$$

The embedded audio signal $\mathrm{f}_\phi(\mathbf{A})$ is used as the queries, and the embedded reference speech $\mathrm{f}_\phi(\mathcal{A}_\mathrm{S})$ is used as the keys and values. d_ϕ denotes the embedding dimension of the speaker recognition model and σ represents the softmax function. This is done to determine how well the speech in the current track's audio correlates with that of the reference speech. ECAPA-TDNN [16] pre-trained on the VoxCeleb dataset [19] is used as f_ϕ due to its previous use in TS-TalkNet and robust performance on various benchmarks [7]. The ECAPA-TDNN model parameters are frozen and therefore do not contribute to model training.

C. Identity-Speech Library Generation

To exploit reference speech information for ASD, a correspondence between candidate speaker identity and reference speech must be established. Following TS-TalkNet [7], this work generates an identity-speech library $\mathcal{E} : \mathbf{i}_\mathrm{S} \rightarrow \mathcal{A}_\mathrm{S}$ which is pre-enrolled offline, where $\mathbf{i}_\mathrm{S}$ is a vector representing the candidate speaker's identity. In ASD datasets, identity annotations are typically not provided. However, by definition, tracks are identity-homogeneous. By clustering the identities of each track, this indirectly clusters each track's corresponding pre-diarised speech signal (if the track contains active speech). This is the premise of identity-speech library generation.

1) Identity Aggregation: To construct the identity-speech library, identity embeddings $\mathbf{i}$ are extracted from the visual component of all tracks in a given dataset. Cosine similarity is used to assess the similarity between a pair of identity embeddings. If the similarity exceeds a static threshold, the identities within each track are considered to be the same, and pre-diarised speech within the track's corresponding audio signal is attributed to this identity in the library.

2) Self-Supervised Video-Based Face Recognition: The performance of the face recognition model used is critical to the quality of the identity-speech library and consequently the utility of the output of SCAN. For example, if $\mathcal{E} : \mathbf{i}_\mathrm{S}$ contains speech not spoken by S when $\mathcal{A}_\mathrm{S}$ is sampled, irrelevant speech could be fed into the speaker recognition module, rendering its output uninformative, or even deleterious. The reliability of existing frame-based face recognition systems [20]–[23], despite their robust performance on other benchmarks, will not be sufficient to withstand the challenges posed by the highly domain-specific nature of egocentric recordings. Subsequently, a method was devised to adapt and finetune an existing pre-trained face recognition model. Firstly, a means of modelling consecutive frames as a sequence was integrated into a frame-based face recognition model, enabling it to effectively encapsulate and leverage the temporal context associated with video data. Secondly, since hard labels describing the trackwise identities of each person within ASD datasets are not typically provided, a self-supervised training objective was employed.

As depicted in Figure 3, input tracks $\mathcal{V}_\mathrm{S}$ are polluted with impostor frames $\mathbf{V}_{\mathrm{I}_x}$ randomly chosen from other tracks. Frames within the polluted track $\mathcal{V}'_\mathrm{S}$ are then individually encoded by the face recognition model [24]. This output is then fed through L transformer encoder layers [25] where the model has the capacity to attend across different frames within the track, thus leveraging the temporal context of video data. The model then determines whether each frame is either an impostor or native frame via binary classification. Learning this classification indirectly conditions the model to assign low weighting to frames not relevant to the track's overall identity (poor quality frames with significant visual distortion or occlusion) and high weighting to crisp frames where the parent track identity is clearly apparent and recognisable.

Once fine-tuned, the output of the last transformer encoder layer is averaged across its temporal dimension to generate a single embedding representative of the candidate speaker's identity $\mathbf{i}_\mathrm{S}$.

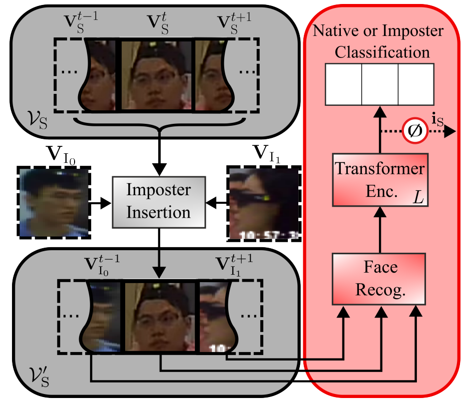

Figure 3. Self-supervised video-based face recognition model. impostor frames are randomly inserted into the parent track, resulting in polluted track $\mathcal{V}'_\mathrm{S}$. The training objective involves the model classifying frames as either native or impostor frames with respect to the parent track. $\varnothing$ denotes mean average.

III. Experiments

This section describes the datasets and experiments used to evaluate SCAN for ASD and the quality of the identity-speech library.

A. Datasets

AVA-ActiveSpeaker [8] is a frequently-used, exocetric, large-scale audiovisual ASD dataset, comprising 262 Hollywood movie clips (120 for training, 33 for validation, and 109 for testing) with 3.65 million human-labeled video frames (38.5 hours of face tracks) and corresponding audio.

Ego4D-AVD [10] records from the egocentric perspective. It comprises 572 distinct video clips. Each video clip is 5 minutes in length, some of which are recorded concurrently. All data is recorded monaurally using a variety of wearable devices. All video is sampled at 30 Hz and uses high-definition resolution. The dataset is stratified as follows: 379 clips for training, 50 clips for validation, and 133 clips for testing. The full validation fold of Ego4D-AVD was annotated by this work in terms of pseudo-identity. This was to provide a robust means of evaluating identity-speech libraries (cf. Table III).

B. Implementation Details

Baselines: All baseline models were implemented using the same input features, optimisers, and learning rates used in each systems's original implementations [1], [7], [13]. Standard ASD augmentation techniques were applied such as negative sampling of the audio signal, and flipping, cropping, and rotating of the video signal.

SCAN: The output of SCAN contributed to each baseline system's loss function as an auxiliary loss, using binary cross entropy to perform framewise classification upon $\mathbf{F}_\mathrm{S}$. Raw waveform audio served as the reference speech input to the speaker recognition model (1024 channel ECAPA-TDNN [16]). The output of SCAN $\mathbf{F}_\mathrm{S}$ used an embedding dimension of 64. 1 second windows of audio was used to extract each speaker embedding from $\mathbf{A}$. Reference speech was randomly selected from the relevant part of the identity-speech library to increase training variability.

Identity-Speech Library: For the finetuning of the face recognition model (cf. Section II-C2), a cross-entropy loss function was used. Input to the system were tracks comprising colour images with a 30% impostor insertion rate. 4 Transformer encoder layers ($L = 4$) each comprising 8 attention heads with a model dimension of 1024 were trained for 10 epochs on a single NVIDIA A100 GPU for 2 hours with a batch size of 1800. To create the identity-speech library a static comparison threshold of 0.9 was used to construct the library and a 2.5 second minimum duration of speech was enforced.

C. Evaluation Metric

Evaluation of each system for ASD is performed using the Cartucho object detection mean Average Precision (mAP) [26], which adheres to the mAP criterion from the PASCAL VOC2012 competition [27]. This approach is consistent with the Ego4D audiovisual diarisation challenge [10] and recent literature [9]. Due to the unavailability of ground truth annotations for the test folds of Ego4D and AVA-ActiveSpeaker, results are reported on the validation folds of each dataset, following the convention in ASD [3], [9], [11], [12], [14], [15].

IV. Results

This section demonstrates the performance of SCAN when used in conjunction with two speaker-embedding-naive systems, TalkNet [1] and Light-ASD [13]. The identity-speech library generation method proposed by this paper (cf. Section II-C) is also evaluated and compared with previous work when applied to egocentric recordings.

A. Audiovisual Active Speaker Detection

The results of incorporating SCAN with two speaker-embedding-naive systems, TalkNet [1] and Light-ASD [13], are shown in Table I.

Table I
Performance comparison on Ego4D-AVD and AVA validation fold. Identity-Speech Library† refers to ground truth identity-speech library. Bold highlights best-performing system with hypothesised identity-speech library, underlined represents best system with ground truth identity-speech library.

Baseline	SCAN used	Identity-Speech Library†	mAP [%] Ego4D	mAP [%] AVA
TS-TalkNet	✗	✗	52.2	**93.9**
	✗	✓	54.0	93.9
TalkNet	✗	-	51.0	92.3
	✓	✗	**58.0**	93.8
	✓	✓	58.4	94.0
Light-ASD	✗	-	54.3	94.1
	✓	✗	57.1	**93.9**
	✓	✓	<u>59.9</u>	<u>94.2</u>

On the Ego4D benchmark, SCAN significantly improves performance of the respective baseline systems for both ground truth identity-speech library and hypothesised identity-speech library configurations. Ground truth identity-speech libraries referring to those which are created directly from the dataset's annotation, hypothesis identity-speech library referring to those created by the method outlined in Section II-C. For the AVA benchmark (exocentric) the improvements are much more modest. This is likely because visually challenging multi-talker scenarios, in which SCAN would be beneficial, are much less prevalent than in Ego4D. Nevertheless, both configurations provide a substantial improvement upon the TalkNet baseline system. Additionally, TalkNet+SCAN outperforms TS-TalkNet, a previous speaker-embedding-informed system by 5.8% and 4.4% mAP for ground truth and non-ground truth identity-speech libraries, respectively. Since a significant improvement upon the TS-TalkNet baseline is apparent when ground truth identity-speech libraries are used, it is fair to deduce SCAN's architectural implementation and method of extracting speaker-specific information from both the candidate audio signal and reference speech is more effective than relying solely on reference speech. Furthermore, the improvement yielded by incorporating SCAN into the baseline systems renders both baseline systems almost competitive with state-of-the-art methods in the context of egocentric data. Specifically, SCAN enhances the TalkNet and Light-ASD baselines by 14.5% and 10.3%, respectively, bridging the gap with state-of-the-art performance, as shown in Table II

Table II
Comparison with the state-of-the-art ASD systems on validation folds of Ego4D and AVA. Values for LoCoNet [14] and SPELL [28] are from their original manuscripts.

System	Spk. Emb. Inf.	Ego4D [%]	AVA [%]
TalkNet [1]	✗	51.0	92.3
TS-TalkNet†	✓	54.0	93.9
Light-ASD [13]	✗	54.3	94.1
LoCoNet [14]	✗	59.7	**95.2**
SPELLL [28]	✗	**60.7**	94.2
TalkNet+SCAN†	✓	58.4	94.0
Light-ASD+SCAN†	✓	59.9	93.9

B. Identity-Speech Library

The results presented in Table III indicate a substantial improvement in the quality of the identity-speech library generated by the proposed method. This improvement is further demonstrated by Figure 4. In the left panel (TS-TalkNet), it is impossible to differentiate same-identity pairs from different identity pairings while in the right panel (SCAN) resolving the two pairings is easier. This is attributed to the face-recognition model's ability to leverage temporal context via self-attention. However, it is noted that the silhouette score of 0.16 indicates only minor cluster separability, suggesting that further refinement of the proposed method might be necessary to achieve more robust future identity-speech library generation.

Table III
Comparison of identity-speech library generation methods.

System	Ego4D-Silhouette
TS-TalkNet	-0.17
SCAN	**0.16**

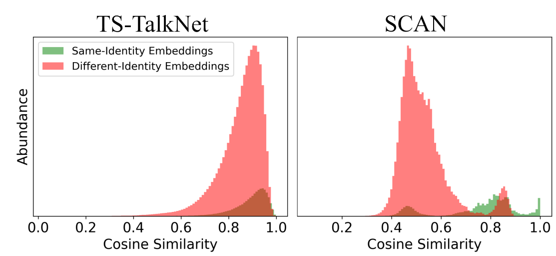

Figure 4. Similarity between same-identity embeddings and different-identity embeddings shown in green and red, respectively, for Ego4D validation fold [10]

V. Conclusion

This work proposes SCAN, a speaker-embedding-informed extension to conventional ASD systems. SCAN assists in disambiguating challenging multi-talker scenarios involving visual noise and physical obfuscations. SCAN builds upon previous work by extracting speaker-specific information from reference speech, but is able to leverage speaker-specific information inherently present in the candidate audio signal itself. Furthermore, SCAN proposes a method to finetune frame-based face-recognition models on video data without hard identity labels by transformer encoder layers and a self-supervised training objective. This approach exhibits a significant performance improvement relative to previous work for identity-speech library generation.

References

[1] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection," in *Proc. 29th ACM Int. Conf. on Multimedia*, 2021.

[2] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Proc. Interspeech 2017*, 2017.

[3] J. L. Alcazar, F. C. Heilbron, L. Mai, F. Perazzi, J.-Y. Lee, P. Arbeláez, and B. Ghanem, "Active Speakers in Context," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] G. Datta, T. Etchart, V. Yadav, V. Hedau, P. Natarajan, and S.-F. Chang, "ASD-Transformer: Efficient Active Speaker Detection Using Self And Multimodal Transformers," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[5] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video," in *British Machine Vision Conference*, 2006.

[6] J. Le'on-Alc'azar, F. C. Heilbron, A. K. Thabet, and B. Ghanem, "MAAS: Multi-modal Assignation for Active Speaker Detection," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[7] Y. Jiang, R. Tao, Z. Pan, and H. Li, "Target Active Speaker Detection with Audio-visual Cues," in *Proc. Interspeech*, 2023.

[8] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, "Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[9] J. Clarke, Y. Gotoh, and S. Goetze, "Improving Audiovisual Active Speaker Detection in Egocentric Recordings with the Data-Efficient Image Transformer," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU23)*, 2023.

[10] "Ego4D: Around the World in 3,000 Hours of Egocentric Video," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[11] "End-to-End Active Speaker Detection, author=Juan Leon Alcazar and Moritz Cordes and Chen Zhao and Bernard Ghanem," in *European Conference on Computer Vision*, 2022.

[12] O. Köpüklü, M. Taseska, and G. Rigoll, "How to Design a Three-Stage Architecture for Audio-Visual Active Speaker Detection in the Wild," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[13] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, and L. Chen, "A Light Weight Model for Active Speaker Detection," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[14] X. Wang, F. Cheng, and G. Bertasius, "LoCoNet: Long-Short Context Network for Active Speaker Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[15] K. Min, S. Roy, S. Tripathi, T. Guha, and S. Majumdar, "Learning Long-Term Spatial-Temporal Graphs for Active Speaker Detection," in *Euro. Conf. on Computer Vision*, 2022.

[16] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech 2020*. ISCA, Oct 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2650

[17] "Personal VAD 2.0: Optimizing Personal Voice Activity Detection for On-Device Speech Recognition, author=Shaojin Ding and Rajeev Rikhye and Qiao Liang and Yanzhang He and Quan Wang and Arun Narayanan and Tom O'Malley and Ian McGraw," 2022. [Online]. Available: https://arxiv.org/abs/2204.03793

[18] B. Zeng, M. Cheng, Y. Tian, H. Liu, and M. Li, "Efficient Personal Voice Activity Detection with Wake Word Reference Speech," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 241–12 245.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech 2017*. ISCA, Aug. 2017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[20] Y. Zheng, D. K. Pal, and M. Savvides, "Ring Loss: Convex Feature Normalization for Face Recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[24] X. An, X. Zhu, Y. Xiao, L. Wu, M. Zhang, Y. Gao, B. Qin, D. Zhang, and Y. Fu, "Partial FC: Training 10 Million Identities on a Single Machine," *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2020.

[25] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *NIPS*, 2017.

[26] J. Cartucho, R. Ventura, and M. Veloso, "Robust Object Recognition Through Symbiotic Deep Learning In Mobile Robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[27] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[28] T. Ishibashi, K. Ono, N. Kugo, and Y. Sato, "Technical Report for Ego4D Long Term Action Anticipation Challenge 2023," 2023. [Online]. Available: https://arxiv.org/abs/2307.01467