
Proceedings Paper:
Clarke, J. orcid.org/0000-0002-1032-6472, Gotoh, Y. and Goetze, S. (Accepted: 2025) Ensembling synchronisation-based and face–voice association paradigms for robust active speaker detection in egocentric recordings. In: Speech and Computer: 27th International Conference, SPECOM 2025 Szeged, Hungary, October 13-14, 2025, Proceedings. SPECOM 2025, 13-14 Oct 2025, Szeged, Hungary. Lecture Notes in Computer Science . Springer Cham ISSN: 0302-9743 EISSN: 1611-3349 (In Press)

Ensembling Synchronisation-Based and Face–Voice Association Paradigms for Robust Active Speaker Detection in Egocentric Recordings

Jason Clarke[1], Yoshihiko Gotoh[1], and Stefan Goetze[1,2]

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

Abstract. Audiovisual active speaker detection (ASD) in egocentric recordings is challenged by frequent occlusions, motion blur, and audio interference, which undermine the discernability of temporal synchrony between lip movement and speech. Traditional synchronisation-based systems perform well under clean conditions but degrade sharply in first-person recordings. Conversely, face-voice association (FVA)-based methods forgo synchronisation modelling in favour of cross-modal biometric matching, exhibiting robustness to transient visual corruption but suffering when overlapping speech or front-end segmentation errors occur. In this paper, a simple yet effective ensemble approach is proposed to fuse synchronisation-dependent and synchronisation-agnostic model outputs via weighted averaging, thereby harnessing complementary cues without introducing complex fusion architectures. A refined preprocessing pipeline for the FVA-based component is also introduced to optimise ensemble integration. Experiments on the Ego4D-AVD validation set demonstrate that the ensemble attains 70.2% and 66.7% mean Average Precision (mAP) with TalkNet and Light-ASD backbones, respectively. A qualitative analysis stratified by face image quality and utterance masking prevalence further substantiates the complementary strengths of each component.

Keywords: Face-voice association, Audiovisual active speaker detection, egocentric recordings

1 Introduction

Audiovisual active speaker detection (ASD) involves identifying the framewise speaking activity of a candidate speaker through the joint analysis of audio signals and temporally aligned face tracks [2, 11, 14, 20, 22, 27, 30]. Traditional ASD systems rely on modelling the temporal correspondence between speech in the audio signal and visual speech-related cues—such as lip movement or cheek posture [11]—in the candidate speaker's face track. These synchronisation-based

approaches assume audiovisual alignment as a prerequisite for detecting speech activity; this assumption dominates modern methods [22, 24, 30, 33]. Extensions to this framework incorporate contextual cues pertaining to inter-speaker relationships [21, 24] and latent information describing the audible context of each scene [9], these extensions help to address multi-talker scenarios and environmental noise, respectively. However, such methods remain fundamentally contingent on the discernibility of audiovisual synchrony, resulting in these approaches still being vulnerable to the challenges posed in egocentric settings [10, 15].

In egocentric recordings, e.g. captured by head-worn recording devices, such as smart or augmented reality (AR) glasses, synchronisation-based ASD performance deteriorates significantly when compared to their performance on exocentric benchmarks [27]. This is largely attributed to the prevalence of visual occlusions, motion blur, and audio interference from overlapping speech or environmental noise [9, 10, 15, 17, 18, 33], all of which are common challenges in egocentric data. Since sychronisation-based methods require sustained discernable audiovisual cues, these challenges significantly degrade their performance.

To circumvent these limitations, recent work by the authors of this paper has explored using face-voice association (FVA) for the task of ASD, as exemplified by the Self-Lifting for Audiovisual Active Speaker Detection (SL-ASD) architecture [3]. Generally, FVA [7, 25, 28, 34] concerns the task of attributing presegmented speaker-invariant utterances to visible identities using cross-modal biometric information rather than temporal alignment. The SL-ASD architecture [3] builds upon this concept by adapting FVA [7] for ASD. This type of approach identifies and leverages transient high-quality facial frames to establish robust voice-face mappings, bypassing the need for fine-grained audiovisual cues being consistently discernable. Prior work [3] has demonstrated robust performance in the context of egocentric recordings achieving mAP scores close to the state-of-the-art despite using significantly less learnable parameters, exclusively for the task of ASD. However, it has been observed [3] that solely relying on face-voice associations introduces two main limitations: face-voice associations falter during speaker-variant utterances (i.e. overlapping speech), and missed speech detections by the speaker-invariant front-end are harshly penalised when the pipeline is evaluated for ASD, holistically. These shortfalls are distinct to the limitations of synchronisation-based methods which struggle more with visual degradation but typically have good recall when the speech signal is audible [9,10,19]. By leveraging the complementary strengths of these two paradigms, this work extends the existing SL-ASD approach [3] and proposes a simple yet effective ensemble approach that combines the benefits of synchronisation-agnostic (i.e. FVA-based) and synchronisation-dependent methods of ASD.

More precisely, the presented system integrates two symbiotic components as an ensemble: (i) a synchronisation-based model that captures temporal audiovisual correspondence [22, 30], and (ii) a speaker-invariant segmentation front-end paired with a FVA module, derived from prior work [3] but with refinements for enhanced ensemble performance. The proposed ensemble aggregates output probability sequences from both systems, via weighted averaging, which miti-

gates each component's divergent failure modes. Although the ensemble mechanism is architecturally lightweight—requiring only a weighted mean fusion of two probability streams—its empirical efficacy demonstrates that synergistic modality insights can be leveraged without complex cross-model attention or gating networks. This simplicity encourages easier deployment on resource-constrained wearable devices.

Contributions:

1. A lightweight late-fusion ensemble method for ASD that combines synchronisation-based and FVA-based models, improving robustness under visual occlusion and audible noise.
2. A refined preprocessing pipeline for SL-ASD to optimise ensemble performance.
3. Empirical validation on Ego4D-AVD: the ensemble achieves 70.2% and 66.7% mAP for two synchronisation-based components (TalkNet and Light-ASD), marking a new state-of-the-art in the domain of egocentric ASD.
4. Qualitative analysis of performance including granular evaluations stratified by Face Image Quality Assessment (FIQA) and randomised utterance masking prevalence to demonstrate the vulnerabilities and strengths of each component of the ensemble.

2 Methodology

This section first provides a brief overview of the typical single-candidate synchronisation-based paradigm used for ASD in Section 2.1, and then describes the FVA-based approach to ASD used by this work in Section 2.2. Finally, the details of the proposed ensemble method, which effectively combines the two synergistic approaches, are presented in Section 2.3.

2.1 Synchronisation-based Approach to Active Speaker Detection

Conventional single-candidate ASD systems operate by assessing the temporal alignment between cues indicative of speech in a given face track signal and the concurrent audio signal as illustrated in Figure 1.

A face track $\mathcal{V}_S = \{\mathbf{V}_{\mathrm{S},1}, \ldots, \mathbf{V}_{\mathrm{S},T}\}$ is defined as a sequence of T contiguous bounding box face crops $\mathbf{V}_{\mathrm{S},t} \in \mathbb{R}^{H \times W}$ of height H and width W, centred on a single candidate speaker S and the concurrent audio signal is defined as a vector of T_A waveform samples $\mathbf{a} \in \mathbb{R}^{T_\mathrm{A}}$ (note that T_A differs from T due to frame rate differences in audio and video modalities).

First, an audio encoder processes the audio signal $\mathbf{a}$, and a video encoder processes face tracks $\mathcal{V}_S$, each producing an embedding with shared dimensions. Specifically, the audio branch yields $\mathbf{F}_\mathrm{A} \in \mathbb{R}^{T \times d}$ and the video branch yields $\mathbf{F}_\mathrm{V} \in \mathbb{R}^{T \times d}$, where d is the embedding dimension of the respective encoders. These two embeddings are then fused to create a single multimodal representation $\mathbf{F}_\mathrm{AV}$. Common fusion operations include channel-wise concatenation,

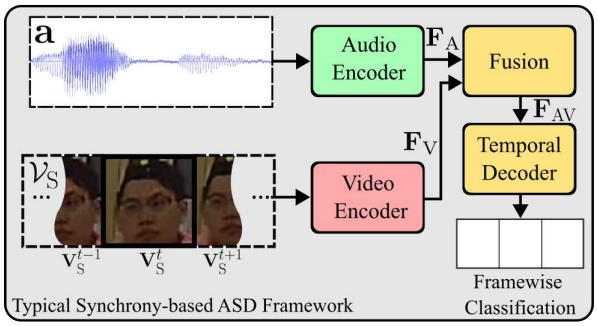

Fig. 1. Typical synchronisation-based single-candidate approach to ASD [22, 30].

element-wise summation [22], or attention-based weighting [30]. Regardless, $\mathbf{F}_\mathrm{AV}$ encodes both audio and visual information at each video-frame.

Finally, a temporal decoder (e.g. a lightweight transformer or temporal convolutional network) is applied along the T dimension of $\mathbf{F}_\mathrm{AV}$ to model longer-range dependencies in speech activity. A frame-wise classification head then produces probabilities indicating whether the candidate is active at each video-frame. This pipeline—embodied by architectures such as TalkNet [30] and Light-ASD [22]—relies fundamentally on audiovisual synchrony, requiring high-quality lip motion and clean audio to be consistently available for accurate detection.

2.2 Face-Voice Association for Active Speaker Detection

The face–voice association approach to ASD replaces the need for explicit audiovisual synchronisation-based modelling by leveraging cross-modal biometric correspondence. This paper follows prior work, specifically the SL-ASD architecture [3], but deviates in terms of preprocessing implementation which has been optimised by this study for the ensemble approach described in Section 2.3. Hence, the method proposed here will be denoted as SL-ASD†, which is outlined as follows.

Front-end Segmentation and Embedding Let $\mathcal{C}$ denote the set of video clips in a given dataset. First, an off-the-shelf speaker-diarisation front-end [4] is applied to the audio signal $\mathbf{a}_c$ of each clip c, segmenting each clip into a set of speaker-invariant utterances. Each utterance $\mathbf{u}_{c,i}$ is then embedded by a pretrained speaker-recognition model [12] yielding an embedding $\mathbf{u}'_{c,i} \in \mathbb{R}^{d_\mathrm{S}}$ for all utterances, where d_S is the embedding dimension of the speaker recognition model. Collectively, these embeddings form $\mathcal{U}' = \{\mathbf{u}'_{c,i} \mid c \in \mathcal{C},\ i = 1, \ldots, N_c\}$, where N_c is the number of utterances in clip c. For this segmentation, the Pyannote Audio diarisation model [4] is used because of its robust performance in the task of audio-only diarisation [32].

Additionally, every face-crop image $\mathbf{V}_{\mathrm{S},T}$ in the dataset is embedded by a pretrained face-recognition model [29] yielding a hierarchical set of face-recognition

embeddings $\mathcal{X} = \{\mathbf{X}_{c,s} \mid s \in \mathcal{S}_c,\ c \in \mathcal{C}\}$, where each matrix $\mathbf{X}_{c,s} = [\mathbf{x}_{c,s,1}, \mathbf{x}_{c,s,1}, ..., \mathbf{x}_{c,s,T_{c,s}}]$ contains face embedding vectors per speaker s in clip c and different $\mathbf{X}_{c,s}$ may be of different size due to variability of frames $T_{c,s}$ per clip and speaker. $\mathcal{S}_c$ is the set of visible identities in clip c, and $T_{c,s}$ is the number of frames for identity s in clip c.

Self-Lifting for Active Speaker Detection During training, the audio component of each batch consists of several speaker-embeddings $\mathbf{u}'_{c,i}$ sampled from $\mathcal{U}'$ ensuring each utterance was taken from the same clip and spoken by the same identity (as per groundtruth annotation). During inference, since groundtruth annotation for utterance identity is not available, the audio component of each batch is simply a single speaker embedding. For both training and inference, the visual component of each batch comprises $\{\mathbf{X}_{c,s} \mid \forall s \in \mathcal{S}_c, \}$, where c refers to the clip from which the speaker embedding(s) in the audio component of the batch were taken from. Each component of the batch is then fed through the relevant branch of the pretrained Self-Lifting [3] model, resulting in $\mathbf{U}'' \in \mathbb{R}^{N_u \times d}$ and $\mathbf{X}'_c \in \mathbb{R}^{|\mathcal{S}_c| \times (\max_{s \in \mathcal{S}_c} T_{c,s}) \times d}$, from the audio and visual branches, respectively. Here, N_u denotes the number of utterances in the audio component of the batch, which is set to 1 during inference. To account for variable visual quality—common in egocentric footage—a lightweight transformer encoder is applied over each sequence dimension (frame dimension) for each visible identity in $\mathbf{X}'_c$. Through its self-attention mechanism, low-quality frames (e.g. blurred or occluded) are down-weighted, and the resulting sequence is mean-pooled to produce a single quality-aware face-recognition embedding for each identity in the visible component of the batch, resulting in $\mathbf{X}''_c \in \mathbb{R}^{|\mathcal{S}_c| \times 1 \times d}$.

Finally, cross-modal association scores are computed by measuring similarity between the embedded utterance and each aggregated face-recognition embedding in the video component of the processed batch, as illustrated in Figure 2. Specifically, scaled dot-product cross-attention is used to produce a matching probability that a given utterance was spoken by each visible identity. This pure face–voice association pipeline thus attributes each speech segment to the most likely visible identity, relying only on biometric consistency rather than audiovisual synchrony.

2.3 Ensembling Synchronisation-Based and FVA-based Approaches to Audiovisual Active Speaker Detection

While synchronisation-based (cf. Section 2.1) and FVA-based approaches (cf. Section 2.2) offer complementary strengths, each exhibits vulnerabilities under challenging audiovisual conditions when used in isolation. To mitigate these limitations, an ensemble strategy is employed which fuses predictions from both paradigms by averaging their respective probability sequences.

Let $\mathbf{p}_{\text{sync}} \in [0,1]^T$ denote the frame-level speaking probabilities predicted by a synchronisation-based model for a given face track. Let $\mathbf{p}_{\text{assoc}} \in [0,1]^T$ denote the probability sequence derived from the FVA-based model for the same

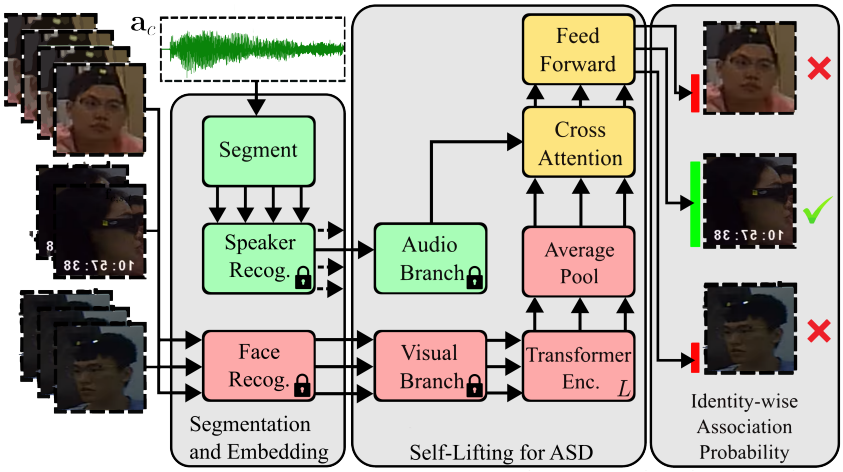

Fig. 2. SL-ASD† framework. Colours indicate modality. Bars adjacent to faces on the right indicate probability of a face-voice match.

hypothesis track as in $\mathbf{p}_{\text{sync}}$. Forming $\mathbf{p}_{\text{assoc}}$ is acheived by projecting the face-voice matching probability uniformly across all concurrent frames in each face track that temporally overlaps with the given utterance. The final ensemble prediction $\mathbf{p}_{\text{ens}}$ is then computed via framewise weighted mean averaging, where α is a mixing coefficient determined empirically:

$$\mathbf{p}_{\text{ens}} = \alpha\,\mathbf{p}_{\text{sync}} + (1-\alpha)\,\mathbf{p}_{\text{assoc}}. \tag{1}$$

This late-fusion scheme requires no additional training and yields a probability sequence that integrates both dynamic synchronisation cues and cross-modal biometric consistency. The resulting ensemble consistently outperforms either constituent method when used in isolation, particularly in scenarios with degraded visual quality or non-frontal faces (cf.Section 4.1).

3 Experiments

This section briefly introduces the egocentric Ego4D dataset used in this work in Section 3.1, the implementation details in Section 3.2, and the evaluation metrics used throughout in Section 3.3.

3.1 Ego4D Dataset for Egocentric Audiovisual Diarisation

The Ego4D dataset [15] comprises egocentric video recordings, totalling 572 unique clips each lasting five minutes in duration, some of which were captured simultaneously. The data was obtained using various wearable devices using

1080p video. The audio signals are standardised to a single-channel in 16 kHz format. Video-frames were recorded at 30 Hz. The dataset reflects real-world conditions – featuring fluctuating lighting, frequent occlusions, and continuously changing viewpoints – making it a particularly demanding testing scenario for ASD. Ego4D-AVD is divided into three non-overlapping folds: 379 clips for training, 50 for validation, and 133 for testing. Because test labels are withheld, the original training set was further split by this work into 110 clips for model training and 23 for development, preserving the reserved validation set for final evaluation. Splits were created to ensure that no individual appears in more than one fold.

3.2 Implementation Details

Synchronisation-Based Models For this component of the ensemble two different ASD systems were used as baselines, namely TalkNet [30] and Light-ASD [22]. These architectures were implemented under the exact configurations and hyperparameters specified in their original manuscripts apart from the training duration. Each model was trained independently 10 times for 30 epochs, and the checkpoint achieving the best performance on the development-set of Ego4D was selected. Finally, the selected checkpoints were employed to generate the synchronisation-based predictions incorporated into the ensemble.

Self-Lifting for Audiovisual Active Speaker Detection The SL-ASD† implementation was similar to that described in [3]. Specifically, the front-end utterance segmentation was performed on a clipwise basis using the Pyannote.audio-speaker-diarization-3.1 system [4] to extract speaker-invariant utterances. Speaker-recognition embeddings were obtained from these utterances using the ECAPA-TDNN [12] model, pretrained on VoxCeleb2 [8]. Face-recognition embeddings were extracted from all face-track frames in the dataset via Inception-V1 [29] pretrained on VGG-Face2 [5]. For finetuning of the Self-Lifting audio and video encoder branches, the model was instantiated with the implementation described in its original manuscript [7], except the number of cluster centroids, which was reduced to 50 to better reflect the number of distinct identities present in Ego4D. In the ASD adaptation (SL-ASD [3]), all original framework parameters were frozen, and only the transformer encoder, the cross-attention module, and the feed-forward layer were trained explicitly for ASD (cf. Figure 2). During training, each batch's audio component comprised all utterances for a single clipwise identity, while its video component included all face-track frames for every visible identity in the clip; during validation and inference, the audio component was limited to single utterances. Optimisation was carried out using Adam with an initial learning rate of 1×10^{-5}, decayed by a factor of 0.2 every 5 epochs, and a single transformer layer with four attention heads was employed for both the encoder and cross-attention.

Face Quality Assessment To perform a granular evaluation of the various approaches to ASD considered by this work (cf. Section 4.2), a method to quantify

the visual quality of the frames in each face-track was employed. In analogy to the well-established domain of Face Image Quality Assessment (FIQA) [16, 23, 31], the per-frame recognisability of the candidate speaker was inferred via the confidence score produced by the pretrained Multi-task Cascaded Convolutional Neural Network (MTCNN) face detector [35]. Specifically, every cropped face image in a groundtruth track was passed through MTCNN, and the resulting detection probabilities were recorded. These per-frame scores were then averaged to yield a single, track-level quality metric.

3.3 Evaluation Metrics

For holistic evaluation, each system is evaluated for ASD using the Cartucho object detection mAP metric [6], which is in alignment with the mAP protocol established by the PASCAL VOC2012 challenge [13]. This evaluation strategy is consistent with the framework adopted by the Ego4D audiovisual diarization challenge [15] and is widely employed in recent ASD research [9, 10]. Owing to the absence of ground truth annotations for the test folds in Ego4D, all results are reported on its validation folds, in accordance with prevailing conventions in the literature [1, 2, 9, 20, 24, 33]. The validation fold is exclusively reserved for testing purposes and are not used during model development. For the evaluations presented in Section 4.2, the problem is reformulated as a binary classification task, with metrics computed using the scikit-learn [26] implementation of average precision.

4 Results

4.1 Comparison with State-of-the-Art Methods

To assess the efficacy of the proposed ensemble, its performance is evaluated holistically against leading ASD systems. Table 1 summarises mAP and parameter counts for each method on the validation fold of the Ego4D-AVD benchmark.

When fused via weighted averaging, the synchronisation-based TalkNet model in conjunction with the face–voice association-based SL-ASD† model yield a combined mAP of 70.2%, outperforming both individual baselines (TalkNet: 51.0%; SL-ASD: 60.7%) by a substantial margin. Crucially, this gain cannot be attributed merely to increased model capacity. This is illustrated by comparing the performance of an ensemble of two synchronisation-based approaches (TalkNet + Light-ASD) of 64.1% with that of Light-ASD + SL-ASD† of 66.7%. While the former yields a significant improvement over its respective baselines, it still exhibits weaker performance than the latter, despite requiring significantly more learnable parameters. This indicates that combining synchronisation-based approaches with FVA-based approaches leverages truly complementary cues.

Moreover, the TalkNet + SL-ASD† ensemble establishes a new state of the art, surpassing the recent LoCoNet [33] model by 1.8% absolute mAP while using fewer than half of its learnable parameters exclusively dedicated to ASD.

Table 1. Comparison with state-of-the-art ASD systems on the validation fold of Ego4D."ASD Params. [M]" denotes the number of learnable-parameters each system uses exclusively for the task of ASD. All values are taken from published literature except ensemble approaches. SL-ASD† indicates the modified implementation of SL-ASD [3] used by this work.

Model	Ensemble	mAP [%]	ASD Params. [M]
TalkNet [15]	✗	51.0	15.1
Light ASD [10]	✗	54.3	1.0
SL-ASD [3]	✗	59.7	**0.4**
SPELL [18]	✗	60.7	>22.5
LoCoNet [33]	✗	68.4	33.5
Light ASD + TalkNet	✓	64.1	16.1
Light ASD + SL-ASD†	✓	67.1	1.4
TalkNet+SL-ASD†	✓	**70.2**	15.5

This demonstrates that simple late fusion of heterogeneous ASD paradigms can yield superior accuracy-efficiency trade-offs compared to monolithic architectures, even those that effectively leverage contextual information.

4.2 Qualitative Analysis

To further investigate the hypothesis that FVA-based models leverage information complementary to that of synchronisation-based approaches, a stratified evaluation was conducted. Face tracks were grouped into discrete bins based on their average face quality scores, enabling a detailed analysis of model performance under varying degrees of visual degradation, including factors such as blur, occlusion, and suboptimal lighting conditions.

The results of this evaluation, shown in Figure 3, reveal that synchronisation-based models, as speculated [3,9,10], exhibit a significant decline in performance as face quality deteriorates. This sensitivity is attributed to their reliance on precise visual cues—particularly lip movements and cheek posture [11]—that must be consistently discernible throughout the duration of the face track. In contrast, the FVA-based model, SL-ASD†, demonstrates a more stable performance across all quality bins. Its robustness stems from the ability to identify and utilise even a limited number of high-quality frames within a sequence. The transformer encoder within SL-ASD† effectively down-weights low-quality frames and emphasizes those that are most informative for identity recognition. This mechanism allows the model to maintain reliable speaker attribution despite transient visual distortions.

Conversely, Figure 4 conveys the effect of audio degradation on each approach. As the probability of randomised utterance masking is increased, only a modest reduction in average precision is exhibited by the synchronisation-based model, owing to its ability to leverage cross-modal information, in this case video, when the audio is obscured. By contrast, a steeper decline is observed for the face–voice association–based SL-ASD†, since uninterrupted utterance segments

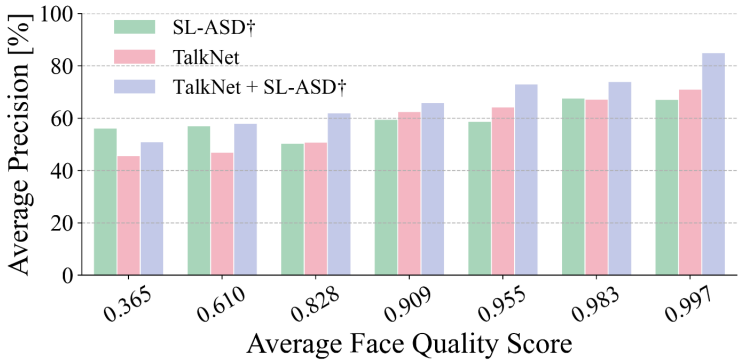

Fig. 3. Comparison of synchronisation-based (TalkNet [30] pink bar), FVA-based (SL-ASD [3], green bar), and ensemble-based (blue bar) approaches to ASD, evaluated on strata of equal size (each comprising tracks with similar average face quality scores). Lower face quality scores indicate tracks with greater visual distortion or occlusion. Irregular face quality score incrementation is due to a non-uniform distribution of trackwise visual quality.

are required by its speaker-invariant front-end for robust speaker embedding extraction. Crucially, higher overall performance across all masking levels is maintained by the ensemble approach, which leverages both streams to compensate for audio distortions that would otherwise impair face–voice association. These findings further substantiate that synchrony-dependent and synchrony-agnostic paradigms leverage complementary information for ASD.

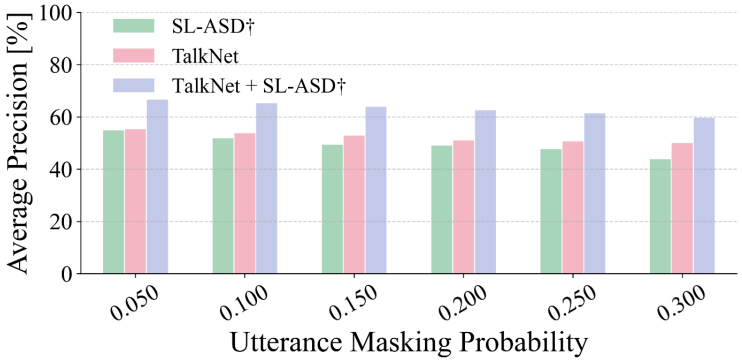

Fig. 4. Comparison of three approaches to ASD on the Ego4D validation set: synchronisation-based (TalkNet [30], pink bar), FVA-based (SL-ASD [3], green bar), and ensemble-based (blue bar) methods. The evaluation is performed with randomised masking applied specifically to utterance regions within the audio signals, simulating various levels of audio signal degradation.

5 Conclusion

In this work, a lightweight late-fusion ensemble for ASD was proposed, combining synchronisation-based and FVA–based models to enhance robustness under visual occlusion and audio interference. The preprocessing pipeline of SL-ASD was refined to optimise its integration within the ensemble, leading to consistent performance gains. Empirical validation on the Ego4D-AVD validation set demonstrated that the ensemble attains 70.2% and 66.7% mAP when paired with TalkNet and Light-ASD backbones, respectively—establishing a new state-of-the-art in ASD. Finally, a qualitative analysis stratified by face quality and utterance masking prevalence was conducted, revealing the complementary strengths and failure modes of each model component. Collectively, these findings substantiate that simple yet principled fusion of synchrony-dependent and synchrony-agnostic streams can reliably mitigate modality-specific degradations in challenging egocentric scenarios.

Acknowledgments. 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.

References

1. Alcazar, J.L., Cordes, M., Zhao, C., Ghanem, B.: End-to-End Active Speaker Detection. In: European Conference on Computer Vision (2022)
2. Alcazar, J.L., Heilbron, F.C., Mai, L., Perazzi, F., Lee, J.Y., Arbeláez, P., Ghanem, B.: Active Speakers in Context. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
3. Authors of this paper (author names redacted, will be added in final version of this paper): Face-Voice Association for Audiovisual Active Speaker Detection in Egocentric Recordings. In: Submitted to European Signal Processing Conference (EUSIPCO) (2025)
4. Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.P.: pyannote.audio: neural building blocks for speaker diarization. In: ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 7124–7128 (2020). https://doi.org/10.1109/ICASSP40776.2020.9054260
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). p. 67–74. IEEE Press (2018). https://doi.org/10.1109/FG.2018.00020, https://doi.org/10.1109/FG.2018.00020
6. Cartucho, J., Ventura, R., Veloso, M.: Robust Object Recognition Through Symbiotic Deep Learning In Mobile Robots. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2018)
7. Chen, G., Zhang, D., Liu, T., Du, X.: Self-lifting: A novel framework for unsupervised voice-face association learning. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. p. 527–535. ICMR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3512527.3531364, https://doi.org/10.1145/3512527.3531364

8. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: Interspeech 2018. ISCA (Sep 2018). https://doi.org/10.21437/Interspeech.2018-1929, https://doi.org/10.21437/Interspeech.2018-1929
9. Clarke, J., Gotoh, Y., Goetze, S.: Improving Audiovisual Active Speaker Detection in Egocentric Recordings with the Data-Efficient Image Transformer. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU23) (2023). https://doi.org/10.1109/ASRU57964.2023.10389764
10. Clarke, J., Gotoh, Y., Goetze, S.: Speaker Embedding Informed Audiovisual Active Speaker Detection for Egocentric Recordings. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2025), https://arxiv.org/abs/2502.06012
11. Datta, G., Etchart, T., Yadav, V., Hedau, V., Natarajan, P., Chang, S.F.: ASD-Transformer: Efficient Active Speaker Detection Using Self And Multimodal Transformers. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2022). https://doi.org/10.1109/ICASSP43922.2022.9746991
12. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In: Interspeech 2020. ISCA (Oct 2020). https://doi.org/10.21437/interspeech.2020-2650, http://dx.doi.org/10.21437/Interspeech.2020-2650
13. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
14. Everingham, M., Sivic, J., Zisserman, A.: Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video. In: British Machine Vision Conference (2006)
15. Grauman, K., et al.: Ego4D: Around the World in 3,000 Hours of Egocentric Video. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
16. Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R., Beslay, L.: Faceqnet: Quality assessment for face recognition based on deep learning. In: 2019 International Conference on Biometrics (ICB). pp. 1–8 (2019). https://doi.org/10.1109/ICB45273.2019.8987255
17. Huh, J., Ortiz, J.A., Kumar, A., Pandey, A., Aroudi, A., Wong, D.D., Nesta, F., Xu, B., Donley, J.: Advancing active speaker detection for egocentric videos. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10888166
18. Ishibashi, T., Ono, K., Kugo, N., Sato, Y.: Technical Report for Ego4D Long Term Action Anticipation Challenge 2023 (2023), https://arxiv.org/abs/2307.01467
19. Jiang, Y., Tao, R., Pan, Z., Li, H.: Target Active Speaker Detection with Audio-visual Cues. In: Proc. Interspeech (2023)
20. Köpüklü, O., Taseska, M., Rigoll, G.: 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). https://doi.org/10.1109/ICCV48922.2021.00123
21. Le'on-Alc'azar, J., Heilbron, F.C., Thabet, A.K., Ghanem, B.: MAAS: Multi-modal Assignation for Active Speaker Detection. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
22. Liao, J., Duan, H., Feng, K., Zhao, W., Yang, Y., Chen, L.: A Light Weight Model for Active Speaker Detection. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (June 2023)

23. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14220–14229 (2021). https://doi.org/10.1109/CVPR46437.2021.01400
24. Min, K., Roy, S., Tripathi, S., Guha, T., Majumdar, S.: Learning Long-Term Spatial-Temporal Graphs for Active Speaker Detection. In: Euro. Conf. on Computer Vision (2022)
25. Ning, H., Zheng, X., Lu, X., Yuan, Y.: Disentangled representation learning for cross-modal biometric matching. IEEE Transactions on Multimedia **24**, 1763–1774 (2022). https://doi.org/10.1109/TMM.2021.3071243
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
27. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., Pantofaru, C.: Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2020). https://doi.org/10.1109/ICASSP40776.2020.9053900
28. Saeed, M.S., Nawaz, S., Yousaf, Khan, M.H., Zaheer, M.Z., Nandakumar, K., Yousaf, M.H., Mahmood, A.: Single-branch network for multimodal training. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2023)
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–9 (2014), https://api.semanticscholar.org/CorpusID:206592484
30. Tao, R., Pan, Z., Das, R.K., Qian, X., Shou, M.Z., Li, H.: Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. In: Proc. 29th ACM Int. Conf. on Multimedia (2021)
31. Terhörst, P., Kolf, J.N., Damer, N., Kirchbuchner, F., Kuijper, A.: Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5650–5659 (2020). https://doi.org/10.1109/CVPR42600.2020.00569
32. Wang, J., Chen, G., Zheng, Y.D., Lu, T.: Exploring detection-based method for speaker diarization @ ego4d audio-only diarization challenge 2022 (2022), https://arxiv.org/abs/2211.08708
33. Wang, X., Cheng, F., Bertasius, G.: 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)
34. Wen, P., Xu, Q., Jiang, Y., Yang, Z., He, Y., Huang, Q.: Seeking the shape of sound: An adaptive framework for learning voice-face association. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16347–16356 (June 2021)
35. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (Oct 2016). https://doi.org/10.1109/LSP.2016.2603342