中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)06-1763-32

论文引用格式:Zhang Y H, Yang S and Shan S G. 2026. Active speaker detection in videos: a survey. Journal of Image and Graphics, 31(6):1763-1794(张远航, 杨双, 山世光. 2026. 视频说话人检测技术综述. 中国图象图形学报, 31(6):1763-1794)[DOI:10.11834/jig.260107]

视频说话人检测技术综述

张远航,杨双,山世光*

1. 中国科学院计算技术研究所,北京　100190; 2. 中国科学院大学,北京　100049

摘　要:视频说话人检测(active speaker detection,ASD)旨在利用音视频信息实现对视频序列中说话人及其发声时段的检测,是人机交互、智能会议系统及媒体内容分析等领域的关键技术。然而,真实场景中的视觉遮挡、环境噪声及多人对话中的语音重叠等复杂因素给该任务带来了严峻挑战。随着深度学习技术的飞速发展,说话人检测领域取得了显著进展。本文回顾该领域的发展历程,并对现有方法进行系统梳理,将其归纳为两类:第1类是基于纯视觉信息的方法,主要解决音频缺失或不可用场景下的检测问题;第2类是音视结合的方法,可进一步细分为:1)基于音视对应的匹配方法,通过学习跨模态对应关系检测音视同步性或建立语音与潜在说话人的身份关联,进而确定说话人;2)基于音视融合的分类方法,将说话人检测任务直接建模为特征融合后的是/非说话人的二分类问题;3)混合方法,结合音视对应与融合分类的互补优势完成检测,以提升鲁棒性。在此基础上,本文还对说话人检测领域常用的数据集与评价指标进行整理。最后,对说话人检测技术的发展趋势进行总结,探讨了当前的若干开放问题,并结合当前的前沿技术进展,展望未来可能的研究方向。相关资源仓库见:https://github.com/VIPL-Audio-Visual-Speech-Understanding/Active-Speaker-Detection。

关键词:说话人检测(ASD);音视频信息;多模态;深度学习;综述

Active speaker detection in videos: a survey

Zhang Yuanhang, Yang Shuang, Shan Shiguang*

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing　100190, China;

2. University of Chinese Academy of Sciences, Beijing　100049, China

Abstract: Active speaker detection (ASD) aims to identify speakers and their active speech intervals within video sequences by leveraging both audio and visual modalities. ASD is a foundational technology for applications such as media content analysis, human-computer interaction, intelligent meeting systems, and audio-visual speech recognition. Despite substantial progress driven by deep learning since 2015, real-world deployment still faces major challenges, including visual occlusion, acoustic interference, overlapping speech, and dynamic camera motion. To address these developments and challenges, this survey reviews ASD research over the past 25 years and categorizes existing methods into vision-based and audio-visual approaches. Vision-based methods infer speech activity solely from visual cues such as lip contours, facial motion, and body gestures, making them useful when audio is unavailable or heavily corrupted. Although immune to acoustic degradation, they inherently struggle to distinguish true speech from non-speech mouth movements and remain highly sensitive to low image resolution, non-frontal head poses, and occlusion. Audio-visual methods, which dominate

收稿日期:2026-02-27;**修回日期**:2026-03-20;**预印本日期**:2026-03-27

*通信作者:山世光　sgshan@ict.ac.cn

基金项目:国家自然科学基金项目(62276247, U24A20332)

Supported by: National Natural Science Foundation of China (62276247, U24A20332)

current research, exploit the complementary strengths of auditory and visual signals. This survey further divides them into three major paradigms. 1) Matching-based methods identify speakers by learning cross-modal correspondences, typically with limited manual annotation. This paradigm includes two routes: synchronization-based and identity-based association. Synchronization-based methods estimate short-term temporal alignment between lip motion and audio, often using contrastive learning to project audio and visual features into a shared embedding space. Although these methods benefit from self-supervised learning, they require tight audio-visual synchronization and can fail under desynchronization or in dubbed videos. Identity-based association methods instead emphasize long-term consistency. They typically cluster speaker embeddings and facial feature sequences separately, then associate voices with faces using co-occurrence statistics or cross-modal face-voice matching networks. This route is more robust to dubbing, off-screen speech, and poor visual quality, such as in egocentric videos, but depends heavily on the accuracy of intermediate clustering. 2) Fusion-based classification methods formulate ASD as a fully supervised binary classification task—speaking versus non-speaking—for each candidate face at each time step. Their pipelines generally include four stages: feature extraction, feature fusion, temporal modeling, and final activity detection. Modern systems use large-scale pre-trained acoustic encoders and deep visual backbones for feature extraction. Dynamic fusion strategies, including cross-attention, gating networks, and uncertainty-aware adaptive fusion, have largely replaced simple static concatenation. Temporal modeling has evolved from local short-term processing to recurrent neural networks (RNNs), and more recently to global spatiotemporal reasoning with Transformers and graph neural networks (GNNs). By explicitly modeling interactions among candidate speakers and the broader scene context, fusion-based classification methods achieve state-of-the-art performance on most benchmarks. However, they require densely annotated data and are vulnerable to domain shift. 3) Hybrid methods combine the complementary strengths of matching and classification paradigms to handle more complex scenarios. By integrating short-term speech behavior, via synchronization or classification, with long-term identity verification through speaker profiles, hybrid systems can suppress interference from non-target speakers, overlapping voices, and off-screen narrators, thereby improving robustness in real-world environments. Beyond this methodological taxonomy, this survey also summarizes benchmark datasets and evaluation metrics commonly used in the ASD community. It traces the evolution of datasets from small, highly constrained laboratory recordings with limited participants to large-scale, in-the-wild benchmarks containing thousands of hours of video from movies, vlogs, egocentric wearable cameras, and surveillance footage. Common evaluation metrics, such as mean average precision (mAP), are also discussed. Finally, this survey highlights emerging technical trends and several persistent open problems. Although current state-of-the-art models achieve near-perfect performance on some benchmarks, they still show limited cross-dataset generalization, particularly across languages, domains, and extreme conditions. Moreover, existing systems still lack deep semantic understanding of conversational dynamics, including turn-taking, interruptions, and non-verbal social cues. Future work should therefore focus on building more inclusive datasets, developing data-efficient learning strategies, integrating large language models and vision-language models (LLMs/VLMs) for higher-level semantic reasoning, and designing lightweight architectures for edge deployment. Additional resources are available at https://github.com/VIPL-Audio-Visual-Speech-Understanding/Active-Speaker-Detection.

Key words: active speaker detection (ASD); audio-visual information; multi-modal; deep learning; survey

0 引　言

视频说话人检测(active speaker detection, ASD)旨在利用视频中的音视信息(尤其是面部运动)判断人物的说话状态,从而实现对画面中说话人及其说话时段的精准检测。如图1所示,ASD系统的输入通常是一段包含若干可能说话人的音视频序列。ASD模型需综合分析面部活动和音频信号等多模态信息,判断画面中每个人物(通常先通过人脸检测和跟踪等前序步骤进行关联)在各时刻的说话状态。利用模型输出的说话概率,可以根据阈值为每个人物分配二值的是/非说话人标签,从而实现对视频中说话人的检测。

ASD技术在多个应用领域中都具有重要作用。在媒体内容智能分析领域,利用ASD技术可以自动对影视剧、新闻、访谈等节目进行对白切分(秦蕾,2009)和剪辑(Girmaji等,2023);结合人脸识别、跟

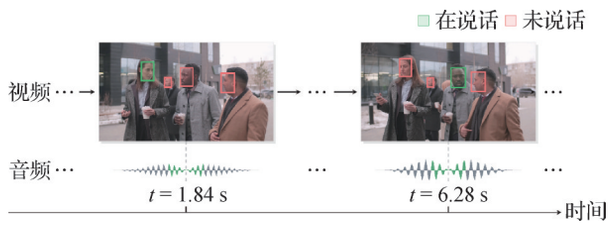

图 1　ASD 任务示意（对话场景）

Fig. 1　Illustration of the ASD task (dialogue scenario)

踪和聚类技术，ASD 的结果可以辅助标注说话人身份（Everingham 等，2006，2009；Hu 等，2015）、统计演员的出镜时长与对白量（Somandepalli 等，2021）等，并为后续的字幕位置调整（Mocanu 和 Tapu，2025；Tapu 等，2019a，b）、字幕（Huh 和 Zisserman，2024；Korbar 等，2024）和剧情梗概生成、人物关系分析以及高光剧情检索等提供基础。在多模态人机交互技术中，ASD 能力是实现流畅、自然交互的前提（Alameda-Pineda 和 Horaud，2015；Sanchez-Riera 等，2012；陶建华 等，2022）。例如，多人交互场景下，机器人或智能助手需要知道当前是谁在对它说话，以做出恰当的回应，如转头、对话等（Cech 等，2013；Ding 等，2024；Stefanov 等，2016，2017）。智能会议系统中，ASD 技术可以实时检测会议中正在发言的与会者，并驱动摄像头追踪发言人或自动切换到发言人对应的画面，实现虚拟导播（virtual cinematography）（Cutler 等，2020；黄星豪，2024）；更进一步，ASD 还可以辅助基于音频的说话人日志（speaker diarization）等模块，实现会议纪要的自动整理、发言者身份标注以及发言时长统计等（Chung 等，2019a）。此外，ASD 作为数据预处理和筛选的关键环节，也为多模态语音识别、唇形合成和视频重配音等下游任务的大规模高质量数据构建提供了重要支撑（Bigata 等，2025；Ephrat 等，2018；Yang 等，2020）。

2015 年以来，深度学习极大推动了计算机视觉与多模态机器学习领域飞速发展，ASD 作为基础性的应用技术得到了学术界和工业界的广泛关注与大力投入。例如在 2019—2023 年，谷歌研究院（Google Research）和 Meta 公司分别在 IEEE 国际计算机视觉与模式识别会议（IEEE Conference on Computer Vision and Pattern Recognition）和欧洲计算机视觉会议上（European Conference on Computer Vision）举办了多届 ActivityNet AVA ActiveSpeaker 挑战赛（https://research.google.com/ava/challenge.html），聚焦于无约束影视剧类视频，以及 Ego4D Talking to Me 挑战赛（https://eval.ai/web/challenges/challenge-page/1625/overview），聚焦于可穿戴设备录制的第一人称视角视频，加速了本领域研究从早期的受限场景向更加复杂和真实的场景迈进。在此背景下，针对 ASD 在真实世界应用中面临的挑战，如对嘈杂环境和视觉遮挡的鲁棒性、多说话人场景中的准确性、跨语言和跨场景的泛化能力以及实时处理能力等，研究者陆续提出了大量创新而有效的方法。及时梳理总结 ASD 领域的最新进展，对于推动该领域发展具有重要意义。

目前，针对该领域的全面系统性综述极为匮乏。此前仅有 Robi 等人（2024）进行过相关梳理，但该文主要将 ASD 视为语音活动检测（voice activity detection，VAD）的多模态扩展，侧重于判断是否存在说话行为，而非解决通常意义下的 ASD 问题。此外，该文对当下主流方法和数据集的覆盖不足，难以全面反映 ASD 领域的最新进展和技术全貌。为填补这一空白，本综述系统梳理了 ASD 领域自 2000 年以来的发展历程，对 ASD 方法进行了全面、系统的综述，并对未来发展趋势进行展望，以期为相关研究与应用提供前沿参考。由于 ASD 与音视语音活动检测（audio-visual voice activity detection，AV-VAD）、说话人定位与跟踪以及说话人日志等相关音视分析任务在方法和目标上密切相关，本综述也会适当纳入相关的研究进展，以展现更完整的技术图景。不涉及视觉绑定的纯音频 VAD 等任务，则不在本综述的讨论范围内。

1　说话人检测方法发展概述

ASD 是一项典型的多模态时序分析任务。如图 2 所示，本文将 ASD 技术近 25 年的发展划分为 4 个阶段，并重点从模态信息利用、时序依赖建模以及实际应用演变 3 个维度进行概述。

1.1　模态利用

ASD 的特征利用经历了从单一模态向多模态横向扩展的过程。传统特征时代（2000—2015 年）多依赖简单几何特征（如唇部几何参数）或浅层表观特征（Cutler 和 Davis，2000；Everingham 等，2006；Slaney 和 Covell，2000）与手工音频特征，信息利用较为单一。进入深度学习时代（2015—2020 年），深度

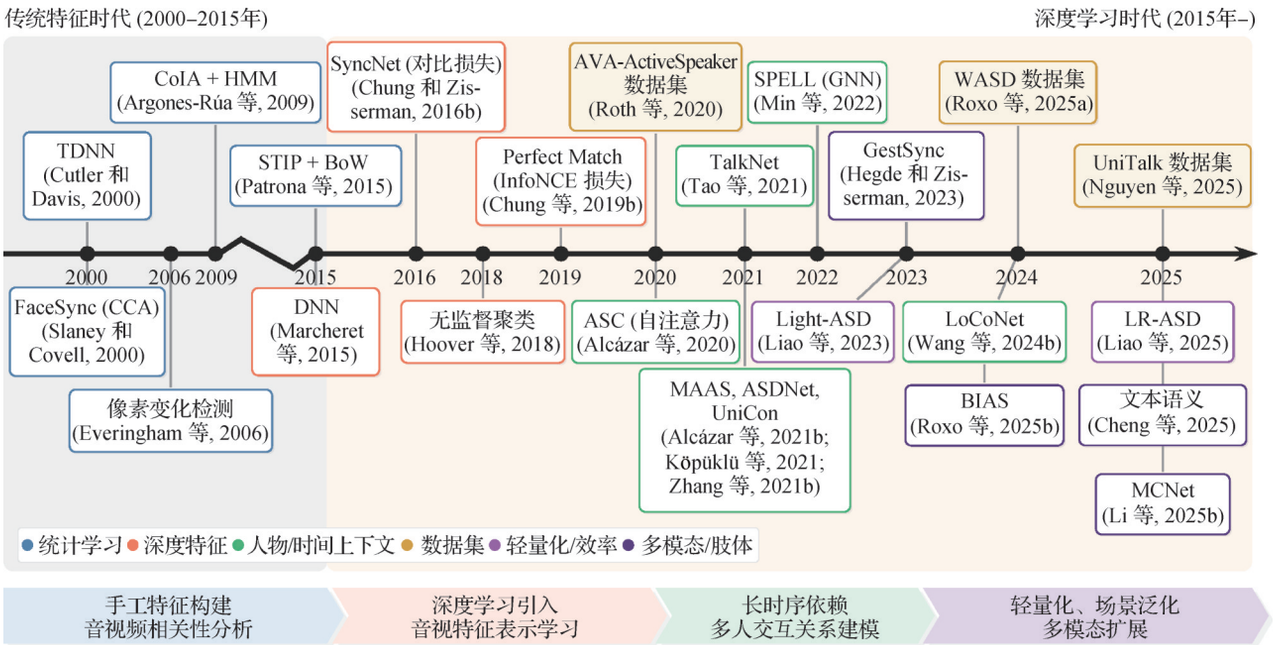

图2　视频说话人检测研究发展时间线

Fig. 2　Timeline of ASD research development

神经网络凭借数据驱动的学习方式，成为更强大的特征提取器（Marcheret 等，2015）。随后，基于双流网络的 SyncNet（Chung 和 Zisserman，2016b）进一步确立了音视双模态结合的方法在 ASD 任务中的基础地位。为应对真实场景下的面部遮挡、低分辨率、噪声等挑战，研究者不再局限于面部和音频特征，引入了肢体语言（Hegde 和 Zisserman，2023；Li 等，2025b；Roxo 等，2025b）、视线模式、场景空间特征（Clarke 等，2023；Zhang 等，2021b）以及文本语义信息（Cheng 等，2025）等辅助模态，显著提升了核心信号受损时模型的判别能力。

1.2　时序建模

在时序处理层面，ASD 技术发展呈现出从短时分析向长时上下文理解演进的趋势。早期方法受模型能力与计算资源限制，多依赖短时（数十至数百 ms）的浅层统计规律。在应对真实对话中的沉默、语速变化及多人交互等挑战时，长时上下文对于准确判定说话状态至关重要。随着循环神经网络的应用（Roth 等，2020；Zhang 等，2019），模型具备了长序列处理能力，能利用上下文弥补局部的信息缺失，实现更鲁棒的连续预测。近年的研究则逐渐转向基于 Transformer（Alcázar 等，2020；Tao 等，2021）和图神经网络（graph neural network，GNN）（Alcázar 等，2021b；Min 等，2022）架构。前者利用自注意力机制捕捉单人的全局时序依赖，后者则通过时空图显式建模多人交互，标志着 ASD 技术实现了从短时判断向对长时对话动态理解的跨越。

1.3　实际应用

在应用与数据层面，ASD 技术正从受控实验室环境向复杂真实场景推进。早期研究多在受控的实验室环境进行，而随着 AVA-ActiveSpeaker（Roth 等，2020）、Ego4D（Grauman 等，2022）、Wilder Active Speaker Detection（Roxo 等，2025a）、UniTalk（Nguyen 等，2025）等大规模基准数据集的发布，研究场景迅速扩展至影视、可穿戴设备及监控等非受控环境。这些数据集涵盖了图像质量、语音重叠以及多样的语言等真实挑战，有力推动了模型泛化能力与鲁棒性的突破。同时，为应对真实世界落地的实时性需求，轻量化设计成为近年的重要新方向，如 Light-ASD（Liao 等，2023）及其后续工作 LR-ASD（light-weight and robust network for active speaker detection）（Liao 等，2025）通过高效架构大幅降低了计算与内存开销。这表明 ASD 技术已从学术探索迈向兼顾“高精度、高效率、强泛化”的实际落地阶段。

2　说话人检测方法

本节将系统梳理当前 ASD 领域的主流方法。

如图 3 所示，ASD 的流程主要包含 4 个关键环节：1)特征提取；2)(可选的)特征融合；3)时序建模；4)最终的说话状态判别。整体而言，现有 ASD 方法可分为基于纯视觉信息的方法和音视结合的方法两类(纯音频无法绑定到画面中说话人)；其中，音视结合的方法又可细分为基于音视对应的匹配方法、基于音视融合的分类方法与混合方法。本节将结合 ASD 技术的演进，详细阐述各环节的关键方法。

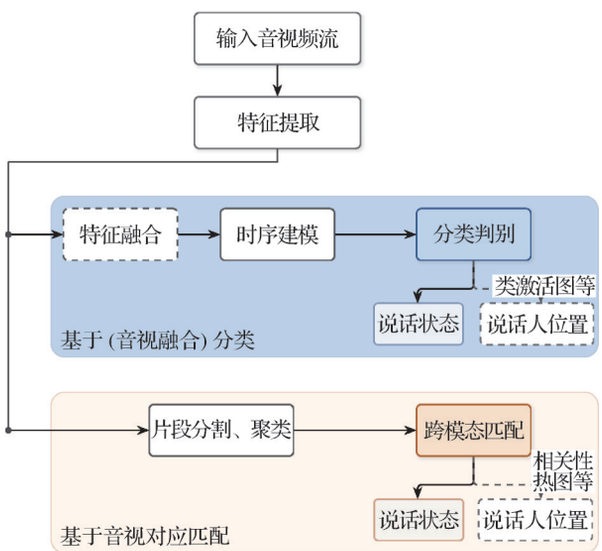

图 3　典型 ASD 方法流程框架示意图

Fig. 3　Framework of typical ASD methods

2.1　特征提取方法

特征提取旨在从高维信号中获取与说话状态相关的紧凑表示。依据模态不同，本文将其分为面部视觉特征、音频特征及其他辅助特征 3 类进行介绍。

2.1.1　面部视觉特征

1)手工特征。早期 ASD 研究主要依赖于手工设计的浅层视觉特征描述面部(尤其是嘴部区域)在说话时的形态和运动变化。基于人脸的视觉特征主要包括几何特征、表观特征、运动特征和时空特征等。几何特征通常是对面部关键点进行定位与跟踪，借助人物嘴唇宽度、高度和面积等参数(Cadavid 等，2009；秦蕾，2009)及其变化的频域分布(Chung 和 Zisserman，2016a)刻画说话状态。表观特征侧重于分析唇部的像素值或纹理模式，常见方法包括像素强度统计(Everingham 等，2006，2009；Siatras 等，2006，2009)、局部二值模式(local binary patterns，LBP)直方图(秦蕾，2009)、离散余弦变换(discrete cosine transform，DCT)变换系数(Thermos 和 Potamianos，2016)及主成分分析(principal component analysis，PCA)降维(何俊和张华，2008)等。运动特征通常利用光流(optical flow)量化说话时的唇部运动信息(Aubrey 等，2010)。时空特征则结合了空间和时间维度的信息，例如采用时空兴趣点或结合词袋模型来表示一段时间内的面部视觉动态(Patrona 等，2015；Tao 等，2015)，以此表示说话行为的时空模式。

2)深度特征。随着深度学习在计算机视觉领域的巨大成功，基于卷积神经网络(convolutional neural network，CNN)如 VGG(Visual Geometry Group)(Chatfield 等，2014；Simonyan 和 Zisserman，2015)、深度残差网络(residual network，ResNet)(He 等，2016)提取的深度面部视觉特征逐渐成为主流(Alcázar 等，2020；Köpüklü 等，2021)。为更好地捕捉说话行为的时空动态，研究者们还广泛采用了动作识别领域的三维卷积网络(Köpüklü 等，2021)、时间移位模块(temporal shift module，TSM)(Lin 等，2019，2022；Min 等，2022；Radman 和 Laaksonen，2024)，以及唇语识别任务中的三维残差网络(3D-ResNet)(Stafylakis 和 Tzimiropoulos，2017)等。值得注意的是，现有方法在输入区域(唇部或人脸)及分辨率(96～224 像素)上选择不一，这对模型性能与实时性影响显著，需在对比与部署时综合考虑。

Transformer 架构(Vaswani 等，2017)(见 2.2 节)因其强大的特征提取和上下文建模能力而应用于 ASD 中的视觉特征提取，如部分模型直接采用视觉 Transformer(vision Transformer，ViT)提取视觉特征(Dosovitskiy 等，2021；Fernandez-Labrador 等，2024)。Prajwal 等人(2022)则提出了与 CNN 配合的视觉 Transformer 池化(visual Transformer pooling，VTP)模块。VTP 模块首先将时空卷积网络提取的逐帧视觉特征图分块，然后用线性 Transformer 编码器提取增强后的特征图。接着，通过可学习的查询向量从增强后的特征图生成视觉注意力权重，对特征图的不同区域进行加权平均，获取各时刻唇部外观和运动的视觉表示，将其用于视觉说话人检测任务。该模块能自动学习和发音相关的面部区域，取得了优异的效果。近期，基于对比语言—图像预训练(contrastive language-image pretraining，CLIP)的 ViT 编码器也被用于 ASD 的视觉特征提取，其在海量数据上学到的强大表征能力能直接捕获与说话行为相关的

显著视觉线索(Appiani和Beyan,2025)。

尽管从视频模态直接提取的表观深度面部特征表现优异,但研究表明,显式地结合面部几何先验信息如唇部关键点(Nguyen等,2026;Wang和Wang,2019)或运动信息如光流(Caus等,2021;Huang和Koishida,2020;Pibre等,2023)仍能为ASD任务带来增益。Geeroms等人(2022)对比发现,稀疏面部关键点特征在理想场景(正面、无遮挡)下精度接近密集图像特征,且计算效率高出2~3个数量级,利于落地;但在非正面姿态或遮挡下,性能有所下降。因此,在实际应用中,视觉特征的选择应综合考量具体场景与效率需求。

最后,除了设计特征提取方式,针对真实视频中常见的图像质量退化问题,专门的输入预处理与数据增广策略也是提升面部特征鲁棒性的关键。例如,针对第一人称视角下(egocentric)摄像机剧烈运动导致的画面抖动与模糊挑战,Yun等人(2024)通过引入球面世界锁定(spherical world-locking)机制将视觉输入与头部朝向对齐,以抵消全局的自身运动干扰;Huh等人(2025)进一步利用面部关键点建立稳定的视觉参考系,实现了唇部区域的精准空间锁定。Huh等人(2025)还通过在训练时随机施加线性模糊核模拟真实运动模糊,从而学习对非稳态视觉输入具有不变性的鲁棒特征。以上策略都在真实场景中取得了显著的性能提升。

2.1.2 音频特征

ASD中的音频特征提取主要借鉴自语音识别和说话人识别领域,其中Mel频率倒谱系数(Mel-frequency cepstrum coefficients,MFCC)和Mel滤波器组(Mel filterbank,Fbank)应用最为广泛。在此基础上,针对特定需求可引入差异化特征,例如在动态建模中引入音高(pitch)特征(Cadavid等,2009);而在说话人聚类这一环节中,基于听觉模型的级联非对称谐振器和快速作用压缩特征(cascade of asymmetric resonators with fast-acting compression,CARFAC)(Lyon,2011)表现优于Mel频谱特征(Hoover等,2018),因此也得到了应用。随着深度学习的发展,研究者通常将上述声学特征或原始音频波形输入深度网络以提取更具判别力的深层表示,并通过时间下采样使其与视频帧率(通常为25帧/s)对齐。常用的网络包括ResNet-18(Alcázar等,2020)、VGG-M(Visual Geometry Group CNN-Medium)(Chung,2019;Chung和Zisserman,2016b;Zhang等,2019)和WaveNet(Ariav和Cohen,2019;van den Oord等,2016)等。近些年,越来越多的工作转向利用基于大规模音频数据预训练的声学编码器,如Wang等人(2024b)尝试加载在音频事件数据集AudioSet(Gemmeke等,2017)上预训练的VGGish模型;Tao(2023)和Zhao等人(2023)则采用基于掩码预测任务在大规模语音数据上自监督学习得到的WavLM(Chen等,2022c)和HuBERT(hidden-unit BERT)(Hsu等,2021)。这些模型提取出的特征包含丰富的语义和时序上下文信息,显著增强了ASD模型对不同人物与场景的泛化能力。由此可见,不同声学特征在特定任务中表现各异,针对具体应用场景和下游模型,仍需对声学特征进行细致的比较与选择。为了进一步提升特征在复杂声学环境下的鲁棒性,部分工作对音频还引入了多样化的数据增广。例如,Tao等人(2021)提出对音频应用负采样(negative sampling)增广,通过叠加同一批次内其他样本的音频作为同领域噪声,学习对噪声和重叠语音更鲁棒的音频特征。Zhang等人(2021a)则引入语音识别中的谱增广(SpecAugment)技术(Park等,2019a),在训练时对语音进行时频掩码以进一步提升模型对噪声、混响等的鲁棒性。

此外,除发音层面的特征外,声音的空间信息(如到达方向、声源位置)在多麦克风场景下也是一类有效的辅助特征。将空间信息与视觉信息(如人脸位置)结合,能显著提升检测精度。例如,Jiang等人(2022)提出一种端到端多通道说话人检测定位模型,通过融合音频的互相关和能量特征与人物视野范围内的视频信息,在360°球面范围内实现了鲁棒的说话人定位,显著提升了真实场景中的检测可靠性。Huang等人(2024)则探索了对会议场景鱼眼镜头的空间频谱特征的利用,通过六通道麦克风阵列生成三维空间频谱热图,结合全景图像的几何变换,将声源定位信息以等距投影方式嵌入场景特征,有效解决了多说话人遮挡场景下的定位偏差问题。

2.1.3 其他特征

为进一步提升ASD在复杂场景下的性能和鲁棒性,研究者们开始探索性地引入面部运动和音频之外的信息和模态提取特征。部分此类特征如下:

1)肢体语言。说话时的头部姿态、手势及上半身动作可作为人脸部分或全部不可见时的有效补

偿，辅助判断说话状态。研究主要通过提取光流或身体关键点等挖掘肢体运动（Beyan 等，2021；Hegde 和 Zisserman，2023；Hegde 等，2025；Hung 和 Ba，2010；Shahid 等，2019，2021），或将其与面部和音频特征融合以提升极端场景下的鲁棒性与可解释性（Carneiro 和 Wermter，2025；Li 等，2025b；Roxo 等，2024a，b，2025a）。此外，轻量的骨架特征也有利于 ASD 的边缘部署（Carneiro 和 Wermter，2025）。

2）视线模式。听众具有倾向于注视说话人的群体行为特性，这一社交信号在会议等场景下可辅助推断人物的说话状态。相关研究通过头部姿态估计会议中的视觉关注焦点（visual focus of attention，VFoA）（Hung 和 Ba，2010），或进一步聚合与会者的注视方向构建场景级视线场（gaze field）（Jiang 等，2025），从而定位说话人。Hradis 等人（2012）探索了在多方视频通话场景下，仅利用单个参与者的眼动追踪数据检测当前活跃说话人的可行性，证明了仅通过听者的注视模式就能以较高准确率判断谁在说话。近期，Durrani 等人（2024）的研究表明，在不依赖音频的情况下，利用所有参与者的连续眼动追踪数据，可以成功训练模型以较高精度识别每个参与者当前扮演的对话角色；在使用 2 s 的输入窗口时，准确率可达 81. 5%。

3）文本。对话内容蕴含丰富的语义，有助于推断对话的上下文和发言轮转逻辑。在语音说话人日志领域，较早就有研究提出可利用自动语音识别（automatic speech recognition，ASR）输出的文本构建语义约束，进而优化声学聚类（Park 和 Georgiou，2018；Park 等，2019b；Shafey 等，2019；Wang 等，2024a）。近期，Cheng 等人（2025）在 ASD 任务中引入文本信息，同样通过分析对话内容检测语义变化点，以此指导说话人的聚类过程。实验表明，在应对复杂对话场景和区分特征相似的说话人时，融合文本信息明显提升了判断的准确率。

除对话内容外，Appiani 和 Beyan（2025）还探索了引入人物行为描述辅助说话状态判别。其 VAD-CLVA 方法利用大型生成式多模态模型如 LLaVA（large language and vision assistant）对视频关键帧进行分析，自动生成描述人物当前行为的描述（例如“该人嘴巴张开，伴有手势，似乎正在参与对话”），并将该描述性文本的特征与视觉特征进行融合，极大地增强了模型区分真实说话行为与干扰性非言语动作的能力。

4）场景空间特征。这类特征旨在通过整合人物位置与整体场景信息，增强模型的空间推理和上下文感知能力。通过将人物的空间位置与其所处的环境相结合，模型不再单纯依赖局部人脸特征，而是能够综合利用全局场景信息评判各个人物在场景和对话中的角色。Zhang 等人（2021b）观察到视频拍摄中镜头常聚焦于说话人，基于此先验设计了二维高斯热力图形式的头部图（head map）表示，将人脸在视频帧中的空间分布与人脸序列特征联合输入模型，建模人物之间的潜在视觉关系，以此辅助完成 ASD 任务；Min 等人（2022）提出一种轻量的空间特征编码方法，将人脸包围框的归一化中心坐标、高度和宽度与面部视觉特征融合后作为图神经网络的节点输入，引入显式的空间位置先验，增强了模型的时空上下文推理能力。Clarke 等人（2023）利用预训练的数据高效图像 Transformer（data-efficient image Transformer，DeiT）模型（Touvron 等，2021），从全场景图像中提取环境声学属性与候选说话人的全局特征，有效克服了第一人称视角视频中运动模糊与语音重叠带来的干扰。Gurvich 等人（2024）针对会议场景引入球面坐标系，有效建模了 360°全景视频中人脸的方位角、高度角及与背景人脸的相对距离。Huang 等人（2024）和黄星豪（2024）引入场景空间谱特征，通过多通道音频生成音频空间谱热图，并将其与全景图像整合，以捕捉场景的空间信息和与会者相对位置关系，有效抑制了无关人物的干扰。

综上所述，ASD 的特征提取已从早期单一的手工音视特征，全面演进为以深度时空网络和预训练大模型为核心，并深度融合肢体、视线、文本及空间先验的多模态特征体系，为提升复杂自然场景下 ASD 鲁棒性奠定了良好的基础。

2. 2 时序建模方法

说话行为是一个连续、动态的过程，涉及丰富的前后时序依赖。准确捕捉和利用这些时序信息对于 ASD 至关重要。越来越多的研究表明，受短暂沉默、语速变化及非语言脸部运动（如大笑、咀嚼）等因素的影响，仅仅依赖短时的音视片段信息（例如几百 ms）可能不足以准确判断说话状态（Alcázar 等，2020；Tao 等，2021）。TalkNet（Tao 等，2021）的研究通过实验证明，将输入模型进行分析的片段长度从 0. 2 s 增加到 2 s，ASD 的性能（平均均值精度）可以

提升超过14%。同样，Alcázar等人（2020）和Wang等人（2024b）的消融研究也表明，增加模型考虑的时间范围（分别达到约2.25 s和8 s）能够带来显著的性能增益。Min等人（2022）甚至设计了能够推理长达数十秒时空上下文的图结构。目前，ASD中常用的时序建模模块包括循环神经网络、Transformer和图神经网络等，下文将分别进行介绍。

2.2.1 循环神经网络

循环神经网络（recurrent neural network，RNN）由于其天然适合处理序列数据的特性，通常被放置在特征提取器之后，用于对音视特征序列进行时序建模，捕捉说话过程中的时间动态和依赖关系。典型的RNN如长短期记忆网络（long-short term memory，LSTM）（Hochreiter和Schmidhuber，1997）和门控循环单元（gated recurrent units，GRU）（Cho等，2014）都在ASD任务中得到了广泛应用（Chung，2019；Liao等，2023，2025；Pouthier等，2021；Roth等，2020；Zhang等，2019）。此外，Zhang等人（2022）注意到在电影等长视频中，人物往往会在不同镜头（场景）中反复出现。为此，该工作引入了一个额外的GRU对不同场景间同一人物的说话行为特征和身份特征进行收集和传递，获取更大范围的时序上下文信息，提升了长视频中ASD的性能。

2.2.2 Transformer

基于注意力机制（attention mechanism）的Transformer模型在各类序列建模任务中取得了突破性进展。凭借其在长距离依赖建模和并行计算方面的显著优势，Transformer已成功引入到ASD任务（Datta等，2022；Truong等，2021）中。Transformer的核心是其注意力机制：在注意力层中，输入的序列首先被投影为查询（query）、键值（key）和值（value）3个矩阵。随后，模型通过计算查询与键特征之间的相似度获取注意力权重。最终，将这些权重与对应的值特征进行加权求和，从而得到融合了上下文信息的输出表示。当查询、键和值特征均提取自同一个输入序列时，构成自注意力（self-attention），在ASD任务中通常用于捕捉同一模态内在时间维度上的长距离依赖关系；当查询特征来自一个序列，而键和值特征来自另一个序列时，则形成交叉注意力（cross-attention）机制，ASD任务中通常用于不同模态间的动态对齐与融合。

ASC（active speakers in context）（Alcázar等，2020）是较早在ASD任务中引入自注意力机制以建模上下文信息的方法。TalkNet（Tao等，2021）明确指出了长时序建模的重要性，并在基于交叉注意力的音视特征融合之后，使用自注意力对单模态特征进行了修正。Truong等人（2021）利用Transformer挖掘和优化音视频信号的时空关系，用于区分对话中的主要发言者和打断者。ASD-Transformer（Datta等，2022）提出了一个结合自注意力和多模态注意力的高效Transformer框架。同时，最早在语音识别领域提出的Conformer模型（Gulati等，2020）作为Transformer的一种变体，通过结合卷积增强局部上下文建模能力，也在ASD任务中得到了广泛应用（Kyoung和Song，2023；Zhang等，2021a，2022）。虽然Transformer性能强大，但其计算和内存开销也是一个需要考虑的问题。一些研究开始关注Transformer模型的效率，提出了更轻量化或适用于流式处理的Transformer结构，探索如何在满足实时性需求的同时，平衡过去和未来上下文信息对ASD性能的影响（Kundu等，2025）。

2.2.3 图神经网络

图神经网络为建模视频中复杂的时空依赖和实体交互提供了统一的计算框架。不同于CNN和RNN处理规则网格或序列数据，图神经网络能够处理非欧几里得空间的图结构数据，这使得它非常适合用于建模ASD任务中多模态特征之间（如音频与视频）、多个人物之间（如对话者交互）以及长时序各帧之间（如上下文依赖）的复杂关系。

图神经网络的基本计算过程通常包含图构建和消息传递两个阶段。首先是图构建，通常将视频中的关键元素（如每一帧的面部特征、音频特征等）定义为图中的节点（node），并依据实体间的关系（如时间相邻、空间邻近或语义相似性）构建边（edge），形成图结构 $\mathcal{G}=(\mathcal{V},\mathcal{E})$。其次是基于消息传递（message passing）的特征更新，其核心思想是让每个节点聚合其邻居节点的信息更新自身的特征表示。在ASD中，图结构能够灵活地建立远距离帧的连接，聚合跨越长距离的时间上下文，从而生成更具判别力的说话人时序特征。例如，MAAS（multi-modal assignment for active speaker detection）（Alcázar等，2021b）将ASD问题建模为多模态分配任务，构建音视关系图（包含局部和时序分配网络），通过图卷积（Kipf和Welling，2017）在节点间传递信息，将语音

事件分配给最可能的视觉源。SPELL（spatial-temporal graph learning）（Min 等，2022）则构建了稀疏的双向时空图，利用双向图神经网络在长达数十秒的窗口内显式地融合跨帧的时序演变信息与同帧内的空间交互特征，在降低计算开销的同时有效捕捉了长时依赖。EASEE（end-to-end active speaker detection）（Alcázar 等，2022）设计了交错图神经网络，将多说话人的空间维度交互与长时序上下文的消息传递过程解耦，实现了端到端的特征学习与上下文聚合优化。Min（2023）针对第一人称视角视频中的说话人日志任务引入了 STHG（spatial-temporal heterogeneous graph）框架，将可见说话人和不可见的相机佩戴者设置为异构图中不同类型的节点，联合建模所有参与者的时空交互，提升了对全员说话状态检测的性能。

总体而言，时序建模旨在捕捉说话行为的时间动态与长程依赖。传统的 RNN 架构在长序列场景下存在记忆瓶颈，而 Transformer 凭借自注意力机制显著增强了对人物说话行为的长时建模能力，成为当前主流；拓扑更为灵活的图神经网络（GNN）则为建模长时多说话人交互提供了新工具。

2.3 特征融合方法

如何有效利用音频、人脸和肢体动作等不同模态特征间的互补性与关联性，进而实现其有效融合是提升 ASD 模型性能的关键。根据模态间的交互方式与融合机制，现有的特征融合策略大致可分为静态与动态融合两种。静态融合中最直接的方式是朴素融合，即通过逐元素相加（加性）（Roxo 等，2024b）、逐元素相乘（乘性）或最常见的拼接（concatenation）操作将两路特征直接合并。动态融合则根据输入特征的语义内容自适应地调整模态间的融合权重或交互方式，以更灵活地捕捉模态间的复杂关联。在众多的动态融合方式中，应用最为广泛的是基于交叉注意力机制的融合。它允许一个模态的特征作为“查询”，动态地关注并聚合另一个模态中的相关信息，从而实现模态对齐与信息增强。Tao 等人（2021）提出的 TalkNet 模型是采用这一方式的代表性工作，该模型通过让音频特征与视觉特征相互查询，并对双向交互特征进行拼接融合，从而有效建模音视模态同步信息。Liao 等人（2023）提出了一种基于门控（gating）机制的轻量自适应融合策略，它通过一个可学习的“门”动态控制两路模态信息在融合特征中的贡献权重，并对两路模态特征进行加权求和得到融合特征。Wang 等人（2026）提出分层门控跨模态融合机制 GateFusion，在 Transformer 的多个层中引入 HiGate（hierarchical gated fusion decoder）解码器，根据音频与视觉特征自适应生成门控权重，将一个模态的隐藏状态逐层注入另一模态表示，从而实现细粒度、上下文感知的双向音视融合。Zhou 等人（2021）和周恒顺（2023）还提出使用因子分解双线性池化（factorized bilinear pooling，FBP）融合音频和唇部特征，通过计算两种模态特征向量的外积刻画它们之间丰富的二阶交互关系，同时保持计算效率。近期，Li 等人（2025b）和 AFs-Net（Yin 等，2025）提出基于异构图神经网络（heterogeneous graph neural network，HGNN）的特征融合策略，将音频、视觉等特征设为异构节点，通过图注意力机制（Veličković 等，2018）动态聚合各模态的互补信息。Li 等人（2025b）还通过设置统一的查询节点引入肢体模态辅助，显著增强了系统在遮挡、低光照等复杂环境下 ASD 的鲁棒性。

现实场景中，单一模态的信号质量可能因环境因素而发生剧烈变化（如强噪声影响音频质量，遮挡影响视觉质量）。为了提高模型的鲁棒性，一些研究探索基于不确定性度量或信号质量的融合策略，在融合时给予更可靠的模态更大的权重，从而实现自适应的、更鲁棒的决策。例如，Pouthier 等人（2021）将 ASD 问题视为多目标学习任务，使用基于自注意力机制和不确定性度量的融合模块对音频和视觉特征进行动态加权融合。其中，对不确定性的估计利用了两类指标：一是数据自身的特性，如候选人数、人脸分辨率等能够间接反映数据质量的高层属性；二是由温度缩放（temperature scaling）（Guo 等，2017）校准后的 softmax 分数给出的单模态预测置信度。He 等人（2024）通过对比学习训练音频和视觉特征提取网络，得到音视同步的置信度，并将该置信度融入音视交叉注意力机制中，根据音频和视频的同步状态，自动选择全融合、部分融合或不融合等不同策略。

在真实场景中，理解个体间的交互关系对判断说话人至关重要。除了不同模态特征的融合外，不同人物特征的全局融合同样是提升检测效果的关键因素。为此，研究者们提出了多种多人上下文建模方法，通过将不同潜在说话人的特征信息进行融合，

显式或隐式地建模人物的相互关系。ASC模型(Alcázar等,2020)开创性地使用自注意力机制进行上下文建模,构建了一个包含所有候选人的长时序特征的“说话人上下文”张量,通过自注意力机制与LSTM实现任意候选人与任意时间步特征的两两交互。这种全连接式的交互使得每个候选人的特征都能感知到全局的对话动态,从而捕捉隐式的多说话人关系。

Zhang等人(2021b)进一步提出了UniCon(unified context network),显式融合空间位置、人物关系与时序上下文线索,通过“非说话人抑制”模块聚合所有候选人的特征生成全局一致性参照,并据此动态调整个体特征,有效增强了潜在说话人的信号特征并抑制了非说话人的干扰。ASDNet(Köpüklü等,2021)采用分阶段优化特征提取、多人关系建模与时序建模的策略,在多人特征融合阶段将目标候选人特征与投影压缩的背景候选人特征拼接,从而利用背景人物信息辅助判别。陈世俊(2022)先通过交叉注意力获取单帧内各候选人音视交互后的特征,再对所有候选人通过简单的卷积操作计算上下文特征。近期,LoCoNet(long-short context network)(Wang等,2024b)指出,在建模短程(如7帧)人物关系时,简单的卷积操作凭借其局部归纳偏置,在效率和性能上均优于复杂的自注意力机制,为轻量化的多人物特征融合提供了新思路。

概括来说,特征融合是连接特征提取与决策判别的桥梁。简单的静态拼接虽然易于实现,但难以捕捉复杂的跨模态交互;基于注意力的动态融合机制通过软对齐实现了更深层次的信息整合,已成为主流范式;而引入图结构则进一步拓展了多模态融合的灵活性。此外,考虑到现实场景中信号质量的波动,基于不确定性估计的自适应融合策略正受到越来越多的关注,以保证模态不可靠场景下的鲁棒性。

2.4 说话状态判别方法

前面几节所介绍的特征提取、特征融合及时序建模等环节旨在从原始音视频信号中提炼出与说话状态相关、具有辨识力的多模态时空特征表示。接下来,本节将聚焦于如何利用这些特征进行最终决策,即判别说话状态。

基于对现有文献的考察和对领域发展趋势的分析,并结合各类ASD方法在核心思想上的差异,本综述将现有方法划分为两大类:基于纯视觉信息的方法和音视结合的方法(图4)。基于纯视觉信息的方法仅利用视觉信号(如面部、唇部或身体的运动模式)来推断说话状态。这类方法的显著优势在于完全不受音频噪声、混响等声学干扰的影响,在无声或音频信号不可靠的场景下具有独特的应用价值。音视结合的方法则同时利用音频和视觉两种模态的信息及其内在关联进行判断,是当前研究的主流。依据其最终判别依据的不同,音视结合的方法进一步细分为3个子类。1)基于音视对应的匹配方法(2.4.2节):通过跨模态度量学习建模音视信息的相关性,用于估计音视同步性或建立音视身份之间的关联,进而判断说话人;2)基于音视融合的分类方法(2.4.3节):将说话人检测任务建模为音视特征融合后的二分类问题;3)混合方法(2.4.4节):结合以上二者的优势,以提升复杂场景下的鲁棒性与泛化能力。

表1中对各类方法的关键特性进行了总结对比,下面将对各类方法进行详细探讨。

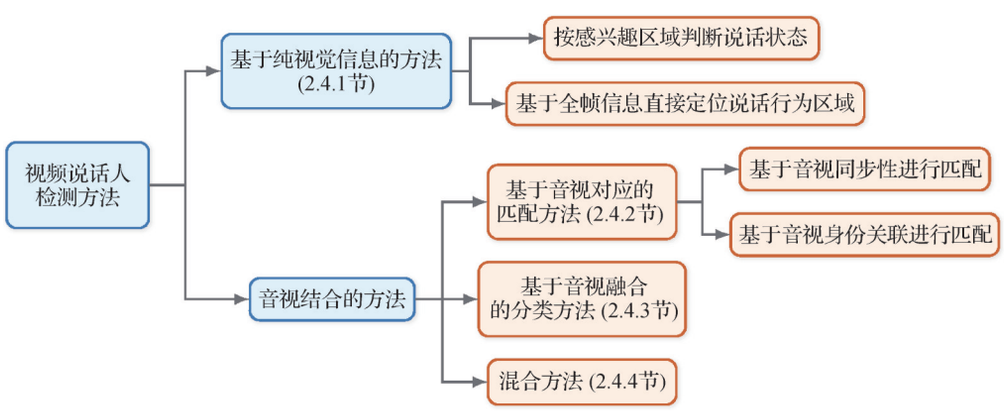

图4 视频说话人检测方法分类

Fig. 4 Taxonomy of ASD methods

表 1　不同视频说话人检测方法和关键特性对比

Table 1　Comparison of different active speaker detection methods and key characteristics

方法类别	核心思想	代表性方法	主要利用信息	优点	缺点	适用场景
基于纯视觉信息的方法	仅依据面部/身体运动推断说话状态	LPN、HiCA、S-VVAD、VTP等	面部几何/表观/光流特征；肢体、手势等运动线索	不受音频噪声/混响影响；适用于无声视频	无法过滤画外音和非言语行为；受视觉质量影响；数据偏见	无音频或强噪环境；V-VAD
基于音视对应的匹配方法（基于同步性）	度量唇动与语音同步性，判定音视信号是否对应	SyncNet、PM、LWTNet、DiVAS等	音频、面部特征等	可自监督训练，不依赖人工标注；显式同步判断	对音视频精确对齐要求高；受短时窗口局限鲁棒性有限	要求强同步；数据无标注；作为分类方法的辅助监督
基于音视对应的匹配方法（基于身份关联）	聚类与跨模态匹配，确定当前声音与人脸的身份一致性	Hoover等人（2018）、SL-ASD、GSCMIA等	声纹、面部/身体特征等	可区分说话人；可处理画外音/背影；可以容忍音视不同步	通常非端到端，对聚类精度要求高；流程复杂	音视说话人日志；需明确人物身份；视频质量不佳
基于音视融合的分类方法	提取音视特征并融合，通过二分类直接判别说话状态	TalkNet、UniCon、SPELL、EASEE等	音频和面部特征、其他候选人特征、其他模态特征	性能优异，可端到端优化；可利用长时/多人物上下文	依赖大量标注；受训练数据域影响大	通用说话人检测；复杂多人交互场景
混合方法	结合基于分类与基于匹配的方法，利用互补性	FaVoA、TS-TalkNet、SCAN等	结合说话行为判断与身份判断	复杂场景；信息互补	模型复杂；计算开销大；可能需注册信息	监控/第一视角等复杂场景；目标说话人检测

2.4.1　基于纯视觉信息的方法

该类方法仅依赖视觉信号（如面部、唇部或身体的运动模式）判断说话状态。相较于音视结合方法，其显著优势在于不受音频噪声、混响及多人语音重叠等声学环境的干扰，因而在无声视频或音频质量不佳的场景中具有独特的应用价值。根据对输入视觉信息处理策略的不同，现有方法主要分为两类：按感兴趣区域（region-of-interest，RoI）逐一判断说话状态的方法和基于全帧直接定位说话行为区域的方法。

1）按 RoI 逐一判断说话状态。该策略遵循自底向上的思想，首先利用人脸检测器定位并裁剪出人脸或唇部候选区域，随后提取该区域的视觉特征并进行帧间时序建模，最后通过分类器根据每个候选区域的特征逐一判定其说话状态。早期研究多利用支持向量机（Chung 和 Zisserman，2016a；Saenko 等，2005）、Adaboost（Liu 等，2011）或隐马尔可夫模型（Patrona 等，2016）等统计学习方法进行判别；也有部分工作探索无监督聚类来挖掘说话模式（Ahmad 等，2013；Tao 等，2015）。此类方法计算量较小，但对视角变化和语速等因素的鲁棒性较差（Joosten 等，2013；Navarathna 等，2011）。此外，为降低对帧级强标注的依赖，部分研究利用语音活动检测获取弱标签指导视觉分类器的训练（Chakravarty 和 Tuytelaars，2016；Chakravarty 等，2015）。随着深度学习的发展，现代方法开始以端到端的方式训练分类器（Aung 和 Ritthipravat，2015；Guy 等，2021）。为了更精准地捕获细微的说话运动并提升鲁棒性，研究者进一步引入了注意力机制。例如，Wang 和 Wang（2019）设计了关键点池化网络（landmark pooling network，LPN），利用唇部关键点引导网络聚焦于嘴部关键区域；Prajwal 等人（2022）则提出了前文介绍的 VTP 模块，利用自注意力机制对空间特征进行加权聚合，使模型在决策时能够自动关注与发音最相关的面部区域（如嘴唇），从而在包含复杂背景或遮挡的场景下做出更准确的判断。近期，LASER（lip landmark assisted speaker detection for robustness）（Nguyen 等，2026）进一步显式引入了唇部轨迹编码（lip track encoding），引导模型依赖精确的唇部运动轨迹而非不稳定的纹理细节进行决策，从而解决了传统方法难以区分微小嘴部运动（如叹气、微笑）与真实说话行为的难题。

2)基于全帧信息直接定位说话行为区域。不同于前一类依赖预先检测人脸的方法,基于全帧信息的视觉ASD方法直接将视频帧序列作为输入,通过学习全局特征判断场景中是否存在说话行为,端到端地预测说话人的位置和状态。这种方法避免了因人脸检测失败或RoI裁剪误差导致的错误累积。此类方法常采用弱监督学习策略,利用音频VAD标签作为近似监督。例如,Sharma等人(2019)提出了分层上下文感知(hierarchical context-aware, HiCA)架构,利用音频VAD标签训练3D-CNN和LSTM预测视频片段的语音活动,其类激活图(class activation map, CAM)显示模型能自发定位到说话的人脸或身体区域。Sharma等人(2023)进一步设计了一种基于多示例学习的框架,将视频片段视为包(bag),其中检测到的各个人脸视为实例(instance),利用电影字幕作为语音活动的代理标签,训练网络用全帧图像判别"包"中是否包含语音,从而自动挖掘出对语音活动贡献最大的关键实例(即真正的说话人)。S-VVAD(visual voice activity detection by motion segmentation)方法(Shahid等,2021)则借鉴运动分割思想,利用"动态图像"(dynamic images)(Bilen等,2016)表示运动信息,结合语音活动标签以弱监督方式训练网络后,通过类激活图生成伪掩码;再训练全卷积网络直接分割出说话区域,从而无需人脸检测。

但是,基于纯视觉信息的方法也面临诸多挑战,如对视觉遮挡、分辨率及非正面视角等视觉因素较为敏感,且难以排除只张嘴不发声(如默念、做嘴型)和非言语行为(如咳嗽、打哈欠等)的干扰。此外,Stefanov等人(2020)还发现数据中的头部姿态分布可能带来意外的空间偏见而影响此类模型泛化。

2.4.2 基于音视对应的匹配方法

基于音视对应的匹配方法通过在跨模态嵌入或统计度量空间中计算音频片段与候选视觉个体之间的相似度对音频流中的语音活动进行归属与判别,可细分为两条主线:1)基于音视同步性:强调短时(数百毫秒至数秒)唇部/头部运动与语音能量/频谱特征的时间对齐,通过音视特征的距离度量形成局部同步判据;2)基于音视身份关联:强调稳定身份特征(人脸外观、声纹)在较长时间尺度上的一致性,常以"音频簇(说话人聚类)—人脸簇(跟踪+聚类)"的匹配或共享嵌入空间最近邻检索实现"是谁在说话"的判断,下面分别进行详细介绍。

1)基于音视同步性进行匹配。该路线通过度量说话过程中音视频信号的相关性确定音频所对应的说话人。该路线的思想可追溯至Hershey和Movellan(1999)的工作。该工作将同步性定义为音频与视频信号间的互信息,在信号服从联合高斯分布的假设下,推导出音频能量与局部像素值之间的Pearson相关系数可用于估计互信息,从而可以通过寻找音视同步的像素位置检测说话人位置,为后续ASD研究奠定了基础。为解决原始特征维度较高的问题,Slaney和Covell(2000)提出FaceSync方法,用典型相关分析(canonical correlation analysis, CCA)对人脸序列和音频之间的相关性进行量化,通过最优线性变换将音视频数据投影至单一轴,再利用Pearson相关系数度量音视同步性。此类方法在独白场景下取得了很好的效果(Iyengar等,2003a,b;Nock等,2002)。Cutler和Davis(2000)还提出可借助时延神经网络(time delay neural network, TDNN)学习音视频特征在时间上的相关性模式,从而检测说话行为的时空位置。

针对前面统计方法通常依赖高斯分布假设的局限性,Fisher和Darrell(2004)提出了一种基于非参数假设检验的框架。该方法利用Parzen窗估计概率密度,在无需预设数据分布的情况下,学习使互信息最大化的线性投影,从而实现了更鲁棒的说话人检测。Argones Rúa等人(2009)则提出利用协惯量分析(coinertia analysis, CoIA)(Dolédec和Chessel,1994)代替CCA度量音视特征间的线性相关性,克服了CCA对样本量的敏感性,并结合耦合隐马尔可夫模型对音视流的动态状态依赖进行建模,基于贝叶斯融合进一步提升了同步性判别的准确率。

后续工作对如何更准确地判断同步性进行了更深入的探索。Cadavid等人(2009)提出通过主动形状模型获取人脸关键点轨迹,借助次谐波—谐波比提取音高(pitch),并运用高斯互信息计算音视频特征向量的同步性。朱铮宇等人(2014)提出了时空融合的一致性检测算法,利用CoIA同时分析唇部几何特征的空域和时域相关性,并引入"自然时延"约束对非生理范围内的相关性进行惩罚,从而实现了更鲁棒的音视一致性判别。Haider等人(2016)则探索了利用视觉韵律信息(如头部运动)在人机多方对话场景中进行ASD,实验结果表明头部运动在检测中起着重要作用,且与唇部运动融合能够进一步提升

性能。但是，以上工作仍依赖于手工设计的特征和传统的信号处理或统计方法，在复杂场景下的鲁棒性和准确性有限。

深度学习的引入为音视同步性估计带来了革命性的突破。Marcheret 等人（2015）率先尝试利用深度神经网络（deep neural network，DNN）解决这一问题。具体而言，该方法设计了一个包含“同步”和 12 个不同时间偏移量（以±30 ms 为步长）的多分类任务，并通过人为平移音频流构造训练样本。该方法证明了 DNN 在建模复杂音视依赖方面的潜力，在性能上显著优于传统的 CCA 和 CoIA 方法。随后，基于度量学习的思想逐渐成为主流，其核心在于学习一个音视模态间共享的嵌入空间（embedding space），在这个空间中，同步的音视频片段的特征表示（嵌入向量）彼此靠近，而不同步的片段则相互远离。SyncNet（Chung 和 Zisserman，2016b）是这一方向的里程碑式工作。该方法采用双流卷积神经网络分别处理音频和裁剪后的唇部视频序列，通过将同一视频片段的对应音频作为正样本，而将该视频片段与其他随机时间点的音频或从其他视频随机选取的音频作为负样本构建训练对，并通过最小化模态间的对比损失（contrastive loss）（Arandjelovic 和 Zisserman，2017；Arandjelović 和 Zisserman，2018；Hadsell 等，2006；Korbar 等，2018；Owens 和 Efros，2018）进行端到端训练。SyncNet 成功地学习到了区分同步与否的能力，通过计算视频中连续多帧（如 10 帧或 100 帧）内的音视特征平均距离即可将其直接应用于 ASD 任务。

此后，针对非正面人脸的音视同步性估计，Chung 和 Zisserman（2017）训练了多视角（multi-view）版本的 SyncNet，将输入区域从唇部拓展到包含全脸的更大区域，并以从正脸开始、逐步加入大角度侧脸的课程学习（curriculum learning）方式进行训练；朱铮宇等人（2023）则提出了一种基于正面唇重构的多视角音唇一致性判别方法，用基于自映射监督循环一致性生成对抗网络（SMS-CycleGAN）的唇重构方法对多视角唇部图像进行角度分类及正面重构后再进行同步性预测。为进一步提升模型在带噪语音上的预测能力，Kim（2021）提出了抗噪声的学习框架，在音频编码器最后一层引出两个卷积分支，分别提取语音和噪声嵌入特征，并提出了新的抗噪声损失函数，最小化噪声与视频的关联并强化语音与视频的同步性。近期，Li 等人（2024）通过广泛的经验性研究，总结了影响 SyncNet 收敛性的关键因素，例如批大小、模型架构、嵌入维度、输入帧数以及数据预处理步骤等，为同步模型的训练提供了宝贵的实践指导。

SyncNet 之后的研究围绕如何学习更具判别力的跨模态嵌入进行了大量探索，使得该路线成为当前 ASD 研究中最活跃和富有成效的方向之一。例如，Chung 等人（2019b，2020b）提出了基于多路匹配损失（multi-way matching loss）的 Perfect Match（PM）模型。PM 的损失函数可视为 *N*-pair 损失（Sohn，2016）或 InfoNCE 损失（van den Oord 等，2018）的一种形式，通过将同步性判断视为从多个候选音频片段中正确匹配唯一同步音频的多分类任务（图 5），迫使模型学习到更细粒度的跨模态对应关系。在原始的 PM 模型中，音视嵌入的相似度以欧氏距离的倒数计算。Chung 等人（2020c）借鉴度量学习中角边际（angular margin）损失函数的思想，进一步改进了相似度得分的计算方式，提出使用基于余弦相似度并结合可学习的尺度和偏置参数的形式，以增强模型的泛化性。

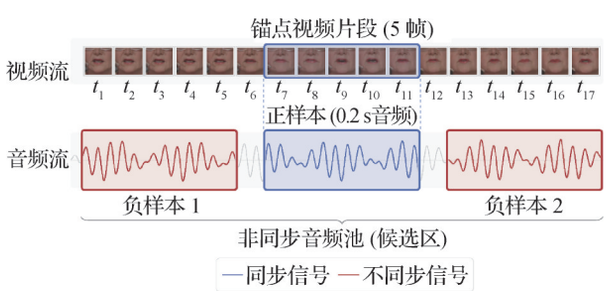

图 5　SyncNet 与 Perfect Match 训练中的负样本采样

Fig. 5　Negative sampling of SyncNet & Perfect Match during training

近期，Park 等人（2024）新设计了平衡二元交叉熵（balanced binary cross-entropy，BBCE）损失函数训练可解释 SyncNet。该方法避免了 InfoNCE 复杂的采样策略，有效平衡了各类正负样本在训练中的贡献，取得了优异的性能。此外，该模型还利用可学习的温度系数将原始输出缩放到合适的范围内，使其具有概率含义（故称可解释）。但是，以上的几种损失函数都不直接区分音视不同步的程度，也没有考虑训练数据本身就可能存在微小不同步的问题。为此，Ding 等人（2020）提出了动态三元组损失

(dynamic triplet loss)和多项式损失(multinomial loss):前者通过对不同偏移程度(如完全同步、有微小偏移、来自不同人)的样本对施加相对边际来学习音视频嵌入空间的细粒度结构;后者则在单次优化中综合比较多类同步与非同步样本,施加差异化惩罚以提升训练效率。两者通过显式区分微小偏移与真实同步,有效提升了说话人检测的性能。类似地,Jawade等人(2025)针对配音(dubbed)视频中存在的部分同步(partial sync)问题提出了一种新的排序监督多相似度(ranking supervised multi-similarity, RSMS)度量学习方法,先粗糙学习区分较大偏移和同步的音视频对,再利用多相似度损失(Wang等,2019)学习区分完美同步、有微小偏移和偏移较大的音视频对。

除上述依赖人脸检测与裁剪的方法外,部分研究直接将完整的视频帧作为输入,端到端地完成说话人的定位与检测(Afouras等,2020; Fernandez-Labrador等,2024; Truong等,2021)。该路线可视为对早期利用互信息寻找同步区域思想的现代深化,借助深度神经网络强大的特征提取能力,克服了人脸检测器计算代价高及在复杂场景下易失效的局限。

此类方法主要包含两种策略:一是显式定位策略,即模型直接计算音频与视频中特定局部区域(如特征图像素或网格块)的相关性确定说话人位置。例如,LWTNet(look who's talking network)(Afouras等,2020)通过计算音视时空特征的余弦相似度生成热图,并结合光流追踪显著峰值;Truong等人(2021)则将图像划分为固定网格,通过计算音频与各网格块间的注意力权重定位高相关性的说话区域。二是隐式定位策略,更侧重于利用全图的全局特征进行判别,而非显式输出定位图。以DiVAS(dynamic video and audio synchronization)模型(Fernandez-Labrador等,2024)为例,该方法利用视觉Transformer编码器中的全局词元(token)[CLS]聚合全图信息,其自注意力机制能够自动学习并聚焦于与音频同步的视觉区域(不局限于嘴部)。在决策时,模型直接基于全局音视特征评估同步置信度,避免了昂贵的逐区域计算,实现了高效推理。

2)基于音视身份关联进行匹配。与基于音视同步性的方法不同,此类方法侧重于利用身份特征的一致性确定说话人,将说话人检测转化为"哪个声音对应哪张人脸"的跨模态关联问题,与音视说话人日志任务紧密相关。其典型流程如下:

(1)单模态特征提取与身份聚类。该环节主要对音频和视频流分别进行特征提取与聚类处理。

在音频端,利用语音活动检测提取语音片段,并通过预训练模型如x-vector、d-vector或局部聚合描述向量(vector of locally aggregated descriptors, VLAD)(Hoover等,2018)等提取声纹特征(speaker embedding),将其聚类为若干个匿名的说话人簇(speaker clusters),每个簇代表一个独特的说话人身份(例如,属于说话人A的片段集合形成一个簇,属于说话人B的片段集合形成另一个簇等)。

在视频端,基于人脸跟踪结果,用预训练的FaceNet或ArcFace模型等人脸特征提取器获取人脸特征(face embedding)并聚类为不同的视觉身份簇。为应对人脸不可见(如背对镜头)的情况,Brown等人(2021)进一步提出了一种多模态高精度聚类算法MuHPC(multi-modal high-precision clustering),综合利用人脸、人体(衣着、体态)和声纹等多种信息辅助人物聚类。

(2)跨模态身份关联与匹配。该环节是基于音视身份关联进行匹配路线的核心,旨在将未知的音频簇与视觉簇进行无监督对齐,主要包含两种策略:

①基于共现统计信息关联。即直接计算音频簇和与人物视觉簇之间的共现(co-occurrence)时长。如果某个音频簇代表的声音与某个人物簇代表的视觉身份在视频中频繁同时出现,则它们很可能属于同一个人(El Khoury等,2014; Hoover等,2018; Vajaria等,2008)。

②基于跨人脸—语音模态匹配模型关联。人脸与语音作为重要的生物特征,存在源于性别、面部结构和发声方式等因素的内在关联性。跨模态人脸—语音匹配(face-voice matching; face-voice association)模型旨在学习这一关联性(Kim等,2018; Li等,2023; Nagrani等,2018; Wen等,2019,2021),通过将人脸特征和语音特征映射到同一嵌入空间,使同一身份的人脸和声音在该空间中距离更近。这类模型可以用于计算音频簇(由声纹表示)和视觉簇(由人脸特征表示)之间的相似度,从而可以直接计算音视簇间的相似度进行匹配。此类方法不依赖连续、清晰的唇部运动信息,因此在第一人称视角等伴随剧烈镜头运动、面部遮挡或模糊的挑战性场景中展现

出更强的鲁棒性。例如，SL-ASD（self-lifting for audiovisual active speaker detection）（Clarke 等，2025a）针对第一人称视频，利用 Transformer 聚合人脸序列中偶尔出现的清晰帧以构建鲁棒的视觉身份特征，并借助 Self-Lifting 关联模型（Chen 等，2022a）完成与语音的匹配，有效克服了画质退化导致的精细唇动特征提取难题。

2.4.3　基于音视融合的分类方法

该路线将 ASD 问题视为一个有监督的二分类任务：对每个候选人，在每个时间点根据融合后的音视特征判断其处于“说话”还是“未说话”状态（考虑到画外音的情况，严格意义上应称“是说话人”还是“非说话人”）。该路线最基本的思路是将经融合和时序建模的逐帧特征序列输入全连接层，以直接预测各时刻该人脸的说话概率或标签（Hu 等，2015；Ren 等，2016；Roth 等，2020），并采用如图 6 所示的主预测损失加单模态辅助预测损失的形式进行有监督学习。这类方法以 TalkNet（Tao 等，2021）为代表，通常结合前文所述的强大有效的深度特征提取器、音视特征融合模块以及时序建模单元。

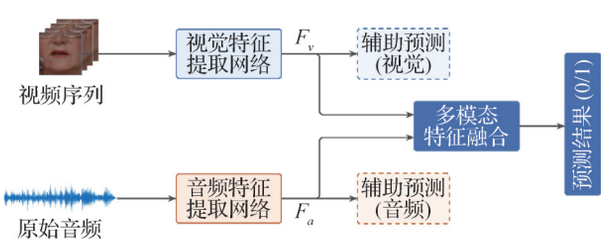

图 6　基于音视融合分类的基本 ASD 模型

Fig. 6　A basic ASD model based on audio-visual fusion and classification

对于基于音视融合分类的方法和前面基于音视同步的方法，Kim 等人（2021）通过构建干净数据集 Active Speakers in the Wild 消除配音干扰，系统地进行了对比。结果证实，基于音视融合的有监督分类方法（如双向 GRU + 全连接层）在性能上显著优于单纯基于音视同步性度量的自监督方法，进一步巩固了前者在实际应用中的主流地位。但是，这并不意味着基于同步的自监督信号在分类模型的有监督框架下无用武之地。相反，近期的研究表明，在有监督分类的基础上引入额外的辅助损失或采用多任务联合训练，能够为模型提供更丰富的跨模态约束，是提升真实复杂场景下 ASD 性能的有效途径。Zhang 等人（2019）率先在二分类损失的基础上，添加了音频和视频特征之间的对比损失进行多任务学习，通过显式约束两模态特征的语义对齐，提取更具判别力的多模态特征，提升了最终的检测准确性。SyncTalkNet 模型（Wuerkaixi 等，2022）延续了这一思路，同样使用多任务学习的形式训练，添加了类似 Perfect Match 的多路匹配损失，并向音视交叉注意力模块中添加时间编码以强化同步性感知。近期提出的 TalkNCE（Jung 等，2024）进一步优化了对比学习的采样策略，仅在人物实际说话的片段计算对比损失，发现能进一步提升有监督 ASD 模型的特征判别力和检测性能。

除了特征层面的自监督对比损失，引入其他维度的辅助任务也能显著提升模型的鲁棒性。针对真实场景中常见的环境噪声和多说话人语音重叠问题，Xiong 等人（2023）提出了一种统一的多任务框架 ADENet（multi-modal correlation learning for active speaker detection and speech enhancement），将 ASD 与音视结合的语音增强联合训练，通过跨模态循环融合使得两者相互促进：精准的检测结果能更好地指导语音增强，而干净的增强特征反过来又促进了更准确的说话人定位。MuSED（multi-modal speech extraction-to-detection）（Tao 等，2025）则将音视结合的目标人语音提取作为预训练的辅助任务，使模型学习到多模态降噪和类似鸡尾酒会效应的选择性听觉能力，随后再针对 ASD 任务进行微调，在极具挑战性的多说话人及低信噪比环境下展现出卓越的鲁棒性。

2.4.4　混合方法

前面介绍的基于音视对应的匹配方法与基于音视融合的分类方法在核心思想与侧重点上存在显著差异。匹配方法侧重于利用长时的身份一致性或局部的音视同步性进行跨模态关联，而分类方法则侧重于逐帧提取音视特征并通过有监督学习进行细粒度的说话状态判别。在复杂的真实场景中，单一的技术路线往往难以满足高准确率与强鲁棒性的需求。因此，当前提升 ASD 模型性能的一个重要研究方向，是尝试将这两大路线进行有机融合，形成兼具“匹配”的全局视野与“分类”的局部精度的混合方法。该路线主要包括以下 3 种策略：

1）级联式方法（两阶段检测）。该策略将分类方法与匹配方法通过前后串联，使二者相互辅助。纯

粹基于共现统计的身份匹配方法在某些场景下(如所有说话人始终同屏出现)容易失效,而基于分类和同步性的ASD模型能提供互补的行为先验信息。例如,Sharma和Narayanan(2022)提出先通过视觉ASD方法提取高置信度的说话人脸片段,再通过聚类为主要说话角色自动生成音视档案(audio-visual profiles),以识别视频中的其他说话片段,本质上是通过分类筛选为身份匹配提供纯净样本。Huh和Zisserman(2024)则利用音视同步性挑选明确的单人片段构建音频范例库(audio exemplars),进而实现后续的角色特征匹配。Sharma和Narayanan(2023)提出一种利用音视活动信息指导跨模态身份关联的无监督框架GSCMIA(AV-activity guided cross-modal identity association),将基于分类的ASD模型(TalkNet)输出的说话概率作为指导信号融入人脸—语音的跨模态匹配关联中。类似地,Huang等人(2025)在第一人称视角的说话人日志任务中利用ASD模型预测场景说话人数量作为先验指导后续基于声纹的聚类与匹配过程,有效解决了纯音频聚类在噪声场景下的人数估计问题。

2)统一式方法(端到端融合)。该策略在单一的分类网络内部显式地引入匹配机制,将跨模态的身份一致性验证作为辅助分类的强判据。例如,FaVoA(face-voice association ambiguous speaker detector)(Carneiro等,2021)借助门控双模态单元(gated bimodal unit)将预训练的人脸—语音匹配网络直接整合到基于分类的ASD主干中,利用身份匹配度抑制非目标说话人的干扰。TS-TalkNet(target speaker TalkNet)(Jiang等,2023)向传统的音视融合分类网络中引入了“目标说话人”概念,不仅融合音视特征进行分类,还提取目标人脸的参考声纹与当前音频进行匹配比对,从而能有效区分目标语音与混叠的无关人语音,显著增强了ASD准确性。受此启发,Clarke等人(2025b)提出的说话人对比辅助网络(speaker comparison auxiliary network,SCAN)在主干分类网络中利用交叉注意力对参考语音和当前音频进行帧级特征对比,从而为分类模型提供了更细粒度的身份一致性判据。

3)集成式方法(决策层融合)。这是一种更为直接的混合策略,旨在决策层面综合匹配与分类两条路线的输出结果。Clarke等人(2025c)指出基于分类和基于音视身份关联的模型在第一人称视角视频等复杂场景下具有天然互补性,通过对这两类独立模型的输出概率(即分类模型的说话概率与匹配模型的身份置信度)进行简单的加权平均,便能有效融合二者优势,提升整体检测精度。

总体而言,说话状态判别是对多模态特征进行综合推理的核心环节。纯视觉方法在强噪或无声等受限场景中具有独特价值,但在通用场景下性能逊于音视结合方法。主流的音视结合路线中,匹配方法擅长利用自监督信号挖掘模态间的内在关联,而分类方法则通过数据驱动的强监督学习追求更高的瞬时判别精度。混合方法通过整合匹配路线的身份一致性验证与分类路线的行为同步性判断,展现出解决多声源干扰、遮挡及配音等复杂挑战的巨大潜力,代表了该领域向真实复杂场景迈进的发展方向。

3 说话人检测方法的评价

3.1 常用数据集

视频说话人检测技术的研究和发展离不开高质量、多样化数据集的支持。近年来,学术界和工业界发布了众多针对不同模场景和模态组合的数据集,为算法模型的训练与评估提供了重要基准。本节对视频说话人检测领域的常用代表性数据集进行简要介绍,主要信息如时长、发布年份和场景等总结于表2,部分示例图像见图7。

3.1.1 受控场景数据

1)AMI(augmented multi-party interaction)数据集(Carletta等,2005)包含100 h的多模态英语会议录音,配备了近讲麦克风(领夹式和头戴式)、远场麦克风阵列、个人特写与全景摄像头以及电子白板等多种同步记录设备,提供了包括文本、发言人身份等在内的精细多层次标注。

2)RAVEL(robots with auditory and visual abilities)(Alameda-Pineda等,2013)是一个较早的面向人机交互场景的公开音视语料库,通过配备双目摄像机和麦克风的机器人头部进行数据采集。该数据集的“交互”(interaction)子集设计了多人聊天、问路以及模拟鸡尾酒会等多种交互场景,但整体规模较小,仅由7名演员在5种特定设定下完成了24条视频录制(总时长约0.4 h)。

3)Columbia(Chakravarty等,2016)数据集来自美国哥伦比亚大学(Columbia University)一场有

表 2　视频说话人检测相关数据集（按年份）

Table 2　Datasets for active speaker detection (by year)

数据集	时长/h	语言	人工标注	场景	特点描述
AMI（Carletta 等，2005）	100	英语	√	多人会议	多模态、多通道同步采集，精细标注
RAVEL (Interaction)（Alameda-Pineda 等，2013）	0.4	英语	√	室内人机交互	双目视觉、面向人机交互，规模小
Columbia（Chakravarty 等，2016）	1.7	英语	√	5 人小组讨论	桌面麦克风，常用简单测试基准
AVDIAR（Gebru 等，2018）	0.7	英语	√	实验室多人交互	自然语音交替与重叠、非固定姿态
AVSpeech（Ephrat 等，2018）	~4 700	多语	×	网络视频	低噪声、准正面为主的说话样本
MELD（Poria 等，2019）	9.3	英语	√	电视剧	附带情感与情绪标签
AVA-ActiveSpeaker（Roth 等，2020）	38.5	多语	√	影视剧片段	大规模精细时序标注、主流评估基准
WildVVAD（Guy 等，2021）	6.1	多语	×	网络视频	自动化标注，多样性高；仅视频
RealVAD（Beyan 等，2021）	0.8	英语	√	小组讨论	多人场景，远场，面部较小
EasyCom（Donley 等，2021）	5.3	英语	√	智能眼镜	多通道音频、宽视场视频，嘈杂环境
ASW（Kim 等，2021）	30.9	英语	√	网络视频	无配音干扰，真实场景
Talkies（Alcázar 等，2021b）	4.2	英语	√	互联网视频/新闻	含频繁遮挡与剧烈运动、画外音
TalkSet（Tao 等，2021）	151.7	英语	×	演讲、访谈	自动构建，正负样本均衡
VPCD（Brown 等，2021）	23.9	英语	√	电影与电视剧	含人体位置和背影标注
AVA-AVD（Xu 等，2022）	29.3	多语	√	影视剧片段	含画外音标注
MSDWild（Liu 等，2022）	80.1	多语	√	网络视频	自然对话，存在频繁交替与语音重叠
Ego4D (AV)（Grauman 等，2022）	47.7	英语	√	智能眼镜	第一人称视角，存在运动模糊
Tragic Talkers（Berghi 等，2022）	3.8	英语	√	戏剧	光场相机与多通道音频，精细标注
MISP2021&2022（Chen 等，2022b）	125.2	汉语	√	家庭电视房间	多设备多通道同步采集，有噪声干扰
EgoCom（Northcutt 等，2023）	38.5	英语	√	智能眼镜	自然社交交互，精细标注
VVAD-LRS3（Lubitz 等，2023）	18.5	英语	×	TED 演讲	自动化生成，存在标签噪声
VoxMM（Kwak 等，2024）	109.3	英语	√	多领域对话	附多维丰富元数据，适合模型评估
WASD（Roxo 等，2025a）	30	英语	√	监控及非受控场景	非协作无约束场景，有人体标注
OpenHumanVid（Li 等，2025a）	~16 700	多语	×	影视剧	超大规模、标注体系全面
UniTalk（Nguyen 等，2025）	44.5	多语	√	网络视频	涵盖长尾稀缺语种与困难多人场景
Seamless Interaction（Agrawal 等，2025）	4 000+	英语	√	面对面双人交互	高质量，含动作、表情、手势标注
SpeakerVid-5M（Zhang 等，2026）	8 743	多语	×	网络视频	超大规模，提供骨架序列

注：“√”表示提供人工标注，“×”表示无标注或自动标注。

5 位发言者、时长 87 min 的小组讨论，音频由桌面上的麦克风录制。画面中最多同时存在 2～3 位发言者。该数据集通常仅用于测试，其中一名说话人被留做验证集（Chung 和 Zisserman，2016b）。该数据集



图7　本综述介绍的部分ASD数据集示例

Fig. 7　Selected examples of ASD datasets covered in this survey

的原始标注中有少量错误，Tao等人(2025)分享了修正后的标签供研究者参考。

4)AVDIAR(Gebru等，2018)是为支持音视说话人日志任务采集的实验室数据集，特别关注挑战性的多人交互场景，但规模较为有限(约0.7 h)。录制在3个不同的房间进行以涵盖不同声学环境，参与者(共12名)在对话时有自然交替和打断，可能会四处走动和转头，而非始终面向相机。

5)AVSpeech(Ephrat等，2018)数据集包含数千小时准正面、低噪声说话片段，是从YouTube的网络视频(如采访、演讲、授课等)中利用Hoover等人(2018)所提方法自动提取的。

6)RealVAD(Beyan等，2021)数据集与Columbia数据集类似，采集自真实的、非扮演式的9人小组讨论场景，但是画面中人脸更小，且麦克风距离发言人较远、存在一定的混响。

7)MISP2021(Chen等，2022b)数据集伴随第1届多模态信息处理与理解(Multimodal Information Based Speech Processing，MISP)挑战赛发布，专注于家庭远场电视房间场景，引入了真实的电视噪声干扰，增加了任务的挑战性。场景为2～6人的中文对话，参与者活动范围相对固定。作为第2届MISP挑战赛的数据集，MISP2022(Wang等，2023)数据集提供了用于音视说话人日志任务的标注。

8)Tragic Talkers数据集(Berghi等，2022)借助光场摄像头、麦克风阵列、领夹麦克风和一阶环境声麦克风采集了两名戏剧专业学生表演莎士比亚著名悲剧《罗密欧与朱丽叶》(Romeo and Juliet)的片段，包含人物的3D嘴部坐标和文本等精细标注。

9)Seamless Interaction(Agrawal等，2025)是一个包含4 000余名参与者、4 000余小时面对面交互的大规模数据集。所有视频均为专业录制的高质量影像，且参与者始终正对镜头。此外，该数据集还提供了涵盖动作、语音、表情和手势的丰富标注。

3.1.2　非受控场景数据

1)MELD(multimodal emotionlines dataset)(Poria等，2019)数据集。最初是为多模态情感识别任务设计的，包含来自电视剧《老友记》(Friends)的多人对话片段，提供了音频、视频和文本(剧本)模态以及情感、情绪标签。Carneiro等人(2023)对其进行重对齐和修正得到了MELD-FAIR(MELD with fixed audiovisual information via realignment)数据集，提升了音视

频模态的时间同步性。

2）AVA-ActiveSpeaker（Roth 等，2020）是 ASD 领域的里程碑式数据集，包含约 38.5 h、源自 YouTube 上影视剧片段的视频。其核心贡献在于提供了 365 万帧的精细时序标注，包括人脸序列的包围框坐标、人物是否说话，以及是否可听到声音，极大地促进了 ASD 算法的开发和标准化评估。该数据集的测试集在 2018—2022 年间用于 CVPR ActivityNet 挑战赛，因此未发布标注，性能评价通常在验证集上进行。该数据集是基于 AVA（atomic visual actions）（Gu 等，2018）和 AVA-Speech（Chaudhuri 等，2018）数据集继续标注的；AVA 数据集中提供了人物动作的时空包围框（spatio-temporal bounding box）标注，将几个数据集的标注进行关联可获得人物的身体包围框（Roxo 等，2024b）。基于该数据集，后续研究衍生出了多个扩展数据集，如 AVA-AVD（Xu 等，2022）在 AVA-ActiveSpeaker 数据集部分视频基础上扩展形成音视说话人日志数据集，包含了对画外音的标注；APES（audiovisual person search）（Alcázar 等，2021a）数据集则对 AVA-ActiveSpeaker 训练和验证集中相同身份的人脸序列进行了关联标注。

3）TalkSet 数据集（Tao 等，2021）是由 LRS3-TED（Afouras 等，2018）和 VoxCeleb2（Chung 等，2018）数据集自动构建的。该数据集包含 15 万个 1～6 s 的视频片段，其中 9 万个在说话样本来自 VoxCeleb2，6 万个未在说话样本来自 LRS3-TED（经 Kaldi 语音端点检测得到），说话和不说话样本总时长大致均衡。

4）VPCD（video person clustering dataset）（Brown 等，2021）数据集。专注于视频中的多模态人物聚类任务，目标是利用人脸、身体和声音信息对视频中的不同人物进行识别和分组。数据来源于电视剧如《吸血鬼猎人巴菲》（Buffy the Vampire Slayer）、《神探夏洛克》（Sherlock）等、电影《隐藏人物》（Hidden Figures）等，覆盖了广泛的人种。

5）WildVVAD（Guy 等，2021）数据集是结合音频 VAD 技术从 YouTube-8M 数据集中的网络视频（主要是访谈和电视节目）自动构建的大规模视觉语音活动检测数据集，包含 13 000 个短视频片段（约 2 s），覆盖了不同族裔、年龄、性别、头部姿态、面部特征、视频分辨率和质量、光照条件及背景等。

6）ASW（active speakers in the wild）（Kim 等，2021）数据集基于 VoxConverse 数据集（Chung 等，2020a）构建，旨在提供更贴近真实场景的 ASD 基准，特别强调避免使用音画不同步的配音视频。数据集包含 212 个视频（总标注时长 30.9 h），涵盖新闻、辩论、访谈以及脱口秀等多种真实场景，提供了 0.2 s 间隔的密集说话人标签和人脸包围框坐标。

7）Talkies（Alcázar 等，2021b）是从 YouTube 自动构建并经人工标注的大规模、无约束 ASD 基准数据集。该数据集包含 10 000 个视频片段和约 80 万个手动标注的人脸包围框，旨在提供比电影数据集更具挑战性的自然场景。其特点在于场景的高度多样性和复杂性，特别强调覆盖多说话人、频繁遮挡、剧烈头部运动以及画外音等 ASD 任务的难点。

8）MSDWild（multi-modal speaker diarization in the wild）（Liu 等，2022）是一个自然场景多模态说话人日志基准数据集，数据主要来源于公共视频平台上的 VLOG（视频博客），覆盖多种语言。MSDWild 中以自发的日常对话为主，因此包含了频繁的交替发言、打断以及大量的语音重叠。该数据集提供了经过人工标注和两轮检查的精确说话人日志标签（包括画外音）以及人脸包围框坐标。

9）VVAD-LRS3（Lubitz 等，2023）是一个大规模的视觉语音活动检测（V-VAD）数据集，基于 LRS3-TED 数据集通过自动标注流程生成，提供了超过 4.4 万个样本（约 18.5h），但存在标签噪声。

10）VoxMM（Kwak 等，2024）是一个包含丰富元数据的多模态、多领域对话数据集。VoxMM 提供了详细的说话人标签、文本、说话人性别、视频领域、质量和年代等信息，适合评估 ASD 模型各维度的泛化能力。

11）WASD（wilder active speaker detection）（Roxo 等，2025a）数据集旨在解决现有 ASD 模型在更复杂、非协作无约束场景下的性能瓶颈问题。该数据集收集了更具挑战性的 YouTube 视频（总时长 30 h），并将其划分为 5 个难度递增的类别，分别针对性地引入了音频质量下降（如语音延迟、重叠、背景人声）、面部遮挡以及模拟监控环境（不保证人脸可见性、音频质量或主体配合度）等挑战。该数据集在确保语言、种族和性别等人口统计学特征均衡的同时，首次为 ASD 任务引入了身体姿态标注，鼓励模型利用除人脸和声音之外更广泛的信息（如肢体语言）判断说话人。

12）OpenHumanVid（Li 等，2025a）是一个大规模、高质量以及以人类中心的视频数据集，包含约1 320万个样本，总时长约1.7万h，数据来源于公开可用的视频。OpenHumanVid 的核心特色在于其精细且全面的标注体系：不仅提供了准确描述人类外观和动作状态的文本描述，还包含了详细的骨架序列和语音音频等辅助模态信息。

13）UniTalk（Nguyen 等，2025）数据集旨在推动说话人检测在真实世界复杂场景下的研究进展。该数据集包含超过44.5 h的视频，涵盖48 693个说话人，并提供了帧级别的说话人标注。数据集涵盖广泛的视频类型，以反映真实世界的多样化条件。与以往数据集如 AVA-ActiveSpeaker 相比，UniTalk 特别强调对挑战性场景的覆盖，包括多种代表性不足的语言、嘈杂的背景环境以及人数较多的场景。

14）SpeakerVid-5M（Zhang 等，2026）是一个时间覆盖2006—2025年的大规模高质量音视双人交互与单人说话数据集：约77万双人片段（1 800 h）与520万单人片段（8 700 h），93%的视频分辨率≥1 080 p。每段包含结构化文本描述、骨架序列、ASR 文本、模糊度评分、身体姿态与视角标签。

15）值得注意的是，近年来第一人称视角的数据集也逐渐兴起。EgoCom（egocentric communications）（Northcutt 等，2023）是首个多模态、多人和自然对话的第一人称视角对话数据集，提供了38.5 h的同步立体声音频、视频流，以及超过24万条带时间戳的词级转录和说话人标签。它为研究第一人称视角下的复杂社交互动、对话动态分析等提供了独特的资源。

16）EasyCom（easy communications）（Donley 等，2021）数据集专为嘈杂环境（模拟“鸡尾酒会效应”）下的对话场景设计，特别关注增强现实（augmented reality，AR）应用。数据通过佩戴AR眼镜采集，包含超过5 h的多通道音频、宽视场视频和文本、人脸位置等丰富的标注信息。

17）Ego4D（Grauman 等，2022）是一个超大规模的第一人称视角视频数据集，用可穿戴设备记录了全球多地参与者日常生活的各种活动。其中，可用于训练和评估音视对话分析及ASD任务的AV子集包含约47.7 h的数据。

总体而言，经过近些年的发展，说话人检测数据集在规模上实现了从小时级到千小时级的跨越，场景也从受控的实验室环境向开放、复杂的真实环境，既涵盖了影视、网络视频及第一人称视角等多样化来源，也引入了包括极端光照、强背景噪声、剧烈运动及非配合行为等在内的现实挑战。同时，数据标注也正朝着细粒度、多维度的方向演进，从单纯的“说话/非说话”二元标签扩展至包含三维空间定位、肢体动作描述甚至语义交互信息的富标注体系，不仅为从多模态、多视角深入理解人类交互行为提供了可能，也推动了ASD算法向着更具鲁棒性和泛化能力的方向不断迭代。

3.2 常见评价指标

对ASD方法的评价主要是在检测精度和效率两方面进行。精度方面，常用的指标包括二分类任务中常用的准确率、F1值、接收者操作特征曲线下面积（area under the receiver operating characteristic curve，AUROC）、等错误率（equal error rate，EER）等。目前评估ASD性能最广泛采用的核心指标是平均精度均值（mean average precision，mAP），即精度—召回率（precision-recall）曲线下的面积；ActivityNet和Ego4D竞赛举办方均提供了mAP指标的标准计算工具。值得注意的是，在大部分ASD评测基准中，mAP通常都是在给定真实人脸边界框的前提下计算的，即仅评估说话状态的分类准确性；而Ego4D挑战赛则采用了mAP@0.5，此指标要求检测结果的人脸边界框与真实框的交并比（intersection over union，IoU）必须大于0.5，评估的是人脸检测与说话状态分类的综合性能。效率方面，常用的指标包括每秒处理帧数（frames per second，FPS）和每秒十亿次浮点运算次数（giga floating point operations per second，GFLOPs）两个指标，以衡量模型在实际应用中的实时性和计算资源需求。

4 说话人检测技术的趋势与展望

当前，ASD领域的研究呈现出结构设计创新、数据利用效率提升与应用边界拓展并进的特点。整体而言，在方法层面，研究的核心在于构建更强大的深度模型以捕捉复杂的音视频时空依赖。模型架构从CNN/RNN向Transformer/GNN的演进显著增强了对长时交互的建模能力；同时，学习范式正从单一任务监督向大规模预训练和多任务协同（如联合ASR、语音分离）转变，旨在通过通用先验和互补知识提升

特征的鲁棒性。在数据层面，鉴于大规模精确标注的高昂成本，降低对人工标签依赖的数据高效范式成为关键。基于音视自然同步性的自监督学习和弱监督学习得到广泛采用，为利用无标注数据提升模型泛化能力奠定了基础。在应用层面，研究重心正从通用基准（如 AVA-ActiveSpeaker）转向更具挑战性的真实场景（如第一人称视角视频），并积极探索利用肢体语言、视线方向等非面部线索辅助检测。同时，面向实际部署，模型的效率（轻量化、流式处理）、鲁棒性和可解释性日益受到关注，成为连接算法研究与实用系统的关键桥梁。本节将围绕上述趋势，总结当下 ASD 仍面临的开放问题，并展望 ASD 领域未来可能的发展方向。

4.1　开放问题

如表 3 所示，当前的最优模型已在 AVA-ActiveSpeaker 等主流基准上取得了非常优异的性能（mAP 超过 95%），这在很大程度上造成了 ASD 任务“接近解决”的假象。然而，近期研究明确指出，这些模型在迁移到更具挑战性的真实世界数据集时，性能会显著下降（Nguyen 等，2025；Roxo 等，2025a）。这一现象暴露了当前研究的核心困境：模型跨数据集和跨场景的泛化能力不足。具体而言，当前 ASD 研究仍面临以下几个关键的开放问题：

1）现有基准数据集的局限性带来的泛化性挑战。（1）语种有限：AVA 等主流数据集大多集中于英语等高资源、印欧语系的语言，导致模型在面对非英语和低资源语种时表现不佳；（2）场景受限：现有数据大多源自影视剧等专业制作内容，缺乏对侧脸、背对镜头（非配合式拍摄视角）、面部遮挡及音频噪声与语音重叠等复杂真实场景的标注，限制了模型的泛化能力。

2）缺乏对交互语义的深层理解。现有模型多局限于“说话/不说话”的二元分类，缺乏对“话轮转换”（turn-taking）等对话结构及深层交互语义的理解。这种理解的缺失，导致模型在面对打断、抢话或非语言交流（如点头示意、大笑）等复杂社交动态时，难以准确判断说话人的意图与状态。

3）高昂的计算成本与低效的系统架构。现有高性能模型多基于 Transformer 或 GNN 等复杂架构，计算代价高昂，难以满足智能会议、人机交互等场景对低延迟流式处理的要求。同时，在多人密集场景下，“先检测人脸、后判断说话”的两阶段流程本身即构成了显著的效率瓶颈，限制了系统在端侧设备上的实时部署能力。

表 3　AVA-ActiveSpeaker 数据集上代表性方法性能汇总
Table 3　Performance of representative ASD methods on AVA-ActiveSpeaker

方法	验证集 mAP/%	测试集 mAP/%	预训练
AV-GRU（Roth 等，2020）	82.2	82.1	×
ASC（Alcázar 等，2020）	87.1	86.7	√
TalkNet（Tao 等，2021）	92.3	90.8	×
UniCon（Zhang 等，2021b）	92.0	90.7	×
ASDNet（Köpüklü 等，2021）	93.5	91.9	√
EASEE-50（Alcázar 等，2022）	94.1	—	√
SPELL+（Min 等，2022）	94.9	—	×
UniCon+（Zhang 等，2022）	94.7	**94.5**	×
ADENet（Xiong 等，2023）	93.2	—	×
Light-ASD（Liao 等，2023）	94.1	—	×
TalkNCE（Jung 等，2024）	95.5	—	×
MuSED（Tao 等，2025）	**95.6**	—	√
LoCoNet（Wang 等，2024b）	95.2	—	×
LR-ASD（Liao 等，2025）	94.5	—	×

注：加粗字体表示当前已知最高成绩；—表示未报告；√表示使用预训练，×表示无预训练。

4.2　未来展望

针对上述开放问题，未来的 ASD 研究正从追求单一基准的性能指标，转向对模型的适应能力、场景理解深度的系统性提升，同时高度关注模型的计算效率。对于 ASD 的未来发展，主要可以从以下几个方面进行展望：

1）扩展。构建兼顾“质”与“量”的真实场景数据集。针对数据的局限性，未来研究需致力于构建更具包容性、覆盖更多样化场景与视角的真实场景数据集。这不仅意味着在“量”上覆盖更多人种、更多语系和更多环境而且更要在“质”上实现突破，提供大侧脸、背对镜头等非配合式拍摄视角的高难度样本，以及如对话结构、对话内容等（Chang 等，2026）更丰富的语义标注，以填补现有基准的空白。

2）内求。充分挖掘与利用数据集已有信息。面对标注成本高昂及真实场景的复杂性，充分挖掘现有数据的潜力至关重要。这包括两个维度：一是减

少对数据标注的依赖，从而扩大可利用的数据范围，如借助弱监督与自监督学习、域适应及测试时适应（test-time adaptation，TTA）技术，探索在无标注数据上提升模型对分布偏移的适应能力；二是充分利用数据中的多类型线索。一方面，针对音视线索明确的数据，可通过数据增广等手段模拟真实的噪声与遮挡，以提升特征鲁棒性；另一方面，可挖掘视线、细微表情、肢体语言等非语言线索，辅助判断话轮转换及社交意图。这不仅能提升模型的交互理解能力（即社会智能），也能确保其在信息缺失条件下的判别稳定性。

3）外援。利用预训练多模态大模型的知识与能力。大模型的兴起提供了强大的外部知识与推理能力，可能推动ASD新范式的产生：一是采用提示驱动（prompt-driven）的方法（如利用目标声纹或文本提示），借助图像与视频分割模型如SAM（segment anything model）系列（Kirillov等，2023；Ravi等，2025；Shi等，2025）直接定位说话人（Wang等，2025）；二是结合大语言模型（large language model，LLM）和视觉语言模型（vision-language model，VLM）的推理能力，更准确地理解“谁在对谁说话”；三是进一步构建具备记忆与推理能力的智能体（agent），在长视频流中动态跟踪并理解多人交互的复杂动态（Long等，2026）。

4）提效。继续设计轻量化与高效架构。ASD在实际场景的落地需求对计算效率提出了严苛挑战。未来的研究可能将从两方面寻求突破：一是在时序建模上，引入Mamba（Dao和Gu，2024；Gu和Dao，2023）、RWKV（receptance weighted key value）（Peng等，2023）等线性复杂度模型替代高计算量的Transformer；二是在系统架构上，探索将ASD与人脸检测、语音识别等上下游任务进行端到端联合设计，以实现系统级效率优化。

5 结 语

视频说话人检测（ASD）作为视听多模态理解的基础任务，是实现自然人机交互和智能内容分析的关键一环。本综述对ASD的发展脉络进行了系统性综述，围绕“基于纯视觉信息”和“音视结合”（包括基于音视对应的匹配方法、基于音视融合的分类方法及混合式的方法）的两大类路径，深入剖析了主流方法、关键技术环节、基准数据集及评价体系。整体而言，尽管深度学习极大推动了ASD技术在主流基准上的性能，但模型在真实、复杂场景下的鲁棒性、泛化能力与系统效率仍是制约其应用落地的核心瓶颈。本综述期望能为相关研究者提供有价值的参考，以推动ASD技术向更智能、更鲁棒的视听感知系统迈进。

参考文献（References）

Afouras T, Chung J S and Zisserman A. 2018. LRS3-TED: a large-scale dataset for visual speech recognition [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/1809.00496.pdf

Afouras T, Owens A, Chung J S and Zisserman A. 2020. Self-supervised learning of audio-visual objects from video//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 208-224 [DOI: 10.1007/978-3-030-58523-5_13]

Agrawal V, Akinyemi A, Alvero K, Behrooz M, Buffalini J, Carlucci F M, et al. 2025. Seamless Interaction: dyadic audiovisual motion modeling and large-scale dataset [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2506.22554.pdf

Ahmad R, Raza S P and Malik H. 2013. Visual speech detection using an unsupervised learning framework//Proceedings of the 12th International Conference on Machine Learning and Applications. Miami, USA: IEEE: 525-528 [DOI: 10.1109/ICMLA.2013.171]

Alameda-Pineda X and Horaud R. 2015. Vision-guided robot hearing. The International Journal of Robotics Research, 34(4/5): 437-456 [DOI: 10.1177/0278364914548050]

Alameda-Pineda X, Sanchez-Riera J, Wienke J, Franc V, Čech J, Kulkarni K, et al. 2013. RAVEL: an annotated corpus for training robots with audiovisual abilities. Journal on Multimodal User Interfaces, 7(1/2): 79-91 [DOI: 10.1007/S12193-012-0111-Y]

Alcázar J L, Caba F, Mai L, Perazzi F, Lee J Y, Arbeláez P, et al. 2020. Active speakers in context//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12462-12471 [DOI: 10.1109/CVPR42600.2020.01248]

Alcázar J L, Cordes M, Zhao C and Ghanem B. 2022. End-to-end active speaker detection//Proceedigs of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 126-143 [DOI: 10.1007/978-3-031-19836-6_8]

Alcázar J L, Heilbron F C, Mai L, Perazzi F, Lee J Y, Arbeláez P, et al. 2021a. APES: audiovisual person search in untrimmed video//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE: 1720-1729 [DOI: 10.1109/CVPRW53098.2021.00188]

Alcázar J L, Heilbron F C, Thabet A K and Ghanem B. 2021b. MAAS: multi-modal assignation for active speaker detection//Proceedings

of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 265-274 [DOI: 10.1109/ICCV48922.2021.00033]

Appiani A and Beyan C. 2025. VAD-CLVA: integrating CLIP with LLaVA for voice activity detection. Information, 16(3): #233 [DOI: 10.3390/INFO16030233]

Arandjelovic R and Zisserman A. 2017. Look, listen and learn//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 609-617 [DOI: 10.1109/ICCV.2017.73]

Arandjelović R and Zisserman A. 2018. Objects that sound//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 451-466 [DOI: 10.1007/978-3-030-01246-5_27]

Argones Rúa E, Bredin H, García Mateo C, Chollet G and González Jiménez D. 2009. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden Markov models. Pattern Analysis and Applications, 12(3): 271-284 [DOI: 10.1007/S10044-008-0121-2]

Ariav I and Cohen I. 2019. An end-to-end multimodal voice activity detection using WaveNet encoder and residual networks. IEEE Journal of Selected Topics in Signal Processing, 13(2): 265-274 [DOI: 10.1109/JSTSP.2019.2901195]

Aubrey A J, Hicks Y A and Chambers J A. 2010. Visual voice activity detection with optical flow. IET Image Processing, 4(6): 463-472 [DOI: 10.1049/iet-ipr.2009.0042]

Aung Z H and Ritthipravat P. 2015. Robust visual voice activity detection using long short-term memory recurrent neural network//7th Pacific-Rim Symposium on Image and Video Technology. Auckland, New Zealand: Springer: 380-391 [DOI: 10.1007/978-3-319-29451-3_31]

Berghi D, Volino M and Jackson P J B. 2022. Tragic Talkers: a Shakespearean sound- and light-field dataset for audio-visual machine learning research//Proceedings of the 19th ACM SIGGRAPH European Conference on Visual Media Production. London, UK: ACM: #5 [DOI: 10.1145/3565516.3565522]

Beyan C, Shahid M and Murino V. 2021. RealVAD: a real-world dataset and A method for voice activity detection by body motion analysis. IEEE Transactions on Multimedia, 23: 2071-2085 [DOI: 10.1109/TMM.2020.3007350]

Bigata A, Mira R, Bounareli S, Stypulkowski M, Vougioukas K, Petridis S, et al. 2025. KeySync: a robust approach for leakage-free lip synchronization in high resolution [EB/OL].[2026-02-20]. https://arxiv.org/pdf/2505.00497.pdf

Bilen H, Fernando B, Gavves E, Vedaldi A and Gould S. 2016. Dynamic image networks for action recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3034-3042 [DOI: 10.1109/CVPR.2016.331]

Brown A, Kalogeiton V and Zisserman A. 2021. Face, body, voice: video person-clustering with multiple modalities//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 3177-3187 [DOI: 10.1109/ICCVW54120.2021.00357]

Cadavid S, Abdel-Mottaleb M, Messinger D S, Mahoor M H and Bahrick L E. 2009. Detecting local audio-visual synchrony in monologues utilizing vocal pitch and facial landmark trajectories//Cavallaro A, Prince S and Alexander D C. Proceedings of the British Machine Vision Conference. London, UK: BMVA Press: 1-11 [DOI: 10.5244/C.23.10]

Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, et al. 2005. The AMI meeting corpus: a pre-announcement//Proceedings of the 2nd International Workshop on Machine Learning for Multimodal Interaction. Edinburgh, UK: Springer: 28-39 [DOI: 10.1007/11677482_3]

Carneiro H, Weber C and Wermter S. 2023. Whose emotion matters? Speaking activity localisation without prior knowledge. Neurocomputing, 545: #126271 [DOI: 10.1016/J.NEUCOM.2023.126271]

Carneiro H and Wermter S. 2025. FabuLight-ASD: unveiling speech activity via body language. Neural Computing and Applications, 37(5): 3561-3579 [DOI: 10.1007/S00521-024-10792-0]

Carneiro H C C, Weber C and Wermter S. 2021. FaVoA: face-voice association favours ambiguous speaker detection//Proceedings of the 30th International Conference on Artificial Neural Networks. Bratislava, Slovakia: Springer: 439-450 [DOI: 10.1007/978-3-030-86362-3_36]

Caus D, Carbajal G, Gerkmann T and Frintrop S. 2021. See the silence: improving visual-only voice activity detection by optical flow and RGB fusion//Proceedings of the 13th International Conference on Computer Vision Systems. [s.l.]: Springer: 41-51 [DOI: 10.1007/978-3-030-87156-7_4]

Cech J, Mittal R, Deleforge A, Sanchez-Riera J, Alameda-Pineda X and Horaud R. 2013. Active-speaker detection and localization with microphones and cameras embedded into a robotic head//Proceedings of the 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids). Atlanta, USA: IEEE: 203-210 [DOI: 10.1109/humanoids.2013.7029977]

Chakravarty P, Mirzaei S, Tuytelaars T and Van Hamme H. 2015. Who's speaking?: Audio-supervised classification of active speakers in video//Proceedings of 2015 ACM on International Conference on Multimodal Interaction. Seattle, USA: ACM: 87-90 [DOI: 10.1145/2818346.2820780]

Chakravarty P and Tuytelaars T. 2016. Cross-modal supervision for learning active speaker detection in video//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 285-301 [DOI: 10.1007/978-3-319-46454-1_18]

Chakravarty P, Zegers J, Tuytelaars T and Van Hamme H. 2016. Active speaker detection with audio-visual co-training//Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan: ACM: 312-316 [DOI: 10.1145/2993148.2993172]

Chang K K, Cramer M H, Ho A, Nguyen T T, Yuan Y L and Bamman D. 2026. Multimodal conversation structure understanding//Pro-

ceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Rabat, Morocco: Association for Computational Linguistics: 7437-7458 [DOI: 10.18653/v1/2026.eacl-long.349]

Chatfield K, Simonyan K, Vedaldi A and Zisserman A. 2014. Return of the devil in the details: delving deep into convolutional nets//Proceedings of 2014 British Machine Vision Conference. Nottingham, UK: BMVA Press

Chaudhuri S, Roth J, Ellis D P W, Gallagher A, Kaver L, Marvin R, et al. 2018. AVA-Speech: a densely labeled dataset of speech activity in movies//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA: 1239-1243 [DOI: 10.21437/Interspeech.2018-2028]

Chen G Y, Zhang D Y, Liu T and Du X Y. 2022a. Self-lifting: a novel framework for unsupervised voice-face association learning//Proceedings of 2022 International Conference on Multimedia Retrieval. Newark, USA: ACM: 527-535 [DOI: 10.1145/3512527.3531364]

Chen H, Zhou H S, Du J, Lee C H, Chen J D, Watanabe S, et al. 2022b. The first multimodal information based speech processing (MISP) challenge: data, tasks, baselines and results//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE: 9266-9270 [DOI: 10.1109/ICASSP43922.2022.9746683]

Chen S J. 2022. Research and implementation of students' speaking behavior detection algorithm in PBL class. Wuhan: Huazhong University of Science and Technology (陈世俊. 2022. PBL课堂中的学生发言行为检测算法研究与实现. 武汉: 华中科技大学) [DOI: 10.27157/d.cnki.ghzku.2022.002590]

Chen S Y, Wang C Y, Chen Z Y, Wu Y, Liu S J, Chen Z, et al. 2022c. WavLM: large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16 (6): 1505-1518 [DOI: 10.1109/JSTSP. 2022. 3188113]

Cheng L Y, Wang H, Deng C, Zheng S Q, Chen Y F, Huang R J, et al. 2025. Integrating audio, visual, and semantic information for enhanced multimodal speaker diarization on multi-party conversation//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics: 19914-19928 [DOI: 10.18653/v1/2025.acl-long.977]

Cho K, van Merriënboer B, Bahdanau D and Bengio Y. 2014. On the properties of neural machine translation: encoder-decoder approaches//Proceedings of SSST-8, the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar: Association for Computational Linguistics: 103-111 [DOI: 10. 3115/v1/W14-4012]

Chung J S. 2019. Naver at ActivityNet challenge 2019 — Task B active speaker detection (AVA) [EB/OL].[2026-02-20]. https://arxiv.org/pdf/1906.10555.pdf

Chung J S, Huh J, Nagrani A, Afouras T and Zisserman A. 2020a. Spot the conversation: speaker diarisation in the wild//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 299-303 [DOI: 10. 21437/INTERSPEECH.2020-2337]

Chung J S, Lee B J and Han I. 2019a. Who said that?: Audio-visual speaker diarisation of real-world meetings//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 371-375 [DOI: 10.21437/ INTERSPEECH.2019-3116]

Chung J S, Nagrani A and Zisserman A. 2018. VoxCeleb2: deep speaker recognition//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA: 1086-1090 [DOI: 10.21437/INTERSPEECH. 2018-1929]

Chung J S and Zisserman A. 2016a. Lip reading in the wild//Proceedings of the 13th Asian Conference on Computer Vision. Taipei, China: Springer: 87-103 [DOI: 10.1007/978-3-319-54184-6_6]

Chung J S and Zisserman A. 2016b. Out of time: automated lip sync in the wild//Proceedings of the ACCV 2016 International Workshops on Computer Vision — ACCV 2016 Workshops. Taipei, China: Springer: 251-263 [DOI: 10.1007/978-3-319-54427-4_19]

Chung J S and Zisserman A. 2017. Lip reading in profile//British Machine Vision Conference 2017. London, UK: BMVA Press: 1-11

Chung S E, Kang H G and Chung J S. 2020c. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 3486-3490 [DOI: 10.21437/INTERSPEECH.2020-1113]

Chung S W, Chung J S and Kang H G. 2019b. Perfect match: improved cross-modal embeddings for audio-visual synchronisation//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE: 3965-3969 [DOI: 10.1109/ICASSP.2019.8682524]

Chung S W, Chung J S and Kang H G. 2020b. Perfect match: self-supervised embeddings for cross-modal retrieval. IEEE Journal of Selected Topics in Signal Processing, 14(3): 568-576 [DOI: 10. 1109/JSTSP.2020.2987720]

Clarke J, Gotoh Y and Goetze S. 2023. Improving audiovisual active speaker detection in egocentric recordings with the data-efficient image transformer//Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop. Taipei, China: IEEE: 1-8 [DOI: 10.1109/ASRU57964.2023.10389764]

Clarke J, Gotoh Y and Goetze S. 2025a. Face-voice association for audiovisual active speaker detection in egocentric recordings//Proceedings of the 33rd European Signal Processing Conference. Palermo, Italy: IEEE: 66-70 [DOI: 10.23919/EUSIPCO63237.2025. 11226795]

Clarke J, Gotoh Y and Goetze S. 2025b. Speaker embedding informed audiovisual active speaker detection for egocentric recordings//Proceedings of 2025 IEEE International Conference on Acoustics,

Speech and Signal Processing. Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10890414]

Clarke J, Gotoh Y and Goetze S. 2025c. Ensembling synchronisation-based and face-voice association paradigms for robust active speaker detection in egocentric recordings//Proceedings of the 27th International Conference on Speech and Computer. Szeged, Hungary: Springer: 289-301 [DOI: 10.1007/978-3-032-07959-6_21]

Cutler R and Davis L. 2000. Look who's talking: speaker detection using video and audio correlation//Proceedings of 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532). New York, USA: IEEE: 1589-1592 [DOI: 10.1109/ICME.2000.871073]

Cutler R, Mehran R, Johnson S, Zhang C, Kirk A, Whyte O, et al. 2020. Multimodal active speaker detection and virtual cinematography for video conferencing//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE: 4527-4531 [DOI: 10.1109/ICASSP40776.2020.9053171]

Dao T and Gu A. 2024. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 10041-10071

Datta G, Etchart T, Yadav V, Hedau V, Natarajan P and Chang S F. 2022. ASD-transformer: efficient active speaker detection using self and multimodal transformers//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE: 4568-4572 [DOI: 10.1109/ICASSP43922.2022.9746991]

Ding B, Kirtay M and Spigler G. 2024. Imitation of human motion achieves natural head movements for humanoid robots in an active-speaker detection task//Proceedings of the 23rd IEEE-RAS International Conference on Humanoid Robots. Nancy, France: IEEE: 645-652 [DOI: 10.1109/HUMANOIDS58906.2024.10769814]

Ding Y F, Xu Y, Zhang S X, Cong Y H and Wang L Q. 2020. Self-supervised learning for audio-visual speaker diarization//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE: 4367-4371 [DOI: 10.1109/ICASSP40776.2020.9054376]

Dolédec S and Chessel D. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology, 31(3): 277-294 [DOI: 10.1111/j.1365-2427.1994.tb01741.x]

Donley J, Tourbabin V, Lee J S, Broyles M, Jiang H, Shen J, et al. 2021. EasyCom: an augmented reality dataset to support algorithms for easy communication in noisy environments[EB/OL].[2026-02-20].https://arxiv.org/pdf/2107.04174.pdf

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview. net

Durrani I A, Liu C R, Ishi C T and Ishiguro H. 2024. Is it possible to recognize a speaker without listening? Unraveling conversation dynamics in multi-party interactions using continuous eye gaze. IEEE Robotics and Automation Letters, 9(11): 9923-9929 [DOI: 10.1109/LRA.2024.3440844]

El Khoury E, Sénac C and Joly P. 2014. Audiovisual diarization of people in video content. Multimedia Tools and Applications, 68(3): 747-775 [DOI: 10.1007/S11042-012-1080-6]

Ephrat A, Mosseri I, Lang O R, Dekel T, Wilson K, Hassidim A, et al. 2018. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG), 37(4): #112 [DOI: 10.1145/3197517.3201357]

Everingham M, Sivic J and Zisserman A. 2009. Taking the bite out of automated naming of characters in TV video. Image and Vision Computing, 27(5): 545-559 [DOI: 10.1016/J. IMAVIS. 2008.04.018]

Everingham M R, Sivic J and Zisserman A. 2006. "Hello! My name is... Buffy" — Automatic naming of characters in TV video//Proceedings of the British Machine Vision Conference 2006. Edinburgh, UK: British Machine Vision Association: 899-908 [DOI: 10.5244/C.20.92]

Fernandez-Labrador C, Akçay M, Abecassis E, Massich J and Schroers C. 2024. DiVAS: video and audio synchronization with dynamic frame rates//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26836-26844 [DOI: 10.1109/CVPR52733.2024.02535]

Fisher III J W and Darrell T. 2004. Speaker association with signal-level audiovisual fusion. IEEE Transactions on Multimedia, 6(3): 406-413 [DOI: 10.1109/TMM.2004.827503]

Gebru I D, Ba S, Li X F and Horaud R. 2018. Audio-visual speaker diarization based on spatiotemporal Bayesian fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5): 1086-1099 [DOI: 10.1109/TPAMI.2017.2648793]

Geeroms W, Allebosch G, Kindt S, Kadri L, Veelaert P and Madhu N. 2022. Audio-visual active speaker identification: a comparison of dense image-based features and sparse facial landmark-based features//Sensor Data Fusion: Trends, Solutions, Applications. Bonn, Germany: IEEE: 1-6 [DOI: 10.1109/SDF55338.2022.9931697]

Gemmeke J F, Ellis D P W, Freedman D, Jansen A, Lawrence W, Moore R C, et al. 2017. Audio set: an ontology and human-labeled dataset for audio events//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA: IEEE: 776-780 [DOI: 10.1109/ICASSP. 2017.7952261]

Girmaji R, Achary S, Deshmukh A and Gandhi V. 2023. Assessing active speaker detection algorithms through the lens of automated editing//Proceedings of 2023 ACM International Conference on Interactive Media Experiences Workshops. Nantes, France: ACM:

123-130 [DOI: 10.1145/3604321.3604373]

Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, et al. 2022. Ego4D: around the world in 3,000 hours of egocentric video//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18973-18990 [DOI: 10.1109/CVPR52688.2022.01842]

Gu A and Dao T. 2023. Mamba: linear-time sequence modeling with selective state spaces [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2312.00752.pdf

Gu C H, Sun C, Ross D A, Vondrick C, Pantofaru C, Li Y Q, et al. 2018. AVA: a video dataset of spatio-temporally localized atomic visual actions//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6047-6056 [DOI: 10.1109/CVPR.2018.00633]

Gulati A, Qin J, Chiu C C, Parmar N, Zhang Y, Yu J H, et al. 2020. Conformer: convolution-augmented Transformer for speech recognition//Proceedings of the 21st Annual Conference of the International Speech Communication Association, Interspeech 2020. Shanghai, China: ISCA: 5036-5040 [DOI: 10.21437/INTERSPEECH.2020-3015]

Guo C, Pleiss G, Sun Y and Weinberger K Q. 2017. On calibration of modern neural networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR: 1321-1330

Gurvich I, Leichter I, Palle D R, Asher Y, Vinnikov A, Abramovski I, et al. 2024. A real-time active speaker detection system integrating an audio-visual signal with a spatial querying mechanism//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea(South): IEEE: 8781-8785 [DOI: 10.1109/ICASSP48485.2024.10446169]

Guy S, Lathuilière S, Mesejo P and Horaud R. 2021. Learning visual voice activity detection with an automatically annotated dataset//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 4851-4856 [DOI: 10.1109/ICPR48806.2021.9412884]

Hadsell R, Chopra S and LeCun Y. 2006. Dimensionality reduction by learning an invariant mapping//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE: 1735-1742 [DOI: 10.1109/CVPR.2006.100]

Haider F, Campbell N and Luz S. 2016. Active speaker detection in human machine multiparty dialogue using visual prosody information//Proceedings of 2016 IEEE Global Conference on Signal and Information Processing. Washington, USA: IEEE: 1207-1211 [DOI: 10.1109/GLOBALSIP.2016.7906033]

He J and Zhang H. 2008. Speech endpoint detection based on lip moving. Microcomputer Information, 24(26): 221-223 (何俊, 张华. 2008. 基于唇动特征的语音端点检测. 微计算机信息, 24(26): 221-223) [DOI: 10.3969/j.issn.1008-0570.2008.26.092]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He M K, Du J, Niu S T, Liu Q F and Lee C H. 2024. Quality-aware end-to-end audio-visual neural speaker diarization [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2410.22350.pdf

Hegde S B, Prajwal K R, Kwon T and Zisserman A. 2025. Understanding co-speech gestures in-the-wild//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 9977-9987 [DOI: 10.1109/ICCV51701.2025.00930]

Hegde S B and Zisserman A. 2023. GestSync: determining who is speaking without a talking head//Proceedigns of the 34th British Machine Vision Conference 2023. Aberdeen, UK: BMVA Press: 506-509

Hershey J and Movellan J. 1999. Audio-vision: using audio-visual synchrony to locate sounds//Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver, USA: The MIT Press: 813-819

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/NECO.1997.9.8.1735]

Hoover K, Chaudhuri S, Pantofaru C, Sturdy I and Slaney M. 2018. Using audio-visual information to understand speaker activity: tracking active speakers on and off screen//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada: IEEE: 6558-6562 [DOI: 10.1109/ICASSP.2018.8461891]

Hradis M, Eivazi S and Bednarik R. 2012. Voice activity detection from gaze in video mediated communication//Symposium on Eye Tracking Research and Applications. Santa Barbara, USA: ACM: 329-332 [DOI: 10.1145/2168556.2168628]

Hsu W N, Bolte B, Tsai Y H H, Lakhotia K, Salakhutdinov R and Mohamed A. 2021. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 3451-3460 [DOI: 10.1109/TASLP.2021.3122291]

Hu Y T, Ren J S J, Dai J W, Yuan C, Xu L and Wang W P. 2015. Deep multimodal speaker naming//Proceedings of the 23rd ACM International Conference on Multimedia. Brisbane, Australia: ACM: 1107-1110 [DOI: 10.1145/2733373.2806293]

Huang C and Koishida K. 2020. Improved active speaker detection based on optical flow//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4084-4090 [DOI: 10.1109/CVPRW50498.2020.00483]

Huang H, Yu H Y, Liu D B, Chen H W and Cai M J. 2025. Egocentric speaker diarization with vision-guided clustering and adaptive speech re-detection//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10889699]

Huang X H. 2024. Research of Active Speaker Detection Method Based on Fisheye Camera for Meeting Scenarios. Wuhan: Huazhong Uni-

versity of Science and Technology（黄星豪. 2024. 基于鱼眼镜头的会议说话者检测方法研究. 武汉：华中科技大学）

Huang X H, Jiang W W, Rao L, Xu W and Cheng W Q. 2024. Active speaker detection in fisheye meeting scenes with scene spatial spectrums//Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024. Kos, Greece: ISCA: 4283-4287 [DOI: 10.21437/INTERSPEECH.2024-1402]

Huh J, Ortiz J A, Kumar A, Pandey A, Aroudi A, Wong D D E, et al. 2025. Advancing active speaker detection for egocentric videos// Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10888166]

Huh J and Zisserman A. 2024. Character-aware audio-visual subtitling in context//Proceedings of the 17th Asian Conference on Computer Vision on Computer Vision — ACCV 2024. Hanoi, Vietnam: Springer: 365-383 [DOI: 10.1007/978-981-96-0908-6_21]

Hung H and Ba S O. 2010. Speech/non-speech detection in meetings from automatically extracted low resolution visual features//Proceedings of 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing. Sheraton Dallas Hotel, USA: IEEE: 830-833

Iyengar G, Nock H J and Neti C. 2003a. Audio-visual synchrony for detection of monologues in video archives//Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. Hong Kong, China: IEEE: V-772 [DOI: 10.1109/ICASSP.2003.1200085]

Iyengar G, Nock H J and Neti C. 2003b. Audio-visual synchrony for detection of monologues in video archives//Proceedings of 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698). Baltimore, USA: IEEE: I-329 [DOI: 10.1109/ICME.2003.1220921]

Jawade B, Gadde R T, Bejjani C and Lan Y H. 2025. Audio-visual representation learning for lip-sync estimation through ranking augmented contrastive training//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10888378]

Jiang H, Murdock C and Ithapu V K. 2022. Egocentric deep multichannel audio-visual active speaker localization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10534-10542 [DOI: 10.1109/CVPR52688.2022.01029]

Jiang W W, Rao L, Dai G L, Wu Y F and Xu W. 2025. Gaze-driven active speaker detection in meetings//Proceedings of the 21st International Conference on Advanced Intelligent Computing Technology and Applications. Ningbo, China: Springer: 235-246 [DOI: 10.1007/978-981-96-9849-3_20]

Jiang Y D, Tao R J, Pan Z X and Li H Z. 2023. Target active speaker detection with audio-visual cues//Proceedings of the 24th Annual Conference of the International Speech Communication Association. Dublin, Ireland: ISCA: 3152-3156 [DOI: 10.21437/INTERSPEECH.2023-574]

Joosten B, Postma E and Krahmer E. 2013. Visual voice activity detection at different speeds//Auditory-Visual Speech Processing. Annecy, France: ISCA: 187-190

Jung C, Lee S, Nam K, Rho K, Kim Y J, Jang Y, et al. 2024. TalkNCE: improving active speaker detection with talk-aware contrastive learning//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea (South): IEEE: 8391-8395 [DOI: 10.1109/ICASSP48485.2024.10448124]

Kim C, Shin H V, Oh T H, Kaspar A, Elgharib M and Matusik W. 2018. On learning associations of faces and voices//Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer: 276-292 [DOI: 10.1007/978-3-030-20873-8_18]

Kim U H. 2021. Noise-tolerant self-supervised learning for audio-visual voice activity detection//Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021. Brno, Czechia: ISCA: 326-330 [DOI: 10.21437/INTERSPEECH.2021-43]

Kim Y J, Heo H S, Choe S, Chung S W, Kwon Y, Lee B J, et al. 2021. Look who's talking: active speaker detection in the wild// Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021. Brno, Czechia: ISCA: 3675-3679 [DOI: 10.21437/INTERSPEECH.2021-2041]

Kipf T N and Welling M. 2017. Semi-supervised classification with graph convolutional networks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Köpüklü O, Taseska M and Rigoll G. 2021. How to design a three-stage architecture for audio-visual active speaker detection in the wild// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1173-1183 [DOI: 10.1109/ICCV48922.2021.00123]

Korbar B, Huh J and Zisserman A. 2024. Look, listen and recognise: character-aware audio-visual subtitling//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea (South): IEEE: 2975-2979 [DOI: 10.1109/ICASSP48485.2024.10446480]

Korbar B, Tran D and Torresani L. 2018. Cooperative learning of audio and video models from self-supervised synchronization//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 7774-7785

Kundu A, Jin Y Z, Sekhavat M H, Horton M, Tormoen D and Naik D. 2025. An efficient and streaming audio visual active speaker detection system//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10887588]

Kwak D, Jung J, Nam K, Jang Y, Jung J W, Watanabe S, et al. 2024. VoxMM: rich transcription of conversations in the wild//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, Korea(South): IEEE: 12551-12555 [DOI: 10.1109/ICASSP48485.2024.10446300]

Kyoung M and Song H J. 2023. Modeling long-term multimodal representations for active speaker detection with spatio-positional encoder. IEEE Access, 11: 116561-116569 [DOI: 10.1109/ACCESS.2023.3325474]

Li C Y, Zhang C, Xu W K, Lin J Y, Xie J H, Feng W G, et al. 2024. LatentSync: taming audio-conditioned latent diffusion models for lip sync with SyncNet supervision [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2412.09262.pdf

Li H, Xu M W, Zhan Y, Mu S, Li J Y, Cheng K H, et al. 2025a. OpenHumanVid: a large-scale high-quality dataset for enhancing human-centric video generation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7752-7762 [DOI: 10.1109/CVPR52734.2025.00726]

Li X, Wen Y D, Yang M Q, Wang J L, Singh R and Raj B. 2023. Rethinking voice-face correlation: a geometry view//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 2458-2467 [DOI: 10.1145/3581783.3611779]

Li Y Q, Luo Y and Zhou X. 2025b. Robust active speaker detection in challenging environments using GNN-fused multi-modal cues and body language//Proceedings of the 31st International Conference on Multimedia Modeling. Nara, Japan: Springer: 326-339 [DOI: 10.1007/978-981-96-2064-7_24]

Liao J H, Duan H H, Feng K H, Zhao W B, Yang Y B and Chen L Y. 2023. A light weight model for active speaker detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22932-22941 [DOI: 10.1109/CVPR52729.2023.02196]

Liao J H, Duan H H, Feng K H, Zhao W B, Yang Y B, Chen L Y, et al. 2025. LR-ASD: lightweight and robust network for active speaker detection. International Journal of Computer Vision, 133(7): 4749-4769 [DOI: 10.1007/S11263-025-02399-2]

Lin J, Gan C and Han S. 2019. TSM: temporal shift module for efficient video understanding//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7082-7092 [DOI: 10.1109/ICCV.2019.00718]

Lin J, Gan C, Wang K and Han S. 2022. TSM: Temporal shift module for efficient and scalable video understanding on edge devices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5): 2760-2774 [DOI: 10.1109/TPAMI.2020.3029799]

Liu Q J, Wang W W and Jackson P. 2011. A visual voice activity detection method with adaboosting//Sensor Signal Processing for Defence (SSPD 2011). London, UK: IET: 1-5 [DOI: 10.1049/ic.2011.0145]

Liu T, Fan S, Xiang X, Song H B, Lin S X, Sun J Q, et al. 2022. MSDWild: multi-modal speaker diarization dataset in the wild//Proceedings of the 23rd Annual Conference of the International Speech Communication Association, Interspeech 2022. Incheon, Korea (South): ISCA: 1476-1480 [DOI: 10.21437/INTERSPEECH.2022-10466]

Long L, He Y C, Ye W T, Pan Y Y, Lin Y, Li H, et al. 2026. Seeing, listening, remembering, and reasoning: a multimodal agent with long-term memory//Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: OpenReview.net

Lubitz A, Valdenegro-Toro M and Kirchner F. 2023. The VVAD-LRS3 dataset for visual voice activity detection//Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2023, Volume 2. Lisbon, Portugal: SCITEPRESS: 39-46 [DOI: 10.5220/0011612900003417]

Lyon R F. 2011. Cascades of two-pole-two-zero asymmetric resonators are good models of peripheral auditory function. The Journal of the Acoustical Society of America, 130(6): 3893-3904 [DOI: 10.1121/1.3658470]

Marcheret E, Potamianos G, Vopicka J and Goel V. 2015. Detecting audio-visual synchrony using deep neural networks//Proceedings of the 16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015. Dresden, Germany: ISCA: 548-552 [DOI: 10.21437/INTERSPEECH.2015-201]

Min K. 2023. STHG: spatial-temporal heterogeneous graph learning for advanced audio-visual diarization [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2306.10608.pdf

Min K, Roy S, Tripathi S, Guha T and Majumdar S. 2022. Learning long-term spatial-temporal graphs for active speaker detection//Proceedings of the 17th European Conference on Computer Vision-ECCV 2022. Tel Aviv, Israel: Springer: 371-387 [DOI: 10.1007/978-3-031-19833-5_22]

Mocanu B and Tapu R. 2025. A lightweight audio-visual speaker detection system for assistive video captioning//Proceedings of the 13th European Workshop on Visual Information Processing (EUVIP). Valletta, Malta: IEEE: 1-6 [DOI: 10.1109/EUVIP66349.2025.11238860]

Nagrani A, Albanie S and Zisserman A. 2018. Seeing voices and hearing faces: Cross-modal biometric matching//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8427-8436 [DOI: 10.1109/CVPR.2018.00879]

Navarathna R, Dean D, Sridharan S, Fookes C and Lucey P. 2011. Visual voice activity detection using frontal versus profile views//

Proceedings of 2011 International Conference on Digital Image Computing: Techniques and Applications. Noosa, Australia: IEEE: 134-139 [DOI: 10.1109/DICTA.2011.29]

Nguyen L T P, Yu Z R, Cao K Q N, Guo Y W, Pham T H M, Nguyen T T, et al. 2025. UniTalk: towards universal active speaker detection in real world scenarios [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2505.21954.pdf

Nguyen L T P, Yu Z R and Lee Y J. 2026. LASER: lip landmark assisted speaker detection for robustness//Proceedings of 2026 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 7291-7300 [DOI: 10.1109/WACV61042.2026.00704]

Nock H J, Iyengar G and Neti C. 2002. Assessing face and speech consistency for monologue detection in video//Proceedings of the 10th ACM International Conference on Multimedia. Juan-les-Pins, France: ACM: 303-306 [DOI: 10.1145/641007.641070]

Northcutt C G, Zha S, Lovegrove S and Newcombe R. 2023. EgoCom: a multi-person multi-modal egocentric communications dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6): 6783-6793 [DOI: 10.1109/TPAMI.2020.3025105]

Owens A and Efros A A. 2018. Audio-visual scene analysis with self-supervised multisensory features//Proceedings of 15th European Conference on Computer Vision-ECCV 2018. Munich, Germany: Springer: 639-658 [DOI: 10.1007/978-3-030-01231-1_39]

Park D S, Chan W, Zhang Y, Chiu C C, Zoph B, Cubuk E D, et al. 2019a. SpecAugment: a simple data augmentation method for automatic speech recognition//Proceedings of the 20th Annual Conference of the International Speech Communication Association, Interspeech 2019. Graz, Austria: ISCA: 2613-2617 [DOI: 10.21437/INTERSPEECH.2019-2680]

Park S, Yun J, Lee D and Park M. 2024. Interpretable convolutional SyncNet [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2409.00971.pdf

Park T J and Georgiou P. 2018. Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks//Proceedings of the 19th Annual Conference of the International Speech Communication Association, Interspeech 2018. Hyderabad, India: ISCA: 1373-1377 [DOI: 10.21437/INTERSPEECH.2018-1364]

Park T J, Han K J, Huang J, He X D, Zhou B W, Georgiou P, et al. 2019b. Speaker diarization with lexical information//Proceedings of the 20th Annual Conference of the International Speech Communication Association, Interspeech 2019. Graz, Austria: ISCA: 391-395 [DOI: 10.21437/INTERSPEECH.2019-1947]

Patrona F, Iosifidis A, Tefas A, Nikolaidis N and Pitas I. 2015. Visual voice activity detection based on spatiotemporal information and bag of words//Proceedings of 2015 IEEE International Conference on Image Processing. Quebec City, Canada: IEEE: 2334-2338 [DOI: 10.1109/ICIP.2015.7351219]

Patrona F, Iosifidis A, Tefas A, Nikolaidis N and Pitas I. 2016. Visual voice activity detection in the wild. IEEE Transactions on Multimedia, 18(6): 967-977 [DOI: 10.1109/TMM.2016.2535357]

Peng B, Alcaide E, Anthony Q, Albalak A, Arcadinho S, Biderman S, et al. 2023. RWKV: reinventing RNNs for the transformer era//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Singapore: Association for Computational Linguistics: 14048-14077 [DOI: 10.18653/V1/2023.FINDINGS-EMNLP.936]

Pibre L, Madrigal F, Equoy C, Lerasle F, Pellegrini T, Pinquier J, et al. 2023. Audio-video fusion strategies for active speaker detection in meetings. Multimedia Tools and Applications, 82(9): 13667-13688 [DOI: 10.1007/S11042-022-13746-7]

Poria S, Hazarika D, Majumder N, Naik G, Cambria E and Mihalcea R. 2019. MELD: a multimodal multi-party dataset for emotion recognition in conversations//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 527-536 [DOI: 10.18653/V1/P19-1050]

Pouthier B, Pilati L, Gudupudi L K, Bouveyron C and Precioso F. 2021. Active speaker detection as a multi-objective optimization with uncertainty-based multimodal fusion//Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021. Brno, Czechia: ISCA: 2381-2385 [DOI: 10.21437/INTERSPEECH.2021-80]

Prajwal K R, Afouras T and Zisserman A. 2022. Sub-word level lip reading with visual attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5152-5162 [DOI: 10.1109/CVPR52688.2022.00510]

Qin L. 2009. Dialogue Unit Segmentation of Films and Television Programs Based on Lip Movement Detection. Harbin: Harbin Institute of Technology (秦蕾. 2009. 基于唇动检测的影视作品对白单元切分. 哈尔滨: 哈尔滨工业大学) [DOI: 10.7666/d.D258647]

Radman A and Laaksonen J. 2024. AS-Net: active speaker detection using deep audio-visual attention. Multimedia Tools and Applications, 83(28): 72027-72042 [DOI: 10.1007/S11042-024-18457-9]

Ravi N, Gabeur V, Hu Y T, Hu R H, Ryali C, Ma T Y, et al. 2025. SAM 2: segment anything in images and videos//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Ren J, Hu Y T, Tai Y W, Wang C, Xu L, Sun W X, et al. 2016. Look, listen and learn — A multimodal LSTM for speaker identification//Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI Press: 3581-3587 [DOI: 10.1609/AAAI.V30I1.10471]

Robi S N A M, Ariffin M A Z M, Izhar M A M, Ahmad N and Kaidi H M. 2024. Active speaker detection using audio, visual, and depth modalities: a survey. IEEE Access, 12: 96617-96634 [DOI: 10.1109/ACCESS.2024.3426670]

Roth J, Chaudhuri S, Klejch O, Marvin R, Gallagher A, Kaver L, et al. 2020. AVA active speaker: an audio-visual dataset for active

speaker detection//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE: 4492-4496 [DOI: 10.1109/ICASSP40776.2020.9053900]

Roxo T, Costa J C, Inácio P R M and Proença H. 2024a. How to squeeze an explanation out of your model//Computer Vision — ECCV 2024 Workshops. Milan, Italy: Springer: 247-262 [DOI: 10.1007/978-3-031-92089-9_16]

Roxo T, Costa J C, Inácio P R M and Proença H. 2024b. ASDnB: merging face with body cues for robust active speaker detection//Proceedings of 2025 IEEE International Joint Conference on Biometrics (IJCB). Osaka, Japan: IEEE: 1-10 [DOI: 10.1109/IJCB65343.2025.11410746]

Roxo T, Costa J C, Inácio P R M and Proença H. 2025a. WASD: a wilder active speaker detection dataset. IEEE Transactions on Biometrics, Behavior, and Identity Science, 7(1): 61-70 [DOI: 10.1109/TBIOM.2024.3412821]

Roxo T, Costa J C, Inácio P R M and Proença H. 2025b. BIAS: a body-based interpretable active speaker approach. IEEE Transactions on Biometrics, Behavior, and Identity Science, 7(3): 410-421 [DOI: 10.1109/TBIOM.2024.3520030]

Saenko K, Livescu K, Siracusa M, Wilson K, Glass J and Darrell T. 2005. Visual speech recognition with loosely synchronized feature streams//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE: 1424-1431 [DOI: 10.1109/ICCV.2005.251]

Sanchez-Riera J, Alameda-Pineda X, Wienke J, Deleforge A, Arias S, Cech J, et al. 2012. Online multimodal speaker detection for humanoid robots//Proceedings of the 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012). Osaka, Japan: IEEE: 126-133 [DOI: 10.1109/HUMANOIDS.2012.6651509]

Shafey L E, Soltau H and Shafran I. 2019. Joint speech recognition and speaker diarization via sequence transduction//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 396-400 [DOI: 10.21437/INTERSPEECH.2019-1943]

Shahid M, Beyan C and Murino V. 2019. Voice activity detection by upper body motion analysis and unsupervised domain adaptation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 1260-1269 [DOI: 10.1109/ICCVW.2019.00159]

Shahid M, Beyan C and Murino V. 2021. S-VVAD: visual voice activity detection by motion segmentation//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2331-2340 [DOI: 10.1109/WACV48630.2021.00238]

Sharma R and Narayanan S. 2022. Audio visual character profiles for detecting background characters in entertainment media [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2203.11368.pdf

Sharma R and Narayanan S. 2023. Audio-visual activity guided cross-modal identity association for active speaker detection. IEEE Open Journal of Signal Processing, 4: 225-232 [DOI: 10.1109/ojsp.2023.3267269]

Sharma R, Somandepalli K and Narayanan S. 2019. Toward visual voice activity detection for unconstrained videos//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 2991-2995 [DOI: 10.1109/ICIP.2019.8803248]

Sharma R, Somandepalli K and Narayanan S. 2023. Cross modal video representations for weakly supervised active speaker localization. IEEE Transactions on Multimedia, 25: 7825-7836 [DOI: 10.1109/TMM.2022.3229975]

Shi B W, Tjandra A, Hoffman J, Wang H L, Wu Y C, Gao L Y, et al. 2025. SAM Audio: segment anything in audio [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2512.18099.pdf

Siatras S, Nikolaidis N, Krinidis M and Pitas I. 2009. Visual lip activity detection and speaker detection using mouth region intensities. IEEE Transactions on Circuits and Systems for Video Technology, 19(1): 133-137 [DOI: 10.1109/TCSVT.2008.2009262]

Siatras S, Nikolaidis N and Pitas I. 2006. Visual speech detection using mouth region intensities//Proceedings of the 14th European Signal Processing Conference. Florence, Italy: IEEE: 1-5

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

Slaney M and Covell M. 2000. FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks//Proceedings of the 14th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press: 784-790

Sohn K. 2016. Improved deep metric learning with multi-class *N*-pair loss objective//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona Spain: 1857-1865

Somandepalli K, Guha T, Martinez V R, Kumar N, Adam H and Narayanan S. 2021. Computational media intelligence: Human-centered machine analysis of media. Proceedings of the IEEE, 109(5): 891-910 [DOI: 10.1109/JPROC.2020.3047978]

Stafylakis T and Tzimiropoulos G. 2017. Combining residual networks with LSTMs for lipreading//Proceedings of the 18th Annual Conference of the International Speech Communication Association, Interspeech 2017. Stockholm, Sweden: ISCA: 3652-3656 [DOI: 10.21437/INTERSPEECH.2017-85]

Stefanov K, Adiban M and Salvi G. 2020. Spatial bias in vision-based voice activity detection//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 10433-10440 [DOI: 10.1109/ICPR48806.2021.9413345]

Stefanov K, Beskow J and Salvi G. 2017. Vision-based active speaker detection in multiparty interactions//GLU 2017 International Workshop on Grounding Language Understanding. [s.l.]: [s.n.]: 47-51 [DOI: 10.21437/glu.2017-10]

Stefanov K, Sugimoto A and Beskow J. 2016. Look who's talking: visual identification of the active speaker in multi-party human-robot interaction//Proceedings of the 2nd Workshop on Advance-

ments in Social Signal Processing for Multimodal Interaction. Tokyo, Japan: ACM: 22-27 [DOI: 10.1145/3005467.3005470]

Tao F, Hansen J H L and Busso C. 2015. An unsupervised visual-only voice activity detection approach using temporal orofacial features// Proceedings of the 16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015. Dresden, Germany: ISCA: 2302-2306 [DOI: 10.21437/INTERSPEECH.2015-446]

Tao J H, Wu Y C, Yu C, Weng D D, Li G J, Han T, et al. 2022. A survey on multi-modal human-computer interaction. Journal of Image and Graphics, 27(6): 1956-1987 (陶建华, 巫英才, 喻纯, 翁冬冬, 李冠君, 韩腾, 等. 2022. 多模态人机交互综述. 中国图象图形学报, 27(6): 1956-1987) [DOI: 10.11834/jig.220151]

Tao R J. 2023. Audio-visual Active Speaker Detection and Recognition. Singapore: National University of Singapore

Tao R J, Pan Z X, Das R K, Qian X Y, Shou M Z and Li H Z. 2021. Is someone speaking? : Exploring long-term temporal features for audio-visual active speaker detection//Proceedings of the 29th ACM International Conference on Multimedia. [s.l.]: ACM: 3927-3935 [DOI: 10.1145/3474085.3475587]

Tao R J, Qian X Y, Das R K, Gao X X, Wang J D and Li H Z. 2025. Enhancing real-world active speaker detection with multi-modal extraction pre-training. IEEE Transactions on Multimedia, 27: 2362-2373 [DOI: 10.1109/TMM.2024.3521791]

Tapu R, Mocanu B and Zaharia T. 2019a. Dynamic subtitles: a multimodal video accessibility enhancement dedicated to deaf and hearing impaired users//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 2558-2566 [DOI: 10.1109/ICCVW.2019.00313]

Tapu R, Mocanu B and Zaharia T. 2019b. DEEP-HEAR: a multimodal subtitle positioning system dedicated to deaf and hearing-impaired people. IEEE Access, 7: 88150-88162 [DOI: 10.1109/ACCESS.2019.2925806]

Thermos S and Potamianos G. 2016. Audio-visual speech activity detection in a two-speaker scenario incorporating depth information from a profile or frontal view//Proceedings of 2016 IEEE Spoken Language Technology Workshop. San Diego, USA: IEEE: 579-584 [DOI: 10.1109/SLT.2016.7846321]

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jégou H. 2021. Training data-efficient image Transformers and distillation through attention//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 10347-10357

Truong T D, Duong C N, Vu T D, Pham H A, Raj B, Le N, et al. 2021. The right to talk: an audio-visual Transformer approach//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1085-1094 [DOI: 10.1109/ICCV48922.2021.00114]

Vajaria H, Sarkar S and Kasturi R. 2008. Exploring co-occurence between speech and body movement for audio-guided video localization. IEEE Transactions on Circuits and Systems for Video Technology, 18(11): 1608-1617 [DOI: 10.1109/TCSVT.2008.2005602]

van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. 2016. WaveNet: a generative model for raw audio//The 9th ISCA Speech Synthesis Workshop, SSW 2016. Sunnyvale, USA: ISCA: #125

van den Oord A, Li Y Z and Vinyals O. 2018. Representation learning with contrastive predictive coding [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/1807.03748.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Veličković P, Cucurull G, Casanova A, Romero A, Liò P and Bengio Y. 2018. Graph attention networks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Wang B Y and Wang X L. 2019. Are you speaking: real-time speech activity detection via landmark pooling network//Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition. Lille, France: IEEE: 1-5 [DOI: 10.1109/FG.2019.8756549]

Wang Q, Huang Y L, Zhao G L, Clark E, Xia W and Liao H. 2024a. DiarizationLM: speaker diarization post-processing with large language models//Proceedings of the 25th Annual Conference of the International Speech Communication Association, Interspeech 2024. Kos, Greece: ISCA: 3754-3758 [DOI: 10.21437/INTERSPEECH.2024-209]

Wang X, Han X T, Huang W L, Dong D K and Scott M R. 2019. Multi-similarity loss with general pair weighting for deep metric learning// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5017-5025 [DOI: 10.1109/CVPR.2019.00516]

Wang X Z, Cheng F and Bertasius G. 2024b. LoCoNet: long-short context network for active speaker detection//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18462-18472 [DOI: 10.1109/CVPR52733.2024.01747]

Wang Y, Ha J, Ramirez F M, Wang Y C and Crandall D J. 2026. GateFusion: hierarchical gated cross-modal fusion for active speaker detection//Proceedings of 2026 IEEE/CVF Winter Conference on Applications of Computer Vision. Tucson, USA: IEEE: 1074-1083 [DOI: 10.1109/WACV61042.2026.00111]

Wang Y J, Xu H R, Liu Y, Li J Z and Tang Y S. 2025. SAM2-LOVE: segment anything model 2 in language-aided audio-visual scenes// Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 28932 [DOI: 10.1109/CVPR52734.2025.02694]

Wang Z, Wu S L, Chen H, He M K, Du J, Lee C H, et al. 2023. The multimodal information based speech processing (MISP) 2022 challenge: audio-visual diarization and recognition//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.

1109/ICASSP49357.2023.10094836]

Wen P S, Xu Q Q, Jiang Y B Y, Yang Z Y, He Y and Huang Q M. 2021. Seeking the shape of sound: an adaptive framework for learning voice-face association//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16342-16351 [DOI: 10.1109/CVPR46437.2021.01608]

Wen Y D, Ismail M A, Liu W Y, Raj B and Singh R. 2019. Disjoint mapping network for cross-modal matching of voices and faces//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Wuerkaixi A, Zhang Y, Duan Z Y and Zhang C S. 2022. Rethinking audio-visual synchronization for active speaker detection//Proceedings of the 32nd IEEE International Workshop on Machine Learning for Signal Processing. Xi'an, China: IEEE: 1-6 [DOI: 10.1109/MLSP55214.2022.9943352]

Xiong J W, Zhou Y, Zhang P, Xie L, Huang W and Zha Y. 2023. Look&listen: multi-modal correlation learning for active speaker detection and speech enhancement. IEEE Transactions on Multimedia, 25: 5800-5812 [DOI: 10.1109/TMM.2022.3199109]

Xu E Z, Song Z Y, Tsutsui S, Feng C, Ye M and Shou M Z. 2022. AVA-AVD: audio-visual speaker diarization in the wild//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 3838-3847 [DOI: 10.1145/3503161.3548027]

Yang Y, Shillingford B, Assael Y, Wang M S, Liu W D, Chen Y T, et al. 2020. Large-scale multilingual audio visual dubbing[EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2011.03530.pdf

Yin Y K, Yang X S, Liang L M, Li X and Zou Y X. 2025. Audio-faces intra-frame alignment with graph attention networks for active speaker detection//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10889368]

Yun H, Gao R H, Ananthabhotla I, Kumar A, Donley J, Li C, et al. 2024. Spherical world-locking for audio-visual localization in egocentric videos//Proceedings of the 18th European Conference on Computer Vision-ECCV 2024. Milan, Italy: Springer: 256-274 [DOI: 10.1007/978-3-031-72691-0_15]

Zhang Y H, Liang S S, Yang S, Liu X, Wu Z Q and Shan S G. 2021a. ICTCAS-UCAS-TAL submission to the AVA-ActiveSpeaker task at ActivityNet challenge 2021 [EB/OL].[2026-02-20]. http://research.google.com/ava/2021/S1_ICTCAS-UCAS-TAL.pdf

Zhang Y H, Liang S S, Yang S, Liu X, Wu Z Q, Shan S G, et al. 2021b. UniCon: unified context network for robust active speaker detection//Proceedings of the 29th ACM International Conference on Multimedia. [s.l.]: ACM: 3964-3972 [DOI: 10.1145/3474085.3475275]

Zhang Y H, Liang S S, Yang S and Shan S G. 2022. UniCon+: ICTCAS-UCAS submission to the AVA-ActiveSpeaker task at ActivityNet Challenge 2022 [EB/OL]. [2026-02-20]. https://arxiv.org/pdf/2206.10861.pdf

Zhang Y H, Xiao J Y, Yang S and Shan S. 2019. Multi-task learning for audio-visual active speaker detection [EB/OL]. [2026-02-20]. https://research.google.com/ava/2019/Multi_Task_Learning_for_Audio_Visual_Active_Speaker_Detection.pdf

Zhang Y L, Li Z Y, Wang D M, Zhang J H, Zhou D Y, Yin Z X, et al. 2026. SpeakerVid-5M: a large-scale high-quality dataset for audio-visual dyadic interactive human generation//Proceedings of the 14th International Conference on Learning Representations. Rio de Janeiro, Brazil: OpenReview.net

Zhao H, Zhang L, Li Y, Wang Y N, Wang H J, Rao W, et al. 2023. Joint training or not: an exploration of pre-trained speech models in audio-visual speaker diarization//Proceedings of the 18th National Conference on Man-Machine Speech Communication. Suzhou, China: Springer: 265-275 [DOI: 10.1007/978-981-97-0601-3_23]

Zhou H S. 2023. Application Research for Audio and Video Recognition Tasks. Hefei: University of Science and Technology of China (周恒顺. 2023. 面向音视频识别任务的应用研究. 合肥: 中国科学技术大学)

Zhou H S, Du J, Chen H, Jing Z J, Xiong S F and Lee C H. 2021. Audio-visual information fusion using cross-modal teacher-student learning for voice activity detection in realistic environments//Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021. Brno, Czechia: ISCA: 341-345 [DOI: 10.21437/INTERSPEECH.2021-592]

Zhu Z Y, He Q H, Feng X H, Ye W L, Li Y X and Yang J C. 2014. Lip motion and voice consistency algorithm based on fusing spatiotemporal correlation degree. Acta Electronica Sinica, 42(4): 779-785 (朱铮宇, 贺前华, 奉小慧, 叶婉玲, 李艳雄, 杨继臣. 2014. 基于时空相关度融合的语音唇动一致性检测算法. 电子学报, 42(4): 779-785) [DOI: 10.3969/j.issn.0372-2112.2014.04.024]

Zhu Z Y, Luo C, He Q H, Peng W F, Mao Z W and Zhang S S. 2023. Multi-view lip motion and voice consistency judgment based on lip reconstruction and three-dimensional coupled CNN. Journal of South China University of Technology (Natural Science Edition), 51(5): 70-77 (朱铮宇, 罗超, 贺前华, 彭炜锋, 毛志炜, 张顺四. 2023. 基于唇重构与三维耦合CNN的多视角音唇一致性判别. 华南理工大学学报(自然科学版), 51(5): 70-77) [DOI: 10.12141/j.issn.1000-565X.220435]

作者简介

张远航,男,博士研究生,主要研究方向为计算机视觉。
E-mail:zhangyuanhang15@mails.ucas.ac.cn
山世光,通信作者,男,研究员,主要研究方向为计算机视觉。
E-mail:sgshan@ict.ac.cn
杨双,女,副研究员,主要研究方向为计算机视觉。
E-mail:shuang.yang@ict.ac.cn