arXiv:2303.06326v1 [cs.MM] 11 Mar 2023

基于多模态信息的语音处理（MISP）2022挑战赛：视听说话人分离与识别

*Zhe Wang[1], Shilong Wu[1], Hang Chen[1],Mao-Kui He[1],Jun Du[1], *, Chin-Hui Lee[2], Jingdong Chen[6], Shinji Watanabe[3],Sabato Siniscalchi[2,4],Odette Scharenborg[7], Diyuan Liu[5], Baocai Yin[5],Jia Pan[5],Jianqing Gao[5],Cong Liu[5]*

[1] 中国科学技术大学，中国 [2] 佐治亚理工学院，美国[3] 卡内基梅隆大学，美国 [4] 恩纳科雷大学，意大利 [5] 科大讯飞，中国[6] 西北工业大学，中国 [7] 代尔夫特理工大学，荷兰

摘要

基于多模态信息的语音处理（MISP）挑战赛旨在通过推动对唤醒词、说话人分离、语音识别及其他技术的研究，拓展信号处理技术在特定场景中的应用。MISP2022挑战赛设有两条赛道：1）音视频说话人分离（AVSD），旨在利用音频和视频数据解决“谁在何时说话”的问题；2）一项新颖的视听说话人分离与识别（AVDR）任务，重点是结合视听说话人分离结果解决“谁在何时说了什么”的问题。两条赛道均聚焦中文，并采用真实家庭电视场景中的远场音频和视频：2至6人彼此交谈，背景中伴有电视噪声。本文介绍了MISP2022挑战赛的数据集、赛道设置和基线系统。实验分析和示例表明，AVDR 基线系统表现良好；同时也揭示了该挑战赛的潜在难点，例如远场视频质量、背景中的电视噪声以及难以区分的说话人。

索引词— MISP挑战赛、说话人分离、语音识别、多模态

1. 引言

现代语音交互系统在真实场景中（例如家庭和会议场景）仍会因不利声学环境及对话中的多说话人交互等因素而导致性能下降。受视觉线索能够帮助人类进行语音感知这一发现的启发 [1], , 许多研究者提出利用视觉模态来提升声学鲁棒性 [2, 3]。MISP2021 挑战赛 [4] 发布了一个大型远场多麦克风中文对话视听语料库，并提出了一些先进的音视频语音识别（AVSR）系统 [5, 6]。然而，这些系统假设语音片段与说话人之间的对应关系是预先已知的，这极大地限制了其在现实应用中的适用范围。在第二届MISP挑战赛中，我们针对家庭电视场景中的视听说话人分离（AVSD）以及视听说话人分离与识别（AVDR）问题展开研究。具体而言，AVDR是从AVSR扩展而来的任务，其将理想说话人分离结果替换为AVSD结果。

在纯音频条件下，针对说话人分离和语音识别，已有许多方法被提出。 [7]采用了x-vector [8], 凝聚层次聚类（AHC）和基于LSTM的重叠检测器来获得说话人分离结果，这些结果可用于引导式源分离（GSS）和深度神经网络-隐马尔可夫模型（DNN-HMM）。 [9] 提出了一种新颖的系统，该系统包括一个配备目标说话人语音活动检测（TS-VAD）的说话人分离模块，以及一个采用自注意力机制的语音识别模块。然而，由于潜在的强背景噪声和较高比例的重叠语音，真实场景中的纯音频说话人分离和语音识别任务仍然面临巨大挑战 [10]。

面部行为与语音活动高度相关 [11],，且视觉模态不会受到恶劣声学环境的干扰。研究人员对视听说话人分离（AVSD）和音视频语音识别（AVSR）表现出浓厚兴趣。针对AVSD系统，已有一些相关研究。 [12]利用互信息融合音频和视频模态，而 [13]则采用贝叶斯方法进行视听说话人分离。近年来，涌现出许多深度学习方法。 [14]采用了音视频同步模型， [15] 提出了一种基于自监督学习的说话人分离方法，并取得了积极成果。针对AVSR系统，[3] 在LRS数据集上提出了‘观看、聆听、关注并拼写’（WLAS）网络， [2] 则采用了基于Transformer的模型。 [16] 基于Conformer模块开发了CTC/注意力模型。带有门控层的DNN-HMM混合音视频语音识别系统 [17] 也表现良好。尽管AVSD和AVSR受到越来越多的关注，且已证明其性能显著优于传统的纯音频方法，但目前针对视听说话人分离与识别（AVDR）的研究仍然很少；这类研究聚焦于结合AVSD结果进行AVSR。

MISP2022挑战赛包含两个赛道：视听说话人分离（AVSD）和视听分离与识别（AVDR）。本文介绍了MISP2022挑战赛、相关数据和赛道，并详细描述了基线音视频分离与识别系统，随后进行了深入分析。此外，我们指出了参与者在该挑战赛中可能遇到的困难，包括远场视频质量低、家庭电视场景中的背景噪声，以及存在难以区分的说话人。据我们所知，我们提出了一项全新的AVDR任务，而所提出的AVDR基线系统是首个将AVSD和AVSR整合为一个大型系统的方案。该系统具有广阔的应用前景。更多挑战赛详情[1] 及基线代码[2] 可在相关网站上找到。

*通讯作者

[1]https://mispchallenge.github.io/mispchallenge2022[2] https://github.com/mispchallenge/misp2022 基线

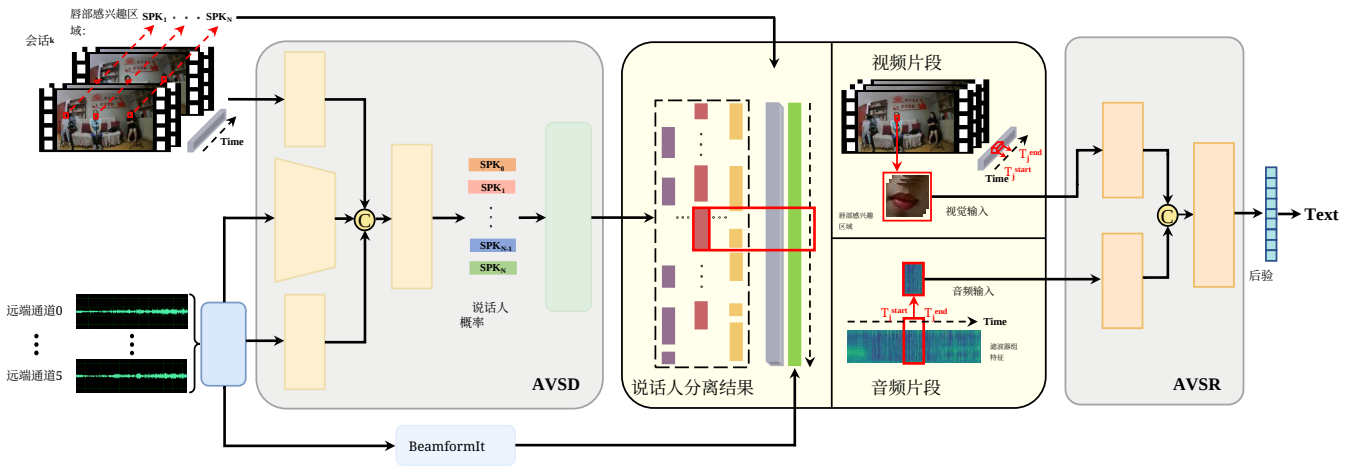


图1。视听说话人分离与识别基线系统的架构

2. 数据集与赛道

2.1. 训练集、开发集和评估集

我们采用了与MISP2021挑战赛更新版AVSR语料库相同的训练集 [18]，并从之前的开发集和评估集中选取了一个新的3小时开发集。新的开发集包含来自8个房间、26名参与者（10名男性和16名女性）的音频和视频录音。未来，为了消除各子集间说话人重叠的情况，我们将发布新的评测集，并将其用于最终排名。评测集仅包含来自远场设备的录音。

2.2. 赛道 1：视听说话人分离

视听说话人分离旨在利用音频和视频数据，将语音时间戳标注为对应说话人身份的类别，从而解决“谁在何时说话”的问题。评测时，仅提供远场音频和视频数据。我们将提供预言语音分段时间戳。参与者需为每个会话提交一个富转录时间标记（RTTM）文件。RTTM 文件是文本文件，每行包含一个轮次 [10]。开始时间（第 4 列）、时长（第 5 列）和说话人 ID（第 8 列）必须保留在相同的列中。

采用说话人分离错误率（DER） [19] 作为评测指标。DER 值越低（0 为满分），排名越高。值得注意的是，我们不设置“无评分”缓冲区，并且会对重叠语音进行评测。

$$\mathrm{DER} = \frac{\mathrm{FA} + \mathrm{MISS} + \mathrm{SPKERR}}{\mathrm{TOTAL}} \tag{1}$$

其中，FA、MISS 和 SPKERR 分别是误报、漏检和说话人错误的总时长，TOTAL 则是所有参考说话人语音时长之和。

在赛道 1 中，可以使用外部音频数据训练音视频说话人分离模型，例如 VoxCeleb 1, 2 [20, 21], CN-Celeb [22], 和其他公开数据集。也允许使用额外的视频数据。不过，参与者应提前将此类数据来源告知组织者，以便所有参赛者都了解这些数据，并有平等的机会使用它们。

2.3. 赛道 2：视听说话人分离与识别

赛道 2 不仅涵盖音视频说话人分离，还包括语音识别任务，即将语音转录为逐字文本。赛道 2 使用与赛道 1 相同的评测集。参与者需要提交 RTTM文件和转录文件。在每个会话中，参与者应按时间顺序合并同一说话人的所有话语，并提供转录文件。转录文件包含两列：话语 ID（第 1 列）和话语（第 2 列）。话语 ID 的格式为 < 说话人 ID > < 会话 ID >。

参照 [23]，中的拼接最小排列词错误率（cpWER），我们在赛道 2 中使用拼接最小排列字符错误率（cpCER）作为评测指标。cpCER 的计算分为三个步骤。首先，在一个会话中，沿时间线拼接属于同一说话人的识别结果和参考转录文本。其次，按如下方式计算说话人排列组合的字符错误率（CER）：

$$\mathrm{CER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \tag{2}$$

其中，S、D、I 分别表示替换错误、删除错误和插入错误的字符数。N 是字符总数。最后，选择最低的 CER 作为 cpCER。

在赛道 2 中，我们对附加数据的使用规则作出限制。允许使用外部音频数据和视频数据。值得注意的是，参与者可以使用时间戳、说话人标签以及除文本内容以外的其他信息。参与者还应提前将此类数据来源告知组织者。

3. 基线音视频分离与识别系统

图1展示了基线音视频分离与识别系统，该系统由一个音视频说话人分离模块和一个音视频语音识别模块依次组成。音视频说话人分离模块也作为赛道 1 的基线系统。本节将详细介绍音视频说话人分离模块和音视频语音识别模块的架构与训练过程，并说明如何将这两个模块结合起来进行解码。

3.1 音视频说话人分离模块的架构与训练

我们沿用先前工作 [24] 作为基线。不同之处在于，先前的工作使用了中场音频和视频数据，而当前的挑战聚焦于远场音频和视频。

如图1中的AVSD模块所示，我们的系统包含三个编码器模块。在视觉编码器模块中，唇部感兴趣区域用作

表 1. 开发集上的说话人分离结果（%）

系统	FA	MISS	说话人错误率	DER
ASD	0.01	19.88	11.36	31.25
VSD	6.64	8.17	3.89	18.69
AVSD	**4.01**	**5.86**	**3.22**	**13.09**

网络的输入，该网络由唇读模型 [25], Transformer模块 [26], 和一个双向长短期记忆层组成。整个网络可视为视觉语音活动检测（V-VAD）模型，用于生成视觉嵌入和初步说话人分离结果。接下来，我们使用经 NARA-WPE [27] 去混响处理的音频，以及 V-VAD 模型的说话人分离结果，计算 i-vector 作为说话人嵌入。此外，通过 FBank 特征提取器和若干二维卷积神经网络层，也可以提取音频嵌入。在解码器模块中，首先将三类嵌入结合起来，再利用若干带投影的双向长短期记忆（BLSTMP）层进一步提取特征，并获得每位说话人的语音或非语音概率。在后处理阶段，我们首先根据概率进行阈值处理，生成初步结果，并采用与 [28] 中相同的方法。此外，使用 DOVER-Lap [29] 融合六通道音频的结果。

训练过程如下：首先，我们使用预训练唇读模型的参数训练 V-VAD 模型。然后，冻结视觉网络参数，训练音频网络和解码器模块。最后，解冻视觉网络参数，联合训练整个网络。

3.2. 音视频语音识别模块的架构与训练

音视频语音识别模型采用深度神经网络-隐马尔可夫模型混合系统 [18]。首先，我们对远场 6 通道语音应用 NARA-WPE [27] 和 BeamformIt [30] 。然后，根据说话人分离结果，对从音频中提取的 FBank 特征和从视频中裁剪的唇部感兴趣区域进行分段。由三维卷积和 ResNet-18 组成的前端模块用于提取视频模态的唇部运动信息，并输出嵌入表示$_{\mathrm{V}}$。与此同时，由一维卷积和 ResNet-18 组成的前端模块用于提取音频特征并获得嵌入表示$_{\mathrm{A}}$。多阶段时间卷积网络（MS-TCN） [31] 模块提取视听特征嵌入表示$_{\mathrm{AV}}$。接下来，其他 MS-TCN 模块计算后验概率。最后，使用高斯混合模型-隐马尔可夫模型、三元语法模型和 DaCiDian 根据后验概率解码文本。

训练阶段使用理想说话人分离结果。使用 Kaldi [32] 在所有远场音频数据上训练 GMM-HMM 系统。基于DNN的声学模型采用交叉熵损失和 Adam优化器进行训练，共训练 100 轮次，初始学习率为 0.0003，并使用余弦调度器。有关该实验的更多详情，请参阅 [18]。

3.3. 联合解码

在推理过程中，音视频说话人分离模块的输出 RTTM文件包含会话、说话人、$\mathrm{T}^{\mathrm{start}}$ 和 $\mathrm{T}^{\mathrm{dur}}$ 的信息。这些信息可用于计算赛道 1 中的对话错误率，并对赛道 2 中的远场视频和远场六通道音频进行预处理。对于会话$_{\mathrm{k}}$，可获取一组话语标识符（说话人$_{\mathrm{i}}$、$\mathrm{T}_{\mathrm{j}}^{\mathrm{start}}$、$\mathrm{T}_{\mathrm{j}}^{\mathrm{dur}}$），其中会话$_{\mathrm{k}}$ 和说话人$_{\mathrm{i}}$ 分别表示第 k 个会话和该会话中的第 i 个说话人，$\mathrm{T}_{\mathrm{j}}^{\mathrm{start}}$和 $\mathrm{T}_{\mathrm{j}}^{\mathrm{dur}}$分别表示说话人$_{\mathrm{i}}$的第 j 个话语的开始时间和时长。对于会话$_{\mathrm{k}}$中的远场视频，我们首先根据 $\mathrm{T}_{\mathrm{j}}^{\mathrm{start}}$ 对整段视频进行分段

表 2. 开发集上的说话人分离和识别结果（%）

系统	S	D	I	cpCER
自动语音识别（理想说话人）	40.84	27.33	0.51	68.68
音视频语音识别(理想说话人)	35.78	27.82	0.36	63.96
音频说话人分离+自动语音识别	31.83	44.34	4.27	80.44
视觉说话人检测+自动语音识别	39.25	31.22	0.66	71.13
视觉说话人检测+音视频语音识别	35.17	31.01	0.61	66.79
音视频说话人分离+音视频语音识别	**35.94**	**29.45**	**0.68**	**66.07**

和 $\mathrm{T}_{\mathrm{j}}^{\mathrm{dur}}$，并在每一帧中裁剪说话人$_{\mathrm{i}}$ 的嘴唇区域，作为音视频语音识别模块的视觉输入。对于会话$_{\mathrm{k}}$中的远场六通道音频，我们首先对原始六通道音频进行加权预测误差处理和 BeamformIt 波束形成，再根据 $\mathrm{T}_{\mathrm{j}}^{\mathrm{start}}$ 和 $\mathrm{T}_{\mathrm{j}}^{\mathrm{dur}}$ 对整段波束形成音频进行分段，作为音视频语音识别模块的音频输入。最后，我们按照时间顺序拼接会话$_{\mathrm{k}}$ 中属于说话人$_{\mathrm{i}}$ 的每个话语的解码文本。

在评估过程中，由于置换不变训练（PIT）和标注片段文本对应关系的问题，我们采用 cpCER 作为最终评估指标。

4. 结果与分析

本节首先介绍基线系统的实验结果。接着，我们通过示例指出 MISP2022挑战赛中的难点，并分析 AVDR 系统表现出色的原因。挑战赛参与者可以利用这些信息，重点解决这些问题，从而使性能优于基线。

4.1. 基线结果

表 1 展示了纯音频说话人日记系统（ASD）、纯视觉说话人分离系统（VSD）和音视频说话人分离系统（AVSD）的虚警（FA）率、漏检（MISS）率、说话人错误（SPKERR）率和对话错误率（DER）；其中，AVSD 系统是 AVSD 赛道的基线系统。对于 ASD 系统，我们采用 VBx 方法 [33]。对于 VSD 系统，我们采用第 3.1 节所述视觉编码器模块的输出结果。ASD 系统的表现较差，这很可能是由于电视背景噪声较大且说话人重叠率较高，导致 MISS 和 SPKERR 率偏高。由于视觉模态不受声学环境干扰，VSD 系统在 MISS、SPKERR 和 DER 方面优于 ASD 系统。然而，VSD 系统的 FA 率较高，这可能是由于静音片段中的嘴唇运动所致。在 AVSD 系统中结合音频和视觉模态后，系统取得了最佳表现，这表明两种模态可以结合起来克服各自的弱点。

如表 2 所示，我们为说话人日记与识别系统设计了 6 项实验。前两项实验采用理想说话人（OS）说话人分离结果作为语音识别模块的输入。其余实验则将说话人分离模块与语音识别模块组合起来，例如 ASD+ASR、VSD+ASR、VSD+AVSR 和 AVSD+AVSR，其中最后一种是 AVDR 赛道的基线系统。对于 ASD+ASR 系统，较高的 MISS 和 SPKERR 比率导致目标说话人出现大量删除错误。此外，较高的 SPKERR 比率还会导致干扰说话人出现插入错误。比较 ASD+ASR 系统和 VSD+ASR 系统可知，说话人分离模块的视觉模态主导着整个说话人日记与识别系统的性能。与 VSD+ASR

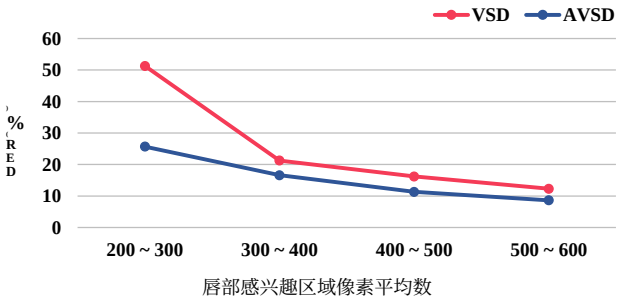


图 2. 对话中不同 LipROI 像素值下 VSD 与音视频说话人分离系统的对话错误率比较

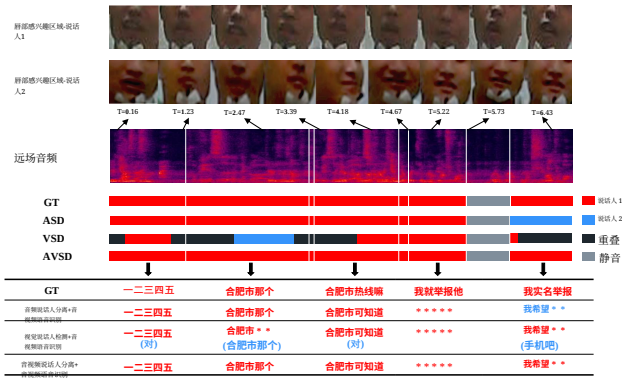


图 3。某个会话中的示例，不同系统结果对比

系统相比，VSD+AVSR 系统的语音识别模块中的视觉模态提供了可区分的信息，可减少替换错误，从而提升整个系统的性能。在所有实验中，只有在两个模块中都结合音频和视觉模态，才能取得最佳系统性能：AVDR。

4.2. 难点分析

为了帮助挑战赛参与者更好地解决问题，我们指出了本次挑战赛中可能存在的难点。同时，我们讨论了不同模块组合的性能，以进一步探究音频和视觉模态的影响。

4.2.1. 远场视频质量

由于摄像机与说话人之间距离较远，远场视频中每位说话人的唇部感兴趣区域在整幅图像中的占比会大幅降低，尤其是在说话人较多的场景中。同时，灯光、位置、角度、遮挡及其他环境因素也可能导致视频质量下降。我们探究了唇部感兴趣区域的像素数量对 VSD 和音视频说话人分离系统性能的影响，如图 2 所示。结果发现，平均像素数量减少时，对话错误率会急剧上升。这也会影响后续的语音识别任务。

从图 3 的示例可以看出，弱光和远距离会导致远场唇部感兴趣区域质量较低，从而造成唇动检测错误或遗漏。视觉说话人检测结果中存在大量重叠片段的误检和说话人混淆。事实上，根据真实标注（真实值），全程只有一位说话人（说话人 1）在讲话。对于使用视觉说话人检测结果的音视频语音识别模块，重叠片段的存在会导致说话人 2 出现更多插入错误，而说话人混淆则会导致删除错误

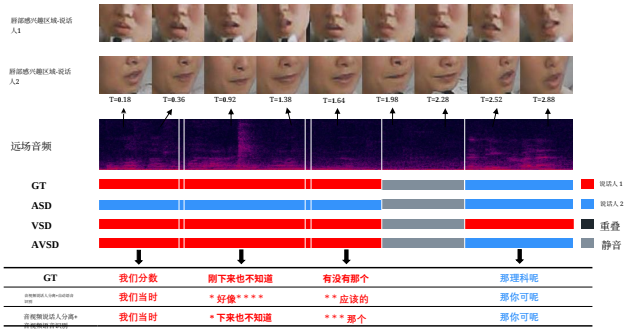


图4。会话中的另一个示例，比较了不同系统的结果

（说话人 1）。由于音视频说话人分离会结合音频模态信息来修正视频模态，因此其结果较视觉说话人检测显著提升，音视频说话人分离与识别系统的结果也有所改善。

4.2.2. 电视背景噪声

由于电视离远场麦克风阵列更近，较大的电视噪声可能会掩盖远场音频中目标说话人的声音。同时，由于电视广播内容丰富多样，音频中可能包含无关说话人的声音，从而干扰说话人分离和语音识别。如图3所示，在第四段话语中，由于电视上的演员大声说话，麦克风接收到的噪声完全掩盖了目标说话人的声音，使音视频说话人分离与识别系统无法识别语音内容。此外，在最后一段中，受电视背景噪声影响，音频说话人分离系统错误地将说话人1的片段分配给说话人2。尽管音视频说话人分离与识别系统的效果优于单模态系统，但电视背景噪声仍是MISP2022面临的一大挑战。

4.2.3. 难以区分的说话人

由于说话人具有多样性，同一会话中可能出现音色相似的说话人。如图4所示，音色相似会导致音频说话人分离结果中出现说话人混淆。此外，视频录音中偶尔会出现嘴唇蠕动，即嘴唇运动但没有话语的情况。模型难以判断说话人是在说话还是只是在动嘴唇。在视觉说话人检测过程中，嘴唇蠕动也会导致说话人混淆，使目标说话人的删除错误增多，干扰说话人的插入错误增多。然而，在音视频说话人分离过程中，通过音频与视觉模态之间的信息互补，我们获得了与真实标注一致的说话人分离结果，从而纠正了错误的说话人分离结果所导致的语音识别错误。

5. 结论

本文介绍了MISP2022挑战赛，这是首次提出视听说话人分离与识别（AVDR）任务的挑战赛。我们对该挑战赛进行了分析，包括基线结果、说话人分离模块与语音识别模块之间的关系，以及挑战赛面临的困难。我们相信，通过MISP数据集和MISP2022挑战赛，能够更好地推动视听说话人分离与识别领域的研究。

6. 致谢

本研究获得中国国家自然科学基金项目（项目编号62171427）的资助。

7. 参考文献

[1] Lawrence D Rosenblum，《作为多模态现象的语音感知》，《心理科学最新进展》，第17卷，第6期，第405–409页，2008年。[2] Triantafyllos Afouras、Joon Son Chung、Andrew Senior等，《深度视听语音识别》，*IEEE*模式分析与机器智能汇刊，2018年。[3] Joon Son Chung、Andrew Senior、Oriol Vinyals等，《真实场景中的句子唇读》，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第6447–6456页。[4] Hang Chen、Hengshun Zhou、Jun Du等，《首届基于多模态信息的语音处理（MISP）挑战赛：数据、任务、基线系统与结果》，载于*ICASSP*论文集。电气与电子工程师协会，2022年，第9266–9270页。[5] Gaopeng Xu、Song Yang、Wei Li等，《用于多通道视听语音识别的通道级视听融合注意力》，载于*ICASSP*论文集。电气与电子工程师协会，2022年，第9251–9255页。[6] Wei Wang、Xun Gong、Yifei Wu等，《上海交通大学在2021年基于多模态信息的语音处理挑战赛中的系统》，载于*ICASSP*论文集。电气与电子工程师协会，2022年，第9261–9265页。[7] Ashish Arora、Desh Raj、Aswin Shanmugam Subramanian等，《约翰斯·霍普金斯大学用于CHiME-6挑战赛的多麦克风多说话人自动语音识别系统》，载于第六届国际日常环境语音处理研讨会论文集，2020年，第48–54页。[8] David Snyder、Daniel Garcia-Romero、Gregory Sell等，《X向量：用于说话人识别的鲁棒深度神经网络嵌入》，载于*ICASSP*论文集。电气与电子工程师协会，2018年，第5329–5333页。[9] Ivan Medennikov、Maxim Korenevsky、Tatiana Prisyach等，《STC系统：CHiME-6挑战赛参赛系统》，载于第六届国际日常环境语音处理研讨会论文集，2020年，第36–41页。[10] Neville Ryant、Kenneth Church、Christopher Cieri等，《第三届DIHARD挑战赛评估方案》，*arXiv*预印本 *arXiv:2006.05815*，2020年。[11] Hani Yehia、Philip Rubin和Eric Vatikiotis-Bateson，《声道与面部行为的定量关联》，*Speech Communication*，第26卷，第1–2期，第23–43页，1998年。[12] Athanasios Noulas、Gwenn Englebienne和Ben JA Krose，《多模态说话人分离》，*IEEE*模式分析与机器智能汇刊，第34卷，第1期，第79–93页，2011年。[13] Israel D Gebru、Sileye Ba、Xiaofei Li等，《基于时空贝叶斯融合的视听说话人分离》，*IEEE*模式分析与机器智能汇刊，第40卷，第5期，第1086–1099页，2017年。[14] Rehan Ahmad、Syed Zubair、Hani Alquhayz等，《使用预训练视听同步模型的多模态说话人分离》，《传感器》，第19卷，第23期，第5163页，2019年。[15]Yifan Ding、Yong Xu、Shi-Xiong Zhang等，《用于视听说话人分离的自监督学习》，载于*ICASSP*论文集。电气与电子工程师协会，2020年，第4367–4371页。[16] Pingchuan Ma、Stavros Petridis和Maja Pantic，《使用Conformer的端到端视听语音识别》，载于*ICASSP*论文集。电气与电子工程师协会，2021年，第7613–7617页。

[17] Fei Tao 和 Carlos Busso，“用于大词汇量视听语音识别的门控神经网络”，*IEEE/ACM*音频、语音与语言处理汇刊，第 26 卷，第 7 期，第 1290–1302 页，2018 年。[18] Hang Chen、Jun Du、Yusheng Dai 等，“MISP2021 挑战赛中的视听语音识别：数据集发布与深度分析”，载于*Interspeech*论文集，2022 年。[19] Jonathan G Fiscus、Jerome Ajot、Martial Michel 等，“2006年春季会议丰富转录识别评测”，载于*MLMI*论文集。Springer，2006 年，第 309–322 页。[20] Arsha Nagrani、Joon Son Chung 和 Andrew Zisserman，“VoxCeleb：大规模说话人识别数据集”，载于*Interspeech*论文集，2017 年，第 2616–2620 页。[21] Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman，“VoxCeleb2：深度说话人识别”，载于 *Interspeech*论文集，2018 年，第 1086–1090 页。[22] Yue Fan、JW Kang、LT Li 等，“CN-Celeb：一个具有挑战性的中文说话人识别数据集”，载于*ICASSP*论文集。电气与电子工程师协会，2020 年，第 7604–7608 页。[23] Shinji Watanabe、Michael Mandel、Jon Barker 等，“CHiME-6 挑战赛：应对未分段录音的多说话人语音识别”，载于第六届日常环境语音处理国际研讨会论文集，2020 年，第 1–7 页。[24] Maokui He、Jun Du 和 Chin-Hui Lee，“端到端视听神经说话人分离”，载于 *Interspeech*论文集，2022 年，第 1461–1465 页。[25] Brais Martinez、Pingchuan Ma、Stavros Petridis 等，“使用时间卷积网络进行唇读”，载于*ICASSP*论文集。电气与电子工程师协会，2020 年，第 6319–6323 页。[26] Anmol Gulati、James Qin、Chung-Cheng Chiu 等，“Conformer：用于语音识别的卷积增强型Transformer”，载于*Interspeech*论文集，2020 年，第 5036–5040 页。[27] Lukas Drude、Jahn Heymann、Christoph Boeddeker 等，“NARA-WPE：一个用于 NumPy 和 TensorFlow 的 Python软件包，可实现在线与离线处理中的加权预测误差去混响”，载于*Speech Communication*；第*13*届*ITG*研讨会。VDE，2018 年，第 1–5 页。[28] Maokui He、Xiang Lv、Weilin Zhou 等，“USTC-Ximalaya系统：用于 ICASSP 2022 多通道多方会议转录（M2MET）挑战赛”，载于*ICASSP*论文集。电气与电子工程师协会，2022 年，第 9166–9170 页。[29] Desh Raj、Leibny Paola Garcia-Perera、Zili Huang 等，“DOVER-Lap：一种结合重叠感知说话人分离输出的方法”，载于 *SLT*论文集。电气与电子工程师协会，2021 年，第 881–888 页。[30]Xavier Anguera、Chuck Wooters 和 Javier Hernando，“用于会议说话人分离的声学波束形成”，*IEEE*音频、语音与语言处理汇刊，第 15 卷，第 7 期，第 2011–2022 页，2007 年。[31] Yazan Abu Farha 和 Jurgen Gall，“MS-TCN：用于动作分割的多阶段时间卷积网络”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019 年，第 3575–3584 页。[32] Daniel Povey、Arnab Ghoshal、Gilles Boulianne 等，“Kaldi语音识别工具包”，载于*ASRU*论文集。电气与电子工程师协会，2011 年。[33] Federico Landini、J´an Profant、Mireia Diez 等，“说话人分离中 x-vector 序列的贝叶斯隐马尔可夫模型聚类（VBx）：理论、实现及标准任务分析”，*Computer Speech & Language*，第 71 卷，文章编号 101254，2022 年。