arXiv:2303.02348v1 [cs.SD] 4 Mar 2023

2021 年 MISP 挑战赛 DKU 挑战赛后视听唤醒词检测系统：深度分析

Haoxu Wang[1,2], *Ming Cheng*[1,2],*Qiang Fu*[3],*Ming Li*[1, 2†]

[1]武汉大学计算机科学学院，中国武汉[2]数据科学研究中心，昆山杜克大学，中国昆山[3]阿里巴巴集团，中国

摘要

本文进一步探讨了我们此前提出的唤醒词检测系统，该系统在 MISP 挑战赛 2021 赛道 1 中排名第二。首先，我们研究了一种基于 3D 和二维卷积的鲁棒单模态方法，并在系统中采用简单注意力模块（SimAM）以提升性能。其次，我们探索了不同的数据增强方法组合，以获得更好的性能。最后，我们研究了多种融合策略，包括分数级融合、级联融合和神经融合。我们提出的多模态系统利用多模态特征，并运用互补的视觉信息，减轻复杂声学场景下纯音频系统的性能下降。在竞赛数据库的评估集上，我们的系统漏检率为 2.15%，误报率为 3.44%，相较于以往系统相对提升 21%，取得了新的最先进性能。相关资源见：https://github.com/Mashiro009/DKU_WWS_MISP。

索引词— 视听唤醒词检测、简易注意力模块、多模态系统、复杂声学场景

1. 引言

随着手机和语音助手等免提设备的快速普及，唤醒词检测（WWS）也越来越受到关注。唤醒词检测是关键词检出（KWS）的一种特定形式，旨在从流式音频中检测预先设定的一个或一组唤醒词。

近年来，已有许多基于深度神经网络的唤醒词检测研究，包括深度神经网络（DNN） [1], 卷积神经网络（CNN） [2], 时间卷积神经网络 [3] 和 Transformer [4]。这些研究在干净环境和近讲场景下表现良好。然而，研究发现，在复杂声学环境中，误报概率会升高，例如存在背景噪声（欢呼声、电视声或尖叫声）、混响以及多人对话交互，且语音重叠比例较高的情况。这会损害实际应用中的用户体验。

尽管视听语音识别（AVSR）领域已有大量多模态研究，[5, 6], 以往的工作主要关注纯音频 WWS 系统，忽略了视觉信息的应用。受人类感知的启发，互补的视觉信息（例如唇部运动）有助于在复杂声学场景中理解语音语义，因为视觉模态不受声学噪声影响。 [7] 在 AVSR 中采用 Conformer [8] 架构，而 [9] 则研究了一种音频增强模块

† 通讯作者，电子邮箱：ming.li369@dukekunshan.ed u.cn

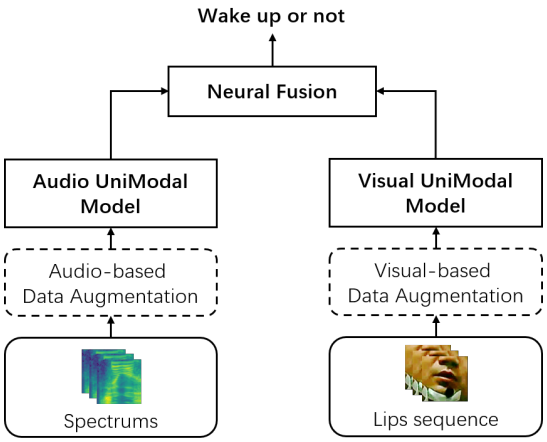


图 1. 我们的视听唤醒词检测系统框架。

称为 V-CAFE，用于 AVSR。在 [10], 中，作者研究了一种简单的基于CNN的音视频关键词检测系统，并证明其在嘈杂音频环境中表现良好。随着首届多模态信息语音处理挑战赛（MISP 挑战赛 2021 [11]）及其数据集的发布， [12], 研究者提出了一些针对视听唤醒词检测（AVWWS）的新方法，包括基于 CNN-3D 的模型 [13] 以及基于Transformer的端到端模型 [14]。

本文将我们先前的工作 [13] 扩展为一个高效的音视频唤醒词检测系统。我们通过将先前的三维卷积神经网络替换为二维与三维卷积神经网络的组合，设计了稳健的单模态系统。我们还将简单注意力模块（SimAM） [15] 应用于 WWS 系统，且不增加额外参数；该模块已用于自动说话人验证（ASV） [16, 17]。为进一步提升 WWS 系统的性能，我们开展实验，探索数据增强方法的更优组合。最后，我们研究了两种单模态系统的融合策略，在先前获得第二名的工作基础上，取得了 21% 的相对提升。此外，我们在 MISP 数据集上取得了最先进水平的成绩（WWS 得分为 5.59%）。

2. 系统描述

如图1所示，我们首先研究适用于单一模态的鲁棒模型，然后研究两种可靠的单模态系统的融合策略，以构建有效的视听 WWS 系统。我们还研究了数据增强策略，并找到了优于基线的组合。

2.1. 单模态模型

2.1.1. 3D-ResNet34

我们之前的工作 [13] 构建了一种类似 ResNet34 的神经网络，基于 3D CNNs（ResNet-3D）来处理单模态唤醒词检测任务。两种模态采用相同的网络架构，但输入维度不同。视频输入是形状为（T，H，W，3）的原始图像序列，其中 T 表示帧数，H，W 表示图像的高度和宽度，3 表示 RGB通道。对于音频输入，

我们首先从一维波形中提取二维声学特征 $X_{spec} \in \mathbb{R}^{T' \times D}$（例如 FBank、梅尔频率倒谱系数），然后使用形状为（D，D）的滑动窗口，以步长 $\frac{T'}{T}$ 沿时间轴对二维声学特征进行切片。由于音频和视频的帧率不同，我们将音频输入 X_{audio} 组织为形状为（T，D，D，1）的形式。因此，我们可以使用输入形状相同的单模态模型进行决策，其输入形状为（T，H，W，C）。

2.1.2. 2D-ResNet34

在以往的纯音频唤醒词检测任务中，[2, 3, 4], 二维频谱 X_{spec} 通常被视为单通道图像，我们也针对这种输入设计了 2D-ResNet34 网络。[18] 中的实验以及第 4.2 节表明，与 2D-ResNet 相比，3D-ResNet 更适合处理时空特征，而构建的 X_{audio} 包含局部短时空间信息，3D-ResNet 能够更好地利用这些信息。

2.1.3. 3D-ResNet18+2D-Resnet18

[18] 这表明，在动作识别任务中，低层网络层的三维卷积神经网络侧重于短期空间建模，而高层网络层的二维卷积神经网络有助于提取语义和时间信息。因此，我们将 3D-ResNet34 拆分为两个级联网络，即 3D-ResNet18 和 2D-ResNet18。将 3D CNN 转换为 2D CNN 的设计可以减少模型的参数数量。同时，由 3D CNN 和 2D CNN 混合构成的网络能够协同进行建模，使不同模块分别关注不同方面。该网络的输入与 3D-ResNet34 相同，但经过 3D-ResNet18 后，特征图会沿空间轴取平均，得到形状为（C，T）的特征图。然后，我们将其视为形状为（1，T，C）的单通道潜在图像，并将其输入 2D-ResNet18。

2.1.4. 3D-ResNet18+2D-Resnet18+SimAM

CNN 的注意力模块已广泛应用，并在计算机视觉和语音处理领域取得了出色的性能。此前的通道压缩激励（SE） [19] 可获取一维通道级权重，并已用于 WWS [20]。卷积块注意力模块（CBAM）分别计算二维和一维注意力权重 [21]。所有这些注意力模块都需要额外的参数。

[15] 基于神经科学中的空间抑制现象提出了 SimAM [22] 。三维权重用于估计各个神经元的重要性，而空间抑制表明，信息量最大的神经元通常表现出与周围神经元不同的放电模式。因此，三维特征图 $X \in \mathbb{R}^{C \times H \times W}$ 中每个目标神经元输出 t 的能量函数定义如下：

$$e_t(w_t,b_t,y,x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2 \quad (1)$$

其中，$\hat{t} = w_t t + b_t$ 和 $\hat{x}_i = w_t x_i + b_t$ 是 t 和 x_i 的线性变换，x_i 表示同一特征图中除 t 以外的神经元输出通道。i 是沿空间维度 $M = H \times W$ 的索引。当 $\hat{t} = y_t$ 和 $\hat{x}_i = y_o$ 时，能量函数取得最小值；由于空间抑制，y_t 应大于 y_o。为简单起见，[15] 使用二值标签设置 $y_t = 1$ 和 $y_o = -1$，并在公式 1 中加入一个正则化项：

$$e_t(w_t,b_t,y,x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t t + b_t))^2 + \lambda w_t^2 \quad (2)$$

求解上述函数的计算开销很大。幸运的是，根据 [15]，方程 2 关于 w_t 和 b_t 存在一个快速的闭式解。然后，将 w_t 和 b_t 代回方程 2，即可得到最终的最小能量：

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (3)$$

其中 $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$ 和 $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu}^2)$。根据空间抑制，使用 $\frac{1}{e_t^*}$ 可以获得每个神经元的重要性。最终的 SimAM 定义如下：

$$\widetilde{X} = \sigma(\frac{1}{E}) \bigotimes X \quad (4)$$

其中，$\bigotimes$ 表示逐元素乘法，E 将三维特征图 X 在通道和空间维度上的所有 e_t^* 分组，而 σ（·）表示 Sigmoid函数。SimAM 基于能量函数计算三维注意力权重，无需额外参数。此外，SimAM 很容易应用于 3D CNN 的四维特征图 $X_{4D} \in \mathbb{R}^{C \times T \times H \times W}$，我们沿时间和空间维度计算每个神经元的能量值，并设置 $M = T \times H \times W$。

2.2. 融合策略

2.2.1. 分数级融合

为了融合视听系统，官方基线仅通过对各个单模态系统的结果进行加权求和来做出最终决策。

$$P_{av} = \alpha \times P_a + \beta \times P_v \quad (5)$$

其中，P_a 和 P_v 分别是由纯音频系统和纯视频系统生成的后验概率。然而，纯音频系统和纯视频系统的后验概率分布差异很大。因此，我们需要探究一种更合适的融合策略。

2.2.2. 级联融合

由于这两个系统的性能存在差距，我们可以采用级联方案：先让性能较差的系统以较低的置信度阈值 th_l 检测唤醒词样本，再让性能较好的系统以较高的阈值 th_h 给出最终结果。这里，我们选择纯视频系统作为第一阶段，纯音频系统作为第二阶段。

2.2.3. 神经融合

从单模态系统中提取的潜在特征蕴含丰富的语义信息，可用于判断这些样本是否为唤醒词样本。因此，我们训练了一个简单的神经网络，进一步利用这些潜在特征。

我们在之前的工作 [13] 中采用了分层模态聚合（HMA）方法。我们对每个

不同层级的单模态神经网络中，分层残差 CNN 模块输出的潜在特征图进行全局平均池化。我们将两个模态在同一层级获得的嵌入向量拼接起来，得到多模态嵌入向量 $\mathbf{c}_l$，其中 l 表示层级。然后，我们使用 HMA 聚合所有$\mathbf{c}_l$ 并生成最终的多模态结果。

2.3. 数据增强

2.3.1. 音频流

对于音频流，我们根据 [9, 13, 11], 尝试了几种基于音频的数据增强方法，包括离线噪声/混响添加、波束形成增强、负子分段、速度扰动、音量扰动、轻微裁剪和 SpecAugment [23]。

- 离线噪声/混响添加（NR）：我们使用 pyroomacoustic 工具生成房间脉冲响应，并将其应用于原始近场数据，以模拟中远场数据；同时，将官方提供的噪声以 -15 至 15 dB 的随机信噪比混入数据，具体参照官方代码。
- 波束形成增强（BE）：我们对多通道音频信号实施波束成形（最小方差无失真响应 [24]），以充分发挥麦克风阵列的潜力，并将经过波束形成增强的音频纳入训练数据。
- 负子分段（NS）：由于正负样本的时长分布差异较大，模型可能仅根据音频长度检测唤醒词，但准确率较低。因此，我们在训练过程中参考 [25] 对负样本进行子分段。
- 音量扰动（VP）：随机选择每段音频并改变其音量。音量变化范围为 $[0.125,2]$。
- 速度扰动（SP）：随机加快或减慢音频的播放速度，比例范围为 $[0.9,1.1]$。
- 轻微裁剪（TS）：随机选择每段音频，在开头或结尾进行轻微裁剪，比例为 0.95。
- SpecAugment（SA）：我们对随机选取的每段音频进行频率掩蔽和时间掩蔽。

2.3.2. 视频流

对于视觉流，我们采用多种基于视频的数据增强方法，包括速度扰动、逐帧旋转、水平翻转、帧级裁剪、颜色抖动和灰度化。具体细节见 [13]。

3. 实验设置

3.1. 数据集和评估指标

我们使用 MISP 挑战赛 2021 提供的数据集。该挑战赛的目标是在远场智能家居场景中检测参与者说出的唤醒词“Xiao T, Xiao T”。

发布的数据集包含两个子集：训练集（47k+个负样本和 5K+ 个正样本）和开发集（dev：2k+ 个负样本和 600+ 个正样本），涵盖近场、中场和远场。此外，我们向竞赛参与者提供了一个没有标注的评估集（eval：8K+），该评估集仅包含远场数据。竞赛结束后，我们从 MISP 委员会获得标注，以便公平地将我们的结果与其他团队进行比较。

在 AVWWS 任务中，正类表示给定样本中存在唤醒词，负类则表示不存在唤醒词。根据评估方案的要求，我们使用漏检率（FRR）、误报率（FAR）和 WWS 得分作为评估指标。令 N_{wake} 表示包含唤醒词的样本数量，$N_{non\ wake}$ 表示不含唤醒词的样本数量。FRR 和 FAR 定义如下：

$$FRR=\frac{N_{FR}}{N_{wake}},\quad FAR=\frac{N_{FA}}{N_{non_wake}} \tag{6}$$

其中，N_{FR} 表示包含唤醒词但未被系统识别的样本数。N_{FA} 表示不包含唤醒词但被系统判定为正例的样本数。因此，唤醒词检测（WWS）的最终得分定义如下：

$$Score^{WWS}=FRR+FAR \tag{7}$$

3.2. 数据预处理

对于视觉流，我们仅使用视频中的 RGB 唇部区域图像。我们使用人脸检测器（RetinaFace [26]）从视频中获取人脸图像和面部关键点。随后，我们部署人脸识别器（ArcFace [27]），从远场视频中选出目标说话人。最后，根据检测到的面部关键点裁剪目标说话人的唇部区域。详情请参见[13]。每个提取出的唇部区域视频都会调整为 112×112 的分辨率，并包含 3 个 RGB 通道。维度 T 设为 64，这意味着每个视频采样后包含 64 帧。因此，视频样本的形状为 (64，112，112，3)。

对于音频流，我们使用 25ms 窗长和 10ms 移位从波形中提取 80 维对数梅尔滤波器组特征（FBank）。维度 T' 设置为 256，这意味着每段音频采样后包含 64 帧。因此，音频样本的形状变为 (256，80)。对于 3D-ResNet 的输入，我们使用形状为 (80，80) 的滑动窗口，以步长 4 对特征进行切片。因此，我们将音频样本组织为形状 (64，80，80，1)。

3.3. 模型详情

对于 SimAM，超参数 λ 设置为 0.001。对于分数级融合，α 和 β 设置为 0.5。在级联融合中，th_l 设置为 0.1，th_h 设置为 0.4。所有模型均在 1 个 GPU 3090 上训练，批量大小为 64。使用 Adam 优化器训练单模态模型时，学习率设置为 0.001；训练 HMA 融合模型时，学习率设置为 0.0001。此外，我们采用加权二元交叉熵（BCE）损失（负样本:正样本为 =5:1）来应对正负样本之间的不平衡问题。

4. 结果与讨论

4.1. 数据增强的消融实验

我们针对基于音频的数据增强进行了消融实验，以找出这些方法的最佳组合。表1报告了远场开发集和评估集上的详细结果。为方便起见，我们在此处的实验中使用2D-ResNet34。我们发现，不使用任何数据增强时，系统性能会下降。我们首先在纯音频系统中应用NR和BE，由于训练样本数量增加，性能有所提升。在此基础上，SP、TS和SA带来了更显著的提升。我们发现，VP和NS会使模型性能略有下降。不过，将NS与其他方法结合使用

表1。 基于纯音频*2D-ResNet34*的不同数据增强方法在远场数据上的性能。*NR*、*BE*、*NS*、*VP*、*SP*、*TS*和*SA*分别表示第*2.3.1*节中的离线噪声/混响添加、波束形成增强、负子分段、音量扰动、速度扰动、轻微裁剪和*SpecAugment*。

ID	DA						开发集[%]			评估集[%]		
	NR+BE	NS	VP	SP	TS	SA	FRR	FAR	WWS	FRR	FAR	WWS
D1	-	-	-	-	-	-	12.5	3.42	15.92	16.62	4.75	21.36
D2	✓	-	-	-	-	-	10.42	5.68	16.09	13.24	6.77	20.01
D3	✓	✓	-	-	-	-	9.78	5.44	15.21	12.08	8.12	20.2
D4	✓	✓	✓	-	-	-	8.33	6.35	14.68	13.12	7.42	20.54
D5	✓	✓	-	✓	-	-	7.53	4.62	12.15	7.66	8.59	16.25
D6	✓	✓	-	-	✓	-	8.65	4.71	13.37	10.79	6.35	17.14
D7	✓	✓	-	-	-	✓	5.45	7.12	12.57	7.05	10.15	17.2
D8	✓	✓	-	✓	✓	✓	5.45	4.76	10.21	5.64	6.05	**11.69**
D9	✓	-	-	✓	✓	✓	6.09	4.43	10.51	4.97	7.33	12.3

表2。 不同纯音频系统的比较。*3D-18+2D-18*表示*3D-ResNet18+2D-ResNet18*。

ID	模型	字段	开发集[%]			评估集[%]		
			FRR	FAR	WWS	FRR	FAR	WWS
A1	2D-ResNet34	Far	5.45	4.76	10.21	5.64	6.05	11.69
A2	3D-ResNet34	Far	5.77	4.28	10.05	5.46	5.88	11.34
A3	3D-18+2D-18	Far	6.76	2.98	9.71	7.05	4.19	11.24
A4	A3+SimAM	Far	5.93	3.61	9.54	6.38	4.65	**11.03**

表3. 不同纯视频系统的比较。*3D-18+2D-18*表示*3D-ResNet18+2D-ResNet18*。

ID	模型	字段	开发集[%]			评估集[%]		
			FRR	FAR	WWS	FRR	FAR	WWS
V1	3D-ResNet34	Far	8.81	8.03	16.85	18.52	9.03	27.54
V2	3D-18+2D-18	Far	9.78	6.64	16.41	14.41	9.62	24.03
V3	V2+预训练	Far	10.1	6.3	16.3	11.83	9.51	21.34
V4	V3+SimAM	Far	6.89	9.09	15.98	9.56	13.37	22.93
V5	V4+预训练	Far	9.13	6.25	15.39	8.03	11.1	**19.13**

有助于提升性能。最终，我们确定了数据增强方法组合，在评估集上取得了11.69%的WWS。

4.2. 不同系统的结果

表2展示了不同架构的纯音频系统的结果。由于能够对短时序特征进行建模，3D-ResNet34 的表现优于 2D-ResNet34。根据 3D 和 2D CNN 的混合，3D-ResNet18+2D-ResNet18 取得了更好的结果。最后，3D-ResNet18+2D-ResNet18+SimAM 在不增加额外参数的情况下取得了最佳结果（11.03% WWS）。

表3展示了不同架构的纯视频系统的结果。根据 3D 和 2D CNN 的混合，3D-ResNet18+2D-ResNet18 取得了更好的结果。3D-ResNet18+2D-ResNet18+SimAM 在不增加额外参数的情况下取得了（22.93% WWS）。此外，我们遵循 [13]，在名为 CAS-VSR-W1k 数据库的唇读数据库 [28]，上对骨干模型进行预训练，然后在 MISP 数据库上对其进行微调。最后，3D-ResNet18+2D-ResNet18+SimAM+预训练取得了最佳结果（19.13% WWS）。

视听融合系统的结果如表 4 所示。此外，我们还将本系统与此前最先进水平的结果进行了比较。我们结合各个模态中表现最佳的模型

表4。 不同视听系统的比较。

ID	模型	字段	开发集[%]			评估集[%]		
			FRR	FAR	WWS	FRR	FAR	WWS
VA1	官方 [12]	Far	7.3	6.8	14.1	10.1	15	25.1
VA2	Xu 等人 [14]	Far	-	-	-	-	-	9.1
VA3	Cheng 等人 [13]	Far	3.85	3.42	7.27	-	-	7.1
VA4	MISP 2021 第 1 名 [11]	Far	-	-	4.1	-	-	5.8
VA5	分数级	Far	1.92	3.37	5.29	1.54	4.82	6.36
VA6	级联融合	Far	5.29	2.3	7.59	7.29	3.5	10.79
VA7	HMA	Far	3.04	2.55	5.59	2.15	3.44	**5.59**

（A4+V5），并测试了多种融合策略。与分数级融合相比，级联系统效果不够理想，因为两个模态的分数分布不一致，筛选作用有限，且需要精细调整阈值。我们发现，分数级融合和级联融合未能充分利用视觉模态的信息。相比之下，HMA 可以进一步利用各个模态的特征，最终得到多模态结果。此外，我们对开发集上损失较低的排名前三的最佳模型进行平均，得到最终模型。最后，与此前的工作相比，我们的 WSS 相对降低了 21%，并取得了最先进的视听结果（5.59% WWS）。

5. 结论

本文在我们先前工作的基础上 [13]，进一步研究了鲁棒的视听 WWS 系统。我们采用混合 3D 和 2D 卷积网络，分别建模低层空间信息和高层语义信息，同时引入无需额外参数的 SimAM，进一步提升单模态网络的性能。我们还探索了不同的多模态融合方案，发现 HMA 能够充分利用两种模态的信息，取得最佳结果。与先前工作相比，系统性能显著提升，并在 MISP 数据集上取得了最先进水平的结果（WWS 为 5.59%）。

6. 致谢

本研究部分由中国国家自然科学基金（62171207）、广州市科技计划项目（202007030011）、苏州市科技计划项目（SYC2022051）和阿里巴巴天池基金资助。谨此感谢先进计算华东分中心提供的计算资源。

7. 参考文献

[1] Ming Sun、D. Snyder、Yixin Gao、Varun K. Nagaraja、Mike Rodehorst、S. Panchapagesan、N. Strom、Spyridon Matsoukas 和 Shiv Vitaladevuni，“ 用于小型化关键词检出的压缩时延神经网络 ”，载于 *Interspeech* 论文集，2017 年，第 3607–3611 页。

[2] Tara N Sainath 和 Carolina Parada，《用于小型关键词检出的卷积神经网络》，载于 *Interspeech* 会议论文集，2015 年。

[3] Seungwoo Choi、Seokjun Seo、Beomjun Shin、Hyeongmin Byun、Martin Kersner、Beomsu Kim、Dongyoung Kim 和 Sungjoo Ha，《用于移动设备实时关键词检出的时序卷积》，载于 *Interspeech* 会议论文集，2019 年，第 3372–3376 页。[4] Yiming Wang、Hang Lv、Daniel Povey、Lei Xie 和 San- jeev Khudanpur，《使用流式Transformer进行唤醒词检测》，载于*ICASSP* 会议论文集。2021 年，第 5864–5868 页，IEEE。[5] Triantafyllos Afouras、Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman，《深度视听语音识别》，*IEEE*模式分析与机器智能汇刊，2018 年。[6] Takaki Makino、Hank Liao、Yannis Assael、Brendan Shilling- ford、Basilio Garcia、Otavio Braga 和 Olivier Siohan，《用于视听语音识别的循环神经网络转导器》，载于*Proc. ASRU*。IEEE，2019 年，第 905–912 页。[7] Pingchuan Ma、Stavros Petridis 和 Maja Pantic，《使用Conformer实现端到端视听语音识别》，载于 *ICASSP* 会议论文集。IEEE，2021 年，第 7613–7617 页。[8] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、YuZhang、Jiahui Yu、Wei Han、ShiboWang、ZhengdongZhang、Yonghui Wu 等，《Conformer：用于语音识别的卷积增强型Transformer》，载于 *Interspeech* 会议论文集，2020 年，第 5036–5040 页。[9] Joanna Hong、Minsu Kim、Daehun Yoo 和 Yong Man Ro，《用于鲁棒端到端视听语音识别的视觉上下文驱动的音频特征增强》，载于 *Interspeech* 会议论文集，2022 年，第 2838–2842 页。

[10] Liliane Momeni、Triantafyllos Afouras、Themos Stafylakis、Samuel Albanie 和 Andrew Zisserman，《看见唤醒词：视听关键词检出》，载于*BMVC* 论文集，2020 年。

[11] Hang Chen、Hengshun Zhou、Jun Du、Chin-Hui Lee、Jingdong Chen、Shinji Watanabe、Sabato Marco Siniscalchi、Odette Scharenborg、Di-Yuan Liu、Bao-Cai Yin 等，《首届基于多模态信息的语音处理（MISP）挑战赛：数据、任务、基线与结果》，载于*ICASSP* 论文集。IEEE，2022 年，第 9266–9270 页。

[12] Hengshun Zhou、Jun Du、Gongzhen Zou、Zhaoxu Nian、Chin-Hui Lee、Sabato Marco Siniscalchi、Shinji Watanabe、Odette Scharenborg、Jingdong Chen、Shifu Xiong 和 Jian-Qing Gao，《MISP2021 挑战赛中的视听唤醒词检测：数据集发布与深度分析》，载于*Interspeech* 论文集，2022 年，第 1111–1115 页。

[13] Ming Cheng、Haoxu Wang、Yechen Wang 和 Ming Li，“面向 2021 年 MISP 挑战赛的 DKU 视听唤醒词检测系统”，载于*ICASSP* 论文集。IEEE，2022 年，第 9256–9260 页。

[14] Yanguang Xu、Jianwei Sun、Yang Han、Shuaijiang Zhao、Chaoyang Mei、Tingwei Guo、Shuran Zhou、Chuandong Xie、Wei Zou 和 Xiangang Li，“面向 MISP 挑战赛 2021 的视听唤醒词检测系统 ”，载于 *ICASSP* 论文集。IEEE，2022 年，第 9246–9250 页。[15] Lingxiao Yang、Ru-Yuan Zhang、Lida Li 和 Xiaohua Xie，“SimAM：一种用于卷积神经网络的简单无参数注意力模块”，载于国际机器学习会议。PMLR，2021 年，第 11863–11874 页。[16] Xiaoyi Qin、Na Li、Chao Weng、Dan Su 和 Ming Li，“基于简单注意力模块和迭代式噪声标签检测的说话人验证”，载于*ICASSP* 论文集。IEEE，2022 年，第 6722–6726 页。[17] Xiaoyi Qin、Danwei Cai 和 Ming Li，“不同域内数据可用性情形下的稳健多通道远场说话人验证”，*IEEE/ACM* 音频、语音与语言处理汇刊，2022 年。[18] Du Tran、Heng Wang、Lorenzo Torresani、Jamie Ray、Yann LeCun 和 Manohar Paluri，“深入探究用于动作识别的时空卷积”，载于*CVPR* 论文集，2018 年，第 6450–6459 页。[19] Jie Hu、Li Shen 和 Gang Sun，“压缩与激励网络”，载于*CVPR* 论文集，2018 年，第 7132–7141 页。[20] Menglong Xu 和 Xiao-Lei Zhang，“用于小型关键词检出的深度可分离卷积 ResNet 与压缩与激励模块”，载于*Interspeech* 论文集，2020 年，第 2547–2551 页。[21] Sanghyun Woo、Jongchan Park、Joon-Young Lee 和 In So Kweon，“CBAM：卷积块注意力模块”，载于*ECCV* 论文集，2018 年，第 3–19 页。[22] Ben S Webb、Neel T Dhruv、Samuel G Solomon、Chris Tailby 和 Peter Lennie，“猕猴纹状皮层周围抑制的早期与晚期机制”，神经科学杂志，第 25 卷，第 50 期，第 11666–11675 页，2005 年。[23] Daniel S. Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D. Cubuk 和 Quoc V. Le，“ SpecAugment：一种用于自动语音识别的简单数据增强方法”，载于*Interspeech 2019* 论文集，2019 年，第 2613–2617 页。、[24] Mehrez Souden、Jacob Benesty 和 SofiEne Affes，“关于用于降噪的最优频域多通道线性滤波”，*IEEE* 音频、语音与语言处理汇刊，第 18 卷，第 2 期，第 260–276 页，2010 年。[25] Yiming Wang、Hang Lv、Daniel Povey、Lei Xie 和 Sanjeev Khudanpur，“采用无对齐无格 MMI 的唤醒词检测”，载于 *Interspeech* 论文集，2020 年，第 4258–4262 页。[26] Jiankang Deng、J. Guo、Yuxiang Zhou、Jinke Yu、Irene Kotsia 和 Stefanos Zafeiriou，“RetinaFace：野外场景下的单阶段密集人脸定位”，*ArXiv*，第 abs/1905.00641 卷，2019 年。[27] Jiankang Deng、Jia Guo、Niannan Xue 和 Stefanos Zafeiriou，“ArcFace：用于深度人脸识别的加性角度间隔损失”，载于*CVPR* 论文集，2019 年，第 4690–4699 页。[28] Shuang Yang、Yuanhang Zhang、Dalu Feng、Mingmin Yang、Chenhao Wang、Jingyun Xiao、Keyu Long、Shiguang Shan 和 Xilin Chen，“LRW-1000：用于野外唇读的自然分布大规模基准”，载于*FG 2019* 论文集，2019 年，第 1–8 页。