MuAViC：用于鲁棒语音识别和鲁棒语音到文本翻译的多语言视听语料库

Mohamed Anwar、Bowen Shi、Vedanuj Goswami、Wei-Ning Hsu、Juan Pino、Changhan Wang

Meta AI{anwarvic,bshi,vedanuj,wnhsu,juancarabina,changhan}
@meta.com

arXiv:2303.00628v2 [cs.CL] 7 Mar 2023

摘要

我们介绍了 MuAViC，这是一个用于鲁棒语音识别和鲁棒语音到文本翻译的多语言视听语料库，包含 9 种语言、总计 1200 小时的视听语音数据。该语料库已全部转写，涵盖 6 个英语到其他语言的翻译方向以及 6 个其他语言到英语的翻译方向。据我们所知，这是首个用于视听语音到文本翻译的开放基准，也是规模最大的多语言视听语音识别开放基准。我们的基线结果表明，MuAViC 有助于构建抗噪语音识别和翻译模型。我们已在 https://github.com/facebookresearch/muavic 提供该语料库。

索引词：语料库、视听、多模态、多语言、语音识别、语音到文本翻译

1. 引言

唇部运动的视觉感知在辅助音频信号、更好地理解口语方面发挥着至关重要的作用 [1]。它不受声学噪声影响，因此在构建抗噪语音处理系统时极具价值。视听语音识别（AVSR）结合音频和视觉输入转写口语话语，已被证明能够提升语音识别的鲁棒性 [2, 3]。深度学习的应用大幅提升了视听语音识别系统的性能，广泛使用的 LRS3-TED 公共基准上的最先进水平（SOTA）模型便是明证。通常，在 0 dB 多人言语噪声等具有挑战性的条件下，SOTA 系统 [3] 能够将词错误率（WER）降低至仅音频系统的三分之一。

尽管前文提到的进展令人鼓舞，现代视听语音识别模型的评测仍主要基于单语言设置，其中大多数以英语为基础。非英语视听数据集的规模要小几个数量级；例如，常用数据集 CMU-MOSEAS [4] 为其涵盖的 4 种非英语语言各提供不到 20 小时的视听语音，规模至少比英语 LRS3-TED 语料库小 20 倍。缺乏大规模视听数据集也阻碍了视听语音到文本翻译（AVST）的发展。此前的研究仅使用合成数据对其进行模拟，但正如本文将展示的那样，预计它也能带来与视听语音识别相似的抗噪效果。随着自监督视听预训练方法日益普及 [5], , 对多语言视听语音识别和视听语音到文本翻译模型进行恰当的基准评测，将有助于开展多语言预训练，从而更有效地利用 AV-Speech [6] 等大规模多语言视听数据。

I本文介绍 MuAViC，这是一个多语言音频 o-

表 1：MuAViC 的时长和说话人统计。† 训练集样例由机器翻译模型 M2M-100 418M [7]进行人工标注（“H”）或伪标注（“P”）。

	小时			说话人		
	训练集（H+P）†	Dev	Test	训练集	Dev	Test
视听语音识别						
英语 [8, 3]	436 + 0	1.1	0.8	4.7K	0.9K	0.4K
Ar	16 + 0	1.5	1.2	95	7	7
De	10 + 0	1.6	1.5	53	9	9
El	25 + 0	2.3	2.0	113	10	8
Es	178 + 0	1.6	1.7	987	16	12
Fr	176 + 0	1.8	1.5	948	12	10
It	101 + 0	1.9	1.9	487	8	8
Pt	153 + 0	1.5	1.8	810	9	13
Ru	49 + 0	1.7	1.8	238	7	9
英语到多语言视听语音到文本翻译						
英语-希腊语	17 + 420	0.2	0.5	4.7K	0.9K	0.4K
英语-西班牙语	21 + 416	0.2	0.6			
英语-法语	21 + 416	0.2	0.6			
英语-意大利语	20 + 417	0.2	0.6			
英语-葡萄牙语	18 + 419	0.2	0.5			
英语-俄语	20 + 417	0.2	0.6			
X-英语视听语音转文本翻译						
希腊语-英语	8 + 17	2.3	2.0	113	10	8
西班牙语-英语	64 + 114	1.6	1.7	987	16	12
法语-英语	45 + 131	1.8	1.6	948	12	10
意大利语-英语	48 + 53	1.9	1.9	487	8	8
葡萄牙语-英语	53 + 100	1.5	1.8	810	9	13
俄语-英语	8 + 41	1.7	1.8	238	7	9

视觉语料库，旨在实现鲁棒语音识别和鲁棒语音转文本翻译。该语料库来源于 TED 和 TEDx 演讲，包含来自 9 种语言（英语（En）、阿拉伯语（Ar）、德语（De）、希腊语（El）、西班牙语（Es）、法语（Fr）、意大利语（It）、葡萄牙语（Pt）和俄语（Ru））的 8000 多位说话人共 1200 小时的转录音视频语音数据。这使 MuAViC 成为迄今为止规模最大的多语种音视频语音识别（AVSR）和唇读开放基准。此外，我们还提供文本翻译，并为 6 个英语到其他语言的翻译方向以及 6 个其他语言到英语的翻译方向建立了基线。据我们所知，MuAViC 是首个公开可用的音视频语音转文本翻译语料库。

2. 相关工作

视听语音识别 视听语音识别（AVSR）长期以来一直是语音

表2：干净环境下的英语语音识别和英到多语种语音翻译结果（A：音频，AV：音频+视频）。

ID	模型	Mode	目标语言							Avg
			En	El	Es	Fr	It	Pt	Ru	
	语音识别，测试词错误率 ↓									
A1	单语 AV-HuBERT [3]	A	2.5	-	-	-	-	-	-	-
		AV	2.3	-	-	-	-	-	-	-
	文本到文本翻译，测试集 BLEU 分数 ↑									
A2	双语 Transformer 基础版	-	-	25.8	29.5	27.0	22.6	23.9	17.2	24.3
A3	M2M-100 418M，多语言、无限制数据 [7]	-	-	24.5	28.7	25.6	21.8	22.2	15.8	23.1
	语音到文本翻译，测试集 BLEU 分数 ↑									
	A1+A2	A	-	25.1	28.9	26.7	21.9	23.8	16.9	23.9
		AV	-	25.5	29.1	26.5	22.3	23.6	16.8	24.0
	A1+A3	A	-	23.8	27.8	25.4	21.3	22.0	15.2	22.6
		AV	-	24.1	28.2	25.2	21.4	22.0	15.3	22.7
	双语 AV-HuBERT	A	-	23.0	27.5	25.1	20.7	20.1	14.7	21.9
		AV	-	23.4	26.6	25.3	20.7	20.5	14.6	21.9

领域的热门研究主题。早期的大多数视听语音识别数据集（如 GRID [9], ）都是在受控环境中采集的，要求受试者在类似实验室的环境中朗读预先设定的短语（例如，采用统一的照明条件）。这些数据集在视觉视角方面也存在局限，而且通常规模较小，词汇量有限，说话人数也不多。首个大规模“自然环境”视听语音识别数据集是 LRW [10], ，其中包含来自英国广播公司节目的 500 个孤立词的说话视频片段，共计 173 小时。近年来，采集句子级视听语音数据的趋势日益明显，例如 LRS3-TED [8, 2], ，该数据集更为自然，也更易获取。LRS3-TED 是目前规模最大的公开视听语音识别数据集，包含来自 5000 多位说话人的约 437 小时英语语音视频。

大多数广泛使用的视听语音识别数据集都是英语数据集，而非英语视听语音识别数据集 [4, 11, 12] 的规模则小得多。例如，规模最大的普通话视听语音识别数据集 CMLR [11], 仅包含 11 位说话人的 86 小时话语。当前最大的多语言视听语音识别数据集 CMU-MOSEAS [4], 共包含西班牙语、葡萄牙语、德语和法语的 68 小时语音。此外，该数据集缺乏跨语言翻译，因此难以用于视听语音到文本翻译（AVST）。

值得注意的是，存在大规模、无标注的多语言视听语音数据集，例如 VoxCeleb2 [13] （2442 小时）和 AV-Speech [6] （约 4700 小时）。它们已广泛用于各种非转录任务，例如说话人验证 [13, 14] 和源分离 [6],，也可作为自监督音视频表征学习的预训练数据来源 [5, 15, 16]。然而，由于缺乏转录文本，这些数据集无法直接用于视听语音识别和视听语音到文本翻译。

语音到文本翻译 近年来，随着开放语音翻译基准的出现，语音到文本翻译（ST）或口语翻译（SLT）已成为语音研究中日益热门的课题。几乎所有语音翻译语料库都是通过为现有语音识别语料库中的转录文本添加译文而创建的，而语音数据则收集自各种来源，例如电话录音 [17],朗读语音录音 [18, 19, 20]和公开演讲录音 [21, 22, 23, 24]。尽管这些语料库的原始数据来源偶尔包含视频轨道，但其创建者只关注音频模态，并且语料库中仅纳入了音频数据。由于缺乏带标注的音视频语料库，视听语音到文本翻译仍是一个鲜少探索的研究领域。

3. 语料库构建

MuAViC 的数据来自 TED 和 TEDx 演讲录制，其中母语或非母语说话人（大多数情况下只有一位说话人）在舞台上进行公开演讲，摄像机则从不同视角切换拍摄舞台场景。我们从录制内容中收集音频和视频轨道，并将其与人工转写及文本翻译对齐。

对于英语演讲，我们重复使用来自 LRS3-TED [8] 的视听数据，并沿用原始数据划分。我们从机器翻译语料 TED2020 [26] 中寻找这些演讲的人工翻译，方法是匹配 LRS3-TED 中的转写与 TED2020 中的源句子。随后，将匹配到的 LRS3-TED 示例与 TED2020 中对应的目标句子配对，作为翻译标签。对于开发集和测试集示例，我们采用精确文本匹配，以确保最高准确率。为了提高训练集的匹配召回率，我们设计了一种模糊文本匹配策略：首先，如果句对两侧包含相同数量的片段，就按标点符号对 TED2020 的源句子和目标句子进行分段。然后，通过移除标点符号并转为小写，对 TED2020 和 LRS3-TED 的文本进行规范化。最后，在两个语料库之间进行精确文本匹配。对于在 TED2020 中找不到匹配项的 LRS3-TED 训练集示例，我们使用机器翻译模型 M 2 100 *418* [7] M-M 的默认解码超参数获取伪翻译标签。

对于非英语演讲，我们复用 mTEDx [24]收集的纯音频数据、转录文本和文本翻译。我们的数据划分也遵循 mTEDx。我们获取原始录音的视频轨道，并将处理后的视频数据与音频数据对齐，形成视听数据，与 LRS3-TED [8]类似。虽然 mTEDx 中的所有音频数据都有转录文本，但只有一部分有翻译。对于训练集中未翻译的样本，我们使用默认解码超参数，从 M2M-100 418M [7] 获取伪翻译标签。

在表 1 中，我们给出了 MuAViC 各数据划分的时长和说话人统计信息。MuAViC 涵盖 9 种语言，提供总计 1200 小时的转录视听数据，来自

表3：干净环境下非英语语音识别和多语言到英语语音到文本翻译的结果（A：音频，AV：音频+视频）

ID	模型	Mode	源语言								Avg
			Ar	De	El	Es	Fr	It	Pt	Ru	
	语音识别，测试词错误率 ↓										
	Transformer，单语 [24]	A	104.4	111.1	109.5	46.4	45.6	48.0	54.8	74.7	74.3
	混合 LF-MMI，单语 [24]	A	80.8	42.3	25.0	16.2	19.4	16.4	20.2	28.4	31.1
	Whisper，多语言，无限制数据 [25]	A	91.5	24.8	25.4	12.0	12.7	13.0	15.5	31.1	28.2
B1	单语 AV-HuBERT	A	99.3	61.1	35.1	16.5	24.4	19.3	23.0	33.3	39.0
		AV	98.5	52.4	26.3	15.9	23.7	18.5	19.4	30.0	35.6
B2	多语种 AV-HuBERT	A	67.7	46.5	40.4	30.6	27.0	19.3	19.8	37.6	36.1
		AV	69.3	47.2	41.2	16.2	19.0	19.8	19.9	38.0	33.8
	文本到文本翻译，测试集 BLEU 分数 ↑										
B3	双语 Transformer 基础版	-	-	-	13.9	28.2	33.8	27.2	31.9	14.2	27.0
B4	M2M-100 418M，多语言，无限制数据 [7]	-	-	-	25.9	29.5	34.9	30.7	33.6	19.1	29.0
	语音到文本翻译，测试集 BLEU 分数 ↑										
	Whisper，多语言，无限制数据 [25]	A	-	-	24.2	28.9	34.5	29.2	32.6	16.1	29.9
	B1+B3	A	-	-	10.7	21.2	24.2	20.3	23.5	10.7	18.4
		AV			11.3	21.1	24.3	20.8	24.9	11.0	18.9
	B2+B3	A	-	-	7.6	9.2	12.8	19.3	23.1	8.7	13.4
		AV			7.5	20.2	24.4	18.6	23.2	8.6	17.1
	B1+B4	A	-	-	15.9	23.1	25.7	23.4	24.2	13.6	21.0
		AV			18.4	23.0	26.1	23.8	26.3	13.9	21.9
	双语 AV-HuBERT	A	-	-	7.3	20.4	26.4	19.5	24.2	9.2	17.8
		AV			7.2	20.2	25.0	19.4	24.1	9.0	17.5
	多语种 AV-HuBERT	A	-	-	9.3	21.0	26.3	21.2	24.3	9.3	18.6
		AV			7.6	20.5	25.2	20.0	24.0	8.1	17.6

8000 多位说话人（说话人数根据 TED/TEDx 演讲 ID 估算）。此外，MuAViC 还提供 9 种语言中 6 种语言与英语之间的文本翻译。所有视听语音识别和视听语音到文本翻译测试集均经过人工标注，每个测试集包含至少 7 位说话人、时长不少于 0.5 小时的视听语音。

4. 实验

4.1. 实验设置

对于视听语音识别和视听语音到文本翻译，我们均使用英语 AV-HuBERT 大型预训练模型 [3] [1], , 该模型在 LRS3-TED [8] 与 VoxCeleb2 [27]的英语部分的组合数据上进行训练。我们遵循 [3] 中的微调超参数设置，但将双语模型微调至 30K 次更新，将多语种音视频语音识别模型微调至 90K 次更新。对于 X-En 音视频语音翻译模型和 En-X 音视频语音翻译模型，我们分别在前 4K 次和 24K 次更新期间冻结预训练编码器。此外，我们随机对 25% 的输入样本添加多种加性噪声进行增强，信噪比为 0。类别为“自然音”“音乐”和“多人嘈杂语音”的噪声音频片段采样自 MUSAN 数据集 [28], ，而重叠“语音”噪声样本则取自 LRS3-TED。在创建“语音”和“多人嘈杂语音”噪声集时，我们确保不同数据划分之间不存在说话人重叠。为提高训练稳定性，我们移除了过短的话语（少于 0.2 秒）和过长的话语（超过 20 秒）。除端到端音视频语音翻译模型外，我们还构建了由视听语音识别模型和基于文本的机器翻译模型（机器翻译模型）组成的级联系统。对于双语机器翻译，我们使用 MuAViC 训练集中的所有转录-翻译对训练 Transformer 基础版 [29] 模型。对于多语言机器翻译，我们利用 M2M-100 418M [7], , 该模型在大规模开放领域挖掘双语语料上进行训练。对于语音识别和 X-En 语音翻译，我们还将 Whisper Large V2 [25] 作为最先进基线进行评估。

推理时，对于视听语音识别/视听语音到文本翻译，我们使用验证准确率最高的检查点；对于机器翻译，则使用验证 BLEU 最高的检查点。束宽设为 5，其他束搜索解码超参数均采用默认值。对于视听语音识别，我们在计算词错误率（WER）前，通过去除标点符号并转换为小写来规范化文本 [30]。对于视听语音到文本翻译和机器翻译，我们使用默认选项的 SacreBLEU [31] ；计算 BLEU [32] 之前，文本会先经过其内置的 13a 分词器处理。

4.2. 视听语音识别（AVSR）

4.2.1. 干净设置

我们在仅音频（“A”）和音视频（“AV”）两种模式下评估视听语音识别模型：前者在微调和推理时仅利用音频模态，后者则同时利用音频和视觉模态。如表2所示，英语视听语音识别模型 [3] 在仅音频和音视频模式下的测试词错误率均较低，分别为 2.5 和 2.3。对于非英语视听语音识别，我们将预训练的英语 AV-HuBERT 模型分别在每种语言上单独微调（得到 8 个单语模型），或在全部 8 种非英语语言上联合微调（得到一个多语模型），其测试词错误率见表3。我们发现，我们的单语视听语音识别模型在音视频

[1]随机初始化的模型表现不佳（见附录 A .1).

表 4：在多人嘈杂语音噪声环境下进行语音识别和语音翻译的结果（信噪比=*0*。A：音频，音视频：音频+视频）。

模型	Mode	源语言/目标语言									Avg
		En	Ar	De	El	Es	Fr	It	Pt	Ru	
		语音识别，测试词错误率 ↓									
Whisper，多语言，无限制数据 [25]	A	202.4	197.9	244.4	113.3	116.3	172.3	172.4	223.6	126.2	174.3
单语 AV-HuBERT	A	25.9	102.3	83.5	74.6	65.1	63.6	71.0	77.3	68.1	70.2
	AV	5.7	101.0	74.6	53.1	45.1	48.1	51.6	47.2	51.9	53.2
多语种 AV-HuBERT	A	-	86.1	76.0	73.7	59.7	54.7	61.0	61.9	61.9	66.9
	AV	-	82.2	66.9	62.2	40.7	39.0	44.3	43.1	43.1	52.7
		英语到其他语言的语音翻译，测试集 BLEU 分数 ↑									
双语 AV-HuBERT	A	-	-	-	15.9	19.2	17.1	12.9	14.4	10.3	15.0
	AV	-	-	-	22.7	24.8	23.8	20.0	20.0	13.7	20.8
		X-En 语音到文本翻译，测试集 BLEU 分数 ↑									
Whisper，多语言，无限制数据 [25]	A	-	-	-	0.1	0.4	0.7	0.1	0.1	0.2	0.3
双语 AV-HuBERT	A	-	-	-	3.9	8.6	13.1	7.7	9.5	5.7	8.1
	AV	-	-	-	4.2	12.4	16.5	12.4	15.4	6.1	11.2
多语种 AV-HuBERT	A	-	-	-	2.9	8.4	12.4	8.1	8.6	0.9	6.9
	AV	-	-	-	4.2	12.8	15.0	12.5	14.8	4.6	10.7

模式下的表现优于可比的自动语音识别基线模型（Transformer，单语），词错误率平均降低 52%。但其表现不及混合 LF-MMI 单语自动语音识别基线模型，词错误率平均高出 14%；这些基线模型在解码时使用 4-gram 语言模型，而我们的模型没有使用。我们的多语种音视频语音识别模型整体略优于单语视听语音识别模型：在一些低资源语言（阿拉伯语和德语）上有所提升，而在其他语言（希腊语和俄语）上有所下降。我们还发现，我们的多语种音视频语音识别模型平均比 Whisper 落后 20% 的词错误率，这归因于训练数据量相差悬殊（0.7K 小时，而 Whisper 使用 680K 小时）。

4.2.2. 噪声设置

表 4 的第一部分展示了我们的音视频语音识别模型在噪声设置下的测试词错误率。在该设置中，我们通过向干净语音输入添加多语种人声嘈杂噪声，并将信噪比（信噪比）设为 0 [2]，来模拟噪声环境。我们观察到，作为最先进的多语言自动语音识别模型，Whis- per 在这一具有挑战性的设置下表现极其糟糕，其在 9 种语言上的平均词错误率高达 174.3。相比之下，我们的单语音视频语音识别模型在纯音频模式下的平均词错误率分别为 70.2 和 66.7。在音视频模式下，我们模型的平均词错误率显著降低了 32%，这表明它们能够有效利用视觉信息来减轻噪声环境的干扰。我们的多语种音视频语音识别模型在纯音频和音视频两种模式下，在每一种非英语语言（希腊语除外）上的表现都优于单语模型。

4.3. 音视频语音到文本翻译（AVST）

4.3.1. 干净设置

表2和表3分别报告了 En-X 音视频语音翻译和 X-En 音视频语音翻译模型的测试集 BLEU 分数。除了端到端 AV-HuBERT 视听语音到文本翻译模型外，我们还设置了双语和多语言机器翻译基线，并通过将 AV-HuBERT 视听语音识别模型与机器翻译模型串联，构建级联式音视频语音转文本翻译基线。在受限数据设置（“A1+A3”和“B1+B3”）下，无论是 En-X 还是 X-En 方向，我们的端到端音视频语音翻译模型都与级联模型表现相当。与非英语视听语音识别的情况类似，我们的六语言多语种多语言到英语音视频语音翻译模型平均表现略优于相应的双语音视频语音翻译模型。

4.3.2. 噪声设置

我们在噪声设置下评估 En-X 音视频语音翻译和 X-En 音视频语音翻译模型，其测试集 BLEU 分数分别见表 4 的第二和第三部分。我们采用与视听语音识别模型相同的方法模拟噪声环境：向干净语音输入中添加信噪比为 0 的多语言多人言语噪声。我们发现，Whisper 是一个最先进的多语言 X-En 语音到文本翻译模型，但在此设置下性能骤降，在 6 个方向上的平均 BLEU 分数仅为 0.3。我们的双语和多语言视听语音到文本翻译模型在纯音频模式下表现远胜于它，平均 BLEU 分数分别提高了 7.8 和 6.6。在音视频模式下，视听语音到文本翻译模型始终优于纯音频模式下的模型；在双语和多语言设置中，平均 BLEU 分数分别提高了 3.2 和 3.3。

5. 结论

我们介绍了一个涵盖 9 种语言、共计 1200 小时语音的多语言视听语料库 MuAViC。它是首个用于视听语音到文本翻译的开放基准，也是规模最大的多语种音视频语音识别开放基准。该语料库可在 https://github.com/facebookresearch/muavic 获取。

6. 参考文献

[1] W. H. Sumby 和 I. Pollack，《噪声中视觉对语音可懂度的贡献》，《美国声学学会期刊》，第 26 卷，第 212–215 页，1954 年。

[2] T. Afouras、J. S. Chung、A. W. Senior、O. Vinyals 和 A. Z 可懂-

[2]不同噪声类型和不同信噪比下的模型性能，详见附录 A.2

man，《深度音视频语音识别》，IEEE模式分析与机器智能汇刊，第 44 卷，第 8717–8727 页，2018 年。

[3] B. Shi, W.-N. Hsu, 和 A. Mohamed, “监督视听语音识别”，载于《第23届国际语音通信协会年会论文集》，2022 年，第 2118–2122 页。

[4] A. Bagher Zadeh、Y. Cao、S. Hessner、P. P. Liang、S. Poria 和 L.-P. Morency，《CMU-MOSEAS：面向西班牙语、葡萄牙语、德语和法语的多模态语言数据集》，载于《2020年自然语言处理实证方法会议论文集》（EMNLP），2020 年 11 月，第 1801–1812 页。

[5] B. Shi, W.-N. Hsu, K. Lakhotia, 和 A. Mohamed, “通过掩码多模态聚类预测学习视听语音表征”，载于《国际学习表征会议》，2021 年。

[6] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, 和 M. Rubinstein，《在鸡尾酒会中寻找聆听目标》，ACM图形学汇刊（TOG），第37卷，第1–11页，2018年。

[7] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary 等，“超越以英语为中心的多语种机器翻译”，《机器学习研究杂志》，第 22 卷，页码 1–48，2021 年。[8] T. Afouras、J. S. Chung 和 A. Zisserman，“LRS3-TED：用于视觉语音识别的大规模数据集”，arXiv 预印本 arXiv:1809.00496，2018 年。

[9] M. Cooke、J. Barker、S. P. Cunningham 和 X. Shao，“用于语音感知和自动语音识别的音频-视觉语料库。”《美国声学学会杂志》，第 120 卷第 5 期，2421–2424 页，2006 年。

[10] J. S. Chung 和 A. Zisserman，《自然场景唇读》，发表于 2016 年亚洲计算机视觉会议。

[11] Y. Zhao, R. Xu, 和 M. Song，“一种级联序列-to-汉语普通话唇读的序列模型，”ACM Multimedia Asia会议论文集，2019年。

[12] D. Ivanko, A. Axyonov, D. Ryumin, A. Kashevnik 和 A. Karpov，《RUSAVIC 语料库：汽车中的俄语视听语音》，载于第十三届语言资源与评估会议论文集，2022年6月，第1555–1559页。

[13] J. S. Chung, A. Nagrani, 和 A. Zisserman，《Voxceleb2：深度说话人识别》，载于Interspeech，2018。

[14] J. S. Chung, S. Yuh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B. J. Lee, 和 I. Han，《关于说话人识别中度量学习的辩护”，ArXiv，第 abs/2003.11982 卷，2020。

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, 和 A. Mohamed，《Hubert：通过掩码预测隐藏单元进行自监督语音表征学习》，arXiv 预印本 arXiv:2106.07447，2021 年。

[16] T. Afouras、J. S. Chung 和 A. Zisserman，“自动语音识别就是你所需要的一切：用于唇读的跨模态蒸馏”，ICASSP 2020 - 2020 IEEE国际声学、语音与信号处理会议（ICASSP），第 2143–2147 页，2019。

[17] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch 和 S. Khudanpur，“借助 Fisher 和 Callhome 西班牙语-英语语音翻译语料库改进语音到文本翻译”，载于《第十届口语翻译国际研讨会论文集》，2013年12月5日至6日。

[18] A. C. Kocabiyikoglu、L. Besacier 和 O. Kraif，《通过法语翻译扩充 LibriSpeech：用于直接语音翻译评估的多模态语料库》，载于《第十一届语言资源与评测会议论文集》（LREC 2018），2018年5月。[19] C. Wang、J. Pino、A. Wu 和 J. Gu，《CoVoST：一个多样化的多语言语音到文本翻译语料库》，载于《第十二届语言资源与评测会议论文集》，2020年5月，第4197–4203页。

[20] C. Wang、A. Wu、J. Gu 和 J. Pino，《CoVoST 2 与大规模多语言语音翻译》，载于 Interspeech 2021 论文集，2021年，第2247–2251页。[21] M. A. Di Gangi、R. Cattoni、L. Bentivogli、M. Negri 和 M. Turchi，《MuST-C：一个多语言语音翻译语料库》，载于《2019年北美计算语言学协会分会会议：人类语言技术》，2019年，第2012–2017页。[22] J. Iranzo-Sánchez、J. A. Silvestre-Cerda、J. Jorge、N. Rosell´o、A. Gim´enez、A. Sanchis、J. Civera 和 A. Juan，《Europarl-ST：用于议会辩论语音翻译的多语言语料库》，载于《ICASSP 2020-2020 IEEE国际声学、语音与信号处理会议论文集》（ICASSP）。IEEE，2020年，第8229–8233页。[23] R. Cattoni、M. A. Di Gangi、L. Bentivogli、M. Negri 和 M. Turchi，《MuST-C：一个用于端到端语音翻译的多语言语料库》，《计算机语音与语言》，第66卷，101155，2021 p. 。[24] E. Salesky、M. Wiesner、J. Bremerman、R. Cattoni、M. Negri、M. Turchi、D. W. Oard 和 M. Post，《用于语音识别与翻译的多语种TEDx语料库》，载于 Interspeech 论文集，2021年。

[25] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，《通过大规模弱监督实现鲁棒语音识别》，arXiv 预印本 arXiv:2212.04356，2022。

[26] N. Reimers 和 I. Gurevych，《构建单语句子嵌入-通过知识蒸馏实现多语言嵌入”，载于《2020年自然语言处理实证方法会议论文集》，2020年11月。

[27] J. S. Chung, A. Nagrani, 和 A. Zisserman，“Voxceleb2：深度说话人识别”，《Interspeech 2018 论文集》，1090，2018，1086页。

[28] D. Snyder、G. Chen、以及 D. Povey，《Musan：一个音乐、语音和噪声语料库》，arXiv 预印本 arXiv:1510.08484，2015 年。

[29] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，《Attention Is All You Need》，《神经信息处理系统进展》，第30卷，2017年。[30] C. Wang、Y. Tang、X. Ma、A. Wu、D. Okhonko 和 J. Pino，《Fairseq S2T：使用 fairseq 进行快速语音转文本建模》，载于《计算语言学协会亚太分会第一届会议暨自然语言处理第十届国际联合会议论文集：系统演示》，2020年12月，第33–39页。[31] M. Post，《呼吁明确报告 BLEU 分数》，载于《机器翻译第三届会议论文集：研究论文》，2018年10月，第186–191页。[32] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，《Bleu：一种机器翻译自动评估方法》，载于《计算语言学协会第40届年会论文集》，2002年，第311–318页。

A. 附录

A.1. 未经预训练的 AV-HuBERT 性能

表5展示了未在未标注视听语音数据上进行预训练的 AV-HuBERT 模型的表现。如表2所示，随机初始化模型的语音识别和语音到文本翻译性能明显低于经过预训练的模型。这一发现与 [3] 仅英语视听语音识别的报告结果一致。

A.2. 噪声环境下，采用其他噪声类型和信噪比的视听语音识别与视听语音到文本翻译模型结果

在表6、表7和表8中，我们给出了视听语音识别和视听语音到文本翻译模型在噪声环境下的更全面结果，涉及 3 种噪声类型（多语言嘈杂噪声、语音噪声和音乐噪声）以及 3 种不同的信噪比（-10、0 和 10）。我们观察到，在所有噪声类型和信噪比下，音视频模式中的视听语音识别模型始终优于纯音频模式中的模型。

表 5：随机初始化的 AV-HuBERT 模型在非英语语音识别和 X-En 语音到文本翻译上的性能

ID	模型	Mode	源语言							Avg
			Ar	De	El	Es	Fr	It	Pt	Ru	
		语音识别，测试词错误率 ↓									
	单语 AV-HuBERT，无预训练	A	99.3	61.1	35.1	16.5	24.4	19.3	23.0	33.3	39.0
		AV	98.5	52.4	26.3	15.9	23.7	18.5	19.4	30.0	35.6
		X-En 语音到文本翻译，测试集 BLEU 分数 ↑									
	双语 AV-HuBERT，无预训练	A	-	-	1.0	2.0	2.3	0.9	1.9	0.7	1.4
		AV			0.6	0.7	1.1	0.6	1.0	0.6	0.7
	多语种 AV-HuBERT，无预训练	A	-	-	0.6	0.9	0.7	0.7	0.8	0.8	0.7
		AV			0.7	0.7	0.7	0.6	0.9	0.8	0.7

表6：噪声环境下多语言嘈杂噪声中的语音识别和语音翻译结果（信噪比=-10、0 和 10）

模型模式		El			Es			Fr			It			Pt			Ru			按信噪比计算的平均值		
		多语言嘈杂噪声，SNR=																				
		-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10
		语音识别，测试词错误率 ↓																				
Mono	A	106.1	74.6	43.1	108.5	65.1	23.6	115.0	63.6	30.3	102.4	71.0	29.8	109.1	77.3	37.3	107.6	68.1	40.4	108.1	69.9	34.1
	AV	97.4	53.1	32.2	101.9	45.1	20.7	106.6	48.1	28.3	98.7	51.6	25.0	95.9	47.2	25.4	98.3	51.9	34.1	99.8	49.5	27.6
Multi	A	97.4	73.7	47.4	94.1	59.7	35.1	92.3	54.7	30.3	94.2	61.0	27.1	92.8	61.9	28.3	96.3	64.1	43.2	94.5	62.5	35.2
	AV	96.0	62.2	46.1	88.6	40.7	20.8	87.0	39.0	21.9	88.0	44.3	24.5	83.9	43.1	24.8	92.9	57.9	42.9	89.4	47.9	30.2
		英语到其他语言的语音翻译，测试集 BLEU 分数 ↑																				
Bi	A	0.2	15.9	22.5	0.3	19.2	26.1	0.3	17.1	23.7	0.2	12.9	19.5	0.1	14.4	19.6	0.3	10.3	14.1	0.2	15.0	20.9
	AV	12.8	22.7	23.0	15.3	24.8	26.1	12.1	23.8	24.7	10.4	20.0	20.4	11.5	20.0	20.6	7.8	13.7	14.6	11.6	20.8	21.6
		X-En 语音到文本翻译，测试集 BLEU 分数 ↑																				
Bi	A	0.6	3.9	6.4	0.3	8.6	17.9	0.6	13.1	24.5	0.6	7.7	17.2	0.6	9.5	20.3	0.4	5.7	8.7	0.5	8.1	15.8
	AV	1.6	4.2	6.5	2.2	12.4	18.4	3.2	16.5	23.5	3.2	12.4	17.6	4.4	15.4	22.0	1.6	6.1	8.6	2.7	11.2	16.1
Multi	A	0.4	2.9	7.6	0.5	8.4	18.3	0.4	12.4	23.8	0.4	8.1	17.7	0.5	8.6	19.9	0.4	3.9	7.8	0.4	7.4	15.9
	AV	0.6	4.2	6.5	2.1	12.8	18.7	2.4	15.0	23.4	2.3	12.5	18.4	3.4	14.8	21.8	0.7	4.6	6.9	1.9	10.7	16.0

表 7：结　果：噪声环境下语音噪声中的语音识别和语音翻译结果（信噪比=-10、0 和 1　　*0)*

模型模式		El			Es			Fr			It			Pt			Ru			按信噪比计算的平均值		
		语音噪声，信噪比=																				
		-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10
		语音识别，测试词错误率 ↓																				
Mono	A	78.6	51.7	39.9	60.5	29.5	20.3	72.2	38.2	27.8	69.8	38.1	24.8	76.3	47.1	32.3	74.8	48.8	37.9	72.0	42.2	30.5
	AV	58.1	36.8	30.1	47.1	25.7	18.9	54.8	32.5	26.2	53.8	30.5	22.7	49.0	29.5	23.5	58.4	38.4	32.6	53.5	32.2	25.7
Multi	A	72.4	53.0	44.2	60.0	39.9	34.3	56.2	37.5	29.6	60.9	34.3	23.9	59.1	33.9	25.2	67.6	50.0	41.3	62.7	41.4	33.1
	AV	63.8	50.5	44.0	42.5	24.3	18.7	42.8	25.4	20.9	45.8	28.6	22.6	44.4	28.5	23.2	60.5	46.6	41.6	50.0	34.0	28.5
		英语到其他语言的语音翻译，测试集 BLEU 分数 ↑																				
Bi	A	3.9	14.2	20.9	5.9	17.9	25.1	5.2	15.3	22.3	3.8	12.2	18.6	3.9	12.7	18.2	3.3	9.0	13.6	4.3	13.6	19.8
	AV	20.6	23.2	22.9	23.5	25.8	26.6	21.8	24.4	24.9	18.1	20.3	20.3	18.8	20.4	20.5	12.4	14.4	14.4	19.2	21.4	21.6
		X-En 语音到文本翻译，测试集 BLEU 分数 ↑																				
Bi	A	3.4	6.0	6.8	8.4	15.8	19.1	11.1	21.3	25.0	7.0	15.1	17.6	8.8	17.6	21.1	4.3	7.5	8.7	7.2	13.9	16.4
	AV	4.0	5.7	6.7	11.1	16.8	19.2	14.0	21.2	24.0	11.0	16.8	18.5	13.7	20.2	22.7	5.4	8.0	8.6	9.8	14.8	16.6
Multi	A	2.8	6.7	8.7	8.3	16.4	19.3	10.7	20.9	24.7	7.6	15.9	19.1	8.5	17.7	21.7	3.4	6.6	8.2	6.9	14.0	16.9
	AV	4.0	6.2	6.8	12.1	17.9	19.5	14.0	21.3	23.9	11.4	16.9	19.0	14.1	20.1	22.5	3.7	6.0	7.5	9.9	14.7	16.5

表 8：结　果：噪声环境中含音乐噪声时的语音识别和语音翻译（信噪比=-10、0 和 1　　*0)*

模型模式		El			Es			Fr			It			Pt			Ru			按信噪比计算的平均值		
		音乐噪声 , SNR=																				
		-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10	-10	0	10
语音识别，测试词错误率 ↓																						
Bi	A	76.2	49.3	39.1	68.8	31.1	19.2	70.9	37.3	26.9	73.3	37.9	24.0	77.8	43.5	28.9	74.8	47.1	37.5	73.6	41.0	29.2
	AV	64.7	37.1	29.1	53.1	25.7	17.8	55.4	32.7	25.7	58.0	31.1	21.7	54.2	29.9	22.2	60.9	39.9	32.1	57.7	32.7	24.8
Multi	A	74.3	52.4	43.8	63.5	40.1	33.3	58.9	36.5	29.7	63.6	34.1	23.0	62.8	35.0	23.4	69.2	48.6	40.4	65.4	41.1	32.3
	AV	66.7	50.1	43.5	45.1	25.3	18.2	45.6	26.1	20.5	49.4	29.1	22.2	47.8	29.3	22.5	63.6	46.8	40.2	53.1	34.4	27.9
英语到其他语言的语音翻译，测试集 BLEU 分数 ↑																						
Bi	A	11.2	21.4	22.3	14.8	24.8	26.6	13.0	22.8	24.5	10.3	18.0	20.0	10.2	19.0	19.8	8.3	13.7	14.7	11.3	19.9	21.3
	AV	20.6	22.9	23.1	23.7	25.9	26.2	22.0	24.4	24.8	18.6	20.6	20.4	18.7	20.5	20.6	12.6	14.5	14.5	19.4	21.4	21.6
多语言到英语语音到文本翻译，测试集 BLEU 分数 ↑																						
Bi	A	3.3	6.0	6.7	7.6	15.9	19.3	11.2	22.1	25.6	7.6	15.3	18.6	8.8	18.1	22.4	4.4	7.5	8.7	7.1	14.1	16.9
	AV	3.5	5.8	6.7	10.2	17.0	19.2	13.4	21.5	24.3	11.0	16.2	18.7	14.0	20.2	22.7	4.9	7.8	8.8	9.5	14.7	16.7
Multi	A	3.0	6.4	8.3	7.5	16.3	19.6	10.4	20.9	25.2	7.8	15.9	19.5	7.8	17.9	22.3	3.1	6.7	8.4	6.6	14.0	17.2
	AV	3.9	6.4	7.1	10.7	17.3	19.5	13.3	21.0	24.4	10.7	16.9	19.2	12.8	20.2	23.2	3.3	6.6	7.4	9.1	14.7	16.8