arXiv:2301.10180v1 [cs.CL] 21 Jan 2023

用于多模态波斯语语音识别的多用途视听语料库：Arman-AV 数据集

Javad Peymanfard *、Samin Heydarian *§、Ali Lashini *§、Hossein Zeinali †、Mohammad Reza Mohammadi * 和 Nasser Mozayani *

*伊朗科技大学计算机工程学院，伊朗德黑兰 电子邮箱：javad_peymanfard@comp.iust.ac.ir、samin_heydarian@comp.iust.ac.ir、a_lashini@comp.iust.ac.ir、mrmohammadi@iust.ac.ir、mozayani@iust.ac.ir

†阿米尔卡比尔理工大学计算机工程系，伊朗德黑兰 电子邮箱：hzeinali@aut.ac.ir

摘要—**近年来，自动唇读取得了显著进展。然而，这些方法需要大规模数据集，而许多低资源语言并不存在此类数据集。本文介绍了一个面向波斯语的新型多用途视听数据集。该数据集包含约 220 小时的视频及其对应的 1760 名说话人。除唇读外，该数据集还适用于自动语音识别、视听语音识别和说话人识别。此外，这是首个大规模波斯语唇读数据集。针对上述每项任务，我们均提供了一种基线方法。此外，我们还提出了一种检测波斯语视素（音素的视觉等价物）的技术。与此前提出的视素相比，通过该方法获得的视素使唇读任务的准确率相对提高了 7%，并且同样可以应用于其他语言。**

索引词——**波斯语数据集、视听语音识别、唇读、视素**

I. 引言

自动语音识别（ASR）旨在从音频信号中理解语音。这项任务经过多年的发展，已足够成熟，可以应用于各种设备。然而，训练模型虽然具备很强的语音识别能力，但在某些特殊条件下仍有不足，例如环境噪声很大的情况。视听语音识别（AVSR）正是为克服这一局限而出现的。AVSR结合视觉信息和音频，以便在汽车内部等嘈杂环境中更有效地解码语音。AVSR更接近人类的理解方式，即通过视听感知来理解彼此。唇读则是另一项仅利用视觉信息的任务。

传统方法 [1]–[3] 采用两阶段算法来处理这类问题。在第一阶段，手工设计的特征提取器从唇部运动中提取有用特征同时，另一个特征提取器从音频信号中提取特征，随后将两者融合。在第二阶段，使用分类器（如隐马尔可夫模型或人工神经网络）对数字、字符等进行分类。然而，近几年来，大型公开数据集的可用性以及深度神经网络的兴起，对这一领域产生了重大影响。深度学习方法通常由前端和后端两部分组成，与传统方法类似，但其特点是可以进行端到端训练。前端部分通常使用卷积神经网络提取视觉和音频特征，后端则使用循环神经网络、注意力模型和 Transformer 等时序网络来建模时序信息。近期，诸如 [4]–[6]等方法采用知识蒸馏来训练唇读和视听语音识别模型。为此，通常将自动语音识别模型用作教师模型，将唇读模型用作学生模型。

最初的数据集是在实验室条件下采集的 [7], [8] 。通常，一个人站在摄像机前，在安静的地方以正常速度读出一些词语或句子。深度学习和自动化流程的兴起，促成了野外条件数据集的构建 [9]–[11] ，这类数据集规模更大，也更具挑战性，并且比以往更接近真实环境。与早期仅限于数字、字符或短语分类的数据集不同，如今的数据集用于词级分类和句子级音视频语音识别。不幸的是，这些数据集大多是英语数据集，而其他语言的数据集寥寥无几 [12]。本文将主要关注句子级音视频语音识别，但我们的数据集并不局限于语音识别任务。我们以能够用于不同任务的方式构建了数据集。接下来的章节将对此作进一步说明。

目前，LRS2 [13] 是使用最广泛的视听数据集。该数据集包含 240 多小时的视频和约 118,000 条话语。此外，他们还使用英国广播公司的新闻节目

§这些作者作出了同等贡献。

和脱口秀作为数据集来源。该数据集向研究界公开。LSVSR [11]是最大的数据集，包含超过 3800 小时的数据，这些数据采自上传至 YouTube 的视频。该数据集由 Google DeepMind 收集，但遗憾的是，他们没有将其公开。

本文提出了一个新颖的大规模数据集，该数据集从 Aparat 网站上的评论、电影等内容中，在“野外环境”下采集而成。该数据集包含来自 1760 位波斯名人的 220 多小时视频。此外，我们还将每个样本中名人的名称作为标签提供，可用于说话人识别等不同用途。据我们所知，这是规模最大的波斯语视听数据集。

由于这是一个视听数据集，因此可用于多种不同的应用，例如自动语音识别、唇读、说话人识别、视听语音合成等。

本文的组织结构如下。第 2 节介绍一些最新的方法和数据集。第 3 节讨论数据统计信息，以及我们构建的数据采集流程。第 4 节将使用一个基准模型评估我们的数据集。最后，第 5 节对全文进行总结。

II. 相关研究

本节将回顾语音识别和说话人识别这两项任务。由于我们引入了一个新数据集，并在该数据集上测试了基线模型的性能，因此将从所提出的方法和数据集的角度探讨相关研究。

A. 方法

1) 语音识别：**唇读**。自动唇读领域已有大量研究。深度学习的兴起以及大规模数据集的可用性，推动该领域的自动化取得了巨大进展。这里我们讨论将唇读视为开放集字符的句子级任务的相关研究。对于句子级识别，近期研究可分为几类方法。第一类方法采用连接时序分类（CTC）。模型逐帧预测字符，并通过最大化和边缘化所有可能路径，找到预测结果与真实标注之间的最佳对齐。LipNet [28]便是采用这一方法的例子。LipNet 是首个句子级唇读模型，它以前端时空卷积神经网络为基础，后端采用双向门控循环单元（Bi-GRU），并使用CTC损失训练网络。Shillingford 等人 [11] 提出了与先前工作最为接近的 V2P模型。该模型输出一系列音素分布，并在训练时使用CTC损失。在推理时，基于有限状态转换器（FST）的解码器会将音素分布序列映射为词序列。

第二种方法采用序列到序列模型，其中输出字符以彼此为条件，这与基于连接时序分类的模型不同。WAS [9]是首个序列到序列模型。该模型由两个关键组件组成，即图像编码器 Watch 和字符解码器 Spell，并采用独特的双重注意力机制。Chung 等人 [15] 将 WAS 模型扩展为 MV-WAS模型，使其能够解码不同姿态下的视觉序列，并证明可以读取侧脸唇语，但其效果不如读取正脸唇语。Peymanfard等人 [29] 提出了外部视素解码，将序列到序列模型分为两个阶段，分别是视频到视素和视素到字符。网络利用外部文本数据，根据视素输出字符序列。Afouras 等人 [30] 提出了视觉 Transformer 池化（VTP）模块，用于学习跟踪并聚合唇部运动表征。本研究还采用 WordPiece分词器（一种子词算法），以便更轻松地学习语言，从而提高时间和内存效率

Petridis 等人 [31] 提出了一种混合架构，并采用了包含循环神经网络、注意力机制和连接时序分类的联合解码器。Afouras 等人 [32] 也采用了混合方法，但建议用 Transformer 替代循环神经网络。这项工作提出在 Transformer 自注意力机制之上采用序列到序列和连接时序分类作为后端。两种模型都使用时空卷积神经网络，随后采用 ResNet 作为通用前端。 [31]的扩展版本是 [33] ，它采用 Transformer 的 Conformer 变体，对图像或音频序列的局部和全局依赖关系进行建模。

唇读领域的一种趋势是借助知识蒸馏，通过从数据集的纯音频对应版本中提取信息来提升性能。在这种方法中，自动语音识别（ASR）模型充当教师模型，而视觉语音识别模型则作为学生模型。 [4], [5] 和 [6]便是这一趋势的例子。

自动语音识别。在自动语音识别中，自监督学习在最先进的模型中发挥着主要作用。这种方法分为两个步骤：预训练和微调。在预训练步骤中，模型通过前置任务从大量无标签数据中学习强大的语音表征。随后，将上一步的输出（语音表征）输入声学模型，以针对下游任务进行微调。在 wav2vec [34], 中，预训练步骤包含两个网络，分别称为编码器和上下文网络。编码器网络将原始音频输入转换为特征表征，上下文网络则将特征表征转换为上下文化特征。这两个网络都是卷积网络。该模型的目标函数由未来时间步与干扰项之间的对比损失定义。vq-wav2vec [35] 是先前工作的改进版本，并借鉴了 BERT 模型。该方法同样由两个阶段组成。在第一阶段，连续的音频特征表征通过 Gumbel-Softmax 或 k均值聚类量化方法离散化。因此，语音数据具有了类似语言的结构，从而可以应用 BERT 模型。在第二阶段，离散化后的表征被输入 BERT 模型，以获得上下文化表征。wav2vec 2.0 [36]

表 I：知名数据集之间的统计比较。

类别	数据集	基本信息							转录信息		视频信息					说话人信息
类别	Name	Year	语言	Task	类别数	话语数	可用性	来源	词汇量	词语	Num.	时长	分辨率	FPS	视角（°）	Num.
音视频语音识别	SFAVD [14]	2013	FA	句子	-	600	-	实验室环境	1000	-	-	-	131×105	30	正面	1
	LRW [9]	2016	ENG	词语	500	400K	可用*[1]	BBC节目	500	400K	-	-	256×256	25	-30 30	1000+
	LRS2 [13]	2017	ENG	句子	-	118K	可用*	BBC节目	174K	807K	-	246h	160×160	25	-30 30	1000+
	MV-LRS [15]	2017	ENG	句子	-	74K	可用*	程序	14K	-	-	165h	160×160	25	-90 90	1000+
	VLRF [16]	2017	SPA	句子	-	600	公开	实验室环境	1374	10K	-	180分钟	1280×720	50	正面	24
	LRS3-TED [10]	2018	ENG	句子	-	165K	可用*	TED演讲	57K	-	-	475h	224×224	25	-90 90	1000+
	LSVSR [11]	2018	ENG	句子	-	2,934K	私有	YouTube	127K	-	-	3,886小时	128×128	30	-30 30	1000+
	LRW-1000 [17]	2019	CHI	词语	1000	718K	可用*	广播新闻	1K	-	-	-	分布式	25	-90 90	2000+
	CMLR [12]	2019	CHI	词语	-	102K	公开	新闻广播	-	-	-	-	-	-	-	11
	LRWR [18]	2021	RUS	词语	235	-	-	YouTube	235	117K	-	-	112×112	25	0 20	135
	GLips [19]	2022	GER	词语	500	-	公开	黑森州议会	500	250K	250K	-	256×256	25	-	100
	RUSAVIC [20]	2022	RUS	-	62	62	可用*	车辆环境	-	-	200	-	1920×1080	60	-30 30	20
	PLRW [21]	2022	FA	词语	500	244K	公开	Aparat	500	-	244K	30h	224×224	25	-	1800
	我们的数据集	2022	FA	句子	-	89K	公开	Aparat	42K	2.5M	89K	220h	224×224	25	-	1760
活跃说话人检测	AVA [22]	2019	-	-	3	-	公开	YouTube	-	-	188	38.5小时	128×128	-	-	-
	主动声呐战 [23]	2021	-	-	2	-	-	VoxConverse 数据集	-	-	212	30.9小时	-	-	-	-
说话人识别	Voxceleb [24]	2017	ENG	-	1251	153K	公开	YouTube	-	-	22K	352h	-	-	-	1,251
	Voxceleb2 [25]	2018	ENG	-	6112	1,128K	公开	YouTube	-	-	150K	2442小时	-	-	-	6,112
	DeepMine [26]	2019	FA&ENG	-	1850	544K	可用*	众包	-	-	540	480h	-	-	-	1,850
	CN-Celeb [27]	2019	CHI	-	1000	130K	公开	哔哩哔哩	-	-	-	273.73小时	-	-	-	1,000

[1]可通过联系获取。

通过联合学习上述两阶段预训练流程，取得了更好的结果。该模型通过上下文化输出与量化表征之间的新型对比损失进行预训练。HuBERT [37]在模型架构上与 wav2vec 2.0 非常相似，但训练过程有所不同。首先，HuBERT 使用交叉熵损失优化模型。其次，数据离散化通过 k 均值聚类算法完成，而非采用量化过程。第三，训练过程在隐藏单元发现与目标预测之间交替进行。该模型在聚类步骤中复用了 BERT 编码器的嵌入向量。

音视频语音识别。视听语音识别任务是在有音频的情况下进行唇读。Shi等人 [38] 提出了 HuBERT 模型的音视频版本 AV-HuBERT。此外，该研究还报告了视觉 HuBERT 在唇读任务中的表现。 [11], [31] 和 [32] 等研究同时探讨了 VSR 和视听语音识别问题，后文将在唇读部分进一步说明。

2) 说话人识别：由于这是一个分类问题，通用模型由两个模块组成。第一个是特征提取器模块，第二个是分类模块。根据输入数据的不同，特征提取器可以是 VGG-M 或 ResNet [25]这样的卷积网络，也可以是任何适用于序列数据的其他神经网络。对于分类，全连接网络是首选，其输出大小为数据集中说话人的数量。由于输入数据是音频文件，网络的输入可以是数据的频谱图 [25]，也可以是梅尔频率倒谱系数（MFCC）等特征提取算法的输出，后者通常是一个一维数据特征向量。

B. 数据集

1) 语音识别：众所周知，深度学习的进步得益于大型数据集。唇读系统的性能会受到位置、光照和说话人多样性的影响，因此数据集的质量和数量都很重要。近年来，一些数据集尝试增加话语数量和视频时长，例如 LRS2 [13], LRS3-TED [10]和 LSVSR [11]，而 MV-LRS [15] 则尝试涵盖更多面部角度。表 I 展示了三类数据集：视听语音识别、活跃说话人检测和说话人识别。表中为每个数据集列出了通用信息、转录信息、视频信息和说话人信息。我们可以从这些信息中看到有用的统计数据，例如数据集所执行的任务类型、所收集视频的来源，以及视频中的说话人数量。表中大多数数据集使用的语言是英语，但值得庆幸的是，近年来也发布了其他语言的数据集，例如 CMLR [12],LRWR [18]和 GLips [19]。自 2013 年以来，直到现在都没有用于唇读的波斯语数据集；我们发布了波斯语音视频语音识别数据集（Arman_AV），包含 8.9 万条话语、220 小时视频和 1760名说话人。

2) 说话人识别：该领域的大多数数据集都是在实验室条件下收集的，不能免费用于研究。最早在“野外环境”下收集的免费数据集之一是野外说话人（SITW） [39] 数据集。该数据集包含 299 位说话人的多媒体内容，以及人工标注的语音样本。VoxCeleb [24] 是一个大型数据集，包含来自 1,000 多位名人的语音样本，共计 100,000 多条话语，这些样本提取自 YouTube 视频；该数据集性别均衡（45% 的说话人为女性），并涵盖具有不同口音、种族背景和年龄的说话人。数据集创建者从维基百科收集了每位说话人的性别和原籍国信息。数据集中的视频是在大量复杂环境下录制的，例如安静和嘈杂的录音室访谈、露天体育场等，而且都混有现实环境中的嘈杂声，包括轻笑声、掩蔽语音、背景闲聊等。此外，数据集作者的另一个目标是提出一套创建全自动数据集的流程。VoxCeleb2 [25] 紧随 VoxCeleb 推出（与其非常相似，但规模大得多），是该领域规模最大的数据集，其中包含六千多位说话人的一百多万条话语。上述三个数据集均为英语数据集。CN-Celeb [27]是另一个大型数据集，包含 130,000 多条话语，与 VoxCeleb 数据集非常相似，但有以下三点不同：首先，CN-Celeb 涵盖 11 种语音类别，例如歌唱、娱乐等，

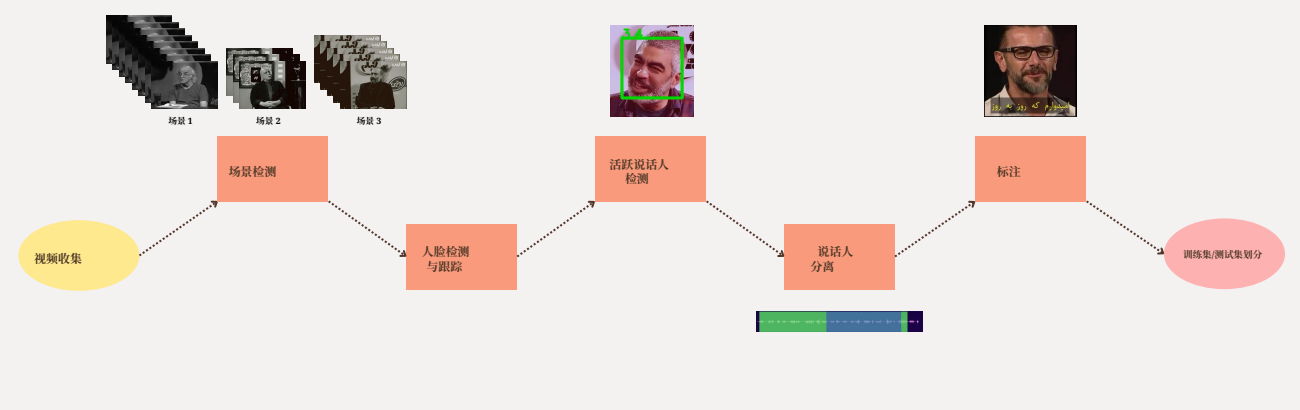


图 1：数据收集流程。



图 2：我们的数据集样本。

比 VoxCeleb 更多。其次，CN-Celeb 专注于中国名人，包含 1,000 位名人的视频。最后但同样重要的是，该数据集并非完全自动化，创建时还采用了人工监督。据我们所知，该领域尚无合适的波斯语数据集。

III. 数据集

A. 自动化流程

在本小节中，我们介绍数据收集的各个步骤。图 1 展示了这些步骤的概览。

步骤 1：视频收集。 我们为数据集挑选了三类视频，下面将分别介绍。

1) **访谈。** 访谈和传记非常适合构建视听数据集。这些视频的主要目的是与艺术家、政治人物或一般名人交谈。这类视频分为两组。第一组节目中有主持人和旁白（在本例中并不理想），第二组则没有主持人和旁白。

2) **电视剧和电影。** 如果经过适当的预处理，这类视频也可以使用。主要挑战是应对摄像机方向和角度的变化（与访谈不同，说话人并非始终在镜头中，或者可能并未出现在镜头里）。此外，整段视频中有大量静音片段或音乐。另一个挑战是多说话人同时语音。总的来说，这类视频并不适合用于收集数据，因为其中冗余数据的比例相对较高。在我们的调查中，只有不到 10% 的输入数据被认为是合适的。

3) **点播视频。** 收集更多数据的另一种方法是在用户生成内容中使用不同的关键词进行搜索。然后，我们会查看搜索结果中的视频，并为数据集挑选最热门的关键词。为获取数据，会使用多个搜索词。

步骤 2：场景检测。 在此阶段，我们希望检测每个场景。为此，我们使用 PySceneDetect[1] 。在该算法中，我们通过计算连续两帧的像素值之差来检测不同场景。该差值是在 HSV颜色空间中计算的。

步骤 3：人脸检测与人脸跟踪。现在，每个视频都被划分为称为场景的较小片段。到目前为止，视频已进行了时间剪辑。请注意，我们需要的是能够看到人脸的帧。因此，S^3FD 模型 [40] 和基于交并比的算法 [41] 分别用于人脸检测和人脸跟踪。根据视频各片段中检测到的人脸数量，每个视频可能会被拆分成一个或多个视频；如果没有人脸，则会被移除。

[1]https://github.com/Breakthrough/PySceneDetect

步骤 4：活跃说话人检测。
在这一步中，我们自动找到了视频中包含说话人面部的部分。

步骤 5：说话人分离。 如前所述，截至第 4 阶段，我们已经选出了包含说话人且拍摄角度保持不变的视频片段。但在访谈中还存在另一个潜在挑战，即主持人和嘉宾同时讲话的情况。为解决这一问题，我们可以使用基于 UIS-RNN 的说话人分离[2] 。

步骤 6：标注。 如前所述，我们的数据集视频来源于 Aparat。该视频托管网站上的视频没有字幕，因此我们利用商业化的 Aipaa 自动语音识别服务[3]，为视频生成粗略的句子级转录文本。

步骤 7：人脸识别与数据集划分。 我们使用 ArcFace [42]并生成用于人脸识别的人脸特征嵌入。此阶段旨在创建一个具有说话人独立属性的数据集。这样的数据集有利于唇读等应用。需要注意的是，我们对人脸识别算法的结果进行了人工检查，并纠正了其中的错误。

B. 统计

首先，该数据集将以人脸裁剪后的 MP4 视频形式提供，分辨率为 224*224 像素，帧率为 25 帧率。我们使用 Aparat（一个类似 YouTube 的波斯语视频分享网站）作为数据集来源。收集的视频来自不同的频道和节目，因此数据集包含各种复杂的环境和情形。我们的工作成果包括 220 小时的视频，其中包含超过 8.9 万条话语，以及 1760 位名人的 250 万个词语。我们将数据集划分为训练集/验证集和测试集，各部分分别包含 211 小时和 9 小时的数据。图 2 展示了一些视频样本示例。表 I 除了列出我们数据集的统计数据外，还列出了其他近期数据集的统计数据。

IV. 视素分析

视素是唇读问题中的一项重大挑战。视素由一组发音时唇形相同的音素构成，而任何语言的视素都尚未得到确切确定。不同说话人之间的这种映射甚至可能有所不同。不过，已有一些研究探讨了英语和波斯语等语言中的这一问题，并为这些语言提出了一些模型。在本研究中，我们提出了一种方法，可在大规模数据上自动识别波斯语视素；只要拥有目标语言的适当数据集，该方法也可应用于其他语言。已有一些研究探讨了波斯语中的视素 [43]。 [43] 中的作者收集了一组实验室数据集，并考察了不同音素发音之间的相似性，据此提出了一种音素分类方法，以识别该语言中的视素。本研究结合深度学习技术自动提取特征，并进行聚类，我们提出了一种音素分类方法。随后，我们将基于这些视素的唇读模型与先前研究中提出的视素进行了比较。在本实验中，我们首先使用 Kaldi 工具确定每个样本的音素级转录。在这份转录中，说话人发出的每个音素均以 30 毫秒为间隔标注。随后，我们使用预训练的 AV-HuBERT 模型获取每个音素的视觉嵌入，并将其作为特征，使用 k 均值算法进行聚类。在这一步中，我们希望具有相同视觉特征（例如唇部运动和唇形相同）的音素能够被归入同一类别。我们使用 [29] 中的方法训练模型。通过采用视频到视素模型和视素到字符模型，并最终将二者结合，我们得到了唇读模型。此方法的一个优点是，可以利用文本数据提升唇读模型的质量。不过，该模型需要一个合适的字素转音素（G2P）模型，将波斯语文本转换为音素，再将音素转换为视素；为此，我们采用了合适的 G2P 模型。对于视频到视素模型的数据集，我们使用音素级转录，并借助相应的映射将音素转换为视素。我们还使用了约 100 万条话语来构建视素到字符转换数据集。为此，我们首先使用 G2P 模型获取每条话语的音素序列，然后使用音素到视素映射提取相应的视素序列。训练视素到字符模型时，我们将视素序列作为输入，并将波斯语文本作为目标输出。我们采用注意力机制实现视频到视素模型，并使用了一个具有两层门控循环单元（GRU）的序列到序列模型。此外，我们采用相同的架构，但减少参数并改变输入类型，以训练视素到字符模型。如表 II 所示，采用新提出的模型识别波斯语视素后，我们在唇读任务中获得了更高的精确率；我们认为，这种音素到视素映射更为合适。此外，该方法可用于识别任何给定语言中的视素。所有数据及其相关转录文本均已公开。

V. 实验

此处将介绍基线方法。我们展示这些方法在我们的数据集上的表现，并提供训练设置的相关信息。在所有基线方法中，图像序列和音频序列输入分别表示为 $x^v_{1:T}$ 和 $x^a_{1:T}$。

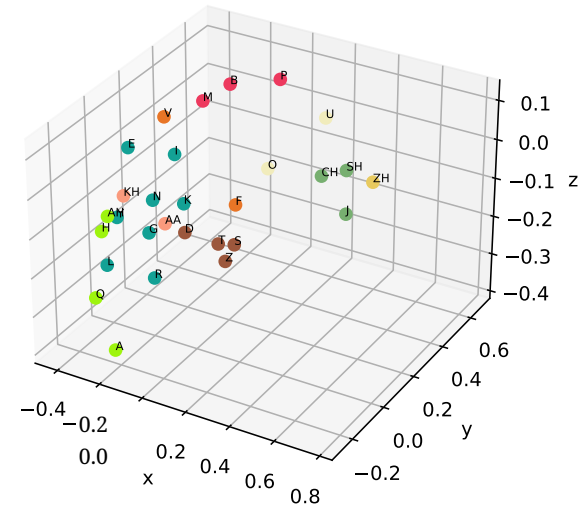
A. 自动语音识别（*ASR*）

本节回顾了使用所生成的数据集进行语音识别的实验。第一项实验仅使用音频数据训练波斯语语音识别。为此，预训练的 AV-HuBERT [44]model uses $x^a_{1:T}$，从每 40 毫秒的音频输入中提取一个 768维特征向量$e^a_{1:T}$。在这项实验中，我们使用预训练的 AV-HuBERT 基准模型，对英语数据进行特征提取。该模型的输出嵌入在这

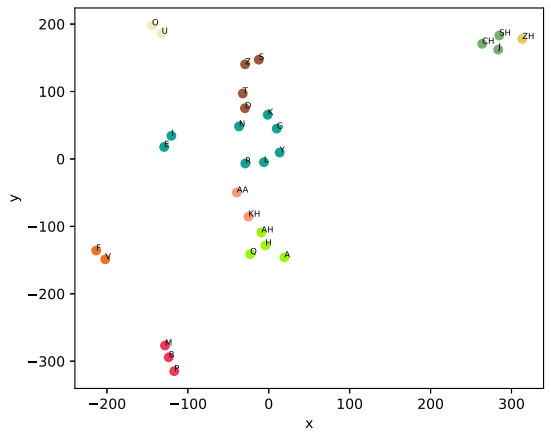
[2]https://github.com/taylorlu/Speaker-Diarization [3] https://aipaa.ir/

表 II：我们数据集上的唇读结果

视素映射	字符错误率（贪心解码）	词错误率（贪心解码）	字符错误率（束搜索解码）	词错误率（束搜索解码）
传统方法 [43]	%54.04	**%76.32**	%52.19	%73.24
AV-HuBERT + K 均值聚类（我们的方法）	**%51.24**	%76.71	**%48.54**	**%72.83**



(a)



(b)

图3：使用我们的数据集对波斯语音素的 AV-HuBERT 进行二维和三维投影。a）使用主成分分析对音素嵌入进行三维投影。b）使用 t-SNE 对音素嵌入进行二维投影。

表 III：使用聚类和 AV-HuBERT 得到的波斯语视素

聚类 ID	音素
1	/F/ /V/
2	/AA/ /KH/
3	/B/ /P/ /M/
4	/ZH/
5	/CH/ /JH/ /SH/
6	/O/ /U/
7	/S/ /Z/ /D/ /T/
8	/A/ /'/ /H/ /GH/
9	/G/ /K/ /E/ /L/ /I/ /Y/ /N/ /R/
10	/sil/

个网络中，每个窗口都表示一个音频片段，包含说话人和语音特征。我们预计模型会在学习过程中使用与语音相关的特征。因此，我们预计与说话人对应的特征会被隐式忽略，因为它们与我们的目的无关。实验中还需考虑的一个问题是：通过英语数据预训练的模型能否用于波斯语。

在下一步中，将 $e^a_{1:T}$ 向量输入双向长短期记忆网络（两层堆叠），得到 $o^a_{1:T}$。该网络使用连接时序分类进行训练。我们假设 $y = (y_1, y_2, \ldots\ldots, y_n)$ 是转录文本，而 π 是函数 B 会映射 π 到 y 的路径。我们通过对 $o^a_{1:T}$应用 Softmax，计算了 52 个字符的 p_t^{CTC} 概率。随着 CTC 损失 L_{CTC} 降低，网络开始学习。

如表 IV 所示，使用所提出的架构，仅凭 211 小时的波斯语语音数据，我们便达到了 84.66% 的字符准确率（CAR）；考虑到数据量，这是一个合适的数值。此外，实验表明，在英语数据上训练的 AV-HuBERT 模型能够为波斯语语音数据提取合适的嵌入。

B. 视听语音识别（视听语音识别）

在第二个实验中，我们训练了一个视听语音识别模型。该模型首先对图像进行预处理，仅裁剪出嘴部感兴趣区域。随后，对每一帧裁剪后的视频使用 AV-HuBERT 模型，得到一个 768维特征向量。本节采用了与前一阶段相似的架构，但也将视觉特征作为模型输入。此外，由于视觉特征包含的信息可能少于语音信息，我们减少了视觉特征的输出维度。该模型如图所示

4. 将输入图像序列 $x^v_{1:T}$ 送入预训练的单模态视觉 HuBERT，得到一个 768维特征向量 $e^v_{1:T}$。同时，将输入音频序列 $x^a_{1:T}$ 送入预训练的单模态音频 HuBERT，得到 768维特征向量 $e^a_{1:T}$。接下来，使用两个分开的双向长短期记忆网络，将 $e^v_{1:T}$ 和 $e^a_{1:T}$ 分别映射为 $o^v_{1:T}$ 和 $o^a_{1:T}$。然后，对 o^v_t 和 o^a_t拼接得到的 c_t 应用两个一维卷积滤波器。此步骤的输出为 f_t。如上所述，使用 CTC损失训练模型。首先，我们计算 p^{CTC}_t概率。该概率通过对 f_t应用 Softmax 和线性投影得到（其中 $W \in$IR$^{52\times(256+1024)}$ 且 $b \in$ IR^{52}）。然后，通过最小化 L_{CTC} 的值来优化模型。

该模型采用与前一个模型完全相同的机制进行训练。如表 IV 所示，借助这一同时提供音频和视觉特征的模型，我们取得了比自动语音识别模型更高的准确率。从这项实验可以看出，在英语数据上训练的预训练 AV-HuBERT 模型所获得的嵌入向量也可用于波斯语，并且能够从中提取出相对理想的唇部运动特征。

C. 说话人识别

我们使用生成的数据测试的另一个问题是说话人识别。为此，我们首先筛选出样本数超过五个的人，然后随机为每位说话人选取五个样本。对于每位说话人，将其中四个样本作为参考，仅用一个样本进行测试。随后，我们使用 TitaNet [45] 预训练模型。该模型在 5 个数据集上训练：Voxceleb [24]和 Voxceleb2 [25], 数据集中的 NIST说话人识别评测 2004—2008年部分（LDC2009E100）、Switchboard-Cellular1 和 Switchboard- Cellular2 [46], Fisher [47] 以及 LibriSpeech [48]。我们将 $x^{a_t}_{1:T}$作为测试样本，将 $x^{a_0}_{1:T}$、……、$x^{a_3}_{1:T}$ 作为参考样本输入 TitaNet [45]模型。模型会生成每个音频样本的 192维嵌入向量，例如 $v^{a_t}_{1:T}$。接下来，通过余弦相似度将测试嵌入向量与每个人的所有参考样本进行比较（此处 $i \in 0$、……、3）。之后，我们对上一步的结果求平均值。然后，根据平均余弦相似度，将测试集中的每个样本分配给与其相似度最高的人。采用这种方法，我们取得了 84.5 的准确率。

VI. 结论

Arman-AV 是一个面向波斯语的大规模多用途数据集，可用于唇读、自动语音识别、视听语音识别和说话人识别等多种任务，并且对公众开放。该数据集包含来自 1760 名说话人的近 220小时视频数据（约 89,000个样本）。尽管收集该数据集的主要目的是用于唇读，但我们仍报告了上述各项任务的基线方法结果。根据所得结果，利用视觉特征后，语音识别的字符错误率相对降低了 14.7%。我们还提出了一种获取波斯语视素的方法。利用这些视素，唇读准确率得到了提高。此外，我们还对该数据集进行了多项分析，例如说话人的年龄和性别，以及各片段中音视频同步情况。所有这些元数据均与数据集一同提供。该数据集还可用于人脸识别、人脸验证和视听说话人识别等任务；本文未涉及这些任务，可作为该数据集未来研究的方向。

VII. 致谢

该项目主要由伊朗的一家人工智能公司 Arman Rayan Sharif 资助。

参考文献

[1] I. Matthews、T. F. Cootes、J. A. Bangham、S. Cox 和 R. Harvey，“唇读视觉特征的提取”，*IEEE* 模式分析与机器智能汇刊，第24卷，213，2002 2，198，第。页码 – 。[2]S. S. Morade 和 S. Patnaik，“一种使用局部主动轮廓模型和隐马尔可夫模型的新型唇读算法：用于数字识别的测试”，*Optik*，第125卷，第18期，页码5181–5186，2014年。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S0030402614005786 [3] Z. Zhou、G. Zhao 和 M. Pietikäinen，“迈向实用的唇读系统”，载于*CVPR 2011*。IEEE，2011年，页码137–144。[4] Y. Zhao、R. Xu、X. Wang、P. Hou、H. Tang 和 M. Song，“听见嘴唇：通过蒸馏语音识别器改进唇读”，载于*AAAI*人工智能会议论文集，第34卷，第04期，2020年，页码6917–6924。[5] T. Afouras、J. S. Chung 和 A. Zisserman，“ASR即是所需：用于唇读的跨模态蒸馏”，载于*ICASSP 2020-2020*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2020年，页码2143–2147。[6] S. Ren、Y. Du、J. Lv、G. Han 和 S. He，“向大师学习：为唇读蒸馏跨模态高级知识”，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2021年，页码13 325–13333。[7] M. Cooke、J. Barker、S. Cunningham 和 X. Shao，“用于语音感知和自动语音识别的视听语料库”，美国声学学会会刊，第120卷，第5期，页码2421–2424，2006年。[8] I. Anina、Z. Zhou、G. Zhao 和 M. Pietikäinen，“OuluVS2：用于非刚性嘴部运动分析的多视角视听数据库”，载于*2015*年第*11*届*IEEE*自动人脸与手势识别国际会议暨研讨会（*FG*），第1卷。IEEE，2015年，页码1–5。[9] J. S. Chung 和 A. Zisserman，“野外唇读”，载于亚洲计算机视觉会议。Springer，2017年，页码87–103。[10] 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。[11] B. Shillingford、Y. Assael、M. W. Hoffman、T. Paine、C. Hughes、U. Prabhu、H. Liao、H. Sak、K. Rao、L. Bennett、M. Mulville、M. Denil、B. Coppin、B. Laurie、A. Senior 和 N. de Freitas，“大规模视觉语音识别”，载于*INTERSPEECH 2019*论文集，2019年，页码4135–4139。[12] Y. Zhao、R. Xu 和 M. Song，“用于中文普通话唇读的级联序列到序列模型”，载于*ACM*亚洲多媒体会议论文集，2019年，页码1–6。[13] J. Son Chung、A. Senior、O. Vinyals 和 A. Zisserman，“野外唇读句子”，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，页码6447–6456。[14] Z. Naraghi 和 M. Jamzad，“SFAVD：Sharif波斯语视听数据库”，载于第五届信息与知识技术会议。IEEE，2013年，页码417–421。[15] J. S. Chung 和 A. Zisserman，“ 侧脸唇读 ”，2017 年。[16] A. Fernandez-Lopez 、 O. Martinez 和 F. M. Sukno，“探索视觉语音识别上限的估计：视觉唇读可行性数据库”，载于*2017*年第*12*届*IEEE*国际自动人脸与手势识别会议（*FG 2017*）。IEEE，2017年，页码208–215。

表 IV：我们数据集上的语音识别结果

方法	字符错误率（贪心解码）	词错误率（贪心解码）	字符错误率（束搜索解码）	词错误率（束搜索解码）
自动语音识别（连接时序分类）	%15.77	%46.43	%15.34	%45.86
视听语音识别(连接时序分类)	**%13.76**	**%44.98**	**%13.08**	**%43.43**

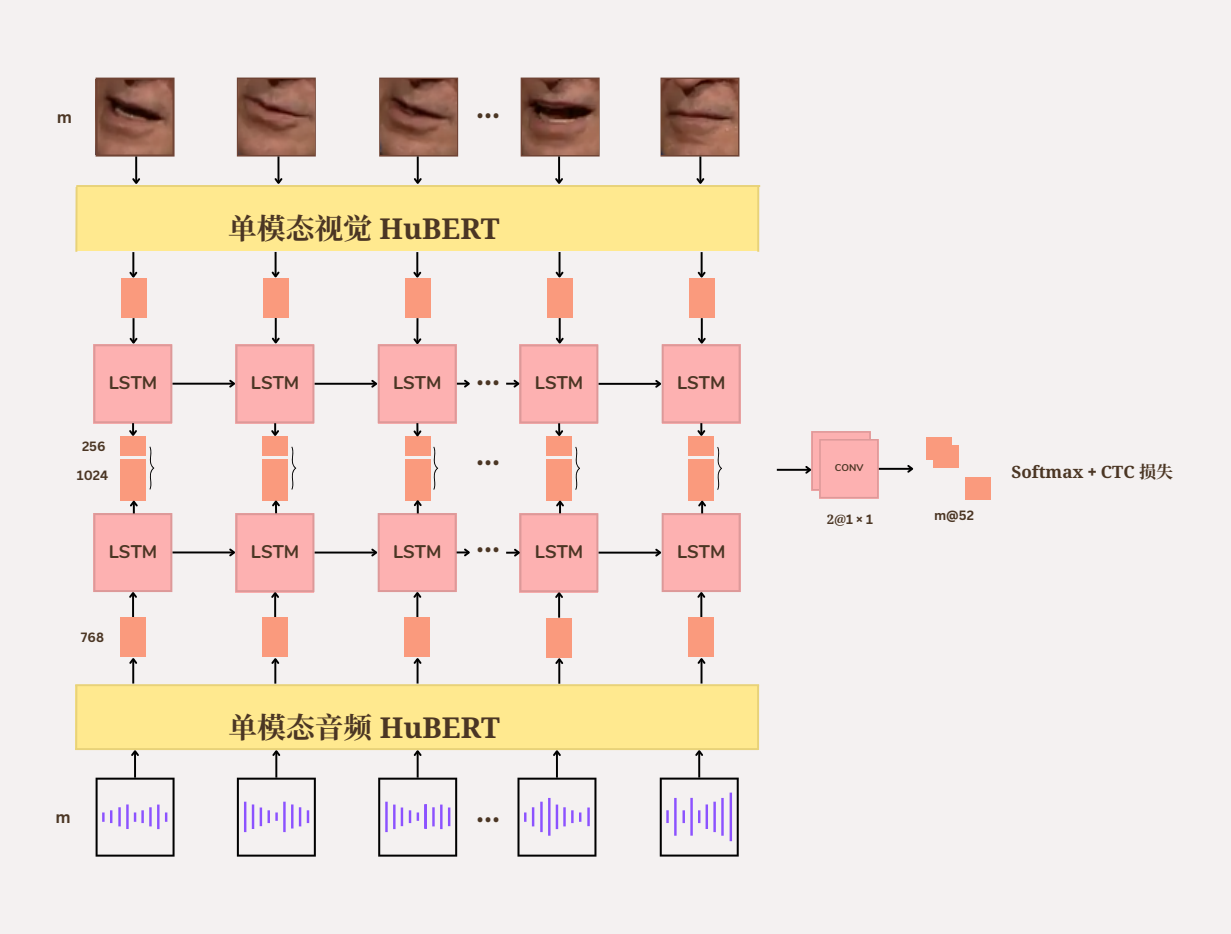


图 4：视听架构。

[17] S. Yang、Y. Zhang、D. Feng、M. Yang、C. Wang、J. Xiao、K. Long、S. Shan 和 X. Chen，“LRW-1000：用于野外唇读的自然分布大规模基准数据集”，载于*2019*年第*14*届*IEEE*国际自动人脸与手势识别会议（*FG 2019*）。IEEE，2019年，页码1–8。[18] E. Egorov、V. Kostyumov、M. Konyk 和 S. Kolesnikov，“LRWR：俄语唇读大规模基准数据集”，*arXiv*预印本 *arXiv:2109.06692*，2021年。[19] G. Schwiebert、C. Weber、L. Qu、H. Siqueira 和 S. Wermter，“用于自动唇读系统和迁移学习的多模态德语数据集”，*arXiv*预印本 *arXiv:2202.13403*，2022年。[20] 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测会议论文集，2022年，页码1555–1559。[21] J. Peymanfard、A. Lashini、S. Heydarian、H. Zeinali 和 N. Mozayani，“词级波斯语唇读数据集”，载于*2022*年第*12*届国际计算机与知识工程会议（*ICCKE*）。IEEE，2022年，页码225–230。[22] J. Roth、S. Chaudhuri、O. Klejch、R. Marvin、A. Gallagher、L. Kaver、S. Ramaswamy、A. Stopczynski、C. Schmid、Z. Xi 等，“Ava主动说话人：用于活跃说话人检测的视听数据集”，载于*ICASSP 2020-2020*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）. IEEE, 2020, pp. 4492–4496. [23] Y. J. Kim, H.-S. Heo, S. Choe, S.-W. Chung, Y. Kwon, B.-J. Lee, Y. Kwon, and J. S. Chung, “看看谁在说话：野外环境中的活跃说话人检测,” *arXiv* 预印本 *arXiv:2108.07640*, 2021. [24] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb：大规模说话人识别数据集,” 载于 *INTERSPEECH*, 2017. [25] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2：深度说话人识别,” 载于 *INTERSPEECH*, 2018. [26] H. Zeinali, H. Sameti, and T. Stafylakis, “Deepmine 语音处理数据库：波斯语和英语中的文本相关及文本无关说话人验证与语音识别。” 载于 *Odyssey*, 2018, pp. 386–392. [27] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, “Cn-celeb：具有挑战性的中文说话人识别数据集,” 载于 *ICASSP 2020-2020*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）. IEEE, 2020, pp. 7604–7608.[28] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, “Lipnet：句级唇读,” *CoRR*, vol. abs/1611.01599, 2016. [在线]。可访问：http://arxiv.org/abs/ 1611 01599 。[29] J. Peymanfard, M. R. Mohammadi, H. Zeinali, and N. Mozayani, “使用外部视素解码进行唇读,” 载于 2022年国际

机器视觉与图像处理会议（*MVIP*）. IEEE, 2022, pp. 1–5. [30] T. Afouras, A. Zisserman 等, “结合视觉注意力的子词级唇读,” *arXiv* 预印本 *arXiv:2110.07603*, 2021. [31] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, “采用混合式 CTC/注意力架构的音视频语音识别,” 载于 *2018*年*IEEE*口语技术研讨会（*SLT*）. IEEE, 2018, pp. 513–520. [32] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “深度音视频语音识别,” *IEEE* 模式分析与机器智能汇刊, 2018. [33] P. Ma, S. Petridis, and M. Pantic, “基于 Conformer 的端到端音视频语音识别,” 载于 *ICASSP 2021-2021*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）. IEEE, 2021, pp. 7613–7617. [34] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec：用于语音识别的无监督预训练,” *arXiv* 预印本 *arXiv:1904.05862*, 2019. [35] A. Baevski, S. Schneider, and M. Auli, “vq-wav2vec：离散语音表征的自监督学习,” *arXiv* 预印本 *arXiv:1910.05453*, 2019. [36] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0：语音表征自监督学习框架,” 神经信息处理系统进展, vol. 33, pp. 12 449– 12 460, 2020. [37] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT：通过掩码预测隐藏单元进行自监督语音表征学习,” *IEEE/ACM* 音频、语音与语言处理汇刊, vol. 29, pp. 3451–3460, 2021. [38] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “通过掩码多模态聚类预测学习音视频语音表征,” *arXiv* 预印本 *arXiv:2201.02184*, 2022. [39] M. McLaren, L. Ferrer, D. Castán, and A. D. Lawson, “野外说话人（SITW）说话人识别数据库,” 载于 *INTERSPEECH*, 2016. [40] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “S3fd：单次尺度不变人脸检测器,” 载于 *IEEE* 国际计算机视觉会议论文集, 2017, pp. 192–201. [41] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, “有人在说话吗？探索用于音视频主动说话人检测的长期时序特征,” 载于 第*29*届 *ACM* 国际多媒体会议论文集, 2021, pp. 3927–3935. [42] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace：用于深度人脸识别的加性角度间隔损失,” 载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集, 2019, pp. 4690– 4699.[43] M. Aghaahmadi, M. M. Dehshibi, A. Bastanfard, and M. Fazlali, “利用音素子空间对波斯语视素进行聚类，以开发视觉语音应用,” 多媒体工具与应用, vol. 65, no. 3, pp. 521–541, 2013. [44] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “通过掩码多模态聚类预测学习音视频语音表征,” *arXiv* 预印本 *arXiv:2201.02184*, 2022. [45] N. R. Koluguri, T. Park, and B. Ginsburg, “TitaNet：采用一维深度可分离卷积和全局上下文的说话人表征神经模型,” 载于 *ICASSP 2022-2022*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）. IEEE, 2022, pp. 8102–8106. [46] J. Godfrey and E. Holliman, “Switchboard-1 Release 2 ldc97s62,” 语言数据联盟, 1993. [47] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, “Fisher 英语训练语音第一部分转录文本,” 费城：语言数据联盟, 2004. [48] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech：基于公有领域有声读物的自动语音识别语料库,” 载于 *2015*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）. IEEE, 2015, pp. 5206–5210。