arXiv:2301.06059v1 [cs.GR] 15 Jan 2023

学习音频驱动的视素动态，实现三维面部动画

LINCHAO BAO，腾讯人工智能实验室，中国 HAOXIAN ZHANG，腾讯人工智能实验室，中国 YUE QIAN，腾讯人工智能实验室，中国 TANGLI XUE，腾讯人工智能实验室，中国 CHANGHAI CHEN，腾讯人工智能实验室，中国 XUEFEI ZHE，腾讯人工智能实验室，中国 DI KANG，腾讯人工智能实验室，中国

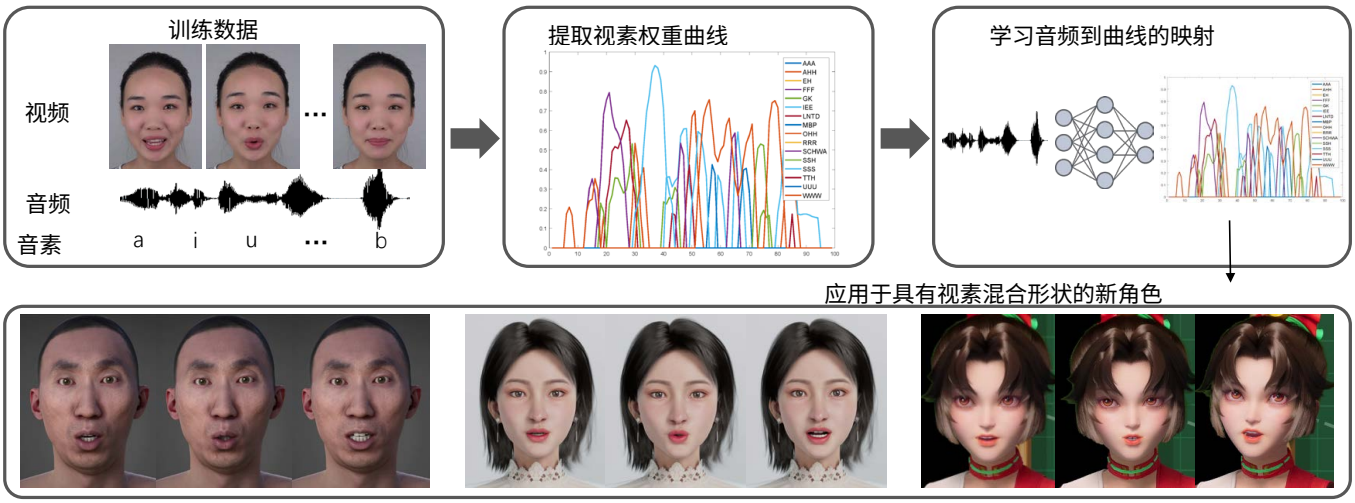


图 1. 我们的音频驱动面部动画方法概览。我们提出了一种新颖的音素引导人脸跟踪算法，用于从视频中提取视素曲线。随后，利用配对的音频和视素曲线数据学习音频到曲线映射模型。该模型可用于根据音频为各种风格的角色生成语音动画。

我们提出了一种新颖的音频驱动面部动画方法，可根据输入音频生成逼真的唇同步三维面部动画。我们的方法从语音视频中学习视素动态，生成便于动画师使用的视素曲线，并支持多语言语音输入。我们方法的核心是一种新颖的参数化视素拟合算法，该算法利用音素先验从语音视频中提取视素参数。在音素的引导下，提取出的视素曲线与音素的相关性更高，因此更易于控制，也更方便动画师使用。为支持多语言语音输入并适应未见过的语音，我们利用在多种语言上预训练的深度音频特征模型，学习从音频到视素曲线的映射。即使输入音频的音量、音高、速度或噪声受到干扰，我们的音频到曲线映射仍能达到最先进的性能。最后，我们提出了一种视素扫描方法，用于获取高保真视素素材，从而高效制作语音动画。我们展示了预测的视素曲线可应用于不同的绑定视素的角色，生成具有逼真、自然面部运动的个性化动画。我们的方法对艺术家友好，并可轻松集成到典型的动画制作工作流中，包括基于混合变形或骨骼的动画。

作者地址：Linchao Bao，腾讯人工智能实验室，中国深圳，linchaobao@ gmail.com；Haoxian Zhang，腾讯人工智能实验室，中国深圳；Yue Qian，腾讯人工智能实验室，中国深圳；Tangli Xue，腾讯人工智能实验室，中国深圳；ChanghaiChen，腾讯人工智能实验室，中国深圳；Xuefei Zhe，腾讯人工智能实验室，中国深圳；Di Kang，腾讯人工智能实验室，中国深圳。

ACM计算分类体系： • **计算方法** → **动画**； **神经网络**。

附加关键词和短语：三维面部动画、数字人、唇形同步、视素、音素、语音、神经网络

1 引言

逼真的数字角色面部动画对于电子游戏和电影等各种沉浸式应用至关重要。传统上，创作与语音音频同步的面部动画对艺术家而言既耗时又困难。为降低这一难度，人们开发了面部表演捕捉系统。然而，由于需要专业的工作流程和演员，许多情况下使用表演捕捉系统进行动画制作仍然具有挑战性。人们迫切需要研究能够仅以音频为输入、便捷地生成高保真面部动画的技术。

语音驱动面部动画技术的最新进展可分为两类：基于顶点的动画 [Cudeiro 等，2019；Fan 等，2022；Karras 等，2017；Liu 等，2021；Richard 等， 2021] 以及基于参数的动画 [Edwards 等，2016；Huang 等，2020；Taylor 等，2017；Zhou 等， 2018]。基于顶点的

动画方法通过预测网格顶点位置，直接学习从音频到三维人脸模型序列的映射；而基于参数的动画方法则生成动画曲线（即动画参数序列）。尽管近年来前一种方法受到越来越多的关注，但后一种方法更适用于动画制作，因为其结果更易于控制，也更便于融入典型的动画师工作流程。

动画曲线的参数空间可以通过算法中心的分析方法来定义，例如主成分分析（PCA） [Huang 等，2020年；Taylor 等，2017], ；也可以通过艺术家定义的动画师中心绑定来定义，例如 JALI [Edwards 等，2016；Zhou 等， 2018]。算法中心的参数空间易于从面部表演数据中推导，但通常难以应用于新角色。动画师中心的参数空间依赖于艺术家制作的特定面部绑定，通常更便于动画师使用。JALI [Edwards 等， 2016] 基于视素形状定义了一种先进的面部绑定，可通过独立控制下颌与嘴唇动作来调整这些形状。然而，这种灵活却复杂的绑定使得利用面部表演数据学习动画变得困难。尽管已有研究提出精细的程序化动画方法[Edwards 等， 2016] ，可以根据音频生成富有表现力的面部动画，但人工生成的动态效果并不自然。基于程序化生成数据训练的深度学习方法 VisemeNet [Zhou 等， 2018] 也存在同样的问题。

本文提出了一种新方法，采用便于艺术家使用的动画参数空间，并从面部表演数据中学习动画曲线。具体而言，动画曲线的每一帧都由一组视素混合形状的权重构成。我们首先使用一种新颖的音素引导的面部跟踪算法，从面部表演视频中提取视素权重。随后，利用神经网络学习从音频到视素曲线的映射。由于每个视素都有明确的含义，为新角色准备视素混合形状十分直接，也便于艺术家操作。艺术家可以通过扫描人类演员的视素来完成这项工作，也可以手动为风格化角色设计个性化视素。由于从视频中提取的视素权重曲线体现了面部运动的动态变化，应用这些曲线生成的最终动画比程序化生成的动画更加自然。最重要的是，在面部跟踪过程中引入音素引导，降低了参数拟合过程中的模糊性，因此得到的视素权重与音素的关联更加准确。这使动画曲线更便于动画师使用；更重要的是，将其应用于新角色时，泛化性能也得到了显著提升。

本文的贡献包括：

- 一种新颖的音素引导三维面部跟踪算法，用于语音驱动的面部动画，能够更准确地激活视素权重，且便于艺术家使用。
- 一种用于将音频映射为动画曲线的先进神经网络模型，支持多语言语音输入，并能很好地泛化到未见过的说话人。
- 通过扫描人类演员的视素，高效制作高保真数字人动画。

2 相关工作

视频面部表演捕捉。 多年来，基于RGB视频的面部表演捕捉一直是一个活跃的研究领域 [Beeler 等，2011；Cao 等，2015、2018；Garrido 等，2013；Laine 等，2017；Shi 等，2014；Thies 等，2016；Zhang 等，2022；Zoss 等， 2019]。全面回顾这一领域超出了本文的范围（请参阅综述 [Egger 等，2020；Zollhöfer 等，2018]）。我们主要关注能够从视频中提取面部表情参数的方法。现有方法通常通过拟合混合形状的线性组合 [Garrido 等，2013；Weise 等， 2011] 或多线性可变形模型 [Shi 等，2014；Thies 等， 2016] 来实现这一目标：针对输入视频的每一帧，通过优化考虑特征对齐、图像一致性、统计正则化等因素的目标函数进行拟合。此外，研究者还开发了基于学习的方法 [Cao 等， 2015, 2013] ，以进一步提高推理效率。混合形状 [Lewis 等，2014] 与多线性模型 [Cao 等，2014] 的关键区别在于，一组混合形状权重所定义的空间通常由艺术家设定，不一定彼此正交；而多线性模型通常由主成分分析（PCA）得到，其参数空间彼此正交。因此，采用不同先验进行混合形状拟合可能会得到不同的解。本文采用视素混合形状空间，并提出一种新颖的音素引导先验用于拟合。

基于参数的音频三维面部动画。 基于视素曲线的音频动画可追溯至 Kalberer 等人在 2000 年代初期的工作 [2002]。他们通过独立成分分析（ICA）构建视素空间，并在该空间内对稀疏的音素到视素映射点拟合样条曲线，以生成平滑的动画。然而，样条曲线无法很好地表现面部运动的真实动态，因此生成的动画不自然。Taylor 等人 [2012] 提出了一种方法，用于发现动态视素的时间单元，从而将逼真的运动融入运动单元。最终动画通过简单拼接这些单元生成，但单元之间仍会出现不自然的过渡。他们后续的工作 [Taylor 等人 2017] 采用深度学习改进了这一问题，直接从数据中学习自然的协同发音运动；但该方法依赖基于主动外观模型（AAM）的参数空间，不便于艺术家使用，且重定向到新角色时会降低动画质量。Edwards 等人 [2016] 提出了一种复杂的程序化动画方法，基于可独立控制下颌与唇部动作的视素绑定系统，即 JALI。动画曲线依据包含经验参数的显式规则生成。尽管 JALI 凭借便于艺术家使用的设计和多功能性，已广泛应用于游戏产业，但生成的面部运动仍带有人工拼凑的感觉。后续工作 VisemeNet [Zhou 等人 2018], 也存在这一问题；该方法采用深度学习，从 JALI 生成的数据中学习动画曲线。

基于顶点的音频驱动三维面部动画。 另一类研究旨在直接从音频生成基于顶点的动画 [Cudeiro 等，2019；Fan 等，2022；Karras 等，2017；Liu 等，2021；Richard 等， 2021], 主要采用深度学习方法。Karras 等， [2017] 提出了一种神经网络，用于直接学习

从音频到人脸模型顶点位置的映射，训练依据是特定角色的 3 至 5 分钟高质量跟踪网格序列。这种方法针对特定主体，因此要为新角色制作动画，就需要新的高成本训练数据，或者进行缺乏个性化控制的变形重定向。VOCA [Cudeiro 等， 2019] 使用独热身份编码对多主体顶点动画进行建模，并基于包含 12 位说话者的四维扫描数据集进行训练。MeshTalk [Richard 等， 2021] 进一步将研究从下半脸扩展到全脸动画，并采用了规模更大的数据集，其中包含 250 位主体的四维扫描数据。然而，这些基于顶点的动画方法不便于艺术家使用，也难以整合进动画制作工作流。另一方面，这些方法中的神经网络模型，包括时序卷积网络 (TCN) [Cudeiro 等，2019；Karras 等，2017；Liu 等，2021；Richard 等， 2021], 长短期记忆网络 (LSTM) [Huang 等，2020；Zhou 等， 2018], Transformer [Fan 等， 2022], 等，为我们开发音频到曲线映射模型提供了宝贵参考。

3 概述

我们首先介绍本方法所使用的数据集和视素。随后，我们简要概述本方法的流程以及本文的组织结构。

数据集。 我们收集了一个包含一名中国女演员16小时讲话视频的数据集。该数据集包含大约12，000条话语（其中10，000条用于训练，2，000条用于测试），内容主要为中文，少量话语中包含英语单词。视频由摄像机和高保真麦克风同步录制。音频的文本转录稿由人工标注。带有起止时间戳的音素由一种类似于 Montreal Forced Aligner [McAuliffe 等人 2017]所使用的内部音素对齐工具生成。

视素与音素。 如图2所示，我们使用一组包含16个视素的集合，与 JALI [Edwards 等人 2016]和 Visemenet [Zhou 等人 2018]相比存在细微差异。图中还展示了对应的英语和中文音素。需要注意的是，我们数据集中音频的音素标签同时包含英语和中文音素，而这些音素被映射到同一组视素。

方法。 在我们的方法中，首先使用一种新颖的音素引导的人脸跟踪方法，从视频的每一帧中提取视素权重
第 4 节详细介绍了该算法。随后，使用视素权重曲线及其对应的音轨训练音频到曲线神经网络模型，相关内容见第 5 节。我们还在第 6 节提出了一种方法，可高效获取高保真视素素材，以便将音频驱动动画制作应用于新的角色。补充视频可在我们的项目页面上找到[1]。

4 音素引导的 3D 面部跟踪

本节介绍一种新颖的算法，用于从带有音素标签的视频中提取视素权重。该算法旨在为每个视频帧拟合视素混合形状的线性组合。具体而言，假设某个主体的中性三维模型为 B_0，

[1]https://linchaobao.github.io/viseme2023/

视素	音素（英语）	音素（中文）	视素	音素（英语）	音素（中文）
AAA	æ, eɪ, aɪ	ai	OHH	əʊ, ɔː, ɔɪ	o
AHH	ɒ, ɑː, ʌ, aʊ	a	RRR	r	r
EH	ɛ, h	ei	SCHWA	ə, ɜː	e, er
FFF	f, v	f	SSH	ʧ, ʃ, ʒ, ʤ	zh, ch, sh
GK	g, k	g, k, h	SSS	ð, s, z	z, c, s
IEE	ɪ, iː, j	i	TTH	θ	j, q, x
LNTD	l, n, t, d	l, n, t, d	UUU	ʊ, uː	iu
MBP	m, b, p	m, b, p	WWW	w	u, v

图 2. 我们的方法中的 16 个视素及其对应的音素。英语音素采用国际音标（IPA）标注，中文音素采用汉语拼音标注。

该主体的视素混合形状集合为 $\mathrm{B} = \{B_1, B_2, \ldots\ldots B_{16}\}$，其中每个混合变形 B_i 表示一个视素形状的三维模型。拟合过程是为每帧 j 求得一组视素权重 $\mathrm{X}^j = \{x_i \mid i \in \{1, 2, \ldots\ldots 16\}\ 0 \leq x_i \leq 1\}$，使得 $S = B_0 + \sum_{i=1}^{16} x_i(B_i - B_0)$与主体在每个视频帧中的三维形状相似。为了在语音信息的指导下解决这一问题，我们先介绍两个准备步骤，再描述拟合算法。

4.1 程序化生成视素权重

第一步是根据音素标签，为音轨的每一帧程序化生成视素权重。我们采用与 JALI [Edwards 等人 2016] 相似的算法，为数据集中的所有音轨生成视素权重曲线。曲线根据明确的规则生成，这些规则涉及每个音素的起始段、持续段和结束段时长，并考虑音素类型和协同发音。尽管音素峰值处各帧的视素权重准确，但人工合成曲线的运动动力学与真实动作相差甚远，因此会导致动画不自然，尤其是对于写实角色。假设生成的每帧视素权重记为 $\mathrm{A}^j = \{a_i \mid i \in \{1、2、\ldots\ldots、16\}\ 0 \leq a_i \leq 1\}$。

4.2 视素混合形状准备

第二步是为数据集中的女演员构建中性三维人脸模型和 16 个视素混合形状。正如第 6.1 节所述，借助相机阵列设备可以大幅简化这一步骤，但我们只能获取大规模语料库数据集中该女演员的视频数据。因此，我们采用算法方法来准备三维模型。首先，我们手动选取该女演员面部处于中性状态的若干视频帧，并使用最先进的三维形变模型拟合算法 [Chai 等人 2022] 重建她的三维人脸模型和纹理贴图。然后，我们使用变形迁移算法 [Sumner 和 Popović 2022]，将艺术家创建的一组视素模板模型的网格形变传递到该女演员的人脸

模型上。对于每个视素模型，我们进一步利用相应音素峰值处的帧，在拉普拉斯约束下优化视素形状 [Sorkine 等人 2004]。最终得到该女演员的一组个性化视素混合形状。

4.3 音素引导的视素参数拟合

我们的拟合算法基于 HiFi3DFace 的 3DMM参数优化框架 [Bao 等人 2021]，，并进行了若干关键调整。首先，我们遍历每个视频片段中的各帧来优化参数，同时考虑光流约束和时间一致性。其次，我们针对音素引导引入了新的损失函数。具体而言，每帧 j 需要最小化的损失函数定义如下：

$$L(\mathcal{X}^j) = w_1 L_{lmk} + w_2 L_{rgb} + w_3 L_{sup} + w_4 L_{act} + w_5 L_{flow} + w_6 L_{diff} + w_7 L_{range}, \quad (1)$$

其中各项定义如下。

地标损失 L_{lmk} 和 *RGB*照片损失 L_{rgb}。这两种损失与 HiFi3DFace [Bao 等人 2021]中的类似。地标损失为

$$L_{lmk}(\mathcal{X}^j) = \frac{1}{|\mathcal{P}|} \sum_{k \in \mathcal{P}} \beta_k ||\text{Proj}(\mathbf{s}_k) - \mathbf{p}_k||_2^2, \quad (2)$$

其中 $k \in$ P 是检测到的地标点，Proj(·) 是将三维顶点 $\mathbf{s}_k$ 投影到二维位置的相机投影，$\mathbf{p}_k$ 是地标点的二维位置 k，而 β_k 是用于控制各点重要性的权重。RGB照片损失定义为

$$L_{rgb}(\mathcal{X}^j) = \frac{1}{|\mathcal{F}|} \sum_{p \in \mathcal{F}} ||I_p - I'_p||_2^2, \quad (3)$$

其中，F 是面部区域内像素的集合，I_p 是像素 p处的 RGB 像素值，$I\prime p$ 是通过可微渲染器 [Genova 等人 2018]得到的渲染像素值。所有像素值均归一化到范围 $[0, 1]$。

音素引导的抑制损失 L_{sup} 和激活损失 L_{act}。我们使用第4.1节中根据音素标签程序生成的视素曲线来抑制和激活相应的视素。需要注意的是，程序生成曲线的运动动力学并非我们所需，但曲线准确描绘了每一帧中需要抑制或激活的视素。为利用这些信息，我们首先为每一帧 j确定两组视素索引 $\mathbf{I}_{sup}^j$ 和 $\mathbf{I}_{act}^j$，如下所示：$\mathbf{I}_{sup}^j$由满足以下条件的视素索引i 组成： a_i 未出现在 j 所有相邻帧的前 m 个最大值中；$\mathbf{I}_{act}^j$由满足以下条件的视素索引 i组成： a_i出现在当前帧 j的前 n个最大值中。我们在方法中使用 $m = 3$和 $n = 2$。其思路是，如果某个视素在当前帧中显著出现，就应当激活它；另一方面，如果某个视素在所有相邻帧中都不显著出现，就应当抑制它。相应的损失项定义如下：

$$L_{sup}(\mathcal{X}^j) = \frac{1}{|\mathcal{I}_{sup}^j|} \sum_{i \in \mathcal{I}_{sup}^j} x_i^2, \quad (4)$$

$$L_{act}(\mathcal{X}^j) = -\frac{1}{|\mathcal{I}_{act}^j|} \sum_{i \in \mathcal{I}_{act}^j} x_i^2. \quad (5)$$

光流损失 L_{flow}和时间一致性损失 L_{diff}。对于每一帧 j，我们计算帧 $j-1$ 和 j之间的双向光流，并通过前向-后向一致性检查筛选有效对应关系 [Bao 等人 2014]。光流损失L_{flow}定义如下：

$$L_{flow}(\mathcal{X}^j) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} ||\text{Proj}(\mathbf{s}_k^j) - (\text{Proj}(\mathbf{s}_k^{j-1}) + \mathbf{u}_k)||_2^2, \quad (6)$$

其中， $k \in$ K 是从帧 $j-1$ 到 j的有效顶点对应关系，Proj(·) 表示三维顶点 $\mathbf{s}_k$ 到二维位置的相机投影，$\mathbf{u}_k$ 是该顶点的二维光流位移。此外，定义一个针对视素权重的显式时间一致性损失 L_{diff}，如下所示：

$$L_{diff}(\mathcal{X}^j) = \frac{1}{16} \sum_{i=1}^{16} (x_i^j - x_i^{j-1})^2. \quad (7)$$

值域损失 L_{range}与优化。待求解的视素权重需要满足范围 $[0, 1]$ 内的盒约束。由于我们采用 HiFi3DFace [Bao 等人 2021], 的优化框架，该框架使用迭代优化器更新参数，因此我们将盒约束转换为值域损失 L_{range}，并在优化后截断最终结果，以满足盒约束。值域损失 L_{range} 定义如下：

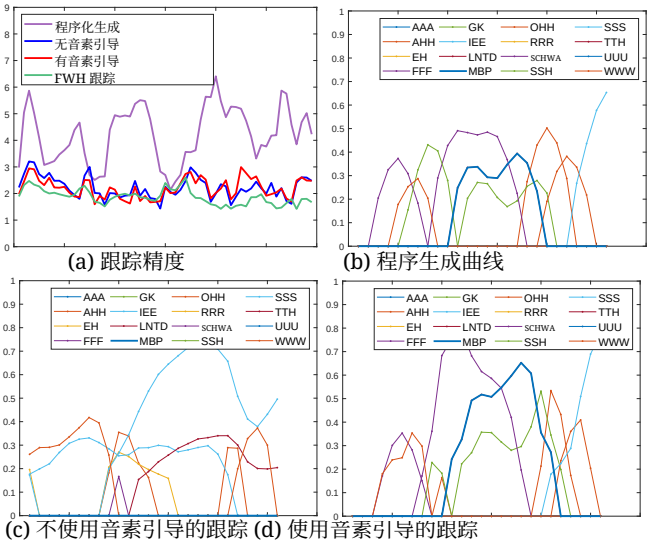
$$L_{range}(\mathcal{X}^j) = \frac{1}{|\mathcal{I}_{upper}|} \sum_{i \in \mathcal{I}_{upper}} (x_i - 1)^2 + \frac{1}{|\mathcal{I}_{lower}|} \sum_{i \in \mathcal{I}_{lower}} (0 - x_i)^2, \quad (8)$$

其中，$\mathbf{I}^{upper}$ 是参数为 $x_i > 1$的视素索引集合，$\mathbf{I}_{lower}$是当前迭代中参数为 $x_i < 0$ 的视素索引集合。请注意，在每次迭代中，都会根据当前估计的参数重新计算这两个集合 $\mathbf{I}^{upper}$ 和 $\mathbf{I}_{lower}$。

4.4 结果与评估

实现细节。 与 HiFi3DFace [Bao 等人 2021]的方法类似，相机姿态和纹理参数与视素权重一同估计。公式（1）中的权重设为 $w_1 = 0.8, w_2 = 1.0,$ $w_3 = 800$ $w_4 = 150$ $w_6 = 300$ $w_7 = 100$, , $w_5 = 1.0,$, 。请注意，直接作用于视素权重的损失项权重比其他项大两个数量级，因为视素权重的平均值通常远小于 1。可微渲染器输出的渲染图像尺寸为 1280×720。第 4.1 节中的程序生成的权重用于初始化视素权重。我们使用 TensorFlow 中的 Adam 优化器[Kingma 和 Ba 2014] ，对每一帧运行 250 次迭代，学习率为 0.1，并每 10 次迭代按指数衰减。对于每个视频片段（1 个话语），我们先按顺序逐帧处理，然后在第二遍中按相反顺序后向处理这些帧。

跟踪精度和曲线动态。 我们首先评估视素曲线提取算法的跟踪精度。图3左上方的图展示了平均值比较的一个示例



(a) 跟踪精度 (b) 程序生成曲线

(c) 不使用音素引导的跟踪 (d) 使用音素引导的跟踪

图3. 跟踪精度与视素曲线对比。左上：嘴部区域面部关键点平均误差的对比。其余三个图展示了使用我们算法的不同版本获得的视素曲线。详细说明请参见第4.4节正文。

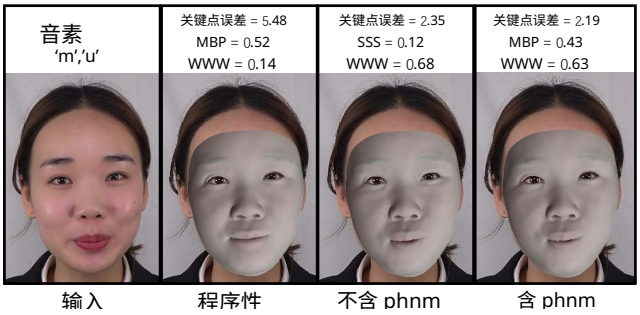


图4. 从视频中提取视素权重时音素引导的影响。图中展示了关键点误差以及激活程度最高的两个视素权重。我们的音素引导跟踪结果误差最低，并能正确激活双唇音素“MBP”。

嘴部关键点误差。我们的跟踪算法使用视素混合形状，无论是否采用音素引导，其精度都与使用双线性面部模型FacewareHouse（简称“fwh”）的跟踪算法相当 [Cao et al. 2014]。虽然我们算法中的音素引导不会影响跟踪精度，但会带来截然不同的视素曲线，如图3底部的两个图所示。请注意，尽管采用音素引导获得的视素曲线（右下）看起来与程序生成的视素曲线（右上）相似，但跟踪精度却有了显著提升（见左上图）。从曲线中还可以看出，我们跟踪得到的视素曲线形状不那么规则，变化也更为细致，因此更贴近真实运动动态，并能减少最终动画的人工合成感。

音素引导的影响。我们进一步研究音素引导对提取的视素曲线的影响。比较图3底部的两幅图可以发现，在某些帧中，激活的视素与有无音素引导时得到的结果有很大差异。图4展示了这样一帧的示例。由于该帧的音素标签表示从“m”过渡到“u”，程序生成规则主要激活视素“MBP”和“WWW”。由此产生的关键点误差很大，因为生成规则完全忽略了图像信息。另一方面，尽管未使用音素引导的跟踪所产生的关键点误差小得多，却错误地激活了视素“SSS”，而非“MBP”。在音素引导下，我们的跟踪算法激活了正确的视素，并取得了令人满意的关键点误差。请注意，虽然在原始跟踪视频中，激活正确视素与错误视素之间的差异难以察觉，但应用于新角色时，效果会大不相同。例如，视素“MBP”是双唇音音素的发音动作，需要先做出准备动作：上下嘴唇相互挤压以闭合口腔，随后释放气流。这与视素“SSS”大不相同；“SSS”是咝音音素的发音动作，需要收窄下颌并让气流从牙齿间通过。仅凭图像可能难以观察到二者的差异，但经验丰富的艺术家会仔细区分它们，尤其是在需要进行艺术夸张的动画中 [Osipa 2010]。

5 学习音频到曲线的映射

我们从数据集中的所有视频（12，000 段话语）中提取视素曲线，并将数据集划分为训练集（10，000 段话语）和验证集（2，000 段话语）。学习的目标是将每条音轨$\mathbf{U}_0 \in \mathbb{R}^{T_U}$ 映射为一系列视素权重$\mathbf{Y} \in \mathbb{R}^{T\times 16}$，使其尽可能接近从视频中提取的视素权重$\mathbf{X} \in \mathbb{R}^{T\times 16}$（即，$\{\mathbf{X}^j\}_{j=1}^{T}$）。我们的数据集存在一个主要问题：所有话语都来自同一个人，且其音调单一。因此，采用利用额外数据预训练的音频特征，对于使所学习的映射能够泛化到未见过的说话人至关重要。我们广泛研究了各种音频特征，发现由 Baevski 等人 2020] 在涵盖 53 种语言的大规模语料库 [上预训练的 Wav2Vec2 [Conneau 等人 2020] 在我们的设置中效果最佳。此外，我们还研究了各种序列到序列映射骨干网络，包括时序卷积网络（TCN）、Transformer、长短期记忆网络等，并发现双向长短期记忆网络的表现最佳。我们在第 5.1 节介绍性能最佳的模型，并在第 5.2 节介绍实验研究。

5.1 映射模型

音频编码器。 我们将 Wav2Vec2 的网络结构表示为 $\{\phi, \varphi\}$，其中 ϕ表示时序卷积网络（TCN） $n_1 = 7$层的集合， φ 表示多头自注意力层 $n_2 = 24$层的集合，它将音频特征转换为上下文化语音表示 $\mathbf{U}_1 \in \mathbb{R}^{T_W\times D_W}$，即，$\mathbf{U}_1 = \varphi(\phi(\mathbf{U}_0))$。我们遵循 Wav2Vec2 的通用微调策略，即，固定 ϕ 的预训练权重，并允许微调 φ的权重。接下来，我们添加一个随机初始化的线性层 f_{proj}，将特征维度从 $D_W = 1024$

表 1. 不同音频特征的消融研究。所列数值为乘以 10^2 后的 L1 重建误差。音量 表示在验证集上对测试音频的音量进行增强后所执行的泛化性测试。音高、速度和噪声的情况也相同，分别对应音高、速度 和噪声。

特征 \ 指标	验证	音量	音高	速度	噪声
FBank	3.21	3.32	4.59	3.39	4.42
LPC	5.26	5.27	6.50	5.35	5.93
PPG [Huang 等人 2021]	3.01	3.01	3.07	7.50	3.13
Wav2Vec2[1] [Baevski 等人 2020]	1.88	1.89	2.48	2.29	2.47
Wav2Vec2[2] [Baevski 等人 2020]	1.82	1.83	2.55	2.19	2.27
Wav2Vec2$_3$ （我们的方法） [Conneau 等人 2020]	**1.77**	**1.77**	**2.44**	**2.06**	**2.09**

表 2. 不同解码器骨干网络的消融研究。所列数值为乘以 10^2后的 L1 重建误差。

解码器 \ 指标	验证	音量	音高	速度	噪声
TCN	1.77	1.78	2.51	2.11	2.11
LSTM	1.77	1.78	2.50	2.12	2.16
Transformer	1.79	1.80	2.54	2.13	2.14
双向长短期记忆网络（我们的方法）	**1.77**	**1.77**	**2.44**	**2.06**	**2.09**

映射到 $D = 512$，并添加一个插值层 f_{interp}，使特征的帧率与视素曲线对齐。因此，音频特征可表示为 $\mathbf{U}_2 = f_{interp}(f_{proj}(\mathbf{U}_1)) \in \mathbb{R}^{T\times D}$。与 FaceFormer [Fan et al. 2022],类似，我们也采用周期位置编码（PPE）来增强音频特征，即 $\mathbf{U}_3 = \mathbf{U}_2 + \mathbf{P}$，其中 P 是 PPE。

视素曲线解码器。 我们采用单层双向长短期记忆网络，后接一个全连接层，将音频特征映射为 16 维的视素权重，即$\mathbf{Y} = f_{FC}(f_{BLSTM}(\mathbf{U}_3))$。模型通过跟踪得到的视素权重$\mathbf{X}$ 与预测视素权重$\mathbf{Y}$之间的 L1 损失进行训练。

训练。 训练阶段采用 SpecAugment [Park 等人 2019] 的方法。我们使用 Adam 优化器，并以固定学习率 10^{-5} 在每次训练迭代中优化一个话语。模型在训练集上训练 100 个轮次。

5.2 结果与评估

鲁棒性评估。 我们在验证集上通过 L1 度量评估预测视素权重的重建损失。此外，为了评估该算法对具有不同音高、音量、速度和噪声的音频的通用性，我们通过随机调整音高（范围为 -3 至 3 个半音）、音量（范围为 -10 dB 至 +10 dB）和速度（范围为 0.7 至 1.3），并添加振幅小于 0.005 的额外高斯噪声，对验证集进行扩充。

音频特征消融研究。 如表 1 所示，我们研究了多种音频特征，包括滤波器组（FBank）、线性预测编码（LPC）、音素后验图（PPG），以及在不同数据集和网络结构上训练的 Wav2Vec2。Wav2vec2$_1$表示基础模型 [Baevski 等人 2020] ($n_1 = 7, n_2 = 12$)，该模型在 Librspeech 数据集上进行了预训练 [Panayotov 等人 2015]，该数据集包含 960 小时的英语语音音频。它也是所采用的预训练

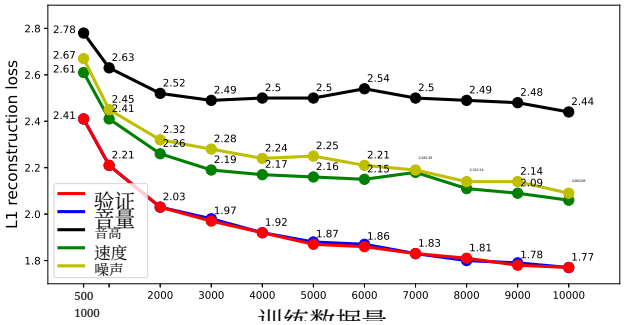


图 5. 我们的模型在不同规模的数据集上训练后，在验证集上的 L1 重建损失。所有实验均训练 10^6 次迭代。

表 3. 与最先进的深度模型的比较。所列数值为 L1 重建误差乘以 10^2。

方法 \ 指标	验证	音量	音高	速度	噪声
Karras 等人 [2017]	4.11	4.15	5.79	7.42	6.34
Huang 等人 [2021]	2.38	2.38	2.45	7.43	2.69
FaceFormer [Fan 等人 2022]	2.10	2.10	2.69	2.51	2.89
Ours	**1.77**	**1.77**	**2.44**	**2.06**	**2.09**

模型，应用于 FaceFormer [Fan 等人 2022]；Wav2Vec2$_2$ 是在同一数据集上使用更大的模型（$n_1 = 7, n_2 = 24$）进行预训练的；Wav2Vec2$_3$ 则是本文采用的模型，该模型在包含 53 种语言的数据集上进行了预训练。维度为 $\mathbb{R}^{T\times 128}$ 的 PPG 特征采用与 Huang 等人 [2021]相同的方法生成。与 FBank 和 LPC 等手工设计的特征相比，从 Wav2Vec2 提取的自监督预训练深度特征能够显著提升性能。具体而言，对比 Wav2Vec2$_2$ 和 Wav2Vec2$_3$的结果，可以发现，在多语言数据集上进行预训练能够进一步提升这项任务的性能。

解码器骨干网络消融研究。 为公平比较不同的解码器骨干网络，我们为每个被比较的解码器保持相同的层数和隐藏状态维度，即 层数 $l = 1$ 和特征维度 $D = 512$。表 2 中的结果表明，任何配备 Wav2Vec2 音频特征的解码器都足以生成高质量的视素曲线输出。同时，双向长短期记忆网络由于在聚合邻域信息方面表现出色，略优于其他解码器。请注意，在本消融研究中，我们采用了 Transformer 的编码器网络结构，利用多头自注意力层提取特征。相比之下，FaceFormer [Fan 等人 2022]，采用基于 Transformer 的解码器，这是一种带掩码以实现跨模态对齐的自回归模型。

训练数据集规模。我们进一步在图 5 中展示了模型在使用不同规模的数据集进行训练时的表现。记录表明，使用全部数据时，我们的模型表现最佳；而使用较少的数据（如 3000 个话语）时，模型也能取得相当不错的结果，并在数据成本与性能之间实现良好平衡。



图 6. 我们简易的摄像机阵列系统由 12 台同步的 4K/60帧视频摄像机和 10 盏方形 LED灯组成。右侧展示了采集到的图像示例。

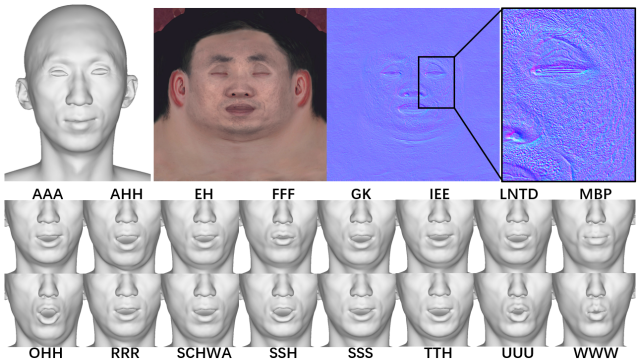


图 7. 通过视素扫描获取的面部资源和视素混合形状示例。动画示例请参见补充视频。

与最先进方法的比较。我们将三种近期基于深度学习的音频驱动语音动画模型 [Fan 等，2022；Huang 等人，2021；Karras 等人， 2017] 纳入我们的设置，并将其表现与我们的模型进行比较。具体而言，我们修改了这些模型网络中的最后一个全连接层，使其预测 16 维视素权重，然后使用我们的训练数据集对模型训练 100 个轮次。结果见表 3。我们的模型表现最佳，尤其是在处理音高和速度方面不同的音频变化时。

泛化到未见过的说话人和其他语言。尽管我们的训练数据集仅包含一位说话人的音频，且大部分为中文，但得益于强大的预训练音频特征模型，我们的模型能够很好地泛化到未见过的说话人和其他语言。补充视频中展示了各种动画示例，其中包括由不同说话人以不同语言驱动的动画。

6 语音动画制作

有了上述训练好的音频到曲线模型，只要为每个角色准备好合适的视素混合形状，就可以根据音轨自动生成语音动画。由于每个视素混合形状都与音素有明确的对应关系，为新角色准备混合形状对美术人员来说十分友好。然而，手工雕刻高质量的视素混合形状仍然需要大量工作。本节介绍一种通过扫描演员的视素形状，高效获取高保真混合形状资源的方法。扫描得到的混合形状资源还可以进一步转换为用于骨骼动画的骨骼姿势资源，这类资源适用于对资源要求敏感的应用场景。

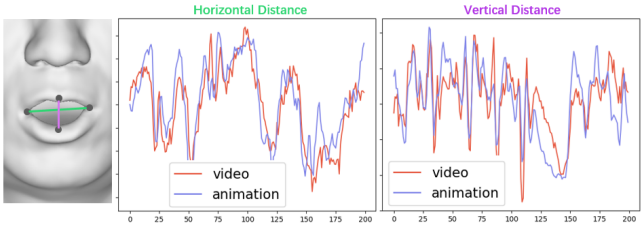


图 8. 真实视频与动画中的唇部运动曲线对比。我们记录了真实视频和动画中两对关键点之间水平及垂直距离随时间变化的序列。对比结果表明，生成的动画中的唇部运动与真实唇部运动十分相似。

6.1 视素混合形状扫描

为了获取演员的高保真视素混合形状，我们使用一个简单的相机阵列系统（如图 6 所示），在演员执行图 2 中音素的发音动作时，采集其多视角视频。我们选取与 16 个视素相对应的视频帧，并使用多视角立体视觉算法 [Cernea 2020] 重建三维人脸模型。随后，使用非刚性配准算法 [Bouaziz 等人 2016] 将三维人脸模型对齐并变形为固定拓扑网格模板。图 7 展示了所得面部资源以及拓扑一致的视素混合形状示例。对于每个视素混合形状，我们进一步采用类似于 Beeler 等人 [2010] 的方法提取纹理贴图和法线贴图，以获得动态纹理贴图 [Li 等人 2020]。我们使用 Bao 等人 [2021] 所述的方法，为每个视素混合形状添加眼球和口腔内部等附件。借助这些面部资源，音频到曲线模型生成的视素权重可以直接应用于生成高质量、逼真的面部动画（参见补充视频）。

我们还可以通过比较演员录制的语音视频与生成的动画中的唇部运动，进一步验证动画的运动准确性。图 8 展示了一个对比片段示例。结果表明，生成的动画中的唇部运动与真实的唇部运动十分相似，包含丰富而逼真的运动细节。请注意，应用于生成的动画的视素曲线是根据我们大规模语料库数据集中的女演员数据学习得到的，而真实视频则由新演员录制。该实验表明，将我们学到的视素动态应用于新角色时，能够生成逼真的唇部动画，并具有良好的泛化性能。

该方法还可进一步用于生成带有面部表情的语音动画。我们扫描演员在以特定表情（如微笑或愤怒）发音时的一组表情视素。随后，将与先前结果相同的视素曲线应用于这些表情视素。图 9 展示了具有两种不同表情的动画快照。动画视频请参阅补充材料。

6.2 重定向与骨骼动画

为了让无法从演员身上扫描的角色动起来，我们使用变形迁移算法 [Sumnerand Popović 2004] 来



图 9. 带表情的语音动画。视素混合形状是从带有不同表情的演员身上扫描得到的，而视素曲线与不带表情的动画相同。视频见补充材料。

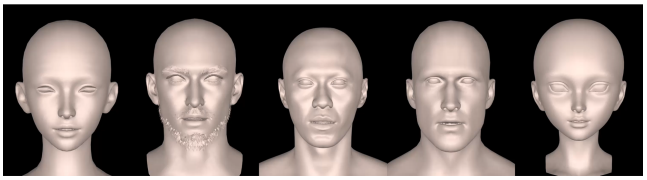


图 10. 语音驱动动画结果快照。这些角色由音频到曲线模型预测的同一组视素曲线驱动。视频见补充材料。

根据艺术模板或扫描得到的角色生成视素混合形状。在 Unity 或虚幻引擎等实时图形引擎中，许多资源受限的应用更倾向于使用基于骨骼的动画，而不是基于混合形状的动画。我们使用 SSDR算法 [Le and Deng 2012] 将每个视素混合形状转换为一组骨骼姿势参数，然后直接应用预测的视素曲线来混合骨骼姿势资源，从而生成骨骼动画。姿势混合技术只需对平移/缩放参数进行线性插值，并对基于四元数的旋转参数进行球面线性插值（SLERP）；实时图形引擎通常都提供这一技术，例如虚幻引擎中的 Pose Asset Animation [EpicGames 2022] 或 Unity 引擎中的 Direct Blend Tree [Technologies 2022]。

6.3 结果

在补充视频中，我们展示了将我们的结果与其他最先进的音频驱动三维面部动画方法所报告结果进行的视觉比较 [Cudeiro 等，2019；Fan 等，2022；Karras 等，2017；Richard 等，2021；Taylor 等，2017]。我们的结果中的唇部运动显然更加逼真、自然。请注意，由于动画机制和面部绑定不同，无法与其他工作进行定量比较，这一点在先前的工作中也曾指出 [Edwards 等，2016；Zhou 等，2018]。此外，我们还在补充视频中展示了不同角色的各种语音驱动动画。图 10 展示了一个动画快照：在为每个角色提供个性化视素混合形状的情况下，使用我们的音频到曲线模型预测的同一组视素曲线驱动不同角色进行动画。

7 结论

我们介绍了一种语音动画方法，可从音频生成逼真的唇形同步三维面部动画。我们的方法以基于视素的方式将输入音频映射为动画曲线参数空间，既便于艺术家使用，也可轻松应用于新角色。我们展示了该方法制作的多种风格的高质量动画，包括写实风格和卡通风格的角色。该方法可作为一种高效的语音动画解决方案，适用于游戏制作或人工智能驱动的数字人应用等多种场景。

局限性与未来工作。 我们的方法无法捕捉说话时舌头的运动动力学。我们为每个视素混合形状关联一个静态的牙齿和舌头姿势，并依靠视素曲线生成牙齿和舌头的动画。尽管在大多数情况下，由于口腔内光线较暗，口腔内部的动画并不明显，但有时不自然的运动会显露出来，尤其是在口腔内部受到照明时。我们注意到，近期有一项工作 [Medina 等人 2022]针对这一问题进行了研究，未来或许值得进一步探讨。

附录

A 网络架构详情

A.1 提出的映射模型

所提出的映射模型的网络架构如图 14 所示。我们省略了每个时序卷积网络和全连接层中的 ReLU 激活函数。网络以形状为 $\mathbb{R}^{T_U}$ 的音频 $\mathbf{U}_0$ 作为输入，并输出预测视素 $\mathbf{Y} \in \mathbb{R}^{T\times 16}$。

A.2 解码器骨干网络的消融研究

在图 11、12 和 13 中，我们还展示了消融研究中所比较的解码器的详细实现。请注意，在本研究中，我们采用多头自注意力机制作为基于 Transformer 的解码器。

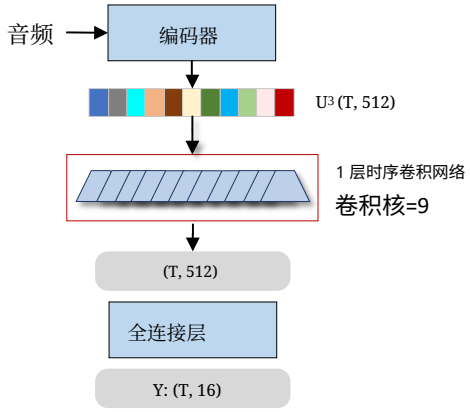


图 11. 消融研究中的基于时序卷积网络的解码器。

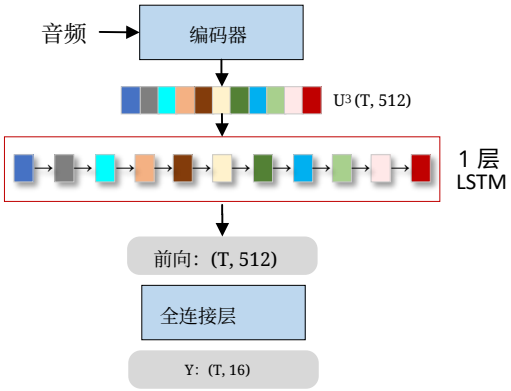


图 12. 消融研究中的基于长短期记忆网络的解码器。

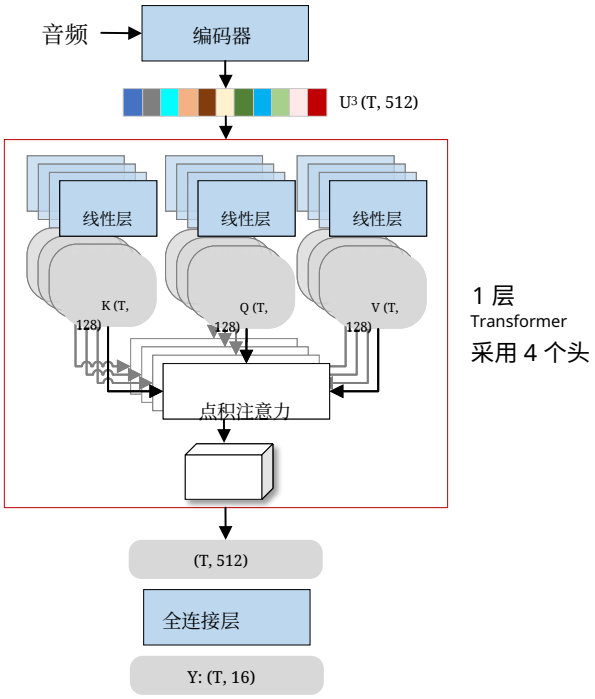


图 13. 消融研究中的基于 Transformer 的解码器。

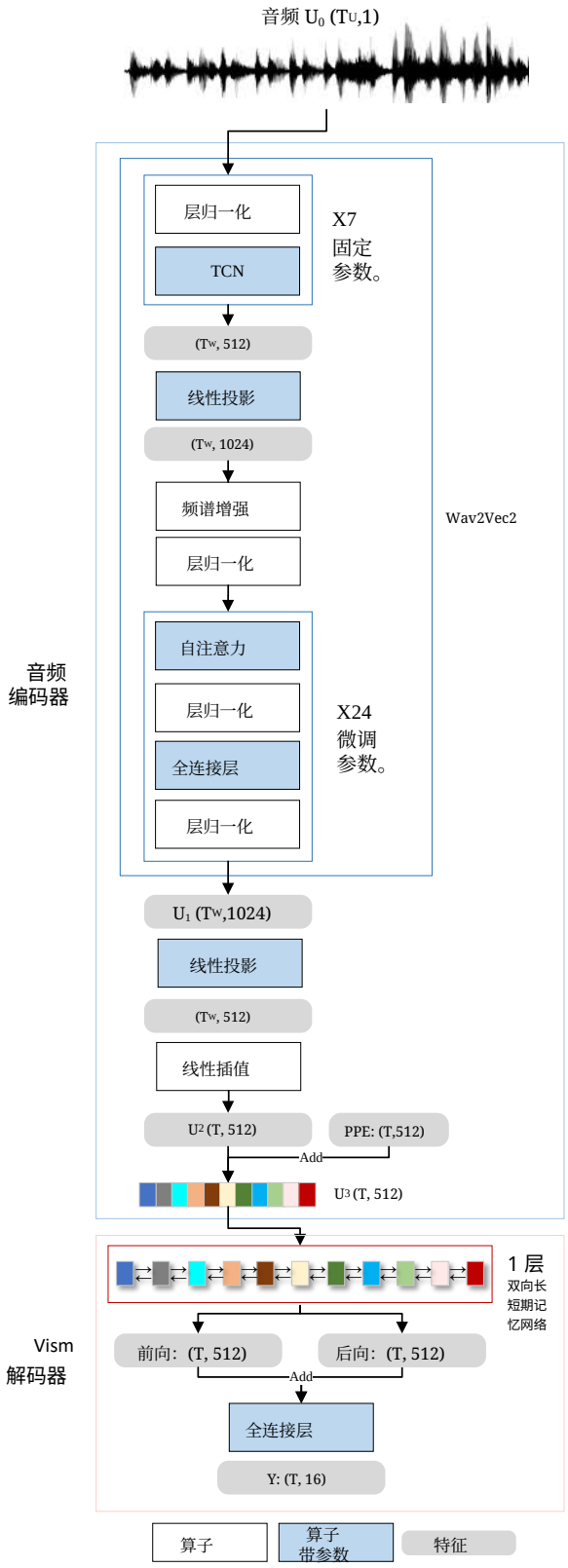


图 14. 我们的音频到曲线映射模型的架构。

参考文献

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020年。wav2vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展 33（2020年），12449–12460。Linchao Bao、Xiangkai Lin、Yajing Chen、Haoxian Zhang、Sheng Wang、Xuefei Zhe、Di Kang、Haozhi Huang、Xinwei Jiang、Jue Wang、Dong Yu 和 Zhengyou Zhang。2021年。基于RGB-D自拍照创建高保真3D数字人头部。*ACM Trans. Graph.* 41，1（2021年），1–21。Linchao Bao、Qingxiong Yang 和 Hailin Jin。2014年。用于大位移光流的快速保边PatchMatch。收录于*CVPR*论文集。3534–3541。Thabo Beeler、Bernd Bickel、Paul Beardsley、Bob Sumner 和 Markus Gross。2010年。高质量的单次面部几何捕捉。*ACM Trans. Graph. (Proc. SIGGRAPH)* 29，4（2010年），1–9。Thabo Beeler、Fabian Hahn、Derek Bradley、Bernd Bickel、Paul Beardsley、Craig Gotsman、Robert W Sumner 和 Markus Gross。2011年。使用锚帧的高质量被动式面部表演捕捉。*ACM Trans. Graph. (Proc. SIGGRAPH)* 30，4（2011年），1–10。Sofien Bouaziz、Andrea Tagliasacchi、Hao Li 和 Mark Pauly。2016年。实时非刚性配准的现代技术与应用。收录于*ACM SIGGRAPH Asia 2016*课程。1–25。Chen Cao、Derek Bradley、Kun Zhou 和 Thabo Beeler。2015年。实时高保真面部表演捕捉。*ACM Trans. Graph. (Proc. SIGGRAPH)* 34，4（2015年），1–9。Chen Cao、Menglei Chai、Oliver Woodford 和 Linjie Luo。2018年。通过学习动态刚性先验实现稳定的实时人脸跟踪。*ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 37，6（2018年），1–11。Chen Cao、Yanlin Weng、Stephen Lin 和 Kun Zhou。2013年。用于实时面部动画的3D形状回归。*ACM Trans. Graph. (Proc. SIGGRAPH)* 32，4（2013年），1–10。Chen Cao、Yanlin Weng、Shun Zhou、Yiying Tong 和 Kun Zhou。2014年。Facewarehouse：用于视觉计算的3D面部表情数据库。*IEEE*可视化与计算机图形学汇刊20，3（2014年），413–425。Dan Cernea。2020年。OpenMVS：多视图立体重建库。（2020年）。https: //cdcseacave.github.io/openMVS。Zenghao Chai、Haoxian Zhang、Jing Ren、Di Kang、Zhengzhuo Xu、Xuefei Zhe、Chun Yuan 和 Linchao Bao。2022年。REALY：重新思考3D人脸重建的评估。收录于欧洲计算机视觉会议论文集。Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 Michael Auli。2020年。用于语音识别的无监督跨语言表征学习。*arXiv*预印本 *arXiv:2006.13979*（2020年）。Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael J Black。2019年。3D说话风格的捕捉、学习与合成。收录于*CVPR*论文集。10101–10111。Pif Edwards、Chris Landreth、Eugene Fiume 和 Karan Singh。2016年。JALI：一种以动画师为中心、用于表现性唇形同步的视素模型。*ACM Trans. Graph. (Proc. SIGGRAPH)* 35，4（2016年），1–11。Bernhard Egger、William AP Smith、Ayush Tewari、Stefanie Wuhrer、Michael Zollhoefer、Thabo Beeler、Florian Bernard、Timo Bolkart、Adam Kortylewski、Sami Romdhani、Christian Theobalt、Volker Blanz 和 Thomas Vetter。2020年。3D可变形人脸模型——过去、现在与未来。*ACM Trans. Graph.*（2020年）。EpicGames。2022年。动画姿势资源。检索于2022年10月31日，网址：https: //
d
o
cs.
unrealengine.com/4.27/en-US/AnimatingObjects/SkeletalMeshAnimation/AnimPose/。Yingruo Fan、Zhaojiang Lin、Jun Saito、Wenping Wang 和 Taku Komura。2022年。FaceFormer：基于Transformer的语音驱动3D面部动画。收录于*IEEE/CVF*计算机视觉与模式识别会议论文集。18770–18780。Pablo Garrido、Levi Valgaerts、Chenglei Wu 和 Christian Theobalt。2013年。从单目视频重建精细的动态面部几何。*ACM Trans. Graph.* 32，6（2013年），158:1–158:9。Kyle Genova、Forrester Cole、Aaron Maschinot、Aaron Sarna、Daniel Vlasic 和 William T Freeman。2018年。3D可变形模型回归的无监督训练。收录于*CVPR*论文集。IEEE，8377–8386。Huirong Huang、Zhiyong Wu、Shiyin Kang、Dongyang Dai、Jia Jia、Tianxiao Fu、Deyi Tuo、Guangzhi Lei、Peng Liu、Dan Su 等。2020年。使用音素后验图进行说话人无关、多语言/混合语言的语音驱动说话头像生成。*arXiv*预印本 *arXiv:2006.11610*（2020年）。Huirong Huang、Zhiyong Wu、Shiyin Kang、Dongyang Dai、Jia Jia、Tianxiao Fu、Deyi Tuo、Guangzhi Lei、Peng Liu、Dan Su 等。2021年。使用音素后验图进行说话人无关、多语言/混合语言的语音驱动说话头像生成。收录于*2021*年亚太信号与信息处理协会年度峰会暨会议（*APSIPA ASC*）。IEEE，1433–1437。Gregor A Kalberer、Pascal Müller 和 Luc Van Gool。2002年。使用视素空间进行语音动画。收录于视觉、建模与可视化研讨会。463–470。Tero Karras、Timo Aila、Samuli Laine、Antti Herva 和 Jaakko Lehtinen。2017年。通过联合端到端学习姿势与情绪实现音频驱动的面部动画。*ACM Trans. Graph. (Proc. SIGGRAPH)* 36，4（2017年），1–12。Diederik P Kingma 和 Jimmy Ba。2014年。Adam：一种随机优化方法。*arXiv*预印本 *arXiv:1412.6980*（2014年）。Samuli Laine、Tero Karras、Timo Aila、Antti Herva、Shunsuke Saito、Ronald Yu、Hao Li 和 Jaakko Lehtinen。2017年。使用深度卷积神经网络进行生产级面部表演捕捉。载于*ACM SIGGRAPH/Eurographics*计算机动画研讨会。ACM，10。Binh Huy Le 和 Zhigang Deng。2012年。结合刚性骨骼的平滑蒙皮分解。*ACM Trans. Graph.*（*SIGGRAPH Asia*论文集）31，6（2012年）。J. P. Lewis、Ken Anjyo、Taehyun Rhee、Mengjie Zhang、Fred Pighin 和 Zhigang Deng。2014年。混合形状面部模型的实践与理论。载于*Eurographics 2014* - 前沿技术报告。https://doi.org/10.2312/egst.20141042 Jiaman Li、Zhengfei Kuang、Yajie Zhao、Mingming He、Karl Bladin 和 Hao Li。2020年。基于单次扫描的动态面部资产与绑定生成。*ACM Trans. Graph.* 39，6（2020年）215 1( )，– 。Jingying Liu、Binyuan Hui、Kun Li、Yunke Liu、Yu-Kun Lai、Yuxiang Zhang、Yebin Liu 和 Jingyu Yang。2021年。用于语音驱动面部动画的几何引导稠密透视网络。*IEEE*可视化与计算机图形学汇刊（2021年）。Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。2017年。Montreal Forced Aligner：使用 Kaldi 的可训练文本-语音对齐。。载于*Interspeech*，第2017卷。498–502。Salvador Medina、Denis Tome、Carsten Stoll、Mark Tiede、Kevin Munhall、Alexander G Hauptmann 和 Iain Matthews。2022年。语音驱动的舌部动画。载于*CVPR*论文集。20406–20416。Jason Osipa。2010年。别再盯着看：面部建模与动画的正确做法。John Wiley & Sons。Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。2015年。LibriSpeech：基于公版有声读物的自动语音识别语料库。载于*2015*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，5206–5210。Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le。2019年。SpecAugment：一种用于自动语音识别的简易数据增强方法。*arXiv*预印本 *arXiv:1904.08779*（2019年）。Alexander Richard、Michael Zollhöfer、Yandong Wen、Fernando De la Torre 和 Yaser Sheikh。2021年。MeshTalk：通过跨模态解耦实现语音驱动的三维面部动画。载于*ICCV*论文集。1173–1182。Fuhao Shi、Hsiang-Tao Wu、Xin Tong 和 Jinxiang Chai。2014年。利用单目视频自动获取高保真面部表演。*ACM Trans. Graph.*（*SIGGRAPH Asia*论文集）33，6（2014年），1–13。Olga Sorkine、Daniel Cohen-Or、Yaron Lipman、Marc Alexa、Christian Rössl 和 H- P Seidel。2004年。拉普拉斯曲面编辑。载于*Eurographics/ACM SIGGRAPH*几何处理研讨会论文集。175–184。Robert W Sumner 和 Jovan Popović。2004年。三角网格的变形传递。*ACM Trans. Graph.*（*SIGGRAPH*论文集）23，3（2004年），399–405。Sarah Taylor、Taehwan Kim、Yisong Yue、Moshe Mahler、James Krahe、Anastasio Garcia Rodriguez、Jessica Hodgins 和 Iain Matthews。2017年。通用语音动画的深度学习方法。*ACM Trans. Graph.*（*SIGGRAPH*论文集）36，4（2017年），1–11。Sarah L Taylor、Moshe Mahler、Barry-John Theobald 和 Iain Matthews。2012年。视觉语音的动态单元。载于*SCA*论文集。275–284。Unity Technologies。2022年。混合树。访问于2022年10月31日，网址：https://docs.unity3d.com/Manual/BlendTree-DirectBlending.html Justus Thies、Michael Zollhofer、Marc Stamminger、Christian Theobalt 和 Matthias Nießner。2016年。Face2face：RGB视频的实时面部捕捉与重演。载于*CVPR*论文集。2387–2395。Thibaut Weise、Sofien Bouaziz、Hao Li 和 Mark Pauly。2011年。实时的基于表演的面部动画。*ACM Trans. Graph.*（*SIGGRAPH*论文集）30，4（2011年），1–10。Longwen Zhang、Chuxiao Zeng、Qixuan Zhang、Hongyang Lin、Ruixiang Cao、Wei Yang、Lan Xu 和 Jingyi Yu。2022年。用于制作的、视频驱动的神经物理学面部资产。*ACM Trans. Graph.*（*SIGGRAPH Asia*论文集）（2022年）。Yang Zhou、Zhan Xu、Chris Landreth、Evangelos Kalogerakis、Subhransu Maji 和 Karan Singh。2018年。VisemeNet：音频驱动的动画师中心语音动画。*ACM Trans. Graph.*（*SIGGRAPH*论文集）37，4（2018年），1–10。Michael Zollhöfer、Justus Thies、Pablo Garrido、Derek Bradley、Thabo Beeler、Patrick Pérez、Marc Stamminger、Matthias Nießner 和 Christian Theobalt。2018年。单目3D人脸重建、跟踪及应用的最新进展。载于*Computer Graphics Forum*，第37卷。Wiley Online Library，523–550。Gaspard Zoss、Thabo Beeler、Markus Gross 和 Derek Bradley。2019年。用于面部表演捕捉的精确无标记下颌跟踪。*ACM Trans. Graph.*（*SIGGRAPH*论文集）38，4（2019年），1–8。