Computer Science and Application 计算机科学与应用, 2024, 14(8), 168-178
Published Online August 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.148174

Hans汉斯

HBF Talk：语音驱动的3D面部动画合成研究

王文祥[1]，王少波[1]，智　宇[1]，陈　昂[1,2]

[1]温州大学计算机与人工智能学院，元宇宙与人工智能研究中心，浙江　温州
[2]温州大学元宇宙与人工智能研究院，浙江　温州

收稿日期：2024年7月10日；录用日期：2024年8月8日；发布日期：2024年8月19日

摘　要

近年来，语音驱动的3D面部动画得到了广泛的研究，虽然先前的工作可以从语音数据中生成连贯的3D面部动画，但是由于视听数据的稀缺性，生成的3D面部动画缺乏真实感和生动性，嘴唇运动的准确性不高。为了提高嘴唇运动的准确性和生动性，本文提出了一种新的模型HBF Talk (端到端的神经网络模型)，通过使用Hu BERT (Hidden-Unit BERT)预训练模型对语音数据进行特征提取和编码，引入Flash模块对提取到的语音特征表示进行进一步的编码，获得更为丰富的语音特征上下文表示，最后使用带偏置的跨模态Transformer解码器进行解码。本文进行了定量和定性实验，并与现有的基线模型进行比较，显示本文HBF Talk模型具有更好的性能，提高了语音驱动的嘴唇运动的准确性和生动性。

关键词

Hu BERT，Flash，Transformer，3D面部动画，嘴唇运动

HBF Talk: Speech-Driven 3D Facial Animation Synthesis Research

Wenxiang Wang[1], Shaobo Wang[1], Yu Zhi[1], Ang Chen[1,2]

[1]Research Center for Metaverse and Artificial Intelligence, College of Computing Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang
[2]Institute of Metaverse and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: Jul. 10th, 2024; accepted: Aug. 8th, 2024; published: Aug. 19th, 2024

Abstract

In recent years, speech-driven 3D facial animation has been widely studied. Previous work on the generation of coherent 3D facial animations was reported from speech data. However, the generated 3D facial animations lacks realism and vividness due to the scarcity of audio-visual data, and the accuracy of lip movements is not sufficient. This work is performed in order to improve the

文章引用: 王文祥, 王少波, 智宇, 陈昂. HBF Talk: 语音驱动的 3D 面部动画合成研究[J]. 计算机科学与应用, 2024, 14(8): 168-178. DOI: 10.12677/csa.2024.148174

accuracy and vividness of lip movement and an end-to-end neural network model, HBF Talk, is proposed. It utilizes the Hu BERT (Hidden-Unit BERT) pre-trained model for feature extraction and encoding of speech data. The Flash module is introduced to further encode the extracted speech feature representations, resulting in more enriched contextual representations of speech features. Finally, a biased cross-modal Transformer decoder is used for decoding. This paper conducts both quantitative and qualitative experiments and compares the results with existing baseline models, demonstrating the proposed HBF Talk model outperforms previous models by improving the accuracy and liveliness of speech-driven lip movements.

Keywords

Hu BERT, Flash, Transformer, 3D Facial Animation, Lip Movements

1. 引言

语音驱动的 3D 面部动画技术具有广阔的应用前景，是一个不断发展但具有挑战性的研究领域，它被广泛地用于娱乐和游戏领域、虚拟现实领域、人机交互领域、教育和培训领域以及影视制作领域。语音驱动的 3D 面部动画研究在科学、技术和应用方面都具有重要的意义和潜力，将为人机交互带来更加自然和丰富的体验，并在多个领域提供创新的解决方案。

由于 3D 面部动画在虚拟现实、电影制作和游戏等领域中的广泛应用，近几十年来一直是一个非常活跃的研究课题。语音和面部动画(尤其是嘴型动画)之间的高度相关性，使得用语音驱动面部动画成为可能。早期人们提出在音素和它们的视觉对应物(即视素)之间建立复杂的映射规则，但是性能有限。随着深度学习的进步，用深度学习的方法来学习语音和面部动画之间的关系成为一种高效且便捷的方法。然而由语音生成准确的面部动画(尤其是嘴形动画)也是一个挑战。本文提出了一个全新的端到端的神经网络模型 HBF Talk，能更好地建模语音和嘴唇运动之间的关系，生成较为准确的嘴唇运动。

2. 相关工作

早期，语音驱动的 3D 面部动画主要是基于语言学的方法，即在音素和它们的视觉对应物(即视素)之间建立了一套复杂的映射规则[1]。Parke [2]在 1972 年首次提出参数化人脸模型，在人脸动画域做出了大量的改进与创新，并提出了很多不同的人脸动画生成方案。在人脸动画生成方案中由于生成思路的不同，其对于人脸模型的构建方式也有很大的区别，主要有三种[3]-[5]：过程式参数模型、生理结构模型和数据驱动参数模型。Paul Ekman [6]创建了人脸动作编码系统，根据解剖学原理将人脸面部区域划分成了 44 个基本运动单元，根据基本动作单元的运动变化来实现相应的表情动作，并将基本单元与生理结构上的肌肉组织进行对应。早期的研究虽然对动画控制很严格，但是程序相当复杂。

2017 年，Zhang 等人[7]设计了一个从大量原始语音数据中学习帧级说话者特征的深度神经网络(Deep Neural Networks, DNN)模型。该模型参数较多、尺寸较大，训练时间较长。Pengcheng 等人[8]通过将卷积神经网络(Convolutional Neural Networks, CNN)直接应用于语音的语谱图，以端到端的方式学习语音情绪特征。该方法没有考虑语音情绪的时序性。这些方法主要是通过神经网络模型从原始语音数据中学习语音中的情绪与风格，没有直观的表示。

近几年来，研究人员更加关注上下文信息，语音驱动的 3D 面部动画更加连贯并且细节越来越好，主要贡献如下。

➢ Daniel Cudeiro 等人[9]提出了一个独特的 4D 人脸数据集，学习模型 VOCA 接受任何语音信号作为输入，甚至是英语以外的语言语音，并逼真地为各种各样的成人面部制作动画，在训练过程中对主题标签进行调节，使模型能够学习各种现实的说话风格。

➢ Speech2Face [10]直接从一个人说话的录音中重建人脸，通过学习将语音的特征空间与使用数百万人说话的自然视频预训练的人脸解码器的特征空间对齐来解决这个问题。

➢ Face Former [11]通过基于 Transformer 的模型对长期音频上下文进行编码，并自回归预测动画 3D 人脸网格序列。

➢ Mesh Talk [12]为面部动画建立了一个分类潜在空间，它基于一种新的交叉模态损失来解开音频相关和音频不相关的信息。该模型确保了高度精确的嘴唇运动，同时还合成了与音频信号不相关的面部部分的可信动画。

➢ Code Talker [13]通过将语音驱动的面部动画作为在学习到的编码本的有限代理空间中的编码查询任务。

➢ Sad Talker [14]提出 Exp Net 通过提取系数和 3D 渲染的面部来从音频中学习准确的面部表情，在头部姿态方面，通过条件 VAE 设计 Pose VAE 来综合不同风格的头部动作。

➢ Diff Talk [15]研究了说话时脸部的控制机制，而不是将音频信号作为单一驱动因素，并将参考面部图像和地标作为人格意识广义合成的条件。通过这种方式，Diff Talk 可以有效地合成高保真音频驱动的广义新身份脸部动画视频。

近些年的方法主要是从语音中重建人脸动画，具有更加直观的表现形式，而且从不同角度探索了语音与 3D 人脸动画之间的关系。与本文工作密切相关的是 Face Former 和 Code Talker，它们在语音驱动表情这一任务上有较好的结果，与他们不同的是，本文采用了自监督预训练语音模型 Hu BERT (Hidden-Unit BERT)，并引入 Flash 模块来提取更为丰富的语音上下文表示。

3. 方法

语音驱动表情这一任务是一个类似序列到序列(Seq2seq)的问题，我们将使用序列到序列的模型来解决这个问题。输入的序列是语音序列，输出的序列是三维人脸网格序列(即四维扫描)。本文的主要目标是提高语音驱动的嘴唇运动的准确性和生动性。

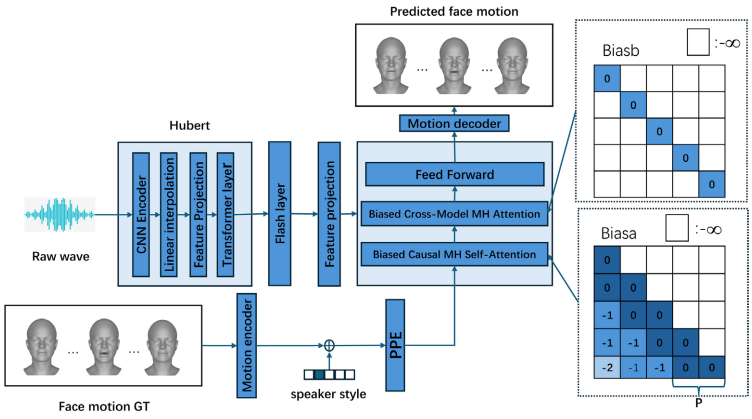

Figure 1. HBF Talk model architecture diagram
图 1. HBF Talk 模型架构图

本文提出的模型 HBF Talk (如图 1 所示)是一个端到端的编码器解码器神经网络架构，其中编码器由 Hu BERT [16]和 Flash [17]模块共同组成，解码器由带偏置的跨模态 Transformer 解码器组成。Hu BERT 是一个预训练的大型语音模型，它可以将连续音频信息编码成离散的时间步长表示。我们使用的是 Hu BERT 架构的预训练版本，并使用发布的 Hu BERT-Base-ls960 版本，该版本在 960 小时的 Libris Speech [18]数据集上进行了训练。语音数据在经过预训练的 Hu BERT 模型处理之后，通过 Flash 模块对语音表示进行再处理，最后通过解码器进行解码。

设 $A_{1:T}=(a_1,\cdots,a_T)$ 是一个真实的 3D 人脸运动序列，T 是真实人脸运动序列的帧数。设 χ 是原始的语音序列。我们的任务是通过原始语音序列 $\chi 1$ 生成一个与真实 3D 人脸运动序列 $A_{1:T}$ 非常接近的预测人脸运动序列 $\hat{A}_{1:T}$。在我们的编码器－解码器的模型架构 HBF Talk 中，如图 1 所示，原始的语音序列 χ 经过 Hu BERT 和 Flash 模块的处理得到语音表示 $B_{1:T'}=(b_1,\cdots,b_{T'})$，$T'$ 为语音表示的帧数，由于 Linear_Interpolation 层的处理，语音表示与 3D 人脸顶点序列已经一一对应。说话人的风格嵌入层包含一组可学习的嵌入，表示说话人身份 $G=(g_1,\cdots,g_N)$。最后，解码器自回归预测面部运动 $\hat{A}_{1:T}=(\hat{a}_1,\cdots,\hat{a}_T)$ 以 $B_{1:T'}$、说话人 n 的风格嵌入 g_n 和过去的面部动作为条件，公式如下。

$$\hat{a}_t=\mathrm{HBFTalk}_\theta(\hat{a}_{<t},g_n,\chi) \tag{1}$$

其中 θ 表示模型参数，t 是序列中的当前时间步长，$\hat{a}_t\in\hat{A}_{1:T}$。

3.1. 编码器

3.1.1. Hu BERT 预训练模型

我们提出的方法在编码器中有效地采用了最先进的自监督预训练语音模型 Hu BERT，用于语音驱动 3D 面部动画生成的下游任务。由于 Hu BERT 模型能够学习并产生结合声学和语言信息的连续音频流的高质量离散隐藏表示，因此 Hu BERT 的作者提出可以将 Hu BERT 预训练的隐藏层表示用于各种下游任务[16]。

Hu BERT 模型架构在 Transformer 层引入了一个类似 BERT [19]的屏蔽语言建模的编码器。与之前的 Wav2vec 2.0 [20]复杂的对比损失相比，它引入了一个简单的交叉熵损失来预测屏蔽单元。此外，与 Wav2Vec2.0 不同，Hu BERT 是通过多次迭代进行训练的。在第一次迭代中，Hu BERT 使用无监督的简单 k 均值聚类进行声学单元发现，以促进在第二次迭代中进行的自监督掩码语言建模学习。在第二次迭代中，训练是在发现的离散隐藏单元上完成的，仅在屏蔽区域上有预测损失，迫使模型使用类似 BERT 的编码器学习声学和语言的组合模型，因此称为 H(idden)-u(nit)-BERT。Hu BERT 由 CNN 编码器、特征投影层、位置卷积嵌入层和 Transformer 层组成，CNN 编码器将连续音频数据离散为 512 维表示。特征投影层将 512 维表示投影为 768 维表示，12 个 Transformer 层用来捕获序列中的上下文信息。

在本文的方法中，采用具有 100 M 个参数的 Hu BERT 下游任务微调模型，在最后一个隐藏状态下产生 768 维的嵌入。初始化预训练权重参数，不冻结 Hu BERT 中的任何层，在预训练权重参数下进行微调。

$$c_t=\mathrm{SpeechEncoder}(Y_t) \tag{2}$$

其中 Y_t 是由原始音频序列 χ 经过处理之后得到的表示。

3.1.2. Linear_Interpolation 层

Linear_Interpolation 层的作用主要是对经特征提取层提取到的语音特征表示进行帧率的转换。在训练和验证阶段，需要传入 3D 面部运动顶点的帧数，从语音提取特征帧率转换为 3D 面部运动顶点的帧率，并对齐顶点的帧数。在推理阶段，根据语音特征表示的帧率和帧数以及 3D 面部运动顶点的帧率计算输出的帧数。该层不包含任何需要训练的参数，只是实现帧率的转换以及保证语音特征表示与 3D 人脸顶点序列的一对一的帧级关系。

3.1.3. Flash

Flash [17]这一模块融合了部分注意力和线性注意力的优点。输入序列被划分为几个大小相同的不重叠的块。首先，对每个块独立施加局部二次注意力，产生部分预门控状态。此外，采用全局线性注意机制来捕捉跨块的远程交互，最后将以上两部分的表示进行相加，再进行门控和后注意力投影。

$$b_{t'} = \text{Flash}\left(c_t\right) \tag{3}$$

其中，$b_{t'} \in B_{1:T'}$，是经过 Speech Encoder 和 Flash 处理后的语音上下文表示。

3.2. HBF Talk 的解码器

3.2.1. Motion Encoder 和 Motion Decoder

Motion Encoder 的作用是将人脸顶点序列从人脸顶点维度转换到特征维度。Motion Decoder 的作用是将预测的人脸顶点序列从特征维度转换到人脸顶点维度。这两部分主要是实现人脸顶点维度与特征维度之间的转换。

3.2.2. 解码器

我们的解码器使用的是 Transformer Decoder 这一模块。在其中引入了 Face Former [11]中使用的周期性位置编码模块(PPE)，这个模块可以注入时间顺序信息。并且与原始 Transformer Decoder 不同，引入了 Face Former [11]中提出的有偏置的跨模态注意力机制和有偏置的自注意力机制。

有偏置的自注意力机制是用于对输入的真实人脸顶点序列进行处理，由图 1 中的 Biasa 可知，该偏置不仅可以屏蔽预测帧之后的信息，而且对于过去的人脸顶点序列，将较高的注意力权重分配给较近的时间段，对于较远的时间段，则分配较低的注意力权重。较近时间段的人脸序列帧对当前帧影响较大。

有偏置的跨模态注意力机制如图 1 中的 Biasb 可知，它的主要作用是实现语音特征表示与经有偏置自注意力机制处理的人脸运动表示之间的对齐。

$$\hat{a}_t = \text{TransformerDecoder}\left(\hat{a}_{<t}, g_n, b_{t'}\right) \tag{4}$$

其中 t 是序列中的当前时间步长，g_n 是说话人 n 的说话风格嵌入，$\hat{a}_t \in \hat{A}_{1:T}$。

3.3. 训练和测试

在训练阶段，我们采用自回归方案代替 Teacher-Forcing 方案。并且在最后做了使用 Teacher-Forcing 方案训练的对比实验。在验证阶段，主要是使预测序列 $\hat{A}_{1:T} = \left(\hat{a}_1, \cdots, \hat{a}_T\right)$ 与真实序列 $A_{1:T} = \left(a_1, \cdots, a_T\right)$ 之间的均方误差(MSE)尽可能的小。

重建损失如公式 5 所示。

$$L_{\text{MSE}} = \sum_{t=1}^{T}\sum_{n=1}^{N}\left\|\hat{a}_{t,n} - a_{t,n}\right\|^2 \tag{5}$$

其中的 N 代表的是三维面部网格的顶点数。

运动速度损失如公式 6 所示。

$$L_{\text{MOE}} = \sum_{t=1}^{T}\sum_{n=1}^{N}\left\|\left(\hat{a}_{t,n} - \hat{a}_{t-1,n}\right) - \left(a_{t,n} - a_{t-1,n}\right)\right\|^2 \tag{6}$$

其中的 N 代表的是三维面部网格的顶点数。

$$L_{\text{total}} = \lambda_1 L_{\text{MSE}} + \lambda_2 L_{\text{MOE}} \tag{7}$$

其中 λ_1 和 λ_2 分别代表重建损失和运动速度损失的权重系数。

在推理阶段，模型自回归地预测 3D 人脸顶点序列。在每个时间步，它预测基于原始音频 χ、之前的面部运动 $\hat{a}_{<t}$ 和说话人的风格嵌入表示 g_n 的面部运动 $\hat{a}_t$，如公式 1 所示。g_n 由说话人的身份来确定，因此可以通过改变 One-Hot 身份向量来操纵不同说话人风格的输出。

4. 实验

4.1. 数据集

本实验使用了一个公开可用的 3D 数据集 VOCASET [9]进行训练和测试。该数据集提供了英语口语的音频-3D 扫描对。VOCASET 包含 255 个独特的句子，其中一些句子在说话者之间共享。

VOCASET 数据集包含从 12 个受试者中捕获的 480 对音频和 3D 面部运动序列。面部运动序列以每秒 60 帧的速度记录，长度约为 4 秒。VOCASET 中的 3D 人脸网格注册到 FLAME [21]拓扑，每个网格有 5023 个顶点。为了和其他几个模型公平的比较，我们采用了与 Face Former [11]和 Code Talker [13]相同的训练(VOCA Train)、验证(VOCA-Val)和测试(VOCA-Test)分割。

4.2. 模型实验

我们将本文提出的 HBF Talk 端到端的神经网络模型与两种先进的方法 Face Former [11]和 Code Talker [13]进行了比较。首先将 Face Former [11]和 Code Talker [13]与 HBF Talk 在同等条件下，用 VOCASET 数据集进行了训练和测试。做了定量评估和在相同说话风格的条件下的定性评估，最后对我们提出的模型进行了消融实验。

所有的模型训练都是在一个装有 Linux 系统的电脑上完成的，该电脑配备 GEFORCE RTX 4070Ti 显卡。模型训练的超参数见表 1 所示。

Table 1. Hyperparameters for model training
表 1. 模型训练的超参数

超参数	HBF Talk
Optimizer	Adam
Learning Rate	$1e^{-4}$
Number of Epochs	100
Feature Dim	64
Flash Layer Dim	768

4.3. 定量评估

Table 2. Quantitative evaluation results on the VOCA-Test
表 2. 在 VOCA-Test 上的定量评估结果

方法	LVE↓ ($\times 10^{-5}$ mm)
Face Former [11]	4.1172
Code Talker [13]	4.1476
HBF Talk (本文模型)	3.6283

对于 3D 人脸的口型评估，本文遵循 Face Former [11]和 Code Talker [13]中使用的唇部同步度量来评

估唇部运动的质量。所有唇形顶点的最大 L2 误差被定义为每帧的唇形误差，误差是通过比较预测和真实的三维人脸几何数据来计算的，取所有序列的所有帧的唇形误差的均值作为最终的 LVE (唇顶点误差)指标。

本文的定量评估实验结果如表 2 所示。

表 2 对比结果表明，本文所提出的 HBF Talk 模型相对于 Face Former [11]和 Code Talker [13]具有更低的唇顶点误差(LVE)，这表明和其他两种方法相比，本文提出的模型可以产生更加准确的嘴唇运动。

4.4. 定性评估

我们直观地比较了本文提出的方法与其他竞争者的方法。为了确保比较的公平性，我们将相同的讲话风格分配给 Face Former、Code Talker 和本文提出的 HBF Talk 作为条件输入。为了比较嘴唇和语音的同步情况，我们举例说明了五个典型的合成人脸动画序列帧，它们在特定的音节上说话。这五个合成人脸动画序列帧如图 2 到图 6 所示。

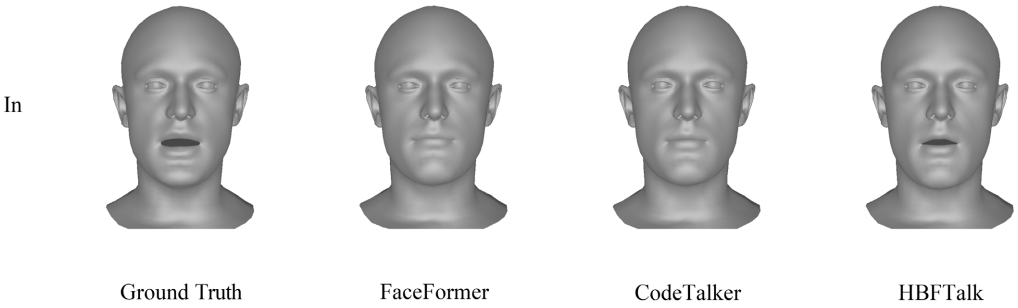

Figure 2. /”**ɪn**”/ pronunciation-related face animation sequence frames
图 2. /”**ɪn**”/发音相关的人脸动画序列帧

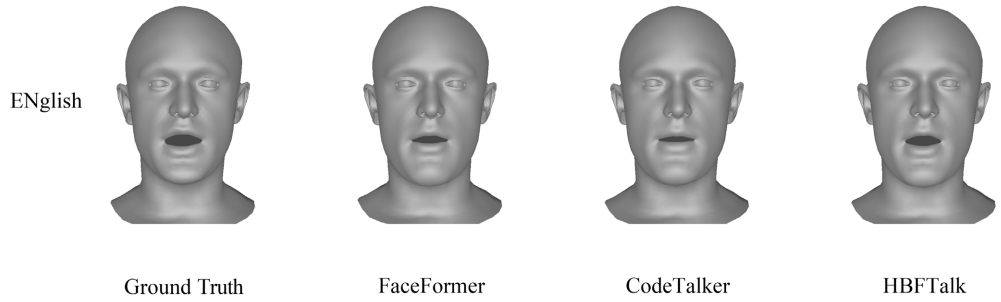

Figure 3. /”**ɪŋ**glɪʃ”/ pronunciation-related face animation sequence frames
图 3. /”**ɪŋ**glɪʃ”/发音相关的人脸动画序列帧

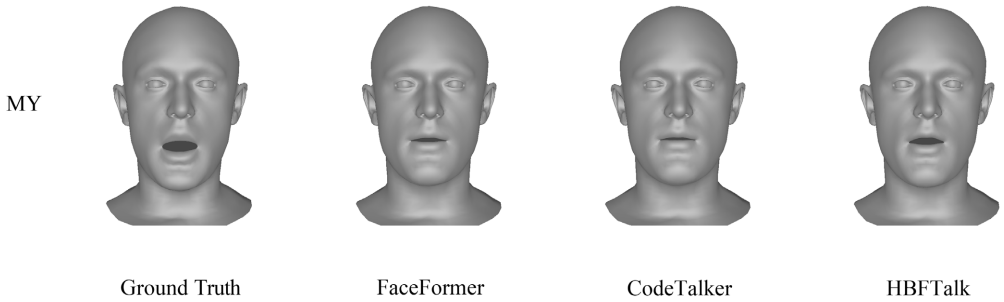

Figure 4. /”m**aɪ**”/ pronunciation-related face animation sequence frames
图 4. /”m**aɪ**”/发音相关的人脸动画序列帧

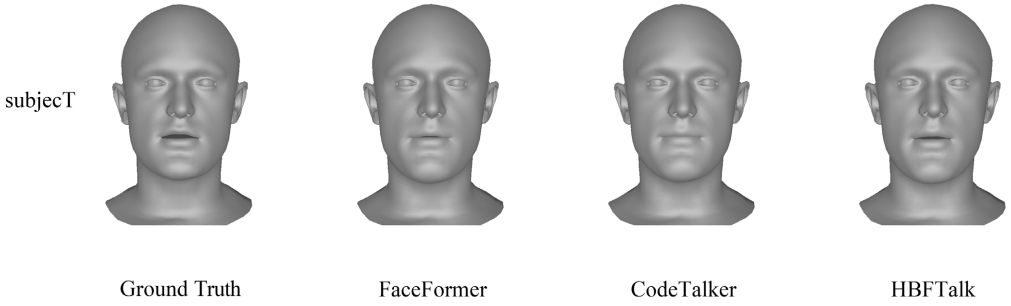

Figure 5. /”sʌbdʒɪk**t**”/ pronunciation-related face animation sequence frames
图 5. /”sʌbdʒɪk**t**”/发音相关的人脸动画序列帧

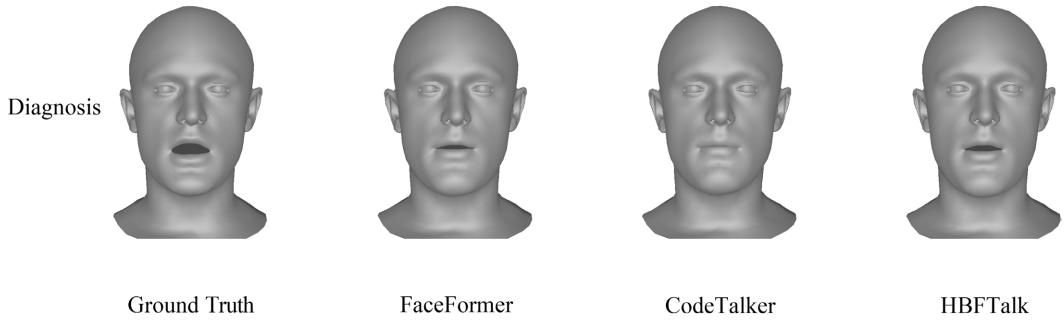

Figure 6. /”**d**aɪəg’noʊsɪs”/ pronunciation-related face animation sequence frames
图 6. /”**d**aɪəg’noʊsɪs”/发音相关的人脸动画序列帧

由图 2 到图 6 可知，与基线模型相比，本文提出的 HBF Talk 产生的唇形更准确地表达了语音信号，也更符合真实的唇形。例如，与 Face Former 和 Code Talker 相比，HBF Talk 在发出/”**ɪn**”/的读音(即图 2)时，有适当的张嘴，可以产生更好的口型同步；在发出 English 的前半部分的读音/”**ɪŋ**glɪʃ”/ (即图 3)、MY 的读音/”m**aɪ**”/ (即图 4)和 diagnosis 的读音/”**d**aɪəg’noʊsɪs”/ (即图 6)时，HBF Talk 可以产生准确的唇形，有适当的张嘴，而其他方法张嘴的幅度较小。即使是 Subject 中的/”**t**”/的发音(图 5)，HBF Talk 相对于其他方法也有一定的张嘴，具有较为准确的唇形。

4.5. 消融实验

为了分析本文提出的模型中的各个模块对模型整体效果的贡献，首先对 Hu BERT 预训练模型各层是否冻结预训练参数进行了研究，然后通过去掉或者替换模型中的某一模块来分析模型中的每一模块对生成的 3D 面部动画质量的影响。

4.5.1. Hu BERT 预训练模型的消融

本文进行了大量的实验来研究和优化 Hu BERT 在 HBF Talk 编码器中的作用。该消融实验是通过冻结 Hu BERT 模型在各个层的预训练参数来研究其作用。从不冻结(即每一层的参数都是可训练的)到所有层的权值都冻结(即所有层都用预训练参数，参数不可训练)。该消融实验研究结果见表 3。

Table 3. Ablation experimental results of Hu BERT pre-trained model on VOCA-Test
表 3. Hu BERT 预训练模型在 VOCA-Test 上的消融实验结果

模型	LVE↓ ($\times 10^{-5}$ mm)	模型可训练的参数量
(i)	3.6283	100,407,549
(ii)	3.7816	96,207,101

续表

(iii)	3.8596	95,812,093
(iv)	3.8609	81,636,349
(v)	3.9271	67,460,605
(vi)	3.9796	53,284,861
(vii)	3.9971	39,109,117
(viii)	4.1975	24,933,373
(ix)	4.3611	17,845,501
(x)	4.6158	10,757,629

根据第 3.1.1 节中描述的 Hu BERT 结构，该消融实验的模型结构配置如下：(i) 没有层被冻结，所有层的参数都可以训练，(ii) CNN 编码器(特征提取层)被冻结，(iii) CNN 编码器(特征提取层)和特征投影层被冻结。对于模型(iv)至(viii)，除了冻结 CNN 编码器(特征提取层)和特征投影层外，每一个模型结构依次多冻结两个 Transformer 层。对于模型(ix)，仅仅保留最后一个 Transformer 层可训练，而对于(x)，整个 Hu BERT 预训练模型在训练期间被冻结。

尽管上述所有的模型配置都可以产生连贯流畅的三维人脸动画，但是我们发现(i)产生的唇顶点误差(LVE)最小，生成的动画与语音的相关性更强，表现力更加丰富。除此之外，当把模型配置从模型(ii)变化到(x)时，其所生成的三维人脸动画变得更加僵硬，并且唇形与语音之间的相关性变弱。本文选择的是没有层被冻结，即所有的参数都可以训练。

除了上述对 Hu BERT 模型的消融实验，本文还做了 HBF Talk 模型中其他部分在 VOCA-Test 上的消融实验。消融实验研究结果见表 4。

Table 4. Ablation results of other parts of the HBF Talk model on VOCA-Test
表 4. HBF Talk 模型其他部分在 VOCA-Test 上的消融实验结果

模型	LVE↓ ($\times 10^{-5}$ mm)
HBF Talk (Autoregressive)	3.6283
HBF Talk (Teacher-forcing)	3.8811
w/o Flash	3.8654
Wav2vec2FlashFormer	3.9339
Hu BERT Flash RNN	4.5979
Hu BERT Flash Bi LSTM	4.6059
L_{total} (9:1)	3.8415

4.5.2. 自回归机制的消融

为了分析训练阶段自回归机制(Autoregressive)的影响，本文在训练阶段使用 Teacher-Forcing 机制做了对比实验。简要解释一下这两种机制，自回归机制是一种生成模型的训练方法，其中模型根据先前生成的内容来逐步生成下一个元素。在自回归过程中，生成的每个元素都依赖于之前生成的元素。Teacher-Forcing 机制在训练过程中，真实的目标输出被用作模型的输入，而不是使用模型自身生成的输出作为下一步的输入。由表 4 可知，在本文的模型中，自回归机制相比于 Teacher-Forcing 机制训练的模型有较好的结果，究其原因，在推理阶段是用模型自身生成的输出作为下一步的输入，所以用在训练阶段使用自回归机制的效果较好。

4.5.3. Flash 模块的消融

为了分析 Flash 模块对模型的影响，本文去除了模型中的 Flash 模块，在其他条件相同的情况下做了对比实验，Flash 模块对语音的隐藏层表示进行了分块的块内局部注意力和块间全局注意力，最后相加经过门控输出，更好的建模了语音的隐藏层表示。由表 4 可知，通过 Flash 模块对语音的隐藏层表示进行再处理之后，会有更好的三维人脸动画效果。并且 Wav2vec2 Flash Former 与 Face Former 相比，相当于加上了 Flash 模块，由表 4 和表 2 可知，在 Wav2vec2 预训练模型下，Flash 模块也能带来更好的三维人脸动画效果。

4.5.4. 音频编码器的消融

在语音驱动 3D 面部动画生成的这项任务中，在 VOCA [9]之后，很多相关的模型都使用 Wav2vec2.0 这一预训练的音频编码器，为了验证我们选用 Hu BERT [16]作为音频编码器的正确性，本文用 Wav2vec2.0 代替了 Hu BERT，在其他条件相同的情况下做了对比实验。从表 4 可知，使用 Hu BERT 预训练模型进行语音驱动 3D 面部动画生成下游任务比使用 Wav2vev2.0 具有更好的效果。

4.5.5. 解码器的消融

为了验证本文所提出的模型中选择的解码器，本文对解码器使用不同的序列建模方法进行了消融实验。我们用更简单的 RNN 解码器和 Bi LSTM 解码器取代了本模型所选择的 Transformer 解码器。实验的结果如表 4 所示，Transformer 解码器在同等参数条件下表现最好，在 VOCA 数据集上的误差更小，主要原因如 3.2.2 节所描述的那样，有偏置的自注意力机制实现了当前帧对过去帧的不同关注程度，而且有偏置的跨模态注意力机制对齐了人脸运动表示与语音表示。

4.5.6. 损失函数的消融

本文使用 L_{MSE} (重建损失，如公式 5 所示)和 L_{total} (组合损失，如公式 7 所示)两种损失函数做了对比试验，由表 4 可知，用单纯的重建损失比组合损失能取得更好的效果。

5. 结论

本文提出的 HBF Talk 显著提高了针对跨模态语义的运动合成质量，HBF Talk 中的编码器 Hu BERT 有效地利用了自监督预训练语音表示，Flash 模块进一步的整合了局部和全局的语音上下文信息，经解码器解码得到面部动画序列。通过与现有的先进技术进行比较，本文提出的方法在实现准确的嘴唇运动方面具有优势。然而，由于视听数据的稀缺性，导致生成的唇形质量仍落后于真实的唇形质量，利用大规模的视听数据来训练模型以提高唇形质量是未来的研究方向。

致　谢

我们感谢 VOCA、Face Former、Code Talker 的作者提供他们的代码和数据集。

基金项目

本课题受到“温州大学元宇宙与人工智能研究院”的“重大课题及项目产业化专项资金”(编号：2023103)的资助。

参考文献

[1] Fisher, C.G. (1968) Confusions among Visually Perceived Consonants. *Journal of Speech and Hearing Research*, **11**, 796-804. https://doi.org/10.1044/jshr.1104.796

[2] Parke, F.I. (1972) Computer Generated Animation of Faces. *Proceedings of the ACM Annual Conference*, **1**, 451-457.

https://doi.org/10.1145/800193.569955

[3] Parke, F.I. and Waters, K. (1996) Computer Facial Animation. A. K. Peters, Ltd., Natick.

[4] Li, L., Liu, Y. and Zhang, H. (2012) A Survey of Computer Facial Animation Techniques. 2012 *International Conference on Computer Science and Electronics Engineering*, Hangzhou, 23-25 March 2012, 434-438. https://doi.org/10.1109/iccsee.2012.129

[5] 李代超. 基于伪肌肉向量的三维人脸动画及其驱动研究与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2011.

[6] Ekman. P. and Friesen, W.V. (1978) Facial Action Coding System (FACS): A Technique for the Measurement of Facial Actions. *Rivista di Psichiatria*, **47**, 126-138.

[7] Zhang, M., Chen, Y., Li, L. and Wang, D. (2017) Speaker Recognition with Cough, Laugh and "Wei". 2017 *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Kuala Lumpur, 12-15 December 2017, 497-501. https://doi.org/10.1109/apsipa.2017.8282083

[8] Li, P.C., *et al*. (2018) An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition. *Proceedings of INTERSPEECH*, Hyderabad, 2-6 September 2018, 3087-3091.

[9] Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A. and Black, M.J. (2019) Capture, Learning, and Synthesis of 3D Speaking Styles. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 10093-10103. https://doi.org/10.1109/cvpr.2019.01034

[10] Oh, T.-H., *et al*. (2019) Speech2Face: Learning the Face behind a Voice.

[11] Fan, Y., Lin, Z., Saito, J., Wang, W. and Komura, T. (2022) FaceFormer: Speech-Driven 3D Facial Animation with Transformers. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 18749-18758. https://doi.org/10.1109/cvpr52688.2022.01821

[12] Richard, A., Zollhofer, M., Wen, Y., de la Torre, F. and Sheikh, Y. (2021) MeshTalk: 3D Face Animation from Speech Using Cross-Modality Disentanglement. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 1153-1162. https://doi.org/10.1109/iccv48922.2021.00121

[13] Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J. and Wong, T. (2023) CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 12780-12790. https://doi.org/10.1109/cvpr52729.2023.01229

[14] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., *et al*. (2023) SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 8652-8661. https://doi.org/10.1109/cvpr52729.2023.00836

[15] Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., *et al*. (2023) DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Animation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 1982-1991. https://doi.org/10.1109/cvpr52729.2023.00197

[16] Hsu, W., Bolte, B., Tsai, Y.H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A. (2021) Hubert: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451-3460. https://doi.org/10.1109/taslp.2021.3122291

[17] Hua, W., Dai, Z., Liu, H. and Le, Q.V. (2022) Transformer Quality in Linear Time.

[18] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. (2015) Librispeech: An ASR Corpus Based on Public Domain Audio Books. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), South Brisbane, 19-24 April 2015, 5206-5210. https://doi.org/10.1109/icassp.2015.7178964

[19] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[20] Baevski, A., Zhou, H., Mohamed, A. and Auli, M. (2020) wav2vec2.0: A Framework for Self-Supervised Learning of Speech Representations.

[21] Li, T., Bolkart, T., Black, M.J., Li, H. and Romero, J. (2017) Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics*, **36**, 1-17. https://doi.org/10.1145/3130800.3130813