Computer Science and Application 计算机科学与应用, 2023, 13(11), 2072-2079
Published Online November 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1311206

Hans汉斯

带有先验的语音驱动三维人脸动画生成方法

吕镇宇，夏方方，刘芳丽，郭润甲，郭子俊

北京信息科技大学计算机学院，北京

收稿日期：2023年10月16日；录用日期：2023年11月15日；发布日期：2023年11月22日

摘　要

语音驱动的三维人脸生成是计算机视觉和图形学中一个非常有吸引力的研究课题。除了有趣之外，它还有广泛的应用，例如游戏动画、3D视频通话和AR/MR的3D化身。由于人脸运动的复杂性和不确定性，以往方法生成的结果有唇形不准确、面部动态性不佳的缺点。不同于以往一阶段的方法，我们使用一种新的两阶段的方法，在模型训练的第一阶段我们使用变分自动编码器将高维的复杂的面部映射进低维的空间，充分学习人脸运动先验。在第二阶段，Transformer根据输入的语音信号在学习到的人脸先验的基础上进行潜在代码查询，以回归的方式生成面部运动序列。这样可以降低生成面部动画的难度，减少了映射的模糊，可以在任意指定音频上得到生动的人脸说话动画，经验证我们的方法与先进的方法相比在唇形和脸部动态性上取得优势。

关键词

语音驱动3D面部动画，3D说话人脸生成，3D动画人

Speech Driven 3D Facial Animation Generation Method with Prior Knowledge

Zhenyu Lv, Fangfang Xia, Fangli Liu, Runjia Guo, Zijun Guo

Computer School, Beijing Information Science and Technology University, Beijing

Received: Oct. 16th, 2023; accepted: Nov. 15th, 2023; published: Nov. 22nd, 2023

Abstract

Speech-driven 3D facial animation is a very attractive research topic in computer vision and graphics. In addition to being interesting, it has a wide range of applications, such as game animation, 3D video calls, and 3D avatars of AR/MR. Due to the complexity and uncertainty of facial movements, previous methods have drawbacks such as inaccurate lip shape and poor facial dynamics. Unlike previous methods, we use a new two-stage approach. In the first stage of model training, we use a

文章引用: 吕镇宇, 夏方方, 刘芳丽, 郭润甲, 郭子俊. 带有先验的语音驱动三维人脸动画生成方法[J]. 计算机科学与应用, 2023, 13(11): 2072-2079. DOI: 10.12677/csa.2023.1311206

variational autoencoder to map high-dimensional complex faces into low-dimensional space, fully learning facial motion priors. In the second stage, the Transformer performs latent code queries based on the learned facial prior based on the input speech signal, and generates facial motion sequences through regression. This can reduce the difficulty of generating facial animation, reduce mapping blur, and obtain vivid facial speech animations on any specified audio. It has been verified that our method has advantages in lip shape and facial dynamics compared to advanced methods.

Keywords

Speech-Driven 3D Facial Animation, 3D Talking Face Generation, 3D Avatar

1. 引言

基于语音驱动的面部动画生成技术是一种根据输入的语音序列生成虚拟人面部动画，再现一个人说话的场景的技术。该技术不仅能提高用户对语音的理解度，还能提供一种真实、友好的人机交互方式。该研究领域主要分为二维说话人脸动画生成方向和三维说话人脸动画生成方向。在本研究中，我们将重点放在三维模型上。三维说话人脸生成目前面临许多问题，首先真实性低，人脸木讷呆滞，动态性差；其次同步性不佳，存在合成嘴型动画不够连续，跳变现象较多的问题。

本文为了提升嘴唇同步性、面部动态性，受无监督表征学习模型 VQ-VAE 的启发，利用它学习的离散编码具有良好的表征能力的优点，提出一种二阶段的方法，利用人脸先验生成三维说话人脸模型，可以得到较好的结果。本文首先介绍了三维人脸表示原理与我们的方法，在实验部分介绍了定量评测指标并进行验证。最后，我们在总结部分对本文的研究内容进行归纳总结，并对未来的研究方向进行展望。

2. 相关工作

2.1. 语音驱动的面部动画生成

生成语音驱动的面部动画的方法可大致分为基于语言学的方法和基于学习的方法。

基于语言学的方法[1]是通过分析输入的语音信号的语音学特征，然后将这些特征转化为对应的面部动画控制参数。首先是要经过语音信号处理，需要对输入的语音信号进行预处理和特征提取。常采用包括语音分帧、短时能量计算、语音共振峰估计等方法提取出语音信号的关键特征，以供后续分析使用；接着是特征转换，需要从语音信号中提取的特征将被转换为与面部动画相关的控制参数，经过此过程可以获得包括声学特征到表情特征的映射，如将音高映射为嘴唇的张合程度，音强映射为眉毛的上下移动等。基于语言学的方法虽然在生成面部动画方面有一定的效果，但仍然存在一些挑战，如语音识别的准确性、特征转换的精确性等。

基于学习的方法[2] [3]是生成语音驱动的面部动画的另一种常用方法，该方法也能够更好地捕捉到语音和面部动画之间的复杂关系。这种方法通过使用大量的输入语音与对应的面部动画数据，通过机器学习算法来学习语音和面部动画之间的映射关系。首先在数据收集方面要搜集大量的包含语音信号和相应面部动画的训练数据；在数据预处理阶段则对收集到的语音和面部动画数据进行预处理，以确保数据的

质量和一致性，便于后续的模型训练；在特征提取时，从语音和面部动画数据中提取出有效的特征。对于语音数据，可以提取声学特征，如梅尔频谱系数(Mel-frequency cepstral coefficients, MFCCs)，音高、音强等[4]。对于面部动画数据，可以提取面部关键点坐标、脸部表情等特征；在模型选择与训练时选择适合的机器学习算法来建立语音和面部动画之间的映射模型。训练模型需要使用大量的语音和面部动画数据来训练深度神经网络，学习从语音信号到面部动画之间的复杂映射，使其能够生成逼真的面部动画。VisemeNet [5]采用三阶段长短期记忆(LSTM)网络来预测下面部嘴唇模型的动画曲线。MeshTalk [6]利用分类潜空间成功地分离了音频相关和非相关的面部信息。然而，由于采用的潜在空间不是最佳的，表现力有限，因此在数据稀缺的环境中应用时，动画质量并不稳定。VOCA [7]采用了强大的音频特征提取模型，可以生成不同说话风格的面部动画；在模型优化与调整方面，通过反复训练和验证，在训练数据上进行模型优化和调整，以提高模型的性能和泛化能力。这包括调整网络结构、改变训练参数、数据增强等方法；最后在动画合成时将新的语音输入到模型中，模型会根据已学习到的映射关系生成对应的面部动画，并根据需求添加额外的细节和效果。

2.2. Attention 机制和 Transformer 架构

Attention [8]机制起源于人类视觉的研究，是深度学习快速发展后广泛应用于自然语言处理、统计学习、图像检测、语音识别等领域的核心技术。本质上，注意力机制是为了实现信息处理资源的高效分配，比如首先关注场景中的一些关键点，而剩下的不重要场景可能会被暂时忽略。注意力机制可以关注权重高的重要信息，而忽略权重低的不相关信息。Attention 机制相比 CNN 和 RNN 的主要优点是参数更少、速度更快、性能更好。我们引入了有效的注意力机制来提高模型性能。

Transformer 最初是针对 NLP 领域提出的。Transformer 利用自注意力机制实现输入序列的编码和表示学习，可以捕获序列中不同位置之间的依赖关系并实现上下文感知。Vision Transformer (ViT)是一种基于 Transformer 模型的图像分类方法，已获得显着的普及和成功。ViT 随后被应用于许多其他计算机视觉领域，例如对象检测、语义分割。Transformer 模型还被用于 3D 面部动画任务，该任务根据输入的音频描述生成逼真且富有表现力的面部动画。我们采用 Transformer 作为我们方法的核心结构。

2.3. 三维人脸的表示方法

Flame [9]是一种强大的 3D 人脸通用模型，它经过训练并使用参数向量作为输入，该向量包含形状参数、姿势参数和表情参数。通过调整这些参数，我们可以精确控制人脸的身份特征、头部的旋转和平移，以及人脸的表情变化。这个模型的输出是一个包含 5023 个顶点和 9976 个面的三维面部网格，这个面部网格非常灵活，可以用来表示各种不同的面部形状和姿势。通过 Flame 模型，我们可以构建个性化的模型，并通过改变表情参数和动作参数来生成全新的 3D 数据，从而进行动画制作。这种模型具有出色的可塑性和表达丰富面部特征的能力，使得它成为生成会说话的人脸的理想选择。通过将 FLAME 模型与语音合成技术结合起来，我们可以实现逼真的、具有口型同步的人脸动画。我们可以根据声音生成相应的面部表情，使得人脸看起来像是在说话。这种模型还可以用于人脸识别、表情分析等应用，提供更准确的结果。如图 1，由于 Flame 是一个表现力丰富的 3D 人脸通用模型，它通过参数化表示实现对人脸形状、姿势和表情的精细控制，它的可塑性和表达能力使其成为生成会说话的人脸的优选模型，为动画制作和虚拟人物的创建带来了无限可能，因此我们的工作选择了驱动 FLAME 模型生成会说话的人脸。

3. 带有先验的语音驱动运动合成

3.1. 面部运动先验

视觉逼真的面部动画应呈现准确的嘴唇运动和自然的表情。为了从语音信号中实现这一目标，需要

额外的运动先验来减少不确定性并补充逼真的运动成分。我们预先训练了一个基于 Transformer 的 VQ-VAE [10]模型，如图 2，该 VQ-VAE 由编码器 E、解码器 D 和上下文丰富的码本 Z 组成，在现实面部动作的自我重构下进行训练。

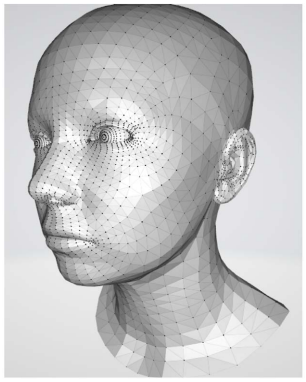

Figure 1. Schematic diagram of vertex mesh in flame model
图 1. Flame 模型顶点网格示意图

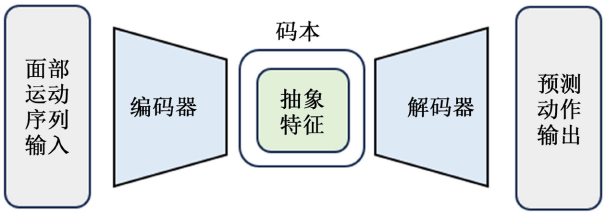

Figure 2. Schematic diagram of variational automatic encoder VQ-VAE
图 2. 变分自动编码器 VQ-VAE 示意图

给定面部运动 $\mathbf{M}_{1:T}$，编码器将其映射到连续潜在变量。然后向量量化函数 $Q(\cdot)$ 映射到离散潜在表示 $\mathbf{Z}_{\mathbf{q}}$，这个过程可以表述如下：

$$\mathbf{Z}_{\mathbf{q}} = Q(Z_e) = \operatorname{argmin}_{e_i} \left\| \mathbf{z}_e - \mathbf{e}_{\mathbf{i}} \right\|_2^2 \tag{1}$$

那么，自重构过程可以表示如下：

$$\hat{\mathbf{M}}_{1:T} = D(\mathbf{Z}_{\mathbf{q}}) = D\left(Q\left(E\left(\mathbf{M}_{1:T}\right)\right)\right) \tag{2}$$

其中 $\hat{\mathbf{M}}_{1:T}$ 是生成的面部动作，$D(\cdot)$ 是运动解码器。

$$L = L_{rec} + \beta L_{vq} \tag{3}$$

我们的损失主要由两部分组成，即重建损失和 vq 损失，并且我们添加了一个系数来平衡两者。第一项是重建损失。它涉及计算序列中每帧的真实运动和预测运动之间的平均绝对误差。第二项是 vq 损失，被用于更新码本项的中间损失。通过最小化码本向量和嵌入特征之间的距离来实现的。我们还添加了停止梯度操作，用作于防止梯度崩溃。

3.2. 语音驱动运动合成

有了学习到的离散动作先验，我们就可以建立一个从输入语音到目标动作代码的跨模态映射，从而

进一步解码为逼真的面部动作，图 3 是我们合成人脸动画模型的示意图。在语音 **A** 和面部模板 **Temp** 的条件下，一个由语音编码器和跨模态解码器组成的自回归模型被用来学习面部动作空间。$\hat{\mathbf{M}}$ 是预测出的人脸运动序列。模型的运作流程可表示为：

$$\hat{\mathbf{M}} = F(\mathbf{A}, \mathbf{Temp}) \tag{4}$$

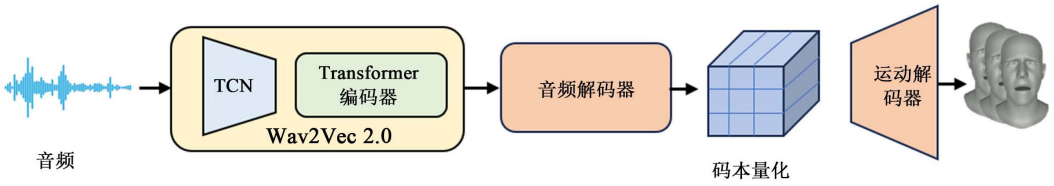

Figure 3. Workflow of speech driven facial animation synthesis
图 3. 语音驱动的人脸动画合成工作流程图

我们的语音编码器采用了最先进的自监督预训练语音模型 Wav2vec 2.0 [11]，该模型由音频特征提取器和多层变换器编码器组成。音频特征提取器通过时序卷积网络(TCN) [12]将原始波形的语音转换成特征向量。转换器编码器利用有效的注意力方案，将音频特征转换为上下文语音表示。音频被转换为 16,000 Hz 频率，并使用滑动窗口剪切成片段。Wav2Vec2.0 使用原始论文提供的模型提取音频特征。

我们训练 Transformer 编码器、音频解码器和，冻结编码本和运动解码器 D。为了从大规模语料库的语音表示学习中获益，我们使用预先训练好的 Wav2Vec2.0 权重。音频解码器输出的特征通过公式(1)进一步量化为 $\mathbf{Z}_q$，并由预先训练好的 VQ-VAE 运动解码器进行解码，最后输出人脸运动序列。

在损失函数设计上，运动合成阶段是在两个损失项的约束下进行训练的，损失函数设计为：

$$L_{\text{syn}} = \lambda_1 * L_{\text{reg}} + \lambda_2 * L_{\text{motion}} \tag{5}$$

正则化损失 L_{reg} 对模型的预测值起到引导作用，促使预测值与量化值更加接近，从而提高预测运动序列的正则性。另一方面，运动损失 L_{motion} 衡量预测运动与实际运动之间的差异，从而增强面部表情的真实性。

4. 实验分析

4.1. 实验环境

实验设备系统为 Linux Ubantu18.04，GPU 为单片 Tesla V100，显存为 16 GB。

4.2. 数据集

本模型使用 VOCASET [7]作为训练数据集。VOCASET 由 12 个受试者录制的 480 个配对视听序列组成。面部动作以 60 fps 的速度捕捉，时长约 4 秒。每个 3D 人脸网格都注册到具有 5023 个顶点的 FLAME 拓扑中。

4.3. 定量评测

评测指标选用平均唇顶点误差 LME (Lip Mean Error)。它测量一个序列的唇形相对于真值的偏差，即计算每个帧的所有唇形顶点的最大 L2 误差，并取所有帧的平均值。该指标越小表明生成的人脸动画在嘴唇方面越接近真实值，说话越自然。

由表 1，在相同的训练数据与相同的实验条件下，评测同一个人的同一段说话序列，与基线方法

VOCA12 相比，我们的方法在平均唇误差上取得显著优势。这表明使用先验码本的二阶段的方法在说话人脸动画生成上较传统的一阶段方法具有优势，在复杂的嘴唇动画上也能很好的处理微小的细节。

Table 1. Comparison of lip errors between different models
表 1. 不同模型的唇误差比较

模型	LME (10^{-4} mm)↓
VOCA	7.642
Ours	**5.574**

由表 2，证明变分自动变分编码器学习到的运动先验有助于嘴唇运动的合成，证明使用码本在三维说话人脸动画生成任务上的有效性。

Table 2. Ablation experiment of facial prior codebook
表 2. 人脸先验码本的消融实验

模型	LME (10^{-4} mm)↓
不带码本	8.138
带有码本	**5.574**

4.4. 定性评价

由图 4，输入指定的人脸模板和一段指定的说话音频，模型可以较好的复原人物面部特征，并且驱动嘴部关键点生成面部说话动画，这表明我们方法的有效性。

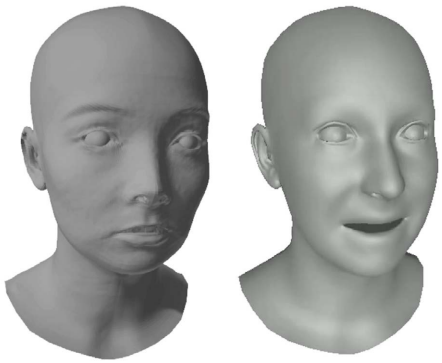

Figure 4. Face template (left) and model generated results (right)
图 4. 人脸模板(左)和模型生成的结果(右)

由图 5，图像序列对应于红色音节文本中的顺序。我们方法的输出具有良好的视觉观感，人物嘴唇逼真。

由图 6，我们的方法与先进的基线方法 FaceFormer [13]在面部动态性上对比(第一行为本文方法，第二行为基线方法)，如红框所示，在眼睛和脸颊的肌肉上，我们的方法明显、更激烈的动作，更具表现力，能传达更多的情感线索。

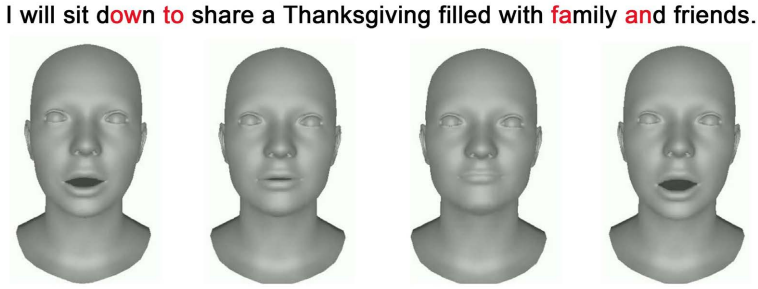

Figure 5. Visualization of facial movements with different pronunciations
图 5. 不同发音的面部动作可视化图

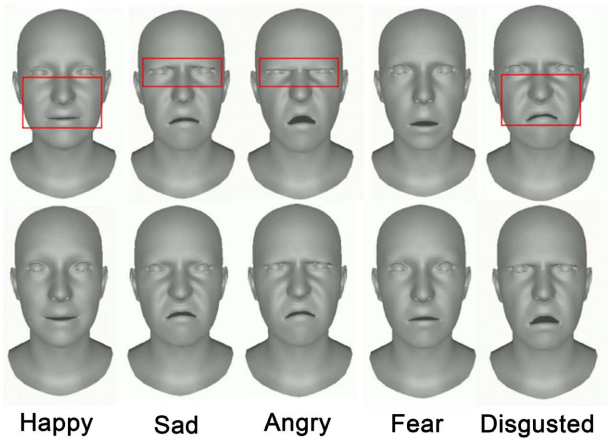

Figure 6. Comparison with SOTA method in facial dynamics (first line of this method)
图 6. 在面部动态性上与 SOTA 方法对比(第一行为本文方法)

4.5. 用户调查

人类可以理解细微的面部动作，因此，这是语音驱动的面部动画任务中的一个可靠指标。与 SOTA 方法 VOCA 和 FaceFormer 相比，我们进行了一项用户研究，以评估动画人脸的面部动态性。此外，为了验证我们提出的方法的有效性，我们对嘴唇同步性和情感表达真实性进行了对比实验。

我们在网上发布了一份问卷进行评估，最终共收到 22 个结果。招聘人员的主要年龄组在 20 至 30 岁之间(21 人)，其余年龄组分布在 30 至 50 岁之间。关于职业分布，21 名是大学生，1 名是教师。我们共设计了 6 个对比案例，每个案例包括 5 个样本，将我们的方法与其他方法进行比较。参与者被要求从几个视频中选择表现最好的一个，最终结果是我们的方法占总数的比例。

如表 3，从这些结果可以看出，受试者更倾向于使用我们的方法。在面部动态性方面，与 VOCA 相比，我们的方法可以获得 84.6%的显著广告优势，与 FaceFormer 相比，可以获得 52.4%的显著广告劣势。我们在动画同步质量和情感传达能力上分别获得了 70.2%、60.2%和 58.6%、50.4%的优势。

Table 3. User study on the quality of facial animation generation
表 3. 关于面部动画生成质量的用户调查

模型	面部动态性	动画同步质量	情感传达能力
Ours vs. VOCA	84.6%	70.2%	58.6%
Ours vs. FaceFormer	52.4%	60.2%	50.4%

5. 结语

我们提出一种基于面部运动先验的语音驱动的三维人脸生成方法，在定量评测中，与先进的方法相比，我们的方法在唇误差上的具有先进性；可视化结果证明了我们方法的可有效性，达到了较好的视觉水平。但我们的方法仍存在一定弊端，我们仍然遵循面部运动与形状无关的假设，其合理性值得进一步研究；此外，总体视觉质量仍然不如真实值，一个主要原因是视听数据的匮乏。在未来的工作中，我们有兴趣利用现有的大规模二维对话头像视频来指导三维面部动画的制作。

基金项目

由北京信息科技大学促进高校分类发展大学生创新创业训练计划项目——计算机学院(5112310855)支持。

参考文献

[1] Edwards, P., Landreth, C., Fiume, E. and Singh, K. (2016) JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization. *ACM Transactions on Graphics*, **35**, 1-11. https://doi.org/10.1145/2897824.2925984

[2] Xing, J.B., Xia, M.H., Zhang, Y.C., Cun, X.D., Wang, J. and Wong, T.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

[3] Peng, Z.Q., Wu, H.Y., Song, Z.B., Xu, H., Zhu, X.Y., Liu, H.Y., He, J. and Fan, Z.X. (2023) EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. arXiv preprint arXiv: 2303.11089.

[4] 宋昕洋, 阎志远, 孙沐毅, 等. 说话人生成研究现状与发展趋势[J]. 计算机科学, 2023, 50(8): 68-78.

[5] Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S. and Singh, K. (2018) Visemenet: Audio-Driven Animator-Centric Speech Animation. *ACM Transactions on Graphics*, **37**, 1-10. https://doi.org/10.1145/3197517.3201292

[6] Richard, A., Zollhofer, M., Wen, Y.D., 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

[7] 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

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, and Polosukhin, I. (2017) Attention Is All You Need. Advances in Neural Information Processing Systems, **30**, 2-3.

[9] 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

[10] van den Oord, A., Vinyals, O. and Kavukcuoglu, K. (2017) Neural Discrete Representation Learning. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Newburyport.

[11] Baevski, A., Zhou, H., Mohamed, A. and Auli, M. (2020) Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477.

[12] Bai, S.J., Kolter, J.Z. and Koltun, V. (2018) An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv:1803.01271.

[13] 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