Computer Science and Application 计算机科学与应用, 2023, 13(12), 2521-2527
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312251

Hans汉斯

一种基于时序损失的语音驱动面部运动方法

王振凯，王承伟，张一帆，李昊渊

河北地质大学信息工程学院，河北 石家庄

收稿日期：2023年11月26日；录用日期：2023年12月22日；发布日期：2023年12月30日

摘　要

语音驱动3D面部运动的研究主要聚焦于拓展多说话人的3D面部运动数据以及获取高质量音频特征上，但采集3D面部运动数据往往需要高昂的成本和繁琐的标注流程，单一说话人的少量数据样本又会导致模型因为数据的稀疏性难以获取高质量音频特征。针对该问题，论文从时间序列任务中获得启发，将可微动态时间规整(Smoothed formulation of Dynamic Time Warping, Soft-DTW)应用到语音特征与面部网格(Mesh)顶点的跨模态对齐中。经过实验表明，采用Soft-DTW作为损失函数在生成面部动画的唇形吻合度方面相较于使用均方误差(Mean Squared Error, MSE)时有所提高，可以合成更高质量的面部动画。

关键词

语音驱动，跨模态对齐，面部动画，Soft-DTW

A Speech-Driven Facial Motion Method Based on Temporal Loss

Zhenkai Wang, Chengwei Wang, Yifan Zhang, Haoyuan Li

College of Information Engineering, Hebei University of Geosciences, Shijiazhuang Hebei

Received: Nov. 26th, 2023; accepted: Dec. 22nd, 2023; published: Dec. 30th, 2023

Abstract

Research on voice-driven 3D facial motion primarily focuses on expanding 3D facial motion data for multiple speakers and obtaining high-quality audio features. However, the collection of 3D facial motion data often entails high costs and a labor-intensive annotation process. Additionally, having a limited amount of data samples for a single speaker can make it challenging for models to obtain high-quality audio features due to data sparsity. To address this issue, this study draws inspiration from temporal tasks and applies the concept of Smoothed Dynamic Time Warping (Soft-DTW) to the cross-modal alignment between speech features and facial mesh vertices. Expe-

文章引用: 王振凯, 王承伟, 张一帆, 李昊渊. 一种基于时序损失的语音驱动面部运动方法[J]. 计算机科学与应用, 2023, 13(12): 2521-2527. DOI: 10.12677/csa.2023.1312251

rimental results have shown that using Soft-DTW as a loss function leads to improved lip synchronization in generating facial animations compared to using Mean Squared Error (MSE). This approach enables the synthesis of higher-quality facial animations.

Keywords

Speech-Driven, Cross-Modal Alignment, Facial Animation, Soft-DTW

1. 引言

近些年来，语音驱动 3D 面部动画已成为学术和工业领域中的重要研究方向之一。该技术能够仅通过语音输入为数字角色生成高质量的面部动画，在 VR、电影特效、游戏以及教育等领域存在着广泛的应用。其中面部动画与语音输入的口型匹配至关重要，因为这一因素直接影响着最终呈现的动画质量，所以面部口型的吻合度是评估面部动画质量时不可忽视的关键标准之一。

3D 面部动画主要通过两类方法生成，一类方法是基于发音规则建立音素与口型动画的映射关系，Pif 等[1]通过显式参数构建出一套音素(Phoneme)与视素(Visemes)间的协同发音规则，Taylor 等[2]则是建立音素到口型的一对多映射，通过视素生成动态离散的口型动画。Xu 等人[3]通过构造一个用于建模协同发音效果的规范集，通过简化音素集中的成对组合来实现文本–语音同步的面部运动。Sako 等[4]将隐马尔可夫模型(Hidden Markov model，HMM)应用到面部动画合成中，较好地还原了面部动画。该类方法主要通过手工方式构建特征，对手工规则依赖性高，因此不适用于高度自动化的应用场景。另一类方法是提取各种多种类型的音频特征，训练出端到端模型完成对面部动画的预测，Zhou 等人[5]在 JALI 的基础上提出了 VisemeNet 模型，通过长短期记忆网络(Long Short-Term Memory, LSTM)取代人工，完成对音素地提取和面部运动地映射。Karras 等[6]在原始音频中提取出线性预测编码(Linear Predictive Coding, LPC)特征并将该特征输入到卷积神经网络(Convolutional Neural Networks)中，实现端到端地预测面部运动。Hochreiter 等[7]为解决传统神经网络无法对大跨度上下文联系进行建模的问题，提出长短期记忆网络策略。Schuster 等[8]提出双向递归神经网络，可同时学到上下文音频信息。Cudeiro 等[9]为解决数据缺乏问题，发布了包含多个说话人的 VOCASET 数据集，并且提出 VOCA 模型将说话风格泛化到不同说话人。Richard 等[10]人则是更进一步，考虑到眨眼动作与语音信息的弱相关性导致上面部静止的问题，通过为面部表情增加可学习的分类潜空间，实现了语音信息与面部动画的解耦。Fan 等[11]人则是首次将预训练语音模型作为特征提取器，缓解了语音数据稀疏的难题，但在口型准确性上仍有待提高。Chen 等人[12]基于 LSTM 模型，首次引入 Soft-DTW 作为损失函数，并通过文字转语音(Text-To-Speech, TTS)技术进行数据增强，提高了说话人面部动作在 52 维混合形状(Blend Shape)上的少样本泛化能力。该类方法引入了深度学习方法获取特征，但并未解决音频特征与人脸模型 Mesh 顶点运动序列间的跨模态对齐问题，导致合成的口型在准确度上仍有所欠缺。

综上所述，提高面部口型动画准确度的关键点在于实现语音特征与面部动作的跨模态对齐，而当前 3D 视听数据集中存在多个说话人在相同内容上的不同语音，使得在拓展数据的同时，加大了语音与口型的对齐难度。考虑到语音信号是一个不规则变长序列，难以通过欧氏距离进行对齐。因此本文从可微的

时序学习损失 Soft-DTW [13]中获得启发，采用 Soft-DTW 取代 MSE 作为模型的损失函数，在扩展视听数据的同时，提高了面部口型动画的准确性。

2. 整体框架

2.1. 改进 Tansformer 模型结构

在语音驱动面部运动任务中，模型接收语音输入，以自回归的方式预测出输入语音所对应的面部 Mesh 顶点的运动序列，属于序列到序列(Sequence-to-sequence, Seq2seq)任务，因此采用改进 Transformer [14]架构的编码器-解码器(Encoder-Decoder)模型。模型的整体结构如图 1 所示。

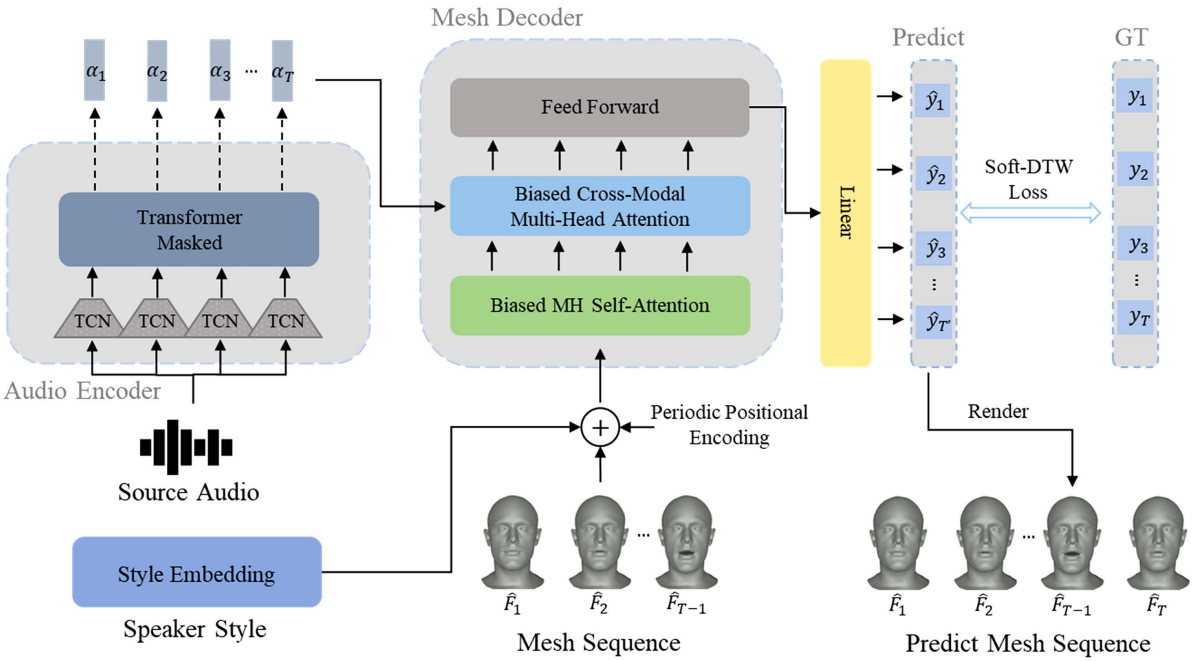

Figure 1. Model network framework
图 1. 模型网络框架

在整个模型中，编码器将输入语音信号转化为上下文相关的语音表示 $A_{1:T}=(a_1,a_2,a_3,\cdots,a_T)$，其中 T 为语音表示的帧长度。风格嵌入层(Style Embedding)通过一组可学习的向量表示不同说话人的说话风格。将语音表示、说话人风格嵌入以及历史的面部运动序列 $\hat{F}_{1:T-1}$ 输入到解码器中，采用 Soft-DTW 损失函数作为惩罚项，自回归的预测并渲染出完整的面部动作动画序列 $\hat{F}_{1:T}=(\hat{F}_1,\hat{F}_2,\cdots,\hat{F}_T)$。

2.2. 编码器 – 解码器设计

由于语音数据的匮乏，整个编码器部分的结构设计参考 wav2vec 2.0 [15]，并采用 wav2vec 2.0 的预训练权重对整个音频编码器部分进行初始化。通过这种方式，预训练模型可以大大减少特征提取时所带来的信息损耗，从而缓解因数据稀疏所带来的口型不准确问题。

在音频编码器(Audio Encoder)中，输入音频的频率被转化为 16 Khz 后，经多个时间卷积层(TCN)从原始信号中提取出对应的离散化特征向量，输出的频率为 49 Hz，每个样本之间的步长为 20 ms，并产生 25 ms 音频的感受野，这些特征向量会被输入到由 Transformer 层堆砌而成的 Encoder 部分，转化为上下文相关的语音表示。整个流程如公式(1)所示。

$$A_{1:T} = \text{AudioEncoder}(S) \tag{1}$$

其中 S 为输入语音，$A_{1:T} = (a_1, a_2, a_3, \cdots, a_T)$ 为 Encoder 层提取出的语音表示。

解码器部分的输入主要由编码器部分输出的语音表示 $A_T \in \mathbb{R}^{64}$、说话人风格嵌入向量 $F_e \in \mathbb{R}^{64}$ 以及历史面部运动序列 $\hat{F}_T \in \mathbb{R}^{64}$ 三部分组成。考虑到 Transformer 对长序列的弱泛化能力，这里引入 FaceFormer 中采用的周期性位置编码(Periodic Positional Encoding, PPE)对输入信息进行处理，以提升模型对长序列的泛化能力，这里将面部运动序列的历史信息与说话人风格嵌入向量沿同一维度进行拼接，如公式(2)所示。

$$F = \text{Concatenate}\left(F_e, \hat{F}_T\right) \tag{2}$$

拼接后的输入特征 F 经过周期性位置编码进行处理，处理后的信息被输入到网格解码器(Mesh Decoder)中，经过自注意力机制与跨模态多头注意力机制后到达前馈层，然后将前馈层的输出结果通过全连接层的解码映射到 5023 维的 3D 面部顶点空间中，最终获得预测出的面部运动序列结果。

2.3. Soft-DTW 损失函数

对于语音序列，因为说话风格上的差异性，通常无法保证在相同说话内容下，语音序列与面部运动序列具有相同的长度。因此在这种变长情况下，采用 MSE 损失函数计算语音预测出的面部运动序列与真实面部运动序列便不再合适。而动态时间规整[16] (Dynamic Time Warping, DTW)中采用动态规划来解决两个时间序列中的最小距离对齐问题，对时间维度上的移位或膨胀具有鲁棒性。Soft-DTW 进一步解决了原始 DTW 因动态规划过程导致的不可微问题，使其可以作为损失函数进行导数运算。

本文从 Soft-DTW 中获得启发，通过引入 Soft-DTW 作为损失函数，通过计算出面部运动序列 $(\hat{y}_1, \hat{y}_2, \hat{y}_3, \cdots, \hat{y}_{T'})$ 与真实面部运动序列 $(y_1, y_2, y_3, \cdots, y_T)$ 之间的 DTW 距离，以实现预测序列与真实序列内数据对之间的对齐。采用 $\omega_{i,j}$ 表示数据对中 $\hat{y}_i$ 与 y_j 之间的欧式距离，如公式(3)所示。

$$\omega_{i,j} = \left\| \hat{y}_i - y_j \right\|_2 \tag{3}$$

然后采用动态规划的方式寻找最优路径，并将预测序列与真实值序列之间的最小成本表示为公式(4)。

$$\varphi_{i,j} = \omega_{i,j} + \min\left\{\varphi_{i-1,j}, \varphi_{i,j-1}, \varphi_{i-1,j}, \varphi_{i-1,j-1}\right\} \tag{4}$$

考虑到 $\varphi_{i,j}$ 中包含的 min 运算无法进行导数计算，因此采用 Soft-DTW 引入最小算子的平滑公式，如公式(5)所示。

$$\min{}^{\gamma}\left\{a_1, a_2, \cdots, a_n\right\} := \begin{cases} \min_{i \le n} a_i & \gamma = 0 \\ -\gamma \log \sum_{i=1}^{n} \mathrm{e}^{-a_i/\gamma} & \gamma > 0 \end{cases} \tag{5}$$

当 $\gamma > 0$ 时，$\varphi_{i,j}$ 可微，结合公式(4)与公式(5)，可得 Soft-DTW 损失函数的计算公式，如公式(6)所示。

$$L_{\text{Soft-DTW}} = \omega_{i,j} - \gamma \log\left(\mathrm{e}^{-\varphi_{i-1,j}/\gamma} + \mathrm{e}^{-\varphi_{i,j-1}/\gamma} + \mathrm{e}^{-\varphi_{i-1,j-1}/\gamma}\right) \tag{6}$$

对公式(6)进行梯度运算，即可完成对参数的更新。

3. 实验与分析

3.1. 实验设置

本文使用公开数据集 VOCASET 进行训练与测试，该数据集提供了英语语音信息以及对应的 3D 面部运动扫描结果。共包含来自 12 名说话人的 480 个面部运动序列，每个序列以 60 FPS 的速度捕获，序列长度在 3 到 4 秒左右。每个 3D 面部网格为 5023 个 3 维顶点的 FLAME 拓扑。为保证实验的公平性，

采用与 VOCA 相同的数据分割方法，从 VOCASET 中分割出对应的训练、验证以及测试集。

本文所提出的模型接收语音信息以及对应面部运动序列作为输入，并在预处理阶段，将音频采样率转化为 16 kHz，动画转化为 30 FPS 进行处理，实现语音信息与面部动画序列的对齐。训练阶段，采用 Adam 优化器，并设置学习率为 1e−4，批量大小为 1，Soft-DTW 中的系数 γ 为 1e−3。整个网络在 NVIDIA V100 32G 上进行训练，训练 100 个周期，时间大约为 6 个小时。

测试环节基于分割后的数据集，在 VOCA、MeshTalk 以及 FaceFormer 三种方法上与本文进行比较，并采用论文作者所提供的预训练权重进行测试。

3.2. 定量评价

为了衡量唇部的同步性，计算预测出的面部序列相较于真实面部运动序列的唇部顶点偏差(Lip Vertex Error, LVE)，即计算每帧中唇部顶点的 L2 误差，然后取所有帧的平均值。LVE 指标越低，说明预测结果相较于真实值的唇部顶点误差越小，生成面部动画的准确度也就越高，否则相反。结果如表 1 所示。

Table 1. Evaluation result on VOCASET-Test
表 1. VOCASET-Test 评估结果

方法	LVE ($\times 10^{-5}$ mm)
VOCA	4.9245
MeshTalk	4.5441
FaceFormer	4.1090
Ours	3.8212

从表 1 中可以看出，本文提出的方法相较于其他方法有着更低的错误率和更高的准确性，这表明了使用 Soft-DTW 损失实现对齐有助于提高面部运动动画的准确性与可理解性，可以产生更高质量的口型效果。

3.3. 定性评价

对于面部运动来说，动画是最直观的表现形式。因此除了进行定量评估外，可视化预测结果也是非常重要的一环。在定性评价中，我们为 VOCA、MeshTalk、FaceFormer 以及本文方法提供了相同的语音内容输入，通过将预测结果与真实值进行可视化对比来直观展示方法的性能表现。结果如图 2 所示：

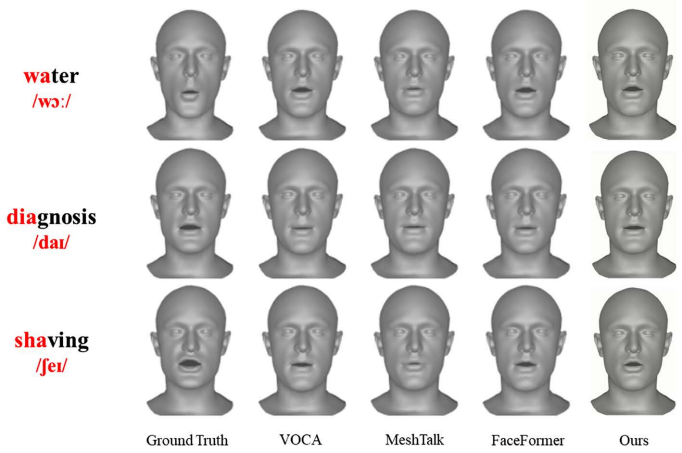

Figure 2. Comparison of visual results
图 2. 可视化结果对比

相较于其他方法，本文的预测结果在口型表现上更贴近于真实值的效果，产生了更加准确的面部运动。例如对于音节“/daɪ/”，本文中口型表现出更加明显的张开动作。当产生音节“/ʃeɪ/”时，本文也具备更加明显的上嘴唇向上运动，进一步表明了 Soft-DTW 对于提高面部运动准确性是有效的。

4. 结论

本文在 FaceFormer 的基础上提出了一种基于时序损失的语音驱动面部运动方法，通过分析 VOCASET 开源数据集的数据内容，引入 Soft-DTW 时序损失函数，实现对不同说话人的相同说话内容进行对齐，在保证说话风格泛化性的同时，有效减少了因对齐误差带来的面部口型不准确问题。定量评估与定性评估显示，本文方法具有更低的唇部顶点误差和更高的面部动画质量，探索了时序损失函数在语音生成面部 Mesh 序列任务中的可靠性，为后续进一步提高面部动画质量铺平了道路。

但是本文中仍存在局限，由于 Transformer 层中自注意力机制以及 Soft-DTW 的引入，导致模型训练时间较长，在时间成本中并不具备优势，对 GPU 设备也有着较高的显存需求，不适合在实时应用场景中使用。未来的研究主要是针对如何在保证面部运动效果的同时，降低时间与部署成本，实现实时面部运动推理。

基金项目

省部级社科类重点项目(项目名称：弘扬地质精神，传承优良学风项目编号：XFCC2023ZZ028)。

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**, Article No. 127. https://doi.org/10.1145/2897824.2925984

[2] Taylor, S.L., Mahler, M., Theobald, B.-J. and Matthews, I. (2012) Dynamic Units of Visual Speech. *Proceedings of the ACM SIGGRAPH/Eurographics Conference on Computer Animation*, Lausanne, 29-31 July 2012, 275-284.

[3] Xu, Y.Y., Feng, A.W., *et al*. (2013) A Practical and Configurable Lip Sync Method for Games. *Proceedings of Motion on Games*, Dublin, 6-8 November 2013, 131-140. https://doi.org/10.1145/2522628.2522904

[4] Sako, S., Tokuda, K., Masuko, T., *et al*. (2000) HMM-Based Text-To-Audio-Visual Speech Synthesis. *Sixth International Conference on Spoken Language Processing*, Beijing, 16-20 October 2000. https://doi.org/10.21437/ICSLP.2000-469

[5] Zhou, Y., Xu, Z., Landreth, C., *et al*. (2018) VisemeNet: Audio-Driven Animator-Centric Speech Animation. *ACM Transactions on Graphics*, **37**, Article No. 161. https://doi.org/10.1145/3197517.3201292

[6] Karras, T., Aila, T., Laine, S., *et al*. (2017) Audio-Driven Facial Animation by Joint End-To-End Learning of Pose and Emotion. *ACM Transactions on Graphics* (*TOG*), **36**, Article No. 94. https://doi.org/10.1145/3072959.3073658

[7] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Schuster, M. and Paliwal, K.K. (1997) Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **45**, 2673-2681. https://doi.org/10.1109/78.650093

[9] Cudeiro, D., Bolkart, T., Laidlaw, C., *et al*. (2019) Capture, Learning, and Synthesis of 3D Speaking Styles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 10101-10111. https://doi.org/10.1109/CVPR.2019.01034

[10] Richard, A., Zollhöfer, M., Wen, Y., *et al*. (2021) MeshTalk: 3d Face Animation from Speech Using Cross-Modality Disentanglement. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Virtual, 11-17 October 2021, 1173-1182. https://doi.org/10.1109/ICCV48922.2021.00121

[11] Fan, Y., Lin, Z., Saito, J., *et al*. (2022) FaceFormer: Speech-Driven 3d Facial Animation with Transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 18770-18780. https://doi.org/10.1109/CVPR52688.2022.01821

[12] Chen, Q., Ma, Z., Liu, T., *et al*. (2023) Improving Few-Shot Learning for Talking Face System with TTS Data Augmentation. *ICASSP* 2023-2023 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Rhodes Island, 04-10 June 2023, 1-5. https://doi.org/10.1109/ICASSP49357.2023.10094777

[13] Cuturi, M. and Blondel, M. (2017) Soft-DTW: A Differentiable Loss Function for Time-Series. *International Conference on Machine Learning*, Sydney, 6-11 August 2017, 894-903.

[14] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017.

[15] Baevski, A., Zhou, Y., Mohamed, A., *et al*. (2020) wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 34*th Conference on Neural Information Processing Systems* (*NeurIPS* 2020), Vancouver, 6-12 December 2020, 12449-12460.

[16] Sakoe, H. (1971) A Dynamic-Programming Approach to Continuous Speech Recognition. https://www.semanticscholar.org/paper/A-Dynamic-Programming-Approach-to-Continuous-Speech-Sakoe-Chiba/2d2eb229c21269ffaa8a85b0961a2bda1116a6c7#citing-papers