基于 Transformer，利用标记磁共振成像从语音过程中的舌运动合成音频

Xiaofeng Liu[a]、Fangxu Xing[a]、Jerry L. Prince[b]、Maureen Stone[c]、Georges El Fakhri[a] 和 Jonghye Woo[a]

[a]戈登医学影像中心，马萨诸塞州总医院和哈佛医学院，美国马萨诸塞州波士顿 02114[b]约翰斯·霍普金斯大学电气与计算机工程系，美国马里兰州巴尔的摩 21218[c]马里兰大学牙科学院神经与疼痛科学系，美国马里兰州巴尔的摩 21201

arXiv:2302.07203v1 [eess.IV] 14 Feb 2023

摘要

研究舌头内部组织点运动与口咽肌肉变形（通过标记磁共振成像测量）和可理解语音之间的关系，有助于推进言语运动控制理论，并开发治疗语音相关障碍的新方法。然而，阐明这两类信息来源之间的关系颇具挑战，部分原因在于时空运动场（即四维运动场）与一维音频波形的数据结构存在差异。本文提出一种高效的编码器-解码器翻译网络，通过二维频谱图作为音频数据的替代表示，探索四维运动场中蕴含的预测信息。具体而言，我们的编码器基于三维卷积空间建模和基于 Transformer 的时间建模。提取出的特征经由非对称二维卷积解码器处理，生成与四维运动场相对应的频谱图。此外，我们将生成对抗训练方法纳入框架，以进一步提升所生成频谱图的合成质量。我们在 63 组配对的运动场序列和语音波形上进行了实验，结果表明，该框架能够根据运动场序列生成清晰的音频波形。因此，该框架有望加深我们对这两种模态之间关系的理解，并为语音障碍治疗方法的开发提供参考。

关键词： 运动场、Transformer、音频合成、磁共振成像。

1. 引言

为了加深我们对健康人群和患病人群（如舌癌患者）言语运动控制的理解，确定动态磁共振成像（MRI）数据与语音音频波形之间的关系至关重要。这有助于我们将舌头和口咽肌肉变形与相应的声学信息联系起来。来自三维（3D）标记磁共振成像[1] 序列的内部组织点跟踪数据，所包含的舌头和口咽运动信息远多于通过电影磁共振成像[2] 和标记磁共振成像获得的更传统的二维（2D）正中矢状面图像序列。[3]然而，将这些四维（4D）变形场与语音音频波形联系起来面临以下挑战：1）如何从复杂且高维的舌头和口咽变形中高效提取特征；2）四维运动场与高频一维（1D）音频波形之间存在异构数据表示。

为应对这些挑战，我们提出一种新颖的框架，利用高效的异构翻译框架从四维运动场合成二维梅尔频谱图。我们采用二维频谱图作为代理表示，这种表示常用于音视频翻译任务；本研究中，它由一维音频波形转换而来，如.[2] 以往将二维 MRI 序列翻译为

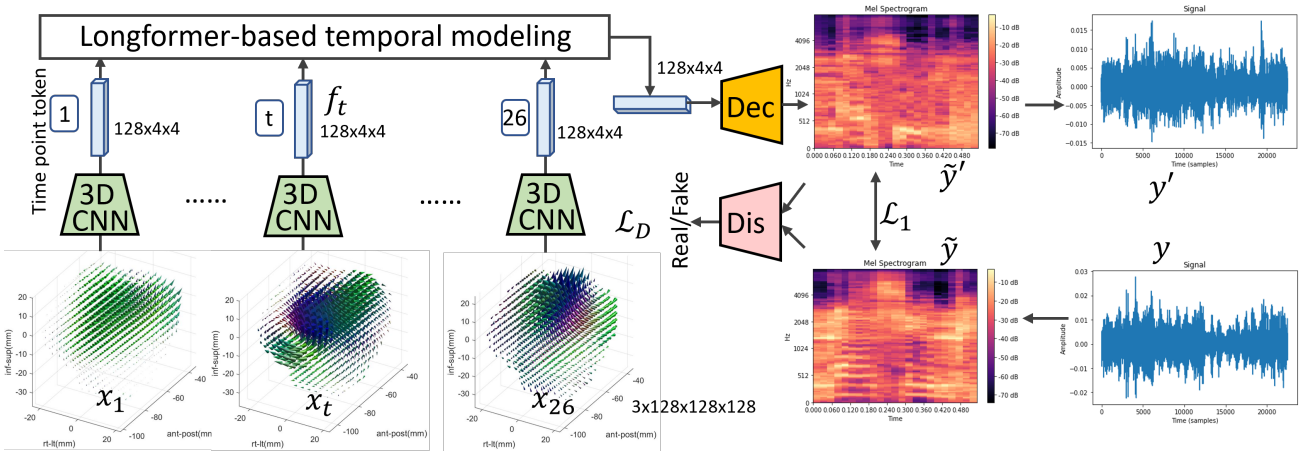

图1. 我们用于从一系列运动场合成音频波形的框架示意图，该框架由两阶段编码器（包含三维卷积神经网络和 Longformer）、二维卷积解码器（Dec）以及判别器（Dis）组成。

音频[2,3] 的研究表明，二维频谱图是此任务的有效表示，因为它能够捕捉声能随时间在不同频率上的分布。[4–7] 为充分利用四维运动场丰富的时空信息，我们提出一种新颖且高效的编码器网络，将三维卷积神经网络（CNN）与基于 Longformer 的 Transformer 模块相结合，从运动场合成频谱图。具体而言，我们使用三维 CNN 对每个时间点的运动场进行空间建模，然后采用 Longformer[8]-基于 Transformer 的模块进行时间建模。与二维 MRI 序列处理中使用的传统时间建模方法（如递归神经网络（RNN）和三维 CNN）相比，[2–4] 我们的 Transformer 模块所需的参数数量减少了 1。$5\times$ 参数更少，并且比传统方法只需更少的训练样本即可完成训练。随后，我们采用二维卷积生成器生成频谱图，再将其转换回相应的音频波形。[9] 我们还引入生成对抗训练，以进一步提升合成频谱图的质量。我们的框架首次尝试学习四维运动场与音频波形之间的映射，有望帮助更好地理解运动与可理解语音之间的关系。

2. 方法

给定一组尺寸为 $3\times N\times H\times W\times T$ 的运动场 x 及其对应的一维波形 y，其中 N，H，W 和 T 分别表示正中矢状切片编号、高度、宽度和时间帧数。值得注意的是，运动场中的每个体素都有三个通道，用于表示三维方向；而电影或标记MRI序列的数据结构则更简单，为 $H\times W\times T$。[2,3] 每段音频波形y 均使用 Librosa*预处理为二维梅尔频谱图 $\tilde{y}\in\mathbb{R}^{64\times 64}$。梅尔频谱图采用梅尔尺度（赫兹尺度的一种非线性变换），以突出 40 至 1000 Hz 范围内的人声频率，并抑制高频乐器噪声。本研究旨在学习一个端到端的异构转换器 $\mathcal{T}$：$x\rightarrow\tilde{y}'$，使其近似于 $\tilde{y}$。

为了利用 x 中丰富的信息，我们采用模块化设计对空间和时间信息进行建模，这与.[10] 类似。首先，在每个时间点 $t\in\{1,\cdots,T\}$ 的三通道三维运动场上应用一个三维卷积神经网络模块，以提取紧凑的表征特征 $f_t\in\mathbb{R}^{128\times 4\times 4}$。每个运动场所用的三维卷积神经网络的详细结构如表 1 所示。在以往的视频到音频合成研究中，RNN 和三维卷积神经网络被广泛用于时间建模。然而，将 RNN 和三维卷积神经网络分别用于时间建模时，都面临各自的挑战，包括在数据集有限的情况下难以训练 RNN[3]，以及三维卷积神经网络难以建模长期相关性。[11]

*Librosa：从音频波形生成梅尔频谱图。

表 1. 使用标记磁共振成像，通过 Transformer 从语音期间的舌运动合成音频的所提网络结构

编码器（三维卷积神经网络）+Longformer		解码器	
层	Size	层	Size
输入	(3, 128, 128, 128)×26	重塑	(128, 4, 4)
三维卷积 (32) & ReLU	(32, 128, 128)×26	二维转置卷积(96) & ReLU	(96, 8, 8)
最大池化	(32, 64, 64)×26		
三维卷积 (32) & ReLU	(32, 64, 64)×26	二维转置卷积(24) & ReLU	(24, 16, 16)
最大池化	(32, 32, 32)×26		
三维卷积(64) & ReLU	(64, 32, 32)×26	二维转置卷积(4) & ReLU	(4, 32, 32)
最大池化	(64, 16, 16)×26		
三维卷积 (64) 和 ReLU	(64, 16, 16)×26	二维转置卷积(1) 和 Sigmoid	(1, 64, 64)
最大池化	(64, 8, 8)×26		
三维卷积 (128) & ReLU	(128, 8, 8)×26		
最大池化	(128, 4, 4)×26	Librosa 转换为音频波形	
Longformer（3 层，滑动窗口尺寸为 3）	(128, 4, 4)		

受视觉 Transformer 领域近期进展的启发，[10] 我们提出采用一个应用注意力机制以探索序列内部全局依赖关系的 Transformer 模块。虽然标准 Transformer 只能处理有限数量的 token 或时间点之间的成对相关性，难以扩展到长序列，但近期发展的 Longformer[8] 采用滑动窗口注意力，其计算复杂度相对于序列长度呈线性增长。与文献[10]，中的做法一样，Longformer 模块同时接收每个时间点的特征和时间点索引 t ，以融合信息并生成序列级表示。值得注意的是，Transformer 中使用的注意力方案具有置换不变性，因此时间点索引对于嵌入序列信息至关重要。我们采用三层 Longformer 作为时序 Transformer 模块。其处理流程如图 1 所示。Longformer 模块的参数量比5× 基于三维卷积神经网络的时间建模少[10] ，因此构成了一个更轻量的网络，在数据集相对有限时可能优于三维卷积神经网络。与传统的基于三维卷积神经网络的时间建模[10] 只能关注相邻帧以进行短期时间建模不同，Transformer 模块能够建模长期时序关系，从而有望获得更具表征能力的音频特征。在生成四维运动场的全局表示后，我们使用 2D 解码器生成 2D 频谱图 $\tilde{y}$，并将其与真实值 $\tilde{y}$ 进行比较，采用 L1 损失 $\mathcal{L}_1$。

我们还加入了一个生成对抗网络（GAN）模块，以进一步提升生成的梅尔频谱图的质量。判别器 $\mathcal{D}$ 同时将真实频谱图 $\tilde{y}$ 和生成频谱图 $\tilde{y}'$作为输入，任务是识别哪些是生成的，哪些是真实的。判别器的二元交叉熵损失可表示为

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{y'}\{\log(\mathcal{D}(y'))\} + \mathbb{E}_{\tilde{y}'}\{\log(1-\mathcal{D}(\tilde{y}'))\}. \tag{1}$$

相比之下，转换器试图通过生成逼真的频谱图来欺骗判别器。[12] 值得注意的是，转换器（$\mathcal{T}$）在 $\log(\mathcal{D}(y'))$ 中不涉及真实频谱图。[13] 因此，可以通过优化以下目标来训练转换器

$$\mathcal{L}_{\mathcal{T}} = \mathbb{E}_{\tilde{y}'}\{-\log(1-\mathcal{D}(\tilde{y}'))\} + \beta\mathcal{L}_1. \tag{2}$$

训练阶段结束后， $\tilde{y}'$ 可以转换回音频波形 y'[†]。

3. 结果

为评估我们的框架，我们使用西门子 3.0T TIM Trio 系统采集了一组配对的磁共振成像序列和音频数据。采集的数据共包含43名受试者，他们执行了“a geese”，另有

[†]Librosa：根据梅尔频谱图生成音频波形。

Table 2. 使用留一法评估进行测试时的数值比较。最佳结果以**粗体**显示 **d**.

方法	使用 GAN	频谱图的 Corr2D ↑	波形的感知语音质量评估 ↑
三维卷积神经网络 +Transformer	√	**0.820**±0.017	**1.646**±0.019
三维卷积神经网络 +Transformer	×	0.818±0.015	1.632±0.022
两阶段三维卷积神经网络	√	0.812±0.018	1.630±0.020
两阶段三维卷积神经网络	×	0.807±0.021	1.625±0.017

20名受试者执行了“a souk”，并跟随周期性节拍器样声音。随后，我们根据标记磁共振成像计算了语音任务期间的一系列体素级运动场。[1,14] 本研究的数据使用配备12通道头线圈和4通道颈线圈的西门子 3.0T TIM Trio 系统采集，采用分段梯度回波序列。[15,16] 四维运动场 x 的尺寸为 $3 \times 128 \times 128 \times 128 \times 26$。相比之下，数据集中配对的一维音频录音长度为21,832至24,175个样本。我们采用滑动窗口将音频波形裁剪至21,000个样本，生成 $100\times$ 个音频波形用于数据增强。随后，我们使用公开可用的 Librosa 库将音频波形转换为 Mel 频谱图，尺寸为 64×64。测试时，我们以受试者独立方式进行留一法评估。

我们的框架使用 PyTorch 实现，并在 NVIDIA V100图形处理器上训练。对于 Longformer，我们使用了一个由三帧组成的注意力窗口，并将其应用于每一层。我们将动量设为 0.5，并分别将编码器-解码器和判别器的学习率设为 10^{-3} 和 10^{-4}。损失项 $\mathcal{L}_T$通过 $\beta = 1$进行平衡。在测试中，对单个受试者的推理时间不到 0.5 秒。我们将所提出的三维卷积神经网络模块应用于每个时间帧的运动场，然后使用三维卷积神经网络或基于 Longformer 的 Transformer 模块进行时间建模，分别称为两阶段三维卷积神经网络或 Transformer。

图 1 展示了预测频谱图和音频波形的示例，表明可以根据运动场序列生成音频。为了量化生成频谱图在频域中的质量，以及音频波形在时域中的质量，我们分别采用了参考文献[3,4] 中所用的二维皮尔逊相关系数（Corr2D）和语音质量感知评估（PESQ）。Corr2D 和 PESQ 的值越高，表示合成性能越好。表 2 展示了数值对比结果，包括三次随机试验的标准差。我们提出的 Transformer框架由三维卷积神经网络和 Longformer 组成，在 Corr2D 和 PESQ 指标上均取得了更优的性能。

与两阶段三维卷积神经网络相比，训练 200 个 epoch 所需的时间和测试时的推理时间分别快了 1.7× 和 1.3× 。我们进行了消融研究，以检验生成对抗网络损失的有效性，结果发现该损失能够提升性能。对参数 β 进行的敏感性分析表明，该参数用于平衡生成对抗网络损失和 L1 损失；当 $\beta \in [0.5, 7]$ 时，系统性能相对稳定。

4. 结论

在本研究中，我们提出了一种新颖的合成框架，可将运动场序列转换为相应的频谱图。我们的模块化两阶段框架结合了基于三维卷积神经网络的空间信息建模与基于 Transformer 的时间建模，从而有效利用规模较小的训练集，并处理复杂的四维运动场结构。我们还成功应用了对抗训练，进一步提升性能。实验结果表明，该框架能够根据四维运动场生成频谱图和清晰可懂的音频；在时间建模方面，其表现优于三维卷积神经网络。该框架还有望用于其他涉及异构时间序列转换的任务。

致谢

本研究由美国国立卫生研究院 R01DC014717、R01DC018511 和 R01CA133015 项目资助。

参考文献

[1] Woo, J., Xing, F., Prince, J. L., Stone, M., Gomez, A. D., Reese, T. G., Wedeen, V. J., and El Fakhri, G., “一种深度联合稀疏非负矩阵分解框架，用于识别语音过程中舌运动的共性功能单元和个体特异性功能单元”, 医学图像分析 72, 102131 (2021)。[2] Liu, X., Xing, F., Stone, M., Prince, J. L., Kim, J., El Fakhri, G., and Woo, J., “Cmri2spec：通过成对异构转换器将电影MRI序列合成为频谱图”，载于 [ICASSP], 1481–1485, IEEE (2022)。[3] Liu, X., Xing, F., Stone, M., Prince, J. L., Kim, J., El Fakhri, G., and Woo, J., “利用成对异构深度转换器将标记MRI合成为音频”，《美国声学学会杂志》151(4), A133–A133 (2022)。[4] Akbari, H., Arora, H., Cao, L., and Mesgarani, N., “Lip2audspec：从无声唇部运动视频重建语音”，载于 [ICASSP], 2516–2520, IEEE (2018)。[5] Ephrat, A. and Peleg, S., “Vid2speech：从无声视频重建语音”，载于 [ICASSP], 5095–5099, IEEE (2017)。[6] He, G., Liu, X., Fan, F., and You, J., “Image2audio：利用面部表情图像促进半监督音频情感识别”，载于 [CVPR], 912–913 (2020)。[7] He, G., Liu, X., Fan, F., and You, J., “分类感知半监督域适应”，载于 [CVPR],964–965 (2020)。[8] Beltagy, I., Peters, M. E., and Cohan, A., “Longformer：长文档 Transformer”，arXiv预印本 arXiv:2004.05150 (2020)。[9] Griffin, D. and Lim, J., “基于改进短时傅里叶变换的信号估计”，IEEE声学、语音与信号处理汇刊 32(2), 236–243 (1984)。[10] Neimark, D., Bar, O., Zohar, M., and Asselmann, D., “视频Transformer网络”，载于 [ICCV], 3163–3172 (2021)。[11] Wang, J., Liu, X., Wang, F., Zheng, L., Gao, F., Zhang, H., Zhang, X., Xie, W., and Wang, B., “基于多视图超声心动图自动解读先天性心脏病”，医学图像分析 69, 101942 (2021)。[12] Liu, X., Xing, F., Prince, J. L., Carass, A., Stone, M., El Fakhri, G., and Woo, J., “用于标记到电影磁共振图像合成的双循环约束双射VAE-GAN”，载于 [ISBI], 1448–1452, IEEE (2021)。[13] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X., “改进生成对抗网络训练的技术”，NIPS 29, 2234–2242 (2016)。[14] Xing, F., Woo, J., Gomez, A. D., Pham, D. L., Bayly, P. V., Stone, M., and Prince, J. L., “用于从标记磁共振图像估计运动的相位向量不可压缩配准算法”，IEEE医学影像汇刊 36(10) (2017)。[15] Lee, J., Woo, J., Xing, F., Murano, E. Z., Stone, M., and Prince, J. L., “利用动态MRI进行三维运动分析的舌部半自动分割”，载于 [ISBI], 1465–1468, IEEE (2013)。[16] Xing, F., Woo, J., Murano, E. Z., Lee, J., Stone, M., and Prince, J. L., “基于标记和电影MR图像的三维舌运动”，载于 [MICCAI], 41–48, Springer (2013)。