利用磁共振成像数据和音标表示从语音重建声道

Sofiane Azzouz*Multispeech*，洛林大学´、法国国家科学研究中心、法国国家信息与自动化研究院，*F-54000*法国南锡
sofiane.azzouz@loria.fr

Pierre-Andr´e Vuissoz*IADI*，洛林大学´、法国国家健康与医学研究院、*IADI U1254*，*F-54000*法国南锡 pa.vuissoz@chru-nancy.fr

Yves Laprie*Multispeech*，洛林大学´、法国国家科学研究中心、法国国家信息与自动化研究院，*F-54000*法国南锡
yves.laprie@loria.fr

arXiv:2603.11847v1 [eess.AS] 12 Mar 2026

摘要—**发音器官声学逆推旨在根据语音信号重建声道的完整几何形状。本文对不同语音分割准确度进行了比较研究，并与我们先前工作中提出的基线方法进行比较；该基线方法基于梅尔频率倒谱系数（MFCCs）。所有考察的方法均基于降噪语音信号，旨在研究通过三个递进层次引入语音信息的影响：未经校正的自动转录、经过时间对齐的语音分割，以及对齐后由专家进行的人工校正。模型经训练后，可利用自动轮廓跟踪方法预测从声道磁共振图像中提取的发音器官轮廓。结果表明，在依赖音标表示的模型中，对齐后进行人工校正的模型性能最佳，接近基线方法。**

索引词—**声学到发音器官逆推，语音生成，声道形状**

I. 引言

声学到发音器官逆推旨在根据音频信号重建声道的形状。为解决这一逆问题，人们提出了多种方法，它们依赖不同类型的数据和方法论。最初，可用数据有限，主要由静态 X 射线图像和面积函数构成。早期方法基于物理方法，采用声学模拟和对声道的简化近似。例如，Coker模型 [1] 代表了将发音器官建模应用于语音合成的早期尝试。

随后， [2], [3]等先驱性研究利用声学传递函数处理逆推问题。具体而言， [4]证明了该问题本质上存在歧义：对于给定的共振峰三元组，可能有无穷多个解。

更丰富数据的出现，例如电影放射摄影（X射线胶片），使得基于发音器官模型的新方法得以探索。由 [5],提出、根据500幅图像构建的模型，代表了一项重大进展：语音合成与发音语音学研究。该模型后来被用于反演，即将问题表述为一个不适定逆问题，通常通过正则化和优化技术求解。

20 世纪 90 年代， [6]引入神经网络，以近似求解声学到发音器官逆推问题。与此同时，人们也探索了基于码本的方法 [7], [8], ：在声学—发音器官词典中搜索与待逆推信号最接近的条目，然后对相应的发音器官构形进行插值。

随着使用电磁 articulography（EMA）和X射线微束系统采集的发音数据集（例如mngu0语料库 [9], ）的出现，人们得以在不依赖显式物理模型 [10] 的情况下，建立声学信号与发音几何之间的直接关系。这些数据集推动了数据驱动统计方法的发展，包括基于高斯混合模型（GMMs）和隐马尔可夫模型（HMMs）的方法，以及基于神经网络的技术。尤其是，循环神经网络架构得到了广泛探索，其中包括双向长短期记忆模型 [11], [12]。此外，[13] 提出了一种基于双向门控循环单元的方法，利用电磁发音数据进行音素到发音的映射。人们还研究了更复杂的架构，例如CNN–BiLSTM [14] 和Bi-GRNN [15], ，以及较新的时间卷积网络（TCNs）[16],。

为进一步提升性能，语音分割也被纳入作为额外输入 [17]。然而，这种策略要求在推理时提供语音分割，从而限制了其适用性。为了解决这一问题，同时仍能利用语音信息，人们提出了多任务学习框架。例如， [18] 联合预测音位标签和发音变量，使模型无需在推理过程中进行显式分割，也能利用语音结构。

然而，电磁发音数据存在若干局限：传感器数量有限

（舌上有三个或四个）、用于粘贴这些传感器的胶水会带来僵硬感（尤其是在舌和嘴唇上），而且嘴唇之间的导线也可能轻微干扰发音。

最近，实时动态磁共振成像（rt-MRI）作为一种替代方案被引入，其首个录制语料库见 [19]。深度学习方法已应用于这些数据 [20]。然而，受多项限制影响，实时磁共振成像的应用仍然有限：难以获取足够大的数据集，去噪后信号质量较低，缺乏稳健的轮廓跟踪工具，以及空间分辨率相对较低（约为 68×68 像素），此外还有磁共振成像相关伪影。 [21]提出了首个基于机器学习的方法，能够针对任意音素序列生成完整的声道形状，涵盖从声门到嘴唇的所有发音器官和静态结构。该方法结合了高度准确的语音分割，以及基于实时磁共振成像数据的可靠轮廓跟踪。

在本研究中，我们使用了一个更高分辨率的实时磁共振成像数据集（136×136 像素），该数据集已完成去噪，质量优于大多数现有的实时磁共振成像数据库。此外，我们没有直接对完整的实时磁共振成像图像进行反演，而是采用自动轮廓跟踪来提取不同发音器官的形状，并仅使用这些轮廓而非整幅图像进行反演。本研究旨在比较不同转写准确度层级的语音分段——从人工校正的对齐，到自动对齐，再到原始转写——与直接使用以梅尔频率倒谱系数表示的语音信号之间的差异。我们希望评估投入大量时间进行人工校正是否值得，或者精度较低的对齐或转写能否取得相近的表现。这项比较将有助于确定预处理工作量与发音器官轮廓建模预测准确度之间的权衡。

II. 数据集

本研究使用的语料库采集于南锡大学医院（南锡大学医院）。语料由一名法语母语女性说话者在五次录音中产生的语音组成。语料库约包含3.5小时的语音，相当于约2,100个句子。总计包含153段序列，每段约持续80秒，并对应4,000帧磁共振图像。

实时磁共振成像数据使用Siemens Prisma 3T扫描仪采集。每张图像对应一个厚度为8 mm的正中矢状切片，空间分辨率为 136 × 136 像素，像素间距为1.62 mm。

音频信号使用光学麦克风以16 kHz的采样率录制，随后使用 [22]进行降噪。

A. 磁共振图像预处理

在磁共振图像预处理中，我们采用一种基于循环卷积神经网络（RCNN）的自动追踪方法，将实时磁共振成像数据分割为八种语音相关发音器官。该分割过程生成每种发音器官的发音器官轮廓，每条轮廓由50个点表示。所考虑的发音器官包括上唇、下唇、舌、软腭中线（软腭）、咽壁、会厌、杓状软骨和声带，如图1所示。

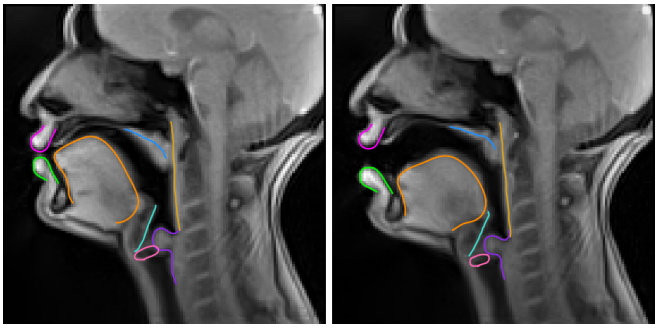


图1. 实时磁共振成像影片中两张图像内追踪的发音器官轮廓分割：杓状软骨，会厌，下唇，声带，软腭中线，舌，上唇，咽壁

B. 输入预处理

在我们之前的工作中 [23], , 模型的输入是降噪语音信号。我们从该信号中提取了前 13 个梅尔频率倒谱系数，以及它们的一阶和二阶差分（一阶差分和二阶差分）。

在本研究中，我们的目标是基于语音分割训练模型。我们仍以降噪语音信号作为起点。为此，我们采用了三种不同的方法。

第一种方法使用经过预训练和微调的 Wav2Vec 2.0 模型进行法语音位转写 [24]。该模型基于自监督学习，并使用通过连接主义时序分类（CTC）损失函数训练的输出层。对于每个时间帧，模型都会生成一个 61 维逻辑值向量，对应于词汇表中各个音素的未归一化分数。随后，这些逻辑值会转换为语音概率分布。沿音素维度应用Softmax函数后，便可为每一帧生成音素出现概率的表示。为稳定发音器官预测模型的训练并降低变异性，我们还按会话对这些概率进行全局归一化，即减去均值并除以标准差。

第二种方法依赖于使用 Astali 工具 [25] 进行强制对齐。该方法同时利用音频信号和经过时间对齐的语音转写。转写以句子为单位提供，带有起始和结束时间，并用于对声学信号与相应的音位序列进行强制对齐。通过这种对齐，可以将每个音素精确关联到音频信号中的一个时间区间。通过这一过程获得了 37 个不同的音素。随后，将提取出的音素编码为独热向量，每个向量表示语音词汇表中一个音素的存在。

第三种也是最后一种方法基于使用 Astali 获得的语音片段划分结果，并由专家进行人工校正。校正主要用于细化声音之间的时间边界。此外，清音爆破音的闭塞部分与爆破部分被分开处理，因为这两部分的发音部位差异显著。校正后获得了 44 个不同的音素，并同样将它们编码为独热向量。

III. 方法

A. 模型架构

所用模型与我们先前的工作相同 [23],[26]。我们对其进行了修改，使其能够接受语音分段作为输入（见图 2），而不是语音信号。该模型由五层组成：前两层是各含 300 个单元的全连接层，之后是两个各含 300 个单元的 Bi-LSTM 层。输出层是一个含 800（8×100）个单元的全连接层，分别对应 8 个发音器官及其各自的 100 个坐标。



图 2. 模型架构

B. 损失函数

由于发音器官声学逆推是一项回归任务，我们采用均方误差（MSE）作为损失函数，这是该领域最常用的方法。

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

其中，n 表示数据集中的观测总数，y_i 和$\hat{y}_i$ 分别表示示例 i输出的真实值和预测值。

C. 评估

我们对原始轮廓和预测轮廓进行反归一化，并评估均方根误差（RMSE）和中位数，二者均以毫米为单位表示。对于每一帧，我们根据代表每个发音器官几何形状的 100 个预测点和原始点，计算每个发音器官的均方根误差和中位数。最后，我们对每个发音器官在所有帧上的均方根误差和中位数取均值，然后再对所有发音器官进行平均。

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

D. 实验

这些实验使我们能够评估不同预处理程度所带来的差异。直接使用原始语音信号无需特别准备，Wav2Vec 2.0 转写甚至不需要任何人工干预。不过，由于它仅是转写，其精度仍然有限。相比之下，使用 Astali 需要先准备口语句子，这可能会耗费不少时间；而人工校正则需要投入更多精力。

我们的目标是评估依靠语音分割是否值得，还是直接使用降噪语音信号更合适。此外，我们还希望确定哪种分割方式最有效：无需投入精力的自动转录、使用 Astali 进行强制对齐，还是强制对齐后再进行人工校正。

我们进行了四项不同的实验。第一项实验作为基线，使用梅尔频率倒谱系数（MFCCs）作为输入来训练模型；我们将其称为基线。第二项实验使用从 Wav2Vec 2.0 模型获得的语音转写，我们将其称为基于 Wav2vec2 的模型。第三项实验依赖于使用 Astali 进行强制对齐所获得的语音分段，称为基于 Astali 的模型。最后，第四项实验使用经过专家手动校正的分段，我们将其称为专家校正模型。

E. 模型参数

所有模型均训练 300 个训练轮次，批大小为 10。训练使用 Adam优化器，学习率为 10^{-3}，并在验证集上采用早停策略，耐心值设为 10 个训练轮次。我们移除了句子之间的静音部分，最终得到共 45,000 帧。数据集经过随机打乱，并划分为 80% 的训练子集、10% 的验证子集和 10% 的测试子集。所有实验均采用相同的随机化方式，并且所有实现均使用 PyTorch 完成。

IV. 结果

表 I 根据所使用的输入特征类型比较了四组实验的性能：基线、自动语音转写（Wav2Vec 2.0）、通过强制对齐获得的语音分割（Astali）以及人工校正切分。所有模型均使用均方根误差和中位数进行评估，两者均以毫米为单位表示。此外，还进行了学生t检验，以评估所观察差异的统计显著性。

所有模型均在 61 个训练轮次内收敛。使用梅尔频率倒谱系数的基线模型取得了最低的均值均方根误差（1.51 mm），优于基于Wav2vec2的模型（1.67 mm）、基于Astali的模型（1.68 mm）和专家校正模型（1.61 mm）。按发音器官考察结果时，基线模型在七个发音器官上表现最佳；唯一例外是软腭，专家校正模型的表现略胜一筹。

中位数也呈现出类似的趋势：基线模型的平均中位数较低，且在七个发音器官上的中位数也较低。专家校正模型排名第二，表现优于基于Wav2vec2和基于Astali的模型；后两者表现相近，不过在某些发音器官上，基于Wav2vec2的模型略优于基于Astali的模型。

图 3 展示了使用基线模型进行推断的示例。左侧音素 /a/ 的均方根误差为 1.36 mm，右侧音素 /t/ 的均方根误差为 1.50 mm。

表 I 不同输入特征的均方根误差（毫米）和中位数（毫米）比较

	基线		基于Wav2vec2的模型		基于Astali的模型		专家校正模型	
	RMSE	中位数	RMSE	中位数	RMSE	中位数	RMSE	中位数
杓状软骨	1.63 ± 1.02	1.38	1.83* ± 1.13	1.55	1.88* ± 1.17	1.58	1.81* ± 1.09	1.54
会厌	1.52 ± 0.88	1.33	1.71* ± 0.99	1.49	1.77* ± 1.04	1.53	1.71* ± 0.99	1.48
下唇	1.53 ± 0.83	1.35	1.67* ± 0.96	1.44	1.64* ± 0.99	1.40	1.54* ± 0.88	1.34
咽部的	1.08 ± 0.58	0.96	1.22* ± 0.67	1.08	1.23* ± 0.70	1.07	1.19* ± 0.65	1.04
软腭	1.34 ± 0.67	1.21	1.42* ± 0.70	1.29	1.37* ± 0.68	1.23	1.34* ± 0.66	1.21
舌	2.33 ± 1.17	2.10	2.53* ± 1.36	2.24	2.57* ± 1.38	2.26	2.45* ± 1.25	2.18
上唇	1.15 ± 0.55	1.03	1.32* ± 0.66	1.18	1.19* ± 0.59	1.07	1.16* ± 0.57	1.04
声带	1.49 ± 0.84	1.31	1.68* ± 0.96	1.48	1.77* ± 1.05	1.53	1.68* ± 0.97	1.46
Mean	1.51 ± 0.91	1.30	1.67* ± 1.03	1.44	1.68* ± 1.07	1.42	1.61* ± 0.99	1.37

*与基线结果（$p < 0.05$）相比存在显著差异（基于t检验）。

V. 讨论

结果表明，基线模型优于其他模型，其均方根误差均值为 1.51 mm，中位数误差均值为 1.30 mm。这说明直接使用语音信号比依赖语音分割更有效。这些结果可以用以下事实来解释：梅尔频率倒谱系数主要捕捉语音信号的频谱和动态结构。

相比之下，语音分割依赖离散的语言单位，这会导致丢失大量对准确预测发音器官轮廓至关重要的声学信息。此外，这种方法容易受到转写或强制对齐误差的影响。即使经过人工校正，语音分段仍受限于这种离散表示，以及音素—发音关系本身存在的歧义；这种关系具有欠定性。

因此，采用语音表示会过度简化语音信号，从而限制模型的性能，使其不如基于梅尔频率倒谱系数的连续声学方法。在仅依赖语音分割的方法中，使用专家校正标注的模型优于另外两种基于语音的模型，这凸显了分段准确性在提升发音预测性能方面的关键作用。此外，依赖自动转录的基于Wav2vec2的模型略优于基于Astali的对齐模型。这种差异可由输入表示的性质解释：基于Wav2vec2的模型使用音素后验概率分布，保留了一定程度的不确定性和时间平滑性；而基于Astali的模型依赖以独热向量编码的硬音素表示，因此输入更加离散，包含的信息也更少。

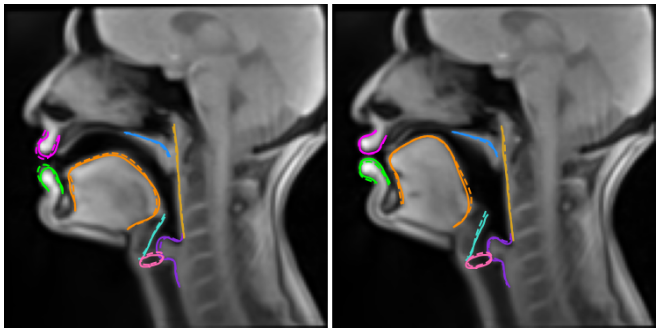


图 3. 基线模型中，实时磁共振成像影片所跟踪的原始轮廓与两种反演示例的对比。虚线表示预测轮廓，实线表示原始轮廓。左侧音素 'a' 的均方根误差为 1.36 mm，右侧 't' 的均方根误差为 1.50 mm。

VI. 结论

在本研究中，我们通过分析从去噪语音信号开始的不同语音分割准确度水平的影响，对比了多种方法。结果表明，基于连续声学表征（尤其是梅尔频率倒谱系数）的基线模型在声道轮廓重建方面仍然最为有效，取得了最低的均方根误差和中位数误差。这证实了语音信号中精细的频谱—时间信息对于发音轮廓预测的重要性。

依赖语音分割的方法虽然能够提供更直接的语言学解释，并减少对原始声学特征的依赖，但其本质上受限于音素的离散性，以及音素内部信息和协同发音信息的丢失。因此，其整体性能仍低于声学基线。

然而，通过人工校正语音分段所观察到的改进，凸显了时间标注和语音标注准确度的关键作用。在基于语音的方法中，对齐后经专家校正的标注，其效果显著优于完全自动分段。此外，结果还强调了概率音素表示的优势：与严格离散的独热编码相比，它能保留语音信号中的部分不确定性和时间连续性。

参考文献

[1] C. H. Coker，“发音动态与控制模型”，IEEE 会议录，第 64 卷，第 4 期，第 452–460 页，2005 年。[2] B. Atal，“直接从语音波确定声道形状”，美国声学学会期刊，第 47 卷，第 1A 增刊，第 65–65 页，1970 年。[3] H. Wakita，“通过对语音声学波形进行逆滤波直接估计声道形状”，IEEE 音频与电声学汇刊，第 21 卷，第 5 期，第 417–427 页，1973 年。[4] B. S. Atal、J. J. Chang、M. V. Mathews 和 J. W. Tukey，“通过计算机排序技术逆转声道中的发音到声学变换”，美国声学学会期刊，第 63 卷，第 5 期，第 1535–1555 页，1978 年。[5] S. Maeda，“声道系统的数字仿真方法”，Speech Communication，第 1 卷，第 3-4 期，第 199–229 页，1982 年。[6] M. G. Rahim 和 C. C. Goodyear，“使用神经网络估计声道滤波器参数”，Speech Communication，第 9 卷，第 1 期，第 49–55 页，1990 年。[7] J. Schroeter 和 M. M. Sondhi，“从语音信号估计声道形状的技术”，IEEE 语音与音频处理汇刊，第 2 卷，第 1 期，第 133–150 页，1994 年。[8] S. Ouni 和 Y. Laprie，“利用超立方体码本对发音空间建模，以实现声学到发音器官逆推”，美国声学学会期刊，第 118 卷，第 1 期，第 444–460 页，2005 年。[9] K. Richmond、P. Hoole 和 S. King，“发布 mngu0 发音语料库中的电磁发音描记术（第 1 天）子集”，载于 Interspeech 2011：国际语音通信协会第12届年会。国际语音通信协会，2011 年，第 1505–1508 页。[10] A. Katsamanis、A. Roussos、P. Maragos、M. Aron 和 M.-O. Berger，“利用多模态数据，从视听语音反演发音信息”，载于第八届国际语音产生研讨会-ISSP’08，2008 年。[11] P. Liu、Q. Yu、Z. Wu、S. Kang、H. Meng 和 L. Cai，“一种用于声学到发音器官逆推的深度循环方法”，载于2015 年 IEEE 国际声学、语音与信号处理会议（ICASSP）。IEEE，2015 年，第 4450–4454 页。[12] M. Parrot、J. Millet 和 E. Dunbar，“说话人无关的声学到发音重建的独立自动评估”，载于 INTERSPEECH 2020 论文集：国际语音通信协会第21届年会，系列：INTERSPEECH 2020 论文集：国际语音通信协会第21届年会，中国上海/线上，2020 年 10 月，第 3740–3744 页。 [在线]。可访问：https://hal.science/hal-03087264

[13] T. Biasutto-Lervat 和 S. Ouni，“使用双向门控循环神经网络进行音素到发音映射”，载于 Interspeech 2018——国际语音通信协会第19届年会，2018 年。[14] A. S. Shahrebabaki、S. M. Siniscalchi、G. Salvi 和 T. Svendsen，“通过对子带频率信号进行时间卷积实现序列到序列发音反演”，database，第 1 卷，第 5 页，2020 年。[15] P. Wu、L.-W. Chen、C. J. Cho、S. Watanabe、L. Goldstein、A. W. Black 和 G. K. Anumanchipalli，“说话人无关的声学到发音语音反演”，载于 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2023 年，第 1–5 页。[16] Y. M. Siriwardena 和 C. Espy-Wilson，“隐秘声源：纳入声源特征以改进声学到发音语音反演”，载于 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2023 年，第 1–5 页。[17] J. Wang、J. Liu、X. Li、M. Yu、J. Gao、Q. Fang 和 L. Liu，“用于说话人无关声学到发音器官逆推的双流联合训练”，载于 IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2023 年，第 1–5 页。[18] Yashish M. Siriwardena、Ganesh Sivaraman 和 Carol Espy-Wilson，“基于多任务学习的声学到发音语音反演”，载于 Interspeech 2022，2022 年，第 5020–5024 页。[19] V. Ramanarayanan、S. Tilsen、M. Proctor、J. T¨oger、L. Goldstein、K. S. Nayak 和 S. Narayanan，“实时磁共振成像语音生成分析”，《计算机语音与语言》，第 52 卷，第 1–22 页，2018 年。[20] T. Csap´o，“使用声道实时磁共振成像进行说话人相关的声学到发音器官逆推”，载于 Proc. INTERSPEECH 2020 – *21st* 国际语音通信协会年度会议，上海，2020年10月，第3720–3724页。[21] V. Ribeiro、K. Isaieva、J. Leclere、P.-A. Vuissoz和Y. Laprie，“根据待发音音素序列自动生成完整声道形状”，*Speech Communication*，第141卷，第1–13页，2022年。[22] A. Ozerov、E. Vincent和F. Bimbot，“处理音频源分离中先验信息的一般灵活框架”，*IEEE*音频、语音与语言处理汇刊，第20卷，2012年5月。[23] S. Azzouz、P.-A. Vuissoz和Y. Laprie，“使用实时磁共振成像数据通过声学到发音逆映射重建完整声道轮廓”，载于*Interspeech 2025*会议论文集，2025年，第978–982页。[24] M. Olivier、J. Hauret和E. Bavu，“wav2vec2-french-phonemizer（修订版e715906）”，2023年。 [在线]。可访问：https://huggingface.co/Cnam-LMSSC/wav2vec2-french-phonemizer [25] D. Fohr、O. Mella和D. Jouvet，“统一口语语料库自动对齐的转写规范的重要性”，载于第*8*届国际语言学´语料库语言学国际会议（*JLC2015*），2015年。[26] S. Azzouz、P.-A. Vuissoz和Y. Laprie，“使用实时磁共振成像数据通过声学到发音逆映射重建完整舌轮廓”，载于*ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2025年，第1–5页。