面向说话人无关声学到发音器官反演的双流联合训练

Jianrong Wang[1] *Jinyu Liu*[2] *Li Liu*[3⋆] *Xuewei Li*[1] *Mei Yu*[1] *Jie Gao*[1] *Qiang Fang*[4]

[1] 天津大学智能与计算学院，中国天津[2] 天津大学天津国际工程师学院，中国天津[3] 香港科技大学（广州），中国广州[4] 中国社会科学院语言研究所，中国北京

arXiv:2302.13273v1 [cs.SD] 26 Feb 2023

摘要

声学到发音器官反演（AAI）旨在根据语音音频估计发音器官的参数。AAI 面临两项常见挑战：数据有限，以及在说话人无关场景下性能不理想。目前大多数研究都专注于直接从语音中提取特征，而忽略了音素信息的重要性，这可能会限制 AAI 的性能。为此，我们提出一种名为 SPN 的新型网络，利用两个不同的流来执行 AAI 任务。首先，为提升说话人无关实验的性能，我们提出一种新的音素流网络，将发音参数估计为音素特征。据我们所知，这是首项从音素中提取说话人无关特征以提升 AAI 性能的研究。其次，为了更好地表征语音信息，我们训练了一个语音流网络，以结合局部特征和全局特征。与最先进水平（SOTA）相比，所提方法在说话人无关实验中使均方根误差（RMSE）降低了 0.18 mm，皮尔逊相关系数提高了 6.0%。代码已发布于 https://github.com/liujinyu123/AAI-Network-SPN。

索引词— 声学到发音器官反演、局部特征、全局特征、说话人无关、音素流

1. 引言

声学到发音器官反演（AAI） [1] 主要解决从语音音频推导关键发音器官发音参数的问题。近年来，这项技术在发音指导[2] 和语音识别 [3, 4, 5], 等诸多领域发挥了重要作用，因此吸引了众多研究人员投身这一领域。

人们提出了多种基于深度学习的模型和声学表征来完成发音动作识别任务。早期，语音反演采用码本 [6] ，但其性能很大程度上取决于码本的质量。后来，人们提出了数据驱动的语音反演模型，例如隐马尔可夫模型（HMM） [7, 8], 、高斯混合模型（GMM） [9], 、深度神经网络（DNN） [10]等。目前，最常用的模型是循环神经网络（RNN）及其变体，例如长短期记忆网络（LSTM） [11, 12]。在 [11, 13, 14], 中，人们使用了不同的语音表征，例如线谱频率（LSF）、梅尔频率倒谱系数（MFCC）和滤波器组能量（FBE）。在我们的工作中，我们将 MFCC 和音素作为模型的输入。

迄今为止，发音动作识别面临两项主要挑战。其一是可用数据集非常有限，因为我们需要同时录制语音和发音参数，这不仅难以采集，而且成本高昂。最常用的公开数据集包括 MOCHA-TIMIT [15], MNGU0 [16], 和 HPRC [17]。另一项挑战是在说话人无关场景中提升性能。

针对第一个挑战， [18] 提出了使用跨语料库数据解决数据量有限问题的方法。针对第二个挑战， [11] 采用声道归一化将每位说话人的发音空间映射到同一空间，以提升模型性能，但这会导致语音个性化信息的丢失。值得注意的是，有研究提出了自监督预训练模型来解决上述挑战，并在 [19]中取得了最佳性能。然而，该研究仅使用梅尔频率倒谱系数作为网络输入，这可能限制发音动作识别的性能；此外，该研究使用一维卷积神经网络提取语音特征，这可能导致语音全局信息丢失。

为了解决上述两个挑战，我们提出了一种由两部分组成的新型网络：语音流网络和音素流网络，简称 SPN。在 [19]中，我们使用一维卷积神经网络提取语音特征。但有研究指出，卷积神经网络只能提取 [20], 中的局部特征，因此我们加入了多头注意力模型，以提取全局特征，从而更好地表征语音信息。此外，我们还提出了一种新的音素流网络。具体而言，我们使用转录得到的音素进行音素反演，然后对

⋆通讯作者：avrillliu@hkust-gz.edu.cn。

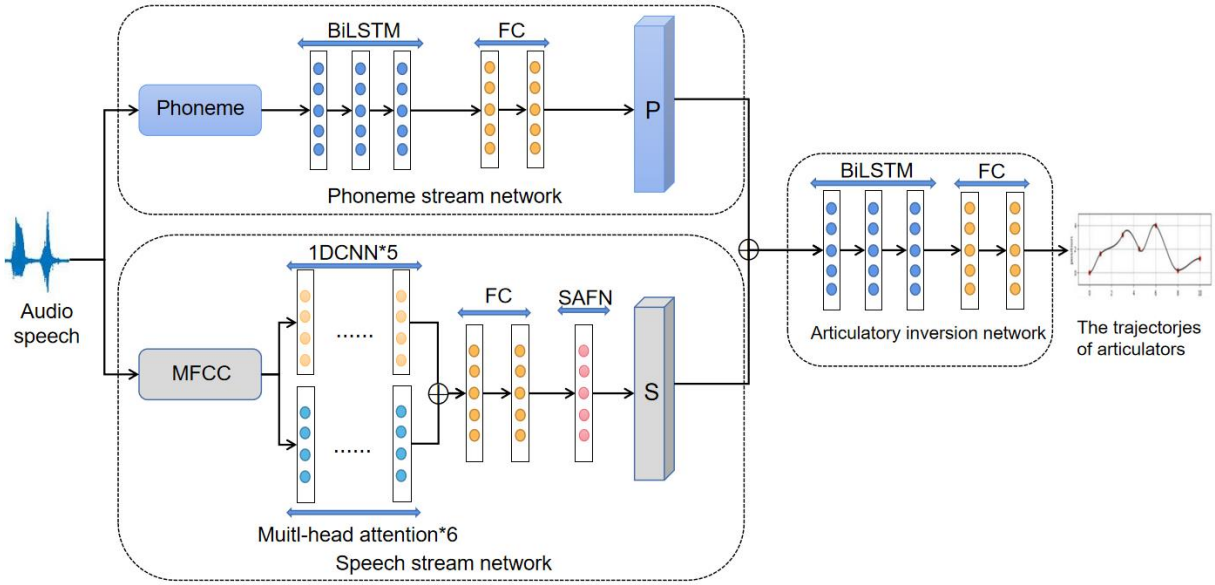


图1. Th SPN的框架。P表示音素流特征，S表示语音流特征 re.

音素流网络的输出进行处理，将其作为音素特征来执行语音反演。其动机在于，音素仅表示语音的内容信息，而不表示身份信息。因此，音素流网络获得的音素特征具有说话人无关性，能够提升说话人无关实验的性能。

总而言之，本研究有三项贡献。

- 为了更好地表示语音信息，我们分别通过一维卷积神经网络和多头注意力模块提取局部特征和全局特征。
- 我们提出了一种新的音素流网络，以获取音素特征，从而提升说话人无关实验的性能。
- 实验结果表明，在公开的HPRC数据集上，所提出的模型明显优于最先进水平，均方根误差降低了0.18毫米，皮尔逊相关系数提高了6%。

2. 所提方法

2.1. 总体框架

如图 1 所示，我们提出的 SPN 由两个核心模块组成：语音流网络和音素流网络。我们通过语音流网络提取全局特征和局部特征，然后将它们输入 SAFN [19] 网络，以获得语音特征；再通过音素流网络获得音素特征。随后，将整合后的语音特征和音素特征输入发音动作反演网络（AIN）[19]，以获得关键发音器官的参数。

2.2. 语音流网络

文献 [20]表明，卷积神经网络只能提取语音的局部特征，因此，为了更好地表征语音信息，我们提出了语音流网络。梅尔频率倒谱系数分别输入级联的一维卷积神经网络和多头注意力模块，以提取局部特征和全局特征，并将其作为语音特征输入 SAFN。如图 2 所示，语音

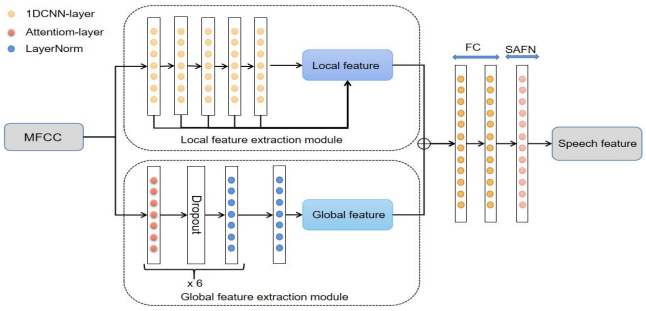


图2。语音流网络的工作流程。

该双流网络主要分为两部分：一部分是局部特征提取模块，另一部分是全局特征提取模块。在局部特征提取模块中，我们采用了 [19] 中使用的级联一维卷积神经网络。我们将梅尔频率倒谱系数输入具有五种不同卷积核的一维卷积神经网络中，这些卷积核的尺寸分别为 1、3、5、7 和 9，以获得局部特征。对于全局特征提取模块，我们使用多头注意力模块，因为它能够关注语音中每一帧与其他帧之间的关系。注意力模块可描述为从一组查询、键和值中获取输出，其中输出、查询、键和值均为对应的向量。具体而言，全局特征提取模块包含六层多头注意力。每层共有八个

注意力头，键和值的维度均为 64。随后，我们将特征输入层归一化层，以获得全局特征。接下来，将局部特征和全局特征输入两个全连接层，每层包含 300 个单元，以获得语音特征。最后，我们将语音特征输入 SAFN。

局部特征的计算表示为：

$$\boldsymbol{y}_{i,j}^{\text{local}} = b_j + \sum_{k=1}^{L_{i-1}} \mathbf{W}_i * \boldsymbol{y}_{i-1,k}^{\text{local}} \tag{1}$$

其中，*表示卷积操作，$y_{i,j}$ 表示第 j个卷积层中第 i个通道的特征图， b_j是第 j个通道的偏置，$\mathbf{W}_i$ 表示第 i个卷积层的权重， L_{i-1} 表示 $\boldsymbol{y}_{i-1}^{\text{local}}$的长度。

2.3. 音素流网络

受 [21], 启发，我们使用音素流网络的输出作为音素特征，以辅助语音反演。原因如下：音素本身只编码语音的内容信息，而不编码说话人无关的身份信息。我们使用宾夕法尼亚大学语音学实验室强制对齐器 [22]提取音素帧。每个音素帧都是一个39维独热向量。

在该模块中，我们使用三层双向长短期记忆网络，最后将其输入两个全连接层，以获得作为音素特征的发音参数。三层双向长短期记忆网络的设置相同，每层包含150个激活单元；全连接层包含300个激活单元。输出为12维发音参数。随后，我们将发音参数作为音素特征，并将从语音流网络模块获得的语音特征一同输入AIN网络，以执行语音反演。音素流网络的核心表示为：

$$\begin{aligned} s_i &= f\left(Ux_i + Ws_{i-1} + b\right), \\ s_i' &= f\left(U'x_i + W's_{i+1} + b'\right), \\ o_i &= g\left(Vs_i + V's_i' + c\right), \end{aligned} \tag{2}$$

其中， o_i 是帧 i处估计的音素特征。 x_i是帧 i处输入的音素序列， s_i 是帧 i处的临时状态，而 U， W， U'， W'是对应的变换矩阵。 b， b' 是偏置。

3. 实验

3.1. 实验设置

数据集。 本文使用的数据集为 HPRC。该数据集包含六个位置：T1、T2、T3、UL、LL 和 LI。为与最先进水平保持一致，我们仅使用舌部器官在 X 和 Z方向上的位置作为实验的预测标签（例如，T1、T2、T3）。HPRC [17] 数据集包含八位以美式英语为母语的说话人（四男四女），每位说话人的数据由 720 个话语组成。

性能指标。 实验性能通过均方误差和皮尔逊相关系数 [23]进行衡量。前者衡量预测标签与真实标签之间的误差，后者则表示两个标签的相似度。

实现细节。 [19] 提出的基于语音分解和辅助特征的网络是发音动作识别领域的最先进水平。此外，网络架构图中 SAFN 和 AIN 模块的设置与最先进水平中的设置相同。在本研究中，我们对模型进行了 20 个训练轮次，并使用Adam优化器。学习率为 1e-4，批量大小为 5。

3.2. 与最先进水平的比较

为检验我们提出的 SPN 网络的有效性，我们根据表 1 设置了三种不同的实验场景。S1 表示仅以音素为输入训练音素流网络，以进行说话人无关实验。S2 表示以音素和梅尔频率倒谱系数作为输入，进行说话人无关实验。在此场景中，我们将音素流网络作为预训练模型，并在训练整个 SPN 时冻结其参数。S3 表示以音素和梅尔频率倒谱系数作为网络输入进行说话人无关实验。与 S2 不同，我们将音素流网络与整个网络一同训练。我们网络的损失函数是两部分的加权和，即 SPN 的 L2损失和音素流网络的 L2损失。公式如下：

$$L_{joint} = \sum_{i=0}^{k} \left(y_m^i - \hat{y}_m^i\right)^2 + \sum_{i=0}^{k} \left(y_n^i - \hat{y}_n^i\right)^2, \tag{3}$$

其中，$\hat{}y_m^i$、$\hat{}y_n^i$ 表示电磁 articulography 的真实值。 y_m^i、 y_n^i 是对应的预测标签。 k 表示语音特征的长度。

更具体地说，我们在场景 3 中同时计算了音素流网络和 SPN 的性能。S3(P) 表示音素流网络的结果，S3（S）表示 SPN 的结果。值得注意的是，我们的实验是在哈斯金斯数据集上进行的，并且在每种场景下，我们都为每位说话人分别训练了一个模型。我们从七位说话人的数据中汇总所有训练集（80%）和验证集（20%），并使用剩余一位说话人的全部数据（100%）进行测试。最后，我们取八位说话人结果的平均值，作为实验的最终结果。

表 2 展示了哈斯金斯数据集三个场景中的均方根误差和皮尔逊相关系数。由于以往最先进水平的工作未提供每位说话人的平均均方根误差，我们在表 2 中相应的位置标记为“-”。总体而言，我们可以清楚地看到，与最先进水平相比，我们提出的 SPN 网络在均方根误差上降低了 0.141 mm，

表 1。三种不同场景的实验设置。* 表示从每位说话人处抽取的数据比例。P 和 M 分别表示音素和梅尔频率倒谱系数。

场景	输入	#说话人数	训练集	验证集	Test
S1	P	N-1 1	80%* ---	20%* ---	--- 100%
S2	P, M	N-1 1	80%* ---	20%* ---	--- 100%
S3	P, M	N-1 1	80%* ---	20%* ---	--- 100%

表2。SPN 和最先进水平的均方根误差与皮尔逊相关系数

场景	F01	F02	F03	F04	M01	M02	M03	M04	均方根误差	皮尔逊相关系数
S1	2.279	2.182	1.681	2.322	1.753	2.546	2.036	1.898	2.087	0.755
S2	2.720	2.740	2.088	2.759	2.352	3.064	2.395	2.528	2.580	0.798
S3(P)	2.169	2.084	1.618	2.172	1.693	2.440	1.961	1.812	1.993	0.773
S3 (S)	2.634	2.665	2.100	2.602	2.235	3.195	2.368	2.504	2.537	0.810
SOTA	-	-	-	-	-	-	-	-	2.721	0.751

在场景 2 中皮尔逊相关系数提高了近 5%；在采用联合训练策略的场景 3 中，均方根误差降低了 0.184 mm，皮尔逊相关系数提高了近 6%。这表明，我们提取的局部特征和全局特征能够更好地表示语音信息。加入音素特征后，还能进一步有效提升该模型的泛化能力。

更值得注意的是，从实验结果可以看出，在场景3中，联合训练下的音素流网络（S3(P)）性能优于单独训练时（S1），均方根误差约降低0.1 mm，皮尔逊相关系数提高近2%。这表明，SPN和音素流网络的性能均有所提升，说明联合训练策略是有效的。

图3展示了SPN与最先进水平的定性比较。显然，与最先进水平相比，SPN生成的舌部预测发音器官更接近真实值，尤其是图3红圈所示区域。这表明，音素流网络和语音流网络对舌部发音运动的重建具有积极作用。

3.3. 消融研究

为证明所提出模型的有效性，我们根据表3进行了消融实验，结果如表4所示。

表3。消融实验验证了各个模块的有效性。SPN-S 表示结合了语音流网络与最先进水平模型的模型。

	语音流网络	音素流网络
SOTA	×	×
SPN-S	✓	×
SPN	✓	✓

显然，采用语音流网络（SPN、SPN-S）的模型优于未采用语音流网络的模型（SOTA）。此外，采用音素流网络的模型

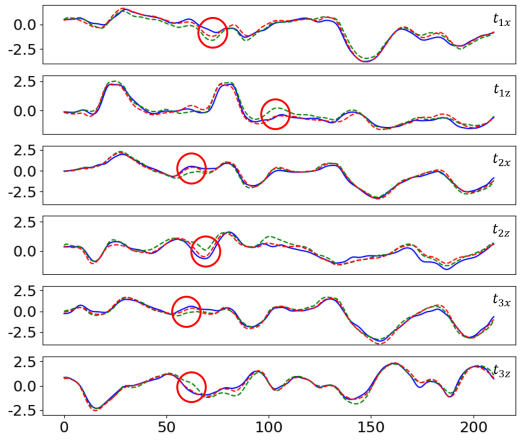


图3。SPN 与最先进方法在舌部运动预测结果上的比较。每个框中，蓝色实线表示真实标签，红色虚线表示 SPN 的预测结果，绿色虚线表示最先进方法的预测结果。

（SPN）优于未采用音素流网络的模型（SPN-S）。从实验结果中可以清楚地观察到，在说话人无关实验中，语音流网络通过提取局部特征和全局特征，更好地表征语音信息，从而提升了模型性能；而加入由音素流网络获得的音素特征后，我们的模型泛化能力进一步大幅提升。

表4。三个场景下的均方根误差和相关系数。

	RMSE	PCC
SOTA	**2.721**	**0.751**
SPN-S	**2.664**	**0.787**
SPN	**2.537**	**0.810**

4. 结论

为提升说话人无关性能并更加关注全局语音信息，我们提出了一种由两部分组成的新型网络。其中一部分是语音流网络，它利用一维卷积神经网络和多头注意力模型提取语音的局部特征和全局特征，从而更好地表征语音信息。同时，我们将音素流网络获得的发音参数作为音素特征，以帮助预测语音反演的发音参数。实验结果证明了所提出网络的有效性。未来，我们将探索元学习在发音动作识别领域的应用。

5. 致谢

本研究获得国家自然科学基金委员会（编号：61977049）、国家自然科学基金委员会（编号：62101351）和广东省基础与应用基础研究基金（编号：2020A1515110376）的资助。

6. 参考文献

[1] Korin Richmond，“从声学语音信号估计发音参数”，附录论文数字化项目，2002年。[2] Li Liu、Gang Feng 和 Denis Beautemps，“基于 CLNF 和混合动态模板的提示语音内唇特征提取”，*EURASIP*图像与视频处理期刊，第 1–15 页，2017年。[3] Li Liu、Gang Feng、Denis Beautemps 和 Xiao-Ping Zhang，“在自动连续提示语音识别中使用手势先行模型实现多模态融合的重新同步”，*IEEE* 多媒体汇刊，第 23 卷，第 292–305 页，2020年。[4] Jianrong Wang、Ziyue Tang、Xuewei Li、Mei Yu、Qiang Fang 和 Li Liu，“用于自动提示语音识别的跨模态知识蒸馏方法”，*arXiv* 预印本 *arXiv:2106.13686*，2021年。[5] Li Liu，《使用先进机器学习方法对法语连续提示语音识别进行建模》，博士论文，格勒诺布尔阿尔卑斯大学，2018年。[6] Sadao Hiroya 和 Masaaki Honda，“使用基于隐马尔可夫模型的语音生成模型，从语音声学中确定发音运动”，载于*ICASSP* 论文集，2002年，第 1 卷，第 433–437 页。[7] Masaaki Honda，“使用基于隐马尔可夫模型的语音生成模型，从语音声学中估计发音运动”，载于 *TSAP* 论文集，第 12 卷，第 175–185 页，2004年。[8]Li Liu、Thomas Hueber、Gang Feng 和 Denis Beautemps，“使用串联卷积神经网络-隐马尔可夫模型方法对连续提示语音进行视觉识别”，载于*Interspeech* 论文集，2018年，第 2643–2647 页。[9] A.W.Black、T. Toda 和 K.Tokuda，“使用高斯混合模型对发音运动与声学频谱进行统计映射”，语音通信，第 50 卷，第 215–227 页，2008年。

[10] K. Zhao、Z.-Y. Wu 和 L.-H. Cai，《基于深度神经网络的实时语音驱动虚拟说话人》，载于 *APSIPA* 会议论文集，第 1–4 页，2013 年。

[11] Ganesh Sivaraman、Vikramjit Mitra 和 Hosung Nam，《面向说话人独立的声道长度归一化声学到发音语音反演》，载于*Interspeech* 会议论文集，2016 年，第 455–459 页。

[12]Maud Parrot、Juliette Millet 和 Ewan Dunbar，“说话人无关声学到发音器官重建的独立自动评估”，载于*Interspeech* 论文集，2020年，第 3740–3744 页。

[13] A. Illa 和 P. K. Ghosh，“使用卷积神经网络进行声学到发音器官反演的表征学习”，载于 *ICASSP* 论文集，2019年，第 5931–5935 页。[14] A. S. Shahrebabaki、N. Olfati、A. S. Imran、S. M. Siniscalchi 和 T. Svendsen，“用于发音器官反演的不同输入特征的音素级分析”，载于*Interspeech* 论文集，2019年，第 3775–3779 页。[15] Alan Wrench，“用于连续语音识别研究的多通道/多说话人发音动作数据库”，*Phonus*，2000年。[16] R. Korin、H. Phil 和 K. Simon，“宣布 mngu0 发音动作语料库中的电磁 articulography 子集”，载于 *Interspeech* 论文集，2011年，第 1505–1508 页。[17] Mark Tiede、Carol Y Espy-Wilson 和 Dolly Goldenberg，“量化对比语速发音任务中弱化的运动学方面”，美国声学学会期刊，第 3580–3580 页，2017年。[18] Sarthak Kumar Maharana、Aravind Illa、Renuka Mannem、Yamini Belur、Preetie Shetty、Veeramani Preethish Kumar、Seena Vengalil、Kiran Polavarapu、Nalini Atchayaram 和 Prasanta Kumar Ghosh，“利用跨语料库声学-发音动作数据对构音障碍语音进行声学到发音器官反演”，载于 *ICASSP* 论文集，2021 年，第 6458–6462 页。[19] Jianrong Wang、Jinyu Liu、Zhao Longxuan、Wang Shanyu、Yu Ruiguo 和 Liu Li，“基于语音分解和辅助特征的声学到发音器官反演”，载于 *ICASSP* 论文集，2022年，第 4808–4812 页。[20] Yulong Li、Dong Zhou 和 Wenyu Zhao，“将局部和全局特征结合到用于句子相似度的孪生网络中”，*IEEE Access*，第 8 卷，第 75437–75447 页，2020年。[21]Ayush Tripathi、Swapnil Bhosale 和 Sunil Kumar Kopparapu，“使用 DeepSpeech 后验概率改进说话人无关的构音障碍语音可懂度分类”，载于*ICASSP* 论文集，2020年，第 6114–6118 页。[22] J. Yuan 和 M. Liberman，“SCOTUS 语料库上的说话人识别”，美国声学学会期刊，第 8 卷，第 3878 页，2008年。[23] Peng Liu、Quanjie Yu 和 Zhiyong Wu，“一种用于声学到发音器官反演的深度循环方法”，载于*ICASSP* 论文集，2015年，第 4450–4454 页。