◆算法研究与应用◆

基于小样本数据统计的双阶段舌位建模研究*

徐正丽[1],肖素芳[1**],简　敏[1],杨明浩[2]
(1.桂林电子科技大学,广西桂林　541004;2.中国科学院自动化研究所,北京　100190)

摘要:舌头是人类重要的发音器官,对发音时其形状的降维分析能有效协助语言学家分析人类的发音模式。主成分分析(Principal Component Analysis,PCA)是目前最常用的舌位轮廓降维分析方法。近年来,基于深度学习的自动编码器在降维方面被证明优于PCA。然而,舌头隐藏于口腔内部,难以获得大量的相关数据,这使得传统自动编码器无法直接用于舌位轮廓建模研究。为此,本文提出一种面向小样本舌位运动轮廓数据的双阶段自动编码器降维方法。首先该方法采用主动形状模型(Active Shape Model,ASM)产生大量舌头轮廓生理变形数据,并构建通用轮廓重建模型;接着,在第一阶段模型上添加降维层,用于对舌位轮廓数据进行压缩和分析。实验选取了从人类发音X光片中获得的240个元音舌形数据,并将该方法与传统PCA方法进行比较。结果表明,所提出方法获得的元音舌位图谱在二维平面上相对于传统PCA方法,区分度更好,具有更好的舌形降维和重建能力。

关键词:深度神经网络;自动编码器;主成分分析;舌位轮廓;隐藏单元

中图分类号:TP389　　**文献标识码**:A　　**文章编号**:1005-9164(2023)04-0745-09
DOI:10.13656/j.cnki.gxkx.20230928.014

舌头是人类重要的发音器官,其形变是人类能够发音的关键,对舌头形状的分析及建模是语音生成领域中的一项重要工作[1-4]。舌头属软组织结构,发音过程中舌头会产生较大变形,从而产生复杂的声道结构。但舌头主要隐藏在口腔内,致使人们难以直接观察舌头发音形状(即舌位),因此对舌位轮廓分析及建模一直是语音分析中的难点之一。在传统的实验语音学领域,人们提出了多种舌位模型来研究舌运动导致的声道结构变化和语音之间的关系。20世纪70年代初,语言学家和言语病理学家从X光片中手动标记舌头轮廓,并使用主成分分析(Principal Component Analysis,PCA)方法获得舌头运动模式[5,6],发现在元音生成中前两个主要成分所占比重为90%以上,即元音对应的舌头变形可通过前两个维度参数进

收稿日期:2023-02-15　　修回日期:2023-04-25
*国家自然科学基金项目(71463010,22180155466),广西科技计划项目(2021GXNSFBA220048,桂科AB21220038)和桂林科技计划项目(2023010123)资助。
【第一作者简介】
徐正丽(1982-),女,在读博士研究生,副教授,主要从事数据统计管理、应用语言学等研究。
【**通信作者】
肖素芳(1990-),女,博士,副教授,主要从事数据统计管理等研究,E-mail:xiaosufang2011@163.com。
【引用本文】
徐正丽,肖素芳,简敏,等.基于小样本数据统计的双阶段舌位建模研究[J].广西科学,2023,30(4):745-753.
XU Z L,XIAO S F,JIAN M,et al.Tongue Shapes Modeling from Small Data Using Two-Stage Autoencoder [J].Guangxi Sciences,2023,30(4):745-753.

行描述。平行因子分析(PARAFAC)也是一种广泛应用的舌位轮廓分析工具[7-11]。通过分析 10 个英语元音的 13 个横截面平行因子,研究者发现发出 10 个英语元音时舌位变化可分解为两个主要运动因素:一是舌根向前运动的同时伴随着舌头前部的向上运动;二是整个舌体的向上和向后运动。然而,PARAFAC 并不具备从低维数据分布中重建舌位轮廓的能力[11,12]。此外,与舌头运动建模相关的研究还包括基于元音的流形表示[13]、舌头轨迹的可视化[14-16]、基于语音驱动的舌面[2,17,18]和基于径向基函数(Radial Basis Function,RBF)的 B 样条拟合[19],基于机器学习的复杂三维有限元生物力学模型[20],基于集总元件模型的舌尖、舌外侧下侧和软腭前侧的平均感知方法[21]、舌苔瘀点的检测方法[22]等,这些方法侧重于从文本、语音记录、舌位受到刺激的反应以及舌噪声图像等方面对舌头运动轨迹和病理进行研究,但并未研究重建舌形以及建立舌位与语音之间的对应关系。

随着深度学习技术的兴起,研究人员将深度神经网络应用到舌位图像分析以及轮廓提取工作中,如 Ruan 等[23]提出了基于 U-Net 的舌头分割模型,从整个舌头图像中准确地分割出舌头主体;Ploumpis 等[24]提出了生成 3D 舌面的新型生成对抗网络(Generative Adversarial Network,GAN),将舌位 3D 模型生成与面部细节重建进行关联;Mansour 等[25]提出了基于深度神经网络的人类舌头图像疾病分类模型。虽然这些方法在舌头图像边缘提取、舌头表面纹理细节处理等方面取得了较好效果,但未能很好地对舌位轮廓进行压缩、重建和分析等[26-28]。

近年来,基于深度学习的自动编码器(Autoencoder)在数据降维和模式挖掘等方面表现良好[29],如面向图像的深度卷积网络自编码器能有效提取低维图像特征[30],降噪自编码器(Denoising Auto Encoder,DAE)在序列数据处理和模式发现等方面表现出良好性能[31-33]等。然而,目前还未见将基于深度学习的自动编码器用于舌位分析的研究,这主要是因为基于深度学习的自动编码器在训练中需大量数据,但由于舌头在口腔中的隐蔽性,真实舌位数据难以大量获得。一些学者通过添加噪声数据或使用 Dropout 技术来增加数据样本的方式提高小样本深度学习 DAE 的性能[29,31,32],这两种方法由于能生成更多有效的训练数据,因此能够提升网络从少量真实数据中提取特征的能力[31,34,35]。一般来说,舌头运动的前部较后部对发音过程的影响更大,因此基于平均随机理念的 Dropout 技术并不适用于舌位数据增强。

描述舌位运动的高性能模型既要维度低又要准确性高[26-28]。低维度表示有利于揭示舌头运动模式以及精准确定舌头运动模式与发音结构间的映射关系。为建立高性能且保持深度结构特征的舌位模型,本文提出一种双阶段自编码器舌位模型,其第一阶段首先利用符合生理特征的舌位变形数据构建大规模的形变舌位轮廓样本,再训练一个 n 层堆叠的舌位轮廓自编码器;第二阶段在舌位轮廓自编码器的基础上添加具有少量隐藏单元的第 $(n+1)$ 层组成最终的自编码器。

1 研究方法

1.1 舌位轮廓标准化

本文采取传统方法对舌位轮廓进行标准化处理。图 1(a)中有 18 条网格线(即图中序号为 1-18 的线条),主要舌形区域对应着网格 4-17 的横截面段,因此使用网格线 4-17(共 13 条横截面)来描述声道结构。首先确定上齿和上腭的尖端,然后将从齿尖点沿腭到会厌的轮廓作为不同声道结构的参考截面,最后将舌头表面和腭之间的归一化横截面范围作为编码器网络的输入。这 13 条网格线从参考截面到舌片表面(与背景正交)的线段长度可用于舌位分析。

本文使用网格线长度的归一化值用于编码器网络训练。归一化函数如式(1)所示,

$$\zeta_{ifj}=\Gamma_{ifj}/\max_{i,f,j}(V,\Gamma_{ifj}), \tag{1}$$

式中,$\Gamma_{ifj}=(G_{ifj}/V_{ifj})\eta$,$i\in\{1,2,\cdots,13\}$ 表示第 i 个舌位轮廓线;f 和 j 是舌位轮廓数据集中第 j 个音素发音阶段的第 f 帧;V 是中矢状面上声道的最宽横截面距离(根据前人研究,本研究将 V 设置为 45 mm),ζ_{ifj} 和 Γ_{ifj} 分别表示归一化和非归一化网格线长度,η 是从舌尖到舌根的实际舌长。通常成年男性的舌长约 175 mm,女性约 140 mm。G_{ifj} 和 V_{ifj} 是以像素为单位的网格线长度和舌位长度,可直接从 X 射线图片中获取[36]。

以元音“a”为例,图 1(b)显示了其发音第 32 帧对应的 13 个横截面长度分布;图 1(c)则以管状模型形式显示了从舌尖到舌根的形状,即舌头运动时对应的声道侧面结构;图 1(d)给出了从舌尖到舌根的 13 条网格线 $\zeta_{i,32,a}$,$i\in\{1,\cdots,13\}$ 所对应的归一化长度值。

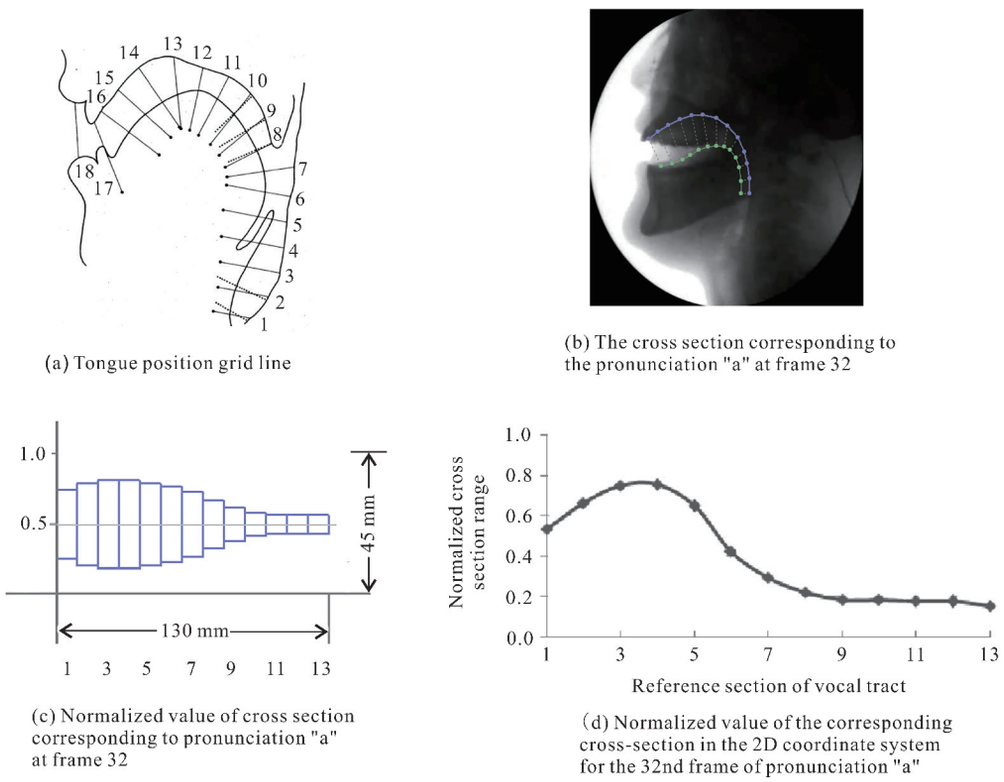


图 1　舌位轮廓标准化

Fig. 1　Tongue shape normalization

1.2　舌位轮廓形变

由于舌头的隐蔽性，舌位数据通常难以大量获取，其真实数据的样本量较小，但本文算法模型需采集大规模舌位轮廓数据才能进行有效训练。为此，本文通过添加噪声到原始的小规模真实舌形数据集来构建大规模的舌位轮廓数据集。考虑到人类发音的舌位不能随机改变，本文采取主动形状模式（Active Shape Mode，ASM）[37,38] 来产生可能存在于用来训练第一阶段舌位轮廓自编码器的舌位轮廓数据的生理变形，如式(2)所示。

$$E(S) = \operatorname{argmin}\sum_{i=1}^{m}\left[E_{\text{int}}(s_i)+E_{\text{edge}}(s_i)+E_{\text{con}}(s_i)\right], \tag{2}$$

式中，S 表示舌头轮廓，$s_i(1\leqslant i\leqslant m)$ 是 S 上的第 i 个控制点，m 为控制点总数。对于 s_i，$E_{\text{int}}(s_i)$ 表示来自相邻点 f_{i-1}^{int}、f_{i+1}^{int} 的内力能量，$E_{\text{edge}}(s_i)$ 是来自原始边缘或轮廓 f_i^{edge} 的边缘力，$E_{\text{con}}(s_i)$ 为来自外部输入的约束力 f_i^{con}。如图 2 所示，s_i 处的随机外部约束力 f_i^{con} 导致了 s_i 与 s_i^{con} 的初始偏差，通过求解式(2)可获得在 s_i^{final} 代表的最终位置。

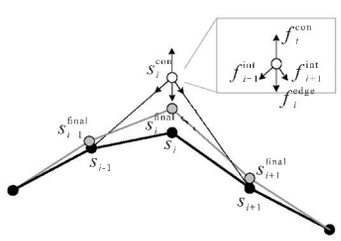


图 2　舌位控制点形变示例

Fig. 2　An example of point deformation

1.3　本文算法模型

本文将第一阶段网络结构定义为舌位轮廓自编码器（Tongue Shapes Denoising Auto Encoder，TS-DAE）[图 3(a)]，第二阶段定义为舌位轮廓降维编码器（Tongue Shape Dimensionality Reduction AutoEncoder，TSDR-AE）[图 3(b)]。为提升 TS-DAE 所需的样本数量，本文采用符合生理特征的舌头形变数据构造大量的舌位轮廓数据，从而扩充第一阶段 n 层网络结构所需的样本，然后再使用真实舌位运动轮廓数据微调 TSDR-AE 网络的第 $(n+1)$ 层。

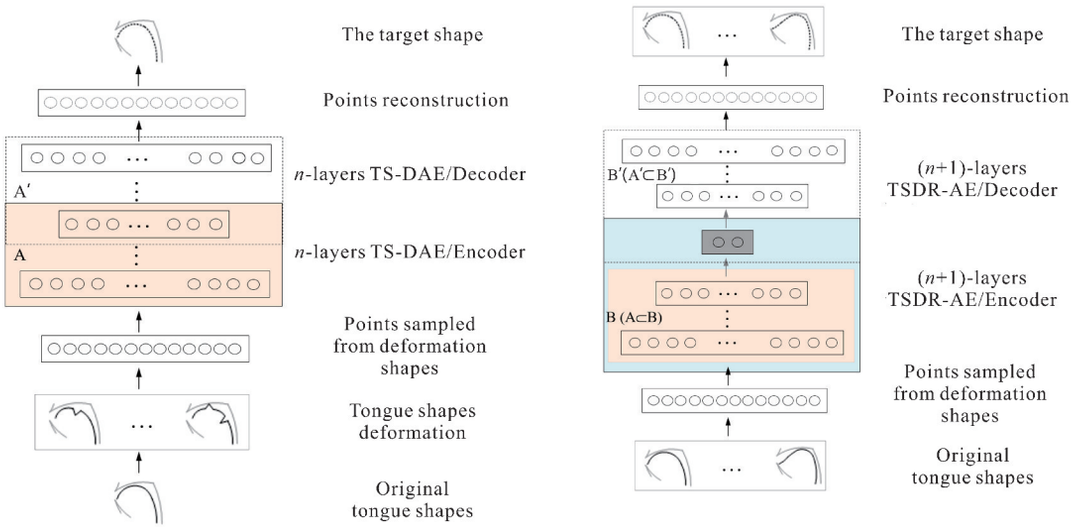


图 3　模型训练和数据处理流程

Fig. 3　Pipeline of the proposed model

1.4　舌位轮廓自编码器(TS-DAE)

基于 ASM 对舌头轮廓的变形能够产生大规模且保持一定生理特征的舌位轮廓数据。如图 3(a)所示,其最底部为真实舌位发音轮廓,在其上方矩形框中的轮廓为 ASM 所产生的变形舌位轮廓。基于这些变形舌位轮廓,TS-DAE 可对舌位轮廓进行有效的自动编码。首先,增强的舌位轮廓通过 TS-DAE 编码器[图 3(a)中的浅橙色框 A]可以获得指定维度的特征表示。然后,将 TS-DAE 编码器连接与之完全对称的 TS-DAE 解码器[图 3(a)中的虚线框 A′],对输入的舌位进行重建。TS-DAE 的舌位轮廓重建性能由生成的舌位重建轮廓[图 3(a)顶部的虚线轮廓]与原始舌位轮廓[图 3(a)下方实线轮廓]的差异值来评估,差异值越小说明 TS-DAE 的舌位轮廓重建性能越好。

1.5　舌位轮廓降维编码器(TSDR-AE)

由于 TS-DAE 输出数据维度较高,为实现舌位轮廓压缩,本文将具有少量隐藏单元的网络层堆叠到 TS-DAE 顶部,进而形成总共有 $(n+1)$ 层的 TSDR-AE。TSDR-AE 对舌位轮廓的编码和解码过程如图 3(b)所示。同 TS-DAE 一样,TSDR-AE 也由结构对称的编码器和解码器构成。TSDR-AE 的编码器[图 3(b)中的浅蓝色实线框 B]包含了 TS-DAE 的编码器,其解码器[图 3(b)中的虚线框 B′]也包含了 TS-DAE 的解码器。TSDR-AE 的最上层添加了维度较小的节点[图 3(b)中灰色部分],并用 TSDR-AE 解码器解码舌位轮廓低维度的特征表示,最终获得重建的舌位轮廓数据。TSDR-AE 的舌位轮廓重建性能由所生成的舌位重建轮廓[图 3(b)上方虚线轮廓]与原始舌位轮廓[图 3(b)下方实线轮廓]之间的差异来评估,差异值越小,TSDR-AE 的舌位轮廓重建性能越好。

2　实验与结果分析

2.1　数据准备

X 光片发音数据在发音观测上具有较好的时间分辨率[37,38],目前被广泛用于语音生成领域。本研究的舌头形状取自中国女性发音 X 光片视频所获得的舌位轮廓视频,包含 20 个音素(包括普通话元音)和 181 个音节。X 射线图像分辨率为 640×480。发音者舌头形状用公式(1)进行归一化处理。每个元音持续 35－50 帧,每帧时长约 30 ms。本研究以 5 个典型元音(“a”、“i”、“u”、“e”、“o”)为对象,选取了对应的 240 个真实舌形及 6 000 个生成的形变轮廓作为训练和测试数据来验证所提出的双阶段自动编码器方法的性能。

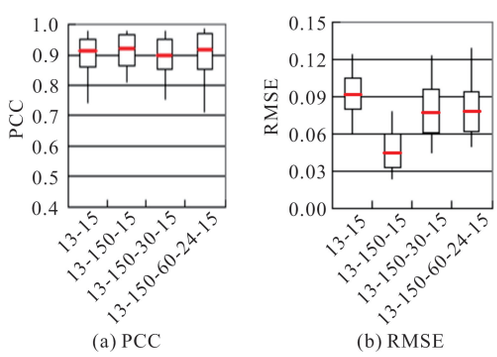
舌位轮廓数据需要转化为一维向量才能输入到神经网络中。首先是在舌尖到舌根的间隔范围内对 13 个节段采样,再将这 13 个节段的归一化横截面距离作为 TSDR-AE 的输入。第 i 层自动编码器的可见和隐藏单元数量分别定义为 l_i^v 和 l_i^h ,其中 $l_i^h = l_{i+1}^v (1 \leqslant i \leqslant n)$,$l_{i+1}^v$ 为自编码器的第$(i+1)$层的神

经节点个数，依次类推，第 $n+1$ 层的神经网络结构为 $l_{n+1}^{v}\times l_{n+1}^{h}$，满足 $l_{n+1}^{h}\ll l_{n}^{h}$ 且 $l_{n+1}^{h}\ll l_{1}^{v}$。第$(n+1)$层的输出可用于舌位轮廓重建，重建的形状应尽可能接近原始输入形状。这样，高维的舌位轮廓可由 TSDR-AE 网络顶层的低维输出值 l_{n+1}^{h} 表示。

由于发音过程中前舌较舌头后部会发生更大形变，本研究的舌位形变单元更多产生在上述 13 个节段的前 6 个。本文通过 120 个真实舌头形状构建了 6 000 个变形轮廓，其中 5 000 个用于第一阶段 TS-DAE 神经网络训练，1 000 个用于第一阶段 TS-DAE 网络性能评估。在舌位轮廓降维编码阶段，从 240 个真实的舌位轮廓中随机抽取 120 个舌形用于微调 TSDR-AE，其余的 120 个舌形则用于 TSDR-AE 网络性能评估。

2.2 自编码器舌位模型网络结构

为实现 TS-DAE 和 TSDR-AE 两阶段在结构及性能等方面的均衡分布，本文还对网络结构进行了优化。通常，隐藏层在获得足够单元输入的前提下，自动编码器能够拟合任意数据分布。在实践中，隐藏单元数为输入单元数的 10 倍左右时，自编码器即能产生较好结果。由于输入矢量包含了 13 个单元节段，且 TSDR-AE 输出层要求节点数较少，因此，TS-DAE 可被构建为“13-15”、“13-150-15”、“13-150-30-15”以及“13-150-60-24-15”等层级网络结构。实验使用原始舌形与模型重建的舌形间的皮尔逊相关系数(Pearson Correlation Coefficient，PCC)和均方误差(Root Mean Square Error，RMSE)来评估模型的性能(图 4)，其中 PCC 的值越大越好，RMSE 的值越小越好。从图 4 可见，“13-150-15”网络结构相对于其他 3 个网络结构能获得较理想的 PCC 和 RMSE，尤其是RMSE明显更优。因此，本文将进一步使用该网络构建$(n+1)$层的 TSDR-AE 网络结构。



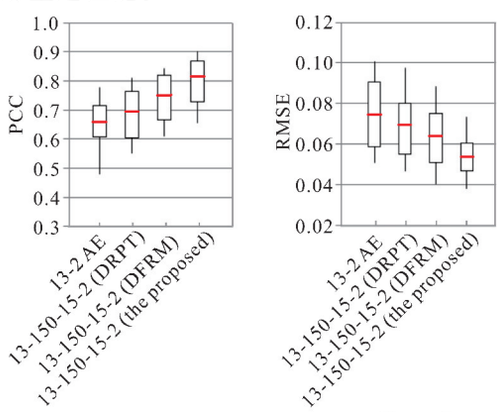
(a) PCC　(b) RMSE

图 4　4 种网络结构的 PCC 和 RMSE 的平均值、最大值、最小值和方差范围比较结果

Fig. 4　The average values, maximal values, minimum values, and variance range of PCC and RMSE of 4 networks

实验将舌位轮廓压缩在 2 个因子内(即 TSDR-AE 的输出层单元为 2)进行分析和比较。因此，本文在第一阶段 TS-DAE 的顶部附加了“15-2”自动编码器，构建了“13-150-15-2”堆叠的 TSDR-AE。为验证本文模型[13-150-15-2 (the proposed)]性能，将其与标准的 2 层“13-2”自编码器模型(13-2 AE)、采用 Dropout 技术进行数据增强的“13-150-15-2”DAE 模型[13-150-15-2 DAE (DRPT)]、使用形变进行舌位增强数据训练的“13-150-15-2”DAE 模型[13-150-15-2 (DFRM)]进行实验比较。其中，将 13-2 AE 模型和 13-150-15-2 DAE (DRPT)模型在 120 个真实舌位轮廓上进行 Dropout 训练；对 13-150-15-2 DAE (DFRM)模型使用了 5 000 个变形舌位轮廓进行训练。本文所提出的 TSDR-AE 模型训练过程与上述 13-150-15-2 DAE (DFRM)类似，但额外随机抽取了 120 个真实舌形数据对其进行训练及微调。

本文采取上述 4 种模型来验证 120 个原始舌形与重建舌形间的 PCC 和 RMSE，图 5 是对比结果的箱线图。由图 5(a)可知，13-150-15-2 (the proposed)模型的 PCC 值高于其他 3 种模型。同时，由图 5(b)可知，13-150-15-2 (the proposed)的 RMSE 比其他 3 种模型更小，说明其误差更小。以上结果充分表明 13-150-15-2 (the proposed)模型所重建的舌位轮廓与真实舌位轮廓更为接近，说明该模型具有更好的舌位轮廓重建性能。



(a) PCC　(b) RMSE

图 5　4 种模型的 PCC 和 RMSE 的平均值、最大值、最小值和方差范围

Fig. 5　The average values, maximal values, minimum values, and variance range of PCC and RMSE for the 4 models

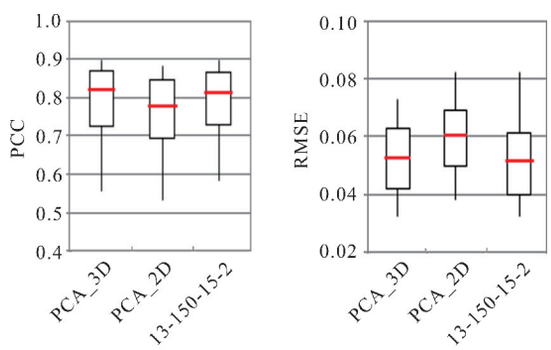
2.3 与 PCA 的重建性能比较

PCA 是语音学领域用于舌位轮廓压缩和重建的常见降维工具[5,37,38]。这里进一步比较 13-150-15-2

(TSDR-AE)网络结构和 PCA 对真实舌位轮廓的重建性能。根据多名学者利用 PCA 模型在舌位轮廓上的分析结果[5,37,38],PCA 舌位模型的前 2－4 个主成分通常占所有成分的 95%以上。图 6 为 120 个测试舌形上 PCA 和 13-150-15-2 (TSDR-AE)模型的 PCC 和 RMSE 的结果,其中 PCA_iD 项中的 i 表示采用前 i 个分量的 PCA 重建结果。

从图 6(a)可以看到,13-150-15-2 (TSDR-AE)模型重建的舌位轮廓与原舌位轮廓的 PCC 平均值为 0.83。相对于 PCA_2D 的 0.77 以及 PCA_3D 的 0.81,13-150-15-2 (TSDR-AE)模型的 PCC 值比 PCA_2D 和 PCA_3D 的更高。这说明与 PCA 相比,该文模型把舌位轮廓压缩到二维后重建的舌位轮廓与原舌位轮廓更相似。

由图 6(b)可知,13-150-15-2 (TSDR-AE)模型重建的舌位轮廓与原舌位轮廓的 RMSE 平均值为 0.05。相对于 PCA_2D 的 0.06 以及 PCA_3D 的 0.05,该模型并不逊色,这说明该模型在把舌位轮廓压缩到二维然后重建的舌位轮廓与原舌位轮廓的误差更小。综上,13-150-15-2 (TS-DAE)模型将舌位轮廓压缩到二维后重建的舌位轮廓比 PCA 压缩舌位轮廓到二维和三维后重建的舌位轮廓更好。



(a) PCC　　(b) RMSE

图 6　PCA (2D,3D)模型和 13-150-15-2 (TSDR-AE)模型在 120 个测试舌位轮廓上重建与原始舌位的 PCC 和 RMSE 的平均值、最大值、最小值和方差范围比较结果

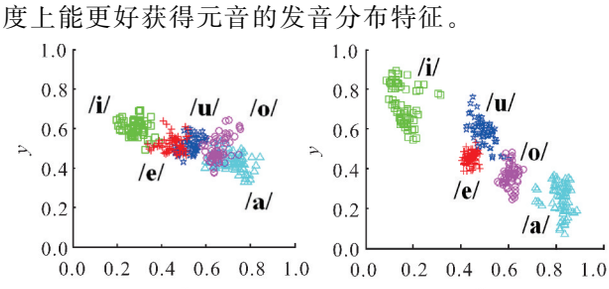
Fig. 6　The average values, maximal values, minimum values and variance range of PCC and RMSE for PCA (2D, 3D) and 13-150-15-2 (TSDR-AE) on 120 test tongue shapes

2.4　元音二维发音图谱分布性能比较

为更直观地验证所提模型性能,实验分别使用 PCA_2D 模型和本文模型将 240 个舌位轮廓压缩为二维变量并投影到 2D 坐标系,相应投影点分布如图 7(a)、图 7(b)所示。其中的不同元音投影点分别采用不同颜色符号进行标识,元音“i”用绿色方块标识、元音“e”用红色加号标识、元音“u”用深蓝色五角星符号标识、元音“o”用紫色圆形符号标识、元音“a”用浅蓝色三角形符号标识。

由图 7(a)可知,不同元音间的二维投影点存在较多重叠,“u”(蓝色区域)和“e”(红色区域)重叠较多,“a”(浅蓝色区域)与“o”(紫色区域)重叠也非常明显,这意味着 PCA_2D 模型所获得的不同元音舌形并不利于区分。

由图 7(b)可知,5 个元音的发音被 13-150-15-2 (TSDR-AE)划分为 5 个簇,其中“i”与“a”的簇间距离比图 7(a)中的“i”与“a”更远,仅有“u”与“e”、“o”存在少量的边界点相邻。因此,13-150-15-2 (TSDR-AE)模型和 PCA_2D 将汉语元音的舌位发音轮廓同时压缩到二维,并将结果在 2D 坐标系进行可视化,前者比后者能获得更好的舌位区分结果,这说明所提出模型相对于被广泛使用的 PCA 方法在二维压缩维度上能更好获得元音的发音分布特征。



(a) PCA_2D　　(b) 13-150-15-2 (TSDR-AE)

图 7　汉语元音发音舌位轮廓降维到二维的可视化结果

Fig. 7　Visualization of the reduction points in 2D coordinate system

3　讨论

将基于舌位形变的 13-150-15-2 (DFRM)模型与标准的两层 13-2 AE 模型、采用 Dropout 技术进行数据增强的 13-150-15-2 DAE (DRPT)模型进行比较可以得知,13-150-15-2 (DFRM)模型相对于其他 2 个模型,其 PCC 值分别提高了 0.09 和 0.05,同时 RMSE 值分别降低了 0.007 和 0.013。这表明基于 ASM 的形变技术能生成更多符合一定发音规律的舌位轮廓数据,使得模型的第一阶段 TS-DAE (舌位轮廓自编码器)受输入数据影响较小,具有更强的鲁棒性,进而有效提高了模型的性能。

模型的第二阶段 TSDR-AE 通过引入带有少量隐藏单元的附加网络层进行微调。该附加网络层能进一步提高对真实舌位轮廓的拟合度,使得本文模型比 13-150-15-2 (DFRM)模型具有更好的舌位重建性

能。从图5可见,本文模型较13-150-15-2(DFRM)模型的PCC值提高0.07,同时其RMSE值降低0.01,表明该模型所提出的第二阶段TSDR-AE能进一步改进舌位轮廓自编码器整体性能。

将所提方法与PCA方法在舌位压缩重建后的效果进行比较,通过对120个真实测试舌形压缩和重建的实验结果表明,采用13-150-15-2(TSDR-AE)将舌位轮廓压缩到二维,其重建的舌位轮廓明显优于采用PCA压缩舌位轮廓到二维重建的舌位轮廓,甚至更优于通过PCA压缩到三维所获得的重建结果。

将舌位轮廓压缩为二维变量并投影到2D坐标系中。由图7可知,本文模型在二维坐标系中的元音舌形压缩和可视化方面均优于传统的PCA_2D模型,其所获得的二维点分布呈现出更好的分类效果,即拥有更好的元音舌位识别能力。究其原因,主要是因为TSDR-AE具有较高的重建性能和良好的降维能力,确保了TSDR-AE模型较传统PCA方法能更直观建立舌关节结构和低维参数之间的双向映射关系。

综上,虽然舌位因其隐蔽性等生理特征而无法产生大量真实样本数据,但本文基于ASM产生的舌位形变数据所提出的两阶段自动编码器舌位模型比PCA舌位模型具有更强的舌位轮廓压缩能力、降维能力以及元音舌位区分能力。

4 结论

针对传统深度学习自动编码器难以直接用于舌位轮廓分析的问题,本文提出了一种基于小样本真实舌位数据统计分析的双阶段自动编码器方法。第一阶段通过引入具有生理特征的大规模变形方法,构建通用轮廓重建模型;第二阶段在前阶段的基础上添加隐藏单元,构建与降维目标维度相等的附加网络层对舌位数据进行压缩。实验在人类真实的小规模元音舌形数据上进行验证,并与传统PCA方法比较了降维、重建性能。实验结果表明,本文所提舌位轮廓重建模型比PCA方法的重建性能更优,所生成的元音舌位图谱在二维平面上也呈现出更好的区分度。

参考文献

[1] ROXBURGH Z,CLELAND J,SCOBBIE J M,et al. Quantifying changes in ultrasound tongue-shape pre- and post-intervention in speakers with submucous cleft palate:an illustrative case study [J]. Clinical Linguistics & Phonetics,2022,36(2/3):146-164.

[2] LI H,YANG M H,TAO J H. Speaker-independent lips and tongue visualization of vowels [C]//2013 IEEE International Conference on Acoustics,Speech,and Signal Processing. Vancouver,Canada:IEEE,2013:8106-8110.

[3] XU K L,YANG Y,LEBOULLENGER C,et al. Contour-based 3D tongue motion visualization using ultrasound image sequences [C]//2016 IEEE International Conference on Acoustics,Speech,and Signal Processing. Shanghai,China:IEEE,2016:5380-5384.

[4] WANG G W,KONG J P. An articulatory model of standard Chinese using MRI and X-ray movie [J]. Journal of Chinese Linguistics,2015,43(1):269-294.

[5] STONE M,JR GOLDSTEIN M H,ZHANG Y Q. Principal component analysis of cross sections of tongue shapes in vowel production [J]. Speech Communication,1997,22(2):173-184.

[6] HEWER A,STEINER I,BOLKART T,et al. A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract [C]//18th International Congress of Phonetic Sciences. Glasgow,United Kingdom:University of Glasgow,2015:136-145.

[7] HARSHMAN R,LADEFOGED P,GOLDSTEIN L. Factor-analysis of tongue shapes [J]. Journal of the Acoustical Society of America,1977,62(3):693-707.

[8] ZHENG Y L,HASEGAWA-JOHNSON M,PIZZA S. Analysis of the three-dimensional tongue shape using a three-index factor analysis model [J]. Journal of the Acoustical Society of America,2003,113(1):478-486.

[9] ISKAROUS K. Patterns of tongue movement [J]. Journal of Phonetics,2005,33(4):363-381.

[10] MAEDA S. An articulatory model of the tongue based on a statistical analysis [J]. Journal of the Acoustical Society of America,1979,65(S1):S22.

[11] BRO R. PARAFAC. tutorial and applications [J]. Chemometrics and Intelligent Laboratory Systems,1997,38(2):149-171.

[12] HARSHMAN R A. Foundations of the PARAFAC procedure:models and conditions for an "explanatory" multimodal factor analysis [J]. UCLA Working Papers in Phonetics,1970,16:1-84.

[13] LU X G,DANG J W. Vowel production manifold:intrinsic factor analysis of vowel articulation [J]. IEEE Transactions on Audio,Speech,and Language Processing,2010,18(5):1053-1062.

[14] LIANG C W,KONG J P,WU X Y. A speech-driven 3-D tongue model with realistic movement in Mandarin

Chinese [C]//Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing (BIC 2021). New York, USA: ACM, 2021: 297-302.

[15] WANG L, CHEN H, LI S, et al. Phoneme-level articulatory animation in pronunciation training [J]. Speech Communication, 2012, 54(7): 845-856.

[16] XU K L, YANG Y, JAUMARD-HAKOUN A, et al. 3D tongue motion visualization based on ultrasound image sequences [C]//Interspeech 2014. Singapore, Singapore: ISCA, 2014: 1482-1483.

[17] BIRKHOLZ P. Modeling consonant-vowel coarticulation for articulatory speech synthesis [J]. PLoS One, 2013, 8(4): e60603.

[18] NARAYANAN S, TOUTIOS A, RAMANARAYANAN V, et al. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC) [J]. Journal of the Acoustical Society of America, 2014, 136(3): 1307-1311.

[19] QIN C, CARREIRA-PERPINÁN M Á, RICHMOND K, et al. Predicting tongue shapes from a few landmark locations [C]//Interspeech 2008. Brisbane, Australia: ISCA, 2008: 2306-2309.

[20] CALKA M, PERRIER P, OHAYON J, et al. Machine-Learning based model order reduction of a biomechanical model of the human tongue [J]. Computer Methods and Programs in Biomedicine, 2021, 198: 105786.

[21] PARK B, BISWAS S, PARK H. Electrical characterization of the tongue and the soft palate using lumped-element model for intraoral neuromodulation [J]. IEEE Transactions on Biomedical Engineering, 2021, 68(10): 3151-3160.

[22] QIAN C Q, GU H Y, YANG Z C, et al. An automatic petechia dots detection method on tongue [C]//The 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. Online: IEEE, 2021: 3362-3365.

[23] RUAN Q S, WU Q F, YAO J F, et al. An efficient tongue segmentation model based on U-Net framework [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2021, 35(16): 2154035.

[24] PLOUMPIS S, MOSCHOGLOU S, TRIANTAFYLLOU V, et al. 3D human tongue reconstruction from single "in-the-wild" images [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2021: 2771-2780.

[25] MANSOUR R F, ALTHOBAITI M M, ASHOUR A A. Internet of things and synergic deep learning based biomedical tongue color image analysis for disease diagnosis and classification [J]. IEEE Access, 2021, 9: 94769-94779.

[26] NIX D A, PAPCUN G, HODGEN J, et al. Two cross-linguistic factors underlying tongue shapes for vowels [J]. Journal of the Acoustical Society of America, 1996, 99(6): 3707-3717.

[27] SALTZMAN E L, MUNHALL K G. A dynamical approach to gestural patterning in speech production [J]. Ecological Psychology, 1989, 1(4): 38-68.

[28] JACKSON P J B, SINGAMPALLI V D. Statistical identification of articulation constraints in the production of speech [J]. Speech Communication, 2009, 51: 695-710.

[29] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[30] DU B, XIONG W, WU J, et al. Stacked convolutional denoising auto-encoders for feature representation [J]. IEEE Transactions on Cybernetics, 2017, 47(4): 1017-1027.

[31] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. Journal of Machine Learning Research, 2010, 11: 3371-3408.

[32] DENG L, YU D. Deep learning: methods and applications [J]. Foundation and Trends in Signal Processing, 2013, 7(3/4): 197-387.

[33] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders [C]//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008: 1096-1103.

[34] HINTON G E, OSINDER S, TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18: 1527-1554.

[35] BENGIO Y. Practical recommendations for gradient-based training of deep architectures [M]//Neural Networks: Tricks of the Trade, 2nd ed. Berlin: Springer, 2012: 437-478.

[36] FANT G. Acoustic theory of speech production [M]. 2nd ed. [S. l.]: Mouton, 1970.

[37] HÖWING F, DOOLEY L S, WERMSER D. Tracking of non-rigid articulatory organs in X-ray image se-

quences [J]. Computerized Medical Imaging and Graphics,1999,23(2):59-67.
[38] KASS M,WITKIN A,TERZOPOULOS D. Snakes:active contour models [J]. International Journal of Computer Vision,1988,1(4):321-331.

Tongue Shapes Modeling from Small Data Using Two-Stage Autoencoder

XU Zhengli[1],XIAO Sufang[1**],JIAN Min[1],YANG Minghao[2]
(1. Guilin University of Electronic Technology,Guilin,Guangxi,541004,China;2. Institute of Automation of the Chinese Academy of Sciences,Beijing,100190,China)

Abstract:The tongue plays a crucial role in human speech production. The dimensionality reduction analysis of tongue pronunciation can effectively assist linguists in analyzing human pronunciation patterns. Traditional methods for tongue position contour compression often relay on Principal Component Analysis (PCA) for dimensionality reduction. In recent years,deep-learning-based autoencoders have been widely used for data compression. However,they require a large number of samples and cannot be directly and effectively used for tongue motion pattern researches. Besides,obtaining a substantial volume of tongue movement data has been challenging due to the tongue's location within the oral cavity. To address these limitations,this paper introduces a two-stage autoencoder dimensionality reduction method designed for small-sample tongue motion contour data. Firstly,Active Shape Model (ASM) is used to generate a large amount of physiological deformation data of tongue contour,and a general tongue contour reconstruction model is constructed based on a conventional automatic encoder. Secondly,on the basis of the automatic encoder in the previous stage,an additional network layer is added to compress and analyze the tongue position data. In experiments,240 vowel and tongue shape datasets obtained from X-ray films of human speech are selected. The tongue position model and traditional PCA methods were compared. The results show that the vowel tongue position map obtained by the proposed method exhibits better discrimination on the two dimensional plane,and has better tongue shape reconstruction performance.

Key words:deep neural network;autoencoder;Principle Component Analysis (PCA);tongue contour;hidden units

责任编辑:陆　雁,陈少凡