三维几何声道建模*

宋婵[1]，方强[2]，王宇光[1]，魏建国[1]，党建武[1,3]

（1.天津大学 计算机科学与技术学院，天津 300072；2.中国社会科学院，北京 100732；
3.北陆无端科学技术大学院大学，日本）

文　摘：生理发音模型可以应用于语音生成，语音分析，语音教学等各个方面。本研究，我们基于汉语普通话核磁共振发音数据库构建了一个汉语普通话发音的几何模型。模型的构建分为两个部分，数据标注和模型控制参数抽取。控制参数抽取是对第一阶段标记的数据进行分析，方法采用线性成分分析法。目的是能够利用较少具有明显物理意义的参数来描述复杂的发音器官的位置和形状，实现对复杂声道形状的控制。分析结果显示，每个发音器官可用三个以内参数来很好的描述，且平均重构误差小于 1mm。

关键词：生理发音模型；发音器官形状；线性成分分析

中图分类号： 分类号 1；分类号 2

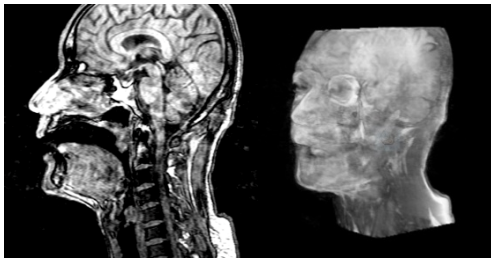
在语音研究领域，语音生成作为一个重要的分支已经受到人们越来越多的关注。基于生理的发音模型有助于我们验证和统合现阶段对于语音产生的生理和物理机制的理解，推动语音产生研究的进一步发展。早期用于生理建模的方法有神经肌肉建模和几何建模。前者的代表是 Dang-Honda 模型[1]等，他们通过不同肌肉力的组合来控制发音器官的运动从而描述整个声道形状的变化。几何模型只需要对不同声道的几何形状进行分析，找出声道形状的主要控制因子，通过调整这些因子的数值达到控制声道形状的目的。

几何模型的构建又有二维中矢面声道形状建模和三维声道形状建模两种。二维模型相对于三维模型来说，构建简单，因而在建模早期被很多人使用。比如 Engwall 和 Badin 在早期的声道建模中都曾构建过二维模型[2]。但是二维中矢面模型在计算声道面积函数的时候只能采用估算的方法；加上一些发音对应的声道形状是中矢面上封闭，两边部分分开,如/l/，无法用二维模型正确描述。因此，三维模型的构建逐渐引起人们的重视。对三维建模的研究也从最初 Engwall 基于瑞典人建立的三维舌头模型[3]，到后来 Badin 基于法语发音构建了三维发音器官模型[4]。

综上所述，国外对发音模型做了大量的工作，而针对汉语发音的三维发音模型的研究很少。本文将基于汉语普通话的核磁共振数据库，建立汉语的发音模型。

本文组织结构如下，第一部分主要介绍所选用数据库的信息；第二部分给出了对数据库图像进行标注的方法，以及每个发音器官标记后的结果形状；对标记结果进行分析的方法在第三部分介绍；分析结果则于第四部分给出并解释；最后第五部分介绍了尚且存在的问题和下一步的研究思路。

1　数据库



a.　　b.

图 1 数据库中的 MRI 图像和三维显示图像，其中 a 是数据库中原始 MRI 图像，b 是三维图像叠加显示的形状

我们选用的数据库是汉语普通话发音 MRI 数据库，数据库共包括 8 个发音人，每个发音人共

*基金项目：国家科学基金面上项目(NO.61175016)、基金重点项目（NO.61233009）
作者简介：宋婵（1989），女（汉），河北，研究生。
通讯联系人：魏建国，副研究员，E-mail：jianguo.fr@gmail.com

有 10 个元音发音和部分辅音以及一组口含蓝莓汁的牙模数据。每组元音发音由 51 帧与矢状位平行的头部切面图像（如图 1.a 所示）组成，所有的切面放到一起可以恢复出三维的头部形状(如图 1.b 所示)。切面大小均为 512*512，单位为像素，其中 1mm 等于两个像素的长度；两个切面的中心距离为 2mm;我们在研究中，选取了其中一个发音人的所有元音和牙模数据作为研究数据。

MRI 原始图像上的牙齿是骨骼结构，而骨骼结构在 MRI 采集过程中与空气一样无法直接显示出来(如图 1.a 所示)。我们想要描述完整的声道形状，就需要同时包括牙齿的形状数据，因此在进行器官的标注之前，首先采用 Takemoto 的补牙方法对 MRI 数据进行牙齿的填补[5]，再对填补之后的图像进行各个器官的标注(如图 3.a 所示)。

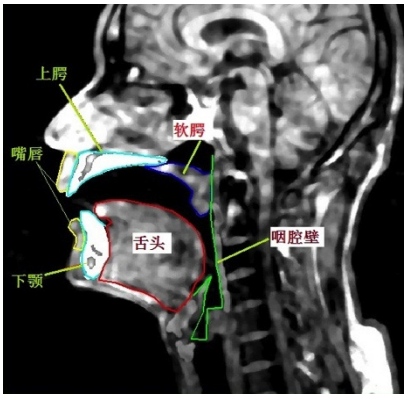
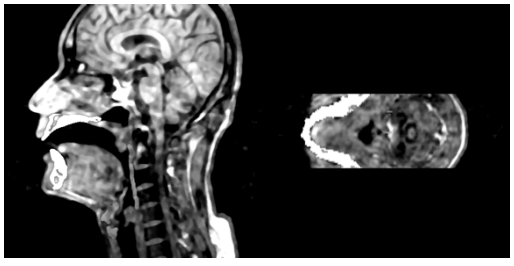


图 2 对数据库中的 MRI 图像标注的器官轮廓

理想情况下，对每个发音的每帧图片上的发音器官进行标注，那么所有 51 帧图像上的器官叠加到一起就是此发音器官的三维形状。而实际上，在靠近两边部位的 MRI 图像上，器官与器官之间的边界并不明显，比如舌与咽腔壁。这是因为当切面与器官的主要伸展方向垂直的时候，显示效果最佳，而舌两边的边界与矢状面几乎平行。



a. b.

图3 原始图像补牙之后的效果和重新分割的横切面图像,其中 a 是.补牙之后的 MRI 图像，b 是重新分割的横切面 MRI 图像

针对这种现象，我们选用一组其他角度的切面图片来标注舌两边的形状。这里所指的是横切面，它是对图 1.b 所示的三维头部进行重新分割得到的，从眼睛部分到声门一共分割成了 50 个水平的横切面，每个切面图像如图 3.b 所示。其中，两个切面的中心点距离为 7.84 个像素点，即 3.92mm，每个切面均由 512*512 像素组成。在此切面上，舌头与咽腔壁之间的界限比较清楚，因此可以对舌头轮廓进行标记，得到舌两边较准确的轮廓。因舌背与横切面几乎平行，横切面无法准确的描述舌背的形状，因此需要将两种切面的轮廓融合来得到最终的舌头形状。

2 数据标注

在我们的标记中，声道器官一共包括 6 部分，分别是上、下嘴唇，下颚，上腭（包括硬腭和软腭），舌和咽腔壁。标记主要以矢状面图像的标记为主，但是舌和咽腔壁还需标记其横切面上的轮廓，然后做融合。其他 4 个声道器官因为在矢状切面上的边界轮廓已经比较清楚，所以用矢状面标注的形状作为最终形状。需要注意的一点是，由于下颚与舌有一部分是连在一起的，所以这部分点在下颚和在舌上必须相同才能保证下颚与舌完全重合，确保器官更好地对应性。

对标注好的数据还需要做预处理，包括上下颚数据的数据校准，器官表面的光滑处理，各个器官上的生理边界统一处理等，预处理前后的结果对比将在第三部分分析结果中给出。

2.1 下颚

从人的生理结构来看，下牙属于骨骼材质，与下颚固定在一起，因此我们将下牙与下颚作为一个整体进行标注和分析，统称下颚。下颚同时与舌和下唇连接在一起，影响着舌和下唇的运动，所以下颚的正确描述至关重要。

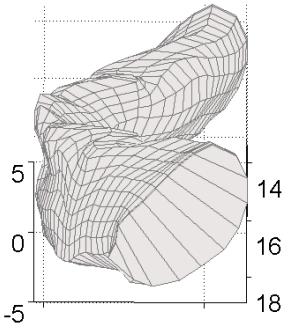


图 4 发/a/音时的下颚网格形状

下颚的标注有两种方法，一种是手动标注的方法，有多少个发音就标注多少次；另外一种是手动加旋转的方法，只需标注其中一个发音的下颚形状，其他发音的下颚数据通过对已标注的下颚进行平移和旋转得到。第二种方法是第一种方

法的改进，依据的原理是下颚在发不同的声音时自身的固定不变性，这么做可以避免引入人为标注误差。平移量和旋转角度是以标注的两个下颚数据中矢面轮廓差异来计算得到的。图 4 是三维下颚形状的网格显示。

2.2 舌

数据库中的矢状面图像上两边部分的舌头轮廓并不清楚，而重新分割的横切面图像上舌头两边部分较清晰，但舌面缺失（如图 5.b 所示），因此舌的描述需要融合两种切面数据得到。我们对舌的处理分为三个层面：

首先，我们需要分别标记矢状面和横切面上的舌形状。之后对标记的数据进行预处理，使得舌上的生理边界点对应。舌上的生理边界点有 3 个，舌尖，舌头固定部分与活动部分的连接处，舌面和舌根的连接处。这样舌就被分为 4 个部分，然后每个部分再进行平分。最终矢状面上的舌由 39 个点组成。横切面的每帧舌头形状也由 39 个点组成。两种切面的舌数据显示形状如图所示，图 5.a 为矢状面的舌头形状，图 5.b 为横切面的舌头形状。 可以看出，矢状面舌头缺少两边的形状，横切面缺少舌面的信息。

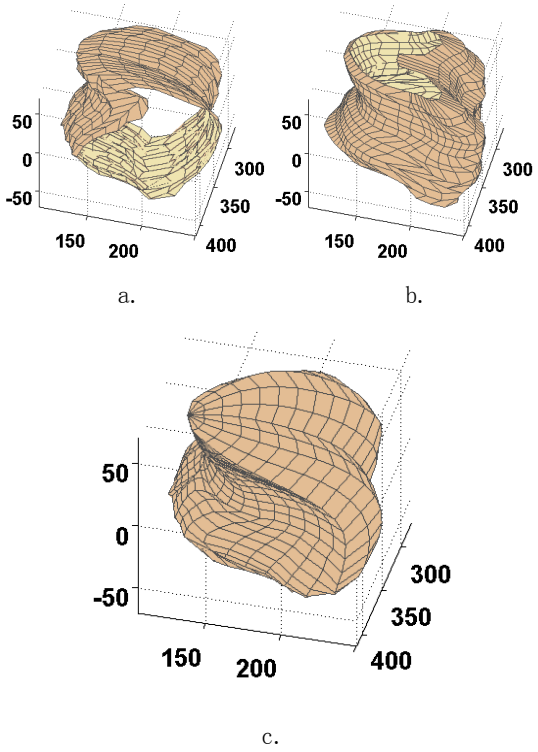


图 5 舌头的网格形状形成过程,a 是矢状位切面标注的舌头，b 是横切位图像标注的舌头，c 是两种切面舌头融合并光滑处理之后的舌头

然后对融合后的舌头进行重新采样和光滑，提高舌头表面点的对应性和规整性，以备后面分析所需。融合后的舌虽然完整了，但是点的分布不规则，且表面粗糙，因此又对其表面进行重新采样和光滑处理，得到如图 5.c 所示的最终舌头形状。

2.3 上下唇

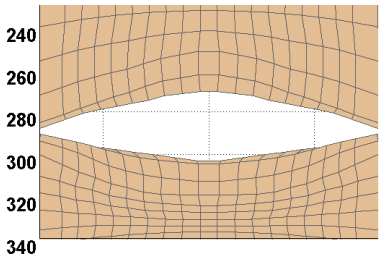


图 6 发/a/音时的上下唇网格形状

嘴唇在人们的语言交流中起着非常重要的作用，人发不同的声音需要不同的嘴唇形状来配合。嘴唇的边界在矢状面的 MRI 图像中已经较清晰，而横切面的嘴唇反而不易辨别，因此将矢状位的轮廓标记作为最终的嘴唇轮廓，上下嘴唇一起显示，结果如图 6 所示。

3 分析方法

发音器官轮廓的网格模型中顶点的位置之间有很强的相关性，我们可以采用线性成分分析的方法抽取出有明显物理意义的发音器官形状控制因子，用来描述各发音器官的形状。

在发音运动过程中，发音器官之间是有关联性的，不仅包括生理上的联系，比如下颚的运行会连带着舌和下唇的运动;还包括功能上的联系，比如发一个声音时各个发音器官之间的协同作用。我们的方法是选取一个发音器官，比如下颚作为基准器官，首先对其进行主成分分析，得出下颚的主要描述因子；然后其他器官先分析其与下颚的关联，然后减去下颚对其的影响之后再做主成分分析，得到各自的描述因子。分析关联的方法是多元线性回归，目的使器官减去下颚的影响得到残差。

所谓线性成分分析法，是多元线性回归和主成分分析法的综合运用。各个器官采用主成分分析法得到其正交化向量，单个器官的所有特征向量之间互相正交；将所有器官的特征向量放到一起的时候，器官和器官之间的控制向量则不一定正交。所有器官的控制参数一起组成整个声道形状的控制参数向量。公式如下：

$$x = \bar{x} + P*b \quad (1)$$

其中，$\bar{x}$是平均器官形状向量。器官的形状向量 x 经分析被转换为一系列的基本形状向量 b，

即线性成分；每个成分的权重系数保存在转换矩阵 P 中。

4 分析结果

按照以上所述方法，每一个发音器官都可以得到一组对应的主要描述参数，用来控制该器官的形状变化。对于任意一个发音器官来说，用它对应的主要描述参数来恢复原始器官，然后计算重构误差，可以得到该器官的描述参数在控制器官变化时的准确程度。在重构的同时，将每个维度的物理意义给出展示，可以直观的看到每一维描述参数对器官变化的影响。下面将依次给出每个器官的降维结果、重构误差以及每一维主控制参数的物理意义。

4.1 下颚

由以上介绍可知，下颚有两种标记方法，下面分别给出两种标记方法得到的分析结果。

对第一种方法得到的数据进行分析，得出第一维贡献率为 60.56%，前 4 维的累计贡献率为 87.15%。而从 EMA 的元音发音运动轨迹中可以看出，下颚的运动几乎可以认为是一个 2 维的运动，即主要是在一个方向上运动，所以下颚的运动规律只需要 1 维或者 2 维即可描述，与此结果矛盾。因此我们做了一个验证实验，分别分析标注的二维中矢面数据和三维的数据。结果显示，2 维分析结果中，前两维已经达到了 92.32%的贡献率，显然与三维结果不同。而下颚是固定不变的，所以 2 维和三维分析结果的不同证实第一种标记方法引入了人为误差。

对第二种改进的方法标记得到的数据进行分析，发现第一维的贡献率可以达到 96.92%，且 2 维和三维的分析结果一致。同时，仅使用第一维分析结果重构下颚，得到误差为：0.0206cm。用二维中矢面的数据可以更清楚的描述下巴的受影响程度，如下图 4 所示。主要表现为下颚在斜向上方向上的平移变化：

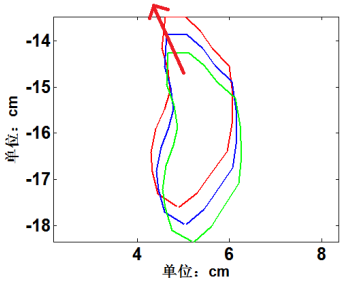


图 7　下颚分析结果中第一维对下颚位置的影响

其中，蓝色是平均下颚的中矢面轮廓，红色是在第一维控制参数基础上再加上$\sqrt{\lambda_1}$重建得到的下颚形状，绿色是减去$\sqrt{\lambda_1}$之后的重建结果。

4.2 舌

舌有一部分是与下颚紧密连接的，在发音过程中这部分点会随着下颚的运动而运动。因此，舌头上与这部分连接在一起的部分需要根据下牙的形状进行调整。

另外，对舌进行分析时，我们希望只分析舌自身的形变规律，所以需要首先减去下颚的影响得到残差，再对残差进行主成分分析，得到只影响舌自身形变的控制参数。分析结果显示，舌第一维的贡献率为 60.60%，前三位的累计贡献率已经达到了 88.06%。用前三位的控制向量进行重构，误差为 0.0644cm。其中每一维的物理意义如图 8 所示。

从图中可以看出，舌的第一维控制参数对舌形状变化的影响主要表现在舌整体在左上方方向上的变化。第二维表现为舌整体在右上方方向上的变化。第三位则主要表现为舌面中间部分的上下变化上。

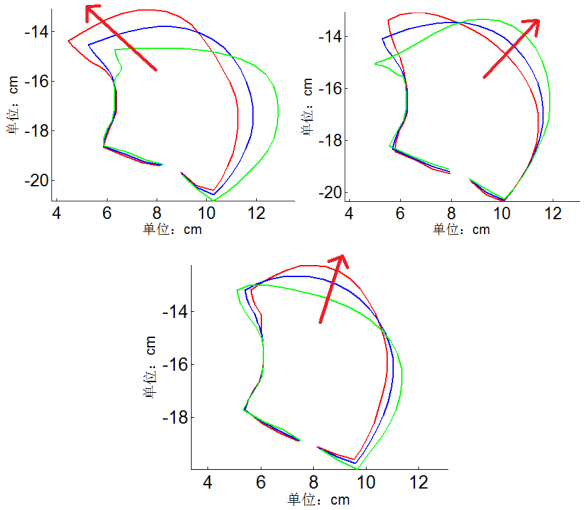
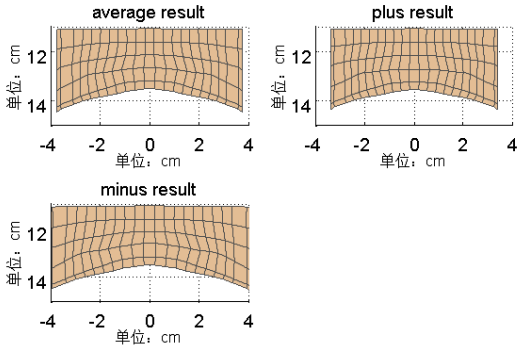


图 8 舌头前三维的物理意义，自上至下依次为第一维到第三维

4.3 上唇



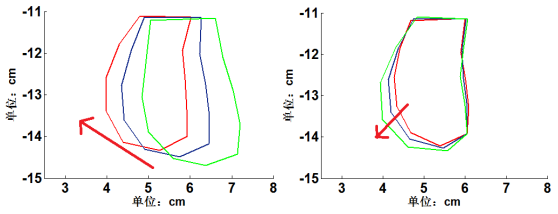


图 9 上唇前两维的物理意义，上面的为第一维的三维形状对比图，下面为前两维贡献率的 2 维中矢面对比图

对 10 个元音发音的上唇数据同样需要首先进行残差计算，然后再做主成分分析。在对嘴唇的数据进行预处理时，我们发现以嘴角为嘴唇的生理边界点进行调整之后，嘴唇的描述有很大的提高。调整前的分析结果，前三位贡献率可累积达到 82.13%，调整之后，前两维的贡献率已经可以达到 94.21%。用前两维的贡献率进行上唇重构，误差为 0.0287cm。前两维的物理意义如图 9 所示。

从图中可以看出第一维物理意义表现在唇的圆－展，随着唇的变窄会同时引起唇的中间部分变厚；第二维表现为上唇的凸起程度。

4.4 下唇

下唇的分析与上唇一致，对嘴角进行调整前的结果中，前三维的累积贡献率达到 91.92%，调整后的分析结果中，前两维的贡献率可以达到 92.58%。我们采用前两维的控制参数对下唇进行重构，重构误差为 0.0275cm，每一维的物理意义如图 10 所示。

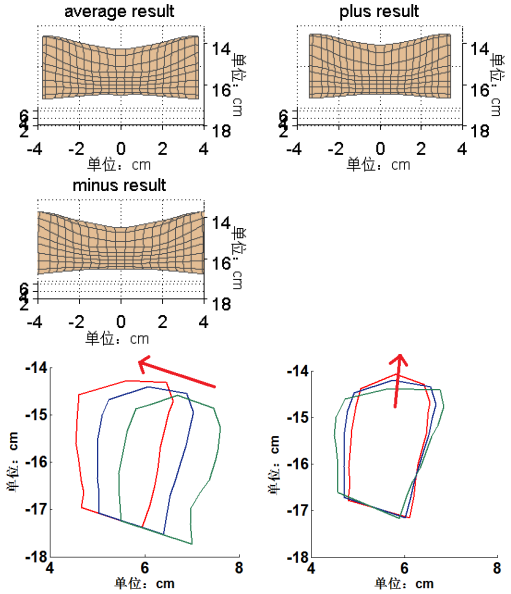


图 10 下唇前三维的物理意义，第一张图为第一维三维形状对比图，下面依次为前两维的 2 维对比图

从图中可以看出，下唇第一维和上唇的第一维物理意义相同，都是描述嘴唇的圆展度；第二维主要表现在描述嘴唇的上下运动上。

5 总结和前景

本研究的最终目的是建立一个生理发音几何模型，借助该模型，可以使用较少的参数来描述整个三维声道的形状。数据库采用中文发音的 MRI 数据库，方法是线性成分分析法。分析步骤有两个，数据标注和数据分析。最后结果显示，每个发音器官的描述因子可以减少到 3 个以内，且重构误差均小于 1mm。

这种方法还存在一些不足，比如器官之间的碰撞问题。由于发音器官是单独进行标注，当舌头向上程度过大的时候可能会出现穿透上腭的情况。Engwall 的团队曾对此做过相关研究，这也是我们下一步工作的思路。下一步的工作还包括对剩余的两个发音器官上腭和咽腔壁进行分析和重构。以及增加数据库，添加一些辅音的发音数据和更多的元音发音数据来扩展发音空间，以期达到更好的训练目的。

6 致谢

本工作得到国家自然科学基金面上项目（No. 61175016）、973 项目 (No. 2013CB329305)、基金重点项目 61233009 及国家科技支撑子课题（No.2011BAH16B01-02）的支持；得到中国社会科学院创新工程“面向语音教学的发音模型研究”资助。

参考文献

[1] Dang, J., Honda, K., 1998. Speech production of vowelsequences using a physiological articulatory model.In:Proceedings of ICSLP98, Vol. 5, pp. 1767 - 1770.

[2] DenisBeautemps, Pierre Badin, and Gerard Bailly, Linear degress of freedom in speech production: Analysis of cineradio- and labio-film data and articulatiry-acoustic modeling, 2001.

[3] OlovEngwall. Combining MRI,EMA and EPG measurements in a three-dimensional tongue model. speech communication,2003.

[4] Pierre Badin, Antoine Serrurier, Three-dimensional modeling of speech organs: Articulatory data and modelss. In IEICE Technical Report, vol. Vol. 106, No.177, SP2006-26, pp. 29-34. Kanazawa, Japan, The Institute of Electronics,Information, and Communication Engineers. 2006

[5] HironoriTakemoto, Tatsuya Kitamura,HironoriNishimoto and Kiyoshi Honda, A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions, Technical report,2004.

[6] AntoineSerrurier and Pierre Baddin,A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data, 2008.

NCMMSC2013: National Conference on Man-Machine Speech Communication 2013

Abstract:. Articulatory model has been widely used in speech research, like speech production, speech analysis and speech teaching. In the research, we will construct a Chinese model, which is a geometry model based on Chinese MRI data. Modeling process includes labelling and extracting control parameters. The method used to extracting is linear component analysis, the purpose is descriping and controlling the organ shaps by less parameters. The result shows, after analysis every organ can be well controlled by less then 3 parameters, with an average reconstructing error of less than 1mm.

Key words: Articulatory model; speech organs shapes; LCA;

（原载 NCMMSC2013 中国贵阳 2013 年 8 月）