汉语元音和日语元音的声道形态归一化

刘红[1]，魏建国[1*]，方强[2]，党建武[1,3]，路文焕[4]，马良[5]
(1. 天津大学 计算机科学与技术学院，天津 300072；)
(2. 中国社会科学院语言所，北京 100732；)
(3. 日本北陆先端科学技术大学院大学 信息科学学院，日本 923-1292；)
(4. 天津大学 软件学院，天津 300072；)
(5. 复旦大学 人文学院，上海 200433；)

文　摘：减小不同研究对象声道的形态差异，将有助于发音器官的数据分析和发音模型的建立。本文采用薄板样条（TPS）的方法：归一化处理二维声道的形态差异。本文的实验数据，由 Electromagnetic Midsagittal Articulographic (EMMA)系统分别采集三位中国人和三位日本人、发/a, i, u/元音时的数据。汉语元音声道形态和日语元音声道形态的两个归一化模板，是分别对两种语言、其中三个研究对象上颚和舌头的发音数据平均后得到的。然后，根据声道上定义的网格线系统，在模板上确定了 44 个参考点。实验结果表明，不同研究对象声道的形态差异，在水平方向和垂直方向都减小，TPS 方法不仅减少说话人之间的形态差异，而且保留了说话人固有的说话特性。最后，文章比较分析不同形态差异归一化的方法，比较结果说明，TPS 方法具有更好的归一化效果。

关键词：声道形态归一化；关节数据；TPS；

发音器官的数据没有声学数据应用广泛，其中的一个原因，是不同研究对象声道形态存在较大地差异，归一化处理比较困难。

器官发音时发生大量的形态变化，不能通过简单的刚性仿射变换完成多个研究对象声道形态的归一化处理。研究中已经存在的归一化方法，如 Bechman et al. [1] 采用拉直声道壁、对 MRI 数据的坐标归一化处理的方法；Hashi et al. [2] 提出的方法，是对 x-射线微束数据库中的元音发音姿态归一化处理。上面两种方法的共同特点是拉直上颚壁、归一化声道的长度，这种方法不能保持归一化之后上颚与舌表面之间的相对位置关系。同时，这种方法不能准确地反映对象之间的非线性关系，尤其是在声道局部发生高度形变的地方。

声道的形状通常反映局部和非线性的形态变化，或者称为弹性变形。我们以前的研究采用了薄板样条翘曲（TPS）[3][4] 的方法，归一化处理不同研究对象的电磁中矢面关节图（EMMA）数据。本文为了进一步评价 TPS 方法的性能，对汉语元音和日语元音的 EMMA 数据归一化处理，并对归一化的结果进行对比分析。两种语言声道形态的归一化模板，是两种数据库中各三个研究对象声道数据分别平均的结果。网格线系统被用来确定模板及各个研究对象的参考点。对比分析归一化之前和归一化之后的数据，表明 TPS 方法不仅减少说话人之间的形态差异，而且保留了说话人固有的说话特性。

1　声道的弹性形变

为了减少不同说话人声道的形态差异，[1，2] 矫直声道长度，然后对声道的长度归一化处理。这种方法主要考虑声道长度的差异，并在归一化之后保持了说话时的缩紧点不变。文献[5]提到，不同说话人声道的形态差异，不仅和声道长度有关，而且和说话时声道前后腔体的大小有关。另外，矫直声道的方法没有考虑声道发生弹性形变的非线性性质。同时，粘附在发音器官（舌头）上的不同传感器，归一化处理之后，它们之间的相对位置关系发生改变，导致发音器官丢失一些运动特性。

薄板样条是一类非刚性的样条映射函数，分为仿射变换和非仿射变换，能够对全局进行平滑处理，同时，根据物理特性，参考点从原点出发映射到目标点。

采用 TPS 方法归一化二维平面上的 n 个点，得到 2（n +3）个参数，其中包括 6 个仿射变换的参数和 2n 个参考点的对应系数。通过求解线性系统[6]得到这些参数。假设$(\hat{x}_i, \hat{y}_i) \in \Re^2$，i=1,2,...,n, n 个控制点在同一个平面内，控制点对应的函数值为$\hat{v}_i \in \Re$，i=1，2，...，

* 中国国家重点基础研究发展计划（973 项目：2013CB329305）资助；中国国家自然科学基金（61175016，重点项目 61233009）资助。魏建国，副研究员，Jianguo.fr@gmail.com，天津大学。

n，薄板样条的插值函数 $f(x,y)$ 定义如下：$f:\Re^2\to\Re$，

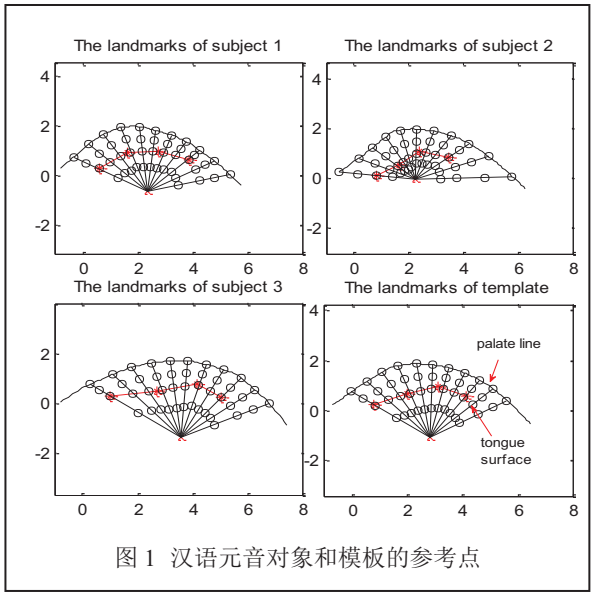

图 1 汉语元音对象和模板的参考点

$$f(x,y)=a_1+a_2x+a_3y+\sum_{i=1}^{n}w_ir_i^2\ln r_i^2 \tag{1}$$

其中 $r_i^2=(x-\hat{x}_i)^2+(y-\hat{y}_i)^2$.

等式(1)，是以点 $(\hat{x}_i,\hat{y}_i)$ 为中心、有限程度的形态变化。薄板样条加上权重值 w_i[6]，会发生不同程度的偏斜。薄板插值函数由两部分组成：仿射变换，函数中包含 a_1,a_2 和 a_3 的部分；非仿射变换--对应函数中的翘曲部分。函数 f 最小的弯曲能量 E_f 定义如下：

$$E_f=\iint_{\Re}\left(\left(\frac{\partial^2 f}{\partial x^2}\right)^2+\left(\frac{\partial^2 f}{\partial x\partial y}\right)^2+\left(\frac{\partial^2 f}{\partial y^2}\right)^2\right)dxdy \tag{2}$$

$$\sum_{i=1}^{n}w_i=0 \tag{3}$$

$$\sum_{i=1}^{n}\hat{x}_iw_i=0 \tag{4}$$

$$\sum_{i=1}^{n}\hat{y}_iw_i=0 \tag{5}$$

等式(3) (4) (5)是三个约束条件，等式(3)约束薄板的权重总和为 0，保持薄板增加权重之后保持固定不变。等式(4)和(5)，约束了 x 轴方向和 y 轴方向增加权重之后，总和分别为 0，保证薄板增加权重之后在 x 轴方向和 y 轴放向不会发生旋转。

TPS 中的参数向量 a 由 a_1,a_2 和 a_3 构成，向量 w 由 w_i 构成，用矩阵形式表示如下：

$$\begin{bmatrix}A & P\\ P^T & O\end{bmatrix}\begin{bmatrix}w\\ a\end{bmatrix}=\begin{bmatrix}v\\ 0\end{bmatrix} \tag{6}$$

其中，$A_{ij}=r_{ij}^2\ln r_{ij}^2$，i=1,2,...,n (n 个参考点)，j=1,2...,m（m 个需要归一化处理的原始点）；P 矩阵的第 i 行坐标值是 $(1,\hat{x}_i,\hat{y}_i)$，O 矩阵是 3*3 的零矩阵。向量 w, a 和 v 分别由 w_i；a_1,a_2,a_3 和 v_i 组成。等式左边是一个大小为(n+3)*(n+3)的矩阵，记为 K。

TPS 方法的基本原理，是将一个研究对象的 EMMA 参考点数据 $(\hat{x}_i,\hat{y}_i)$，映射到模板的参考点 $(\hat{x}_i',\hat{y}_i')$，然后求解插值函数的系数向量 w 和 a，最后，再次应用插值函数，将所有研究对象的原始数据归一化处理。TPS 的原理描述如下：

$$\begin{bmatrix}w_x & w_y\\ a_x & a_y\end{bmatrix}=K^{-1}\begin{bmatrix}\hat{x}' & \hat{y}'\\ 0 & 0\end{bmatrix} \tag{7}$$

其中，$\hat{x}'$ 和 $\hat{y}'$ 分别由 $\hat{x}_i'$ and $\hat{y}_i'$ 构成。此外，w_x 和 a_x 是 x 轴方向的参数向量，w_y 和 a_y 是 y 轴方向的参数向量。经过下面的等式求解得到点 (x_j,y_j) 归一化的坐标 (x_j',y_j')：

$$\begin{bmatrix}x' & y'\end{bmatrix}=\begin{bmatrix}B & Q\end{bmatrix}\begin{bmatrix}w_x & w_y\\ a_x & a_y\end{bmatrix} \tag{8}$$

其中，$B_{ji}=((x_j-\hat{x}_i)^2+(y_j-\hat{y}_i)^2)\ln((x_j-\hat{x}_i)^2+(y_j-\hat{y}_i)^2)$，i=1,2,...n,j=1,2,...m。Q 矩阵的第 j 行坐标值是 $(1,x_j,y_j)$，第 j 行的向量 x' 和 y' 分别由 x 与 y 插值后的坐标 x_j' 和 y_j' 构成。

2 确定参考点

EMMA 系统采集的发音器官数据，由于空间分辨率低，很难找到声道上对应明显地形态学的位置。为了解决这一问题，本文根据网格线系统确定声道上的标注点[7]。标注点被用作插值函数映射的参考点。

归一化的两个声道模板，是分别平均 EMMA 汉语元音和日语元音的各 3 个研究对象的声道数据，得到的平均形状。首先，在模板中定义一组标注点，然后用同样的方法确定每个数据库中各个研究对象发音器官的标注点。然后，计算沿着舌面位置的三个元音舌头数据的平均值（从舌尖到舌后部，采集数据时粘附 4 个传感器，T1-舌尖，T2-舌面，T3-舌背，T4-舌后）。归一化处理中，为了更好的将原点映射到模板，这里定义一个网格线系统。网格线覆盖整个舌头的运动，首先根据舌面上 4 个传感器的位置确定一个质心，连接质心与上牙、质心与上颚线的右端点，形成一个扇区，然后将扇区划分为角度相等的 10 个小扇区，网格线系统包括上颚线、上颚与舌上表面中点的连线、舌头上表面和舌头下表面四条线。因此，得到网格线系统的 44 个插值点，作为归一化模板和各个研究对象发音器官的参考点。

3 实验

实验的 EMMA 数据，由汉语元音和日语元音两部分组成。其中，汉语数据由三位中国人录制，包含/a/i/u/三个元音；日语数据由三位日本人录制，包含/a/i/u/e/o/五个元音，但是为了和汉语

元音数据作对比，实验中只用到/a/i/u/三个日语元音。本文一共进行 2 次归一化处理，分别是归一化汉语元音数据、归一化日语元音数据，最后，

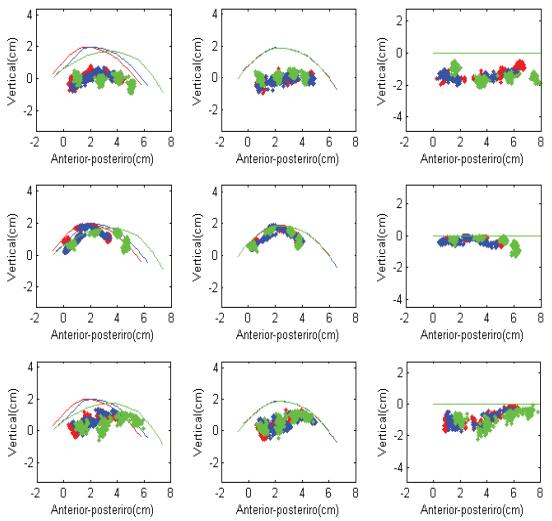

图 2 汉语元音 EMMA 数据，三列分别对应原始数据、TPS 方法归一化的数据、拉直声道壁方法归一化的数据；三行分别对应元音/a//i//u/

对比分析汉语元音和日语元音归一化处理的结果。

汉语的 EMMA 数据，是从不同的发音组合中分别提取/a/i/u/元音各 100 个发音构成。归一化的模板，如图 1 所示。

图 2，三列分别表示三个研究对象的原始数据、TPS 方法归一化的数据以及矫直声道壁方法归一化的数据，三行分别对应元音/a/，/i/，/u/。三种颜色分别代表三个研究对象。

日语的 EMMA 数据，是从不同的上下文中分别提取/a/i/u/元音各 64 个发音构成。采用同样的方法，得到归一化模板和标注点。原始数据和归一化的数据如图 3 所示。

原始数据与 TPS 方法归一化的数据比较，比较结果表明不同研究对象的形态差异均减小。

4 评价

本文以矫直声道归一化的方法[1]作为基准，通过对比分析进一步评价 TPS 方法。图 2 和图 3 的第三列，分别是汉语元音数据和日语元音数据采用矫直上颚方法归一化的结果。实验结果与 TPS 方法的结果比较，比较结果是 TPS 方法的方差分布差异更小，表明 TPS 方法得到了更好的归一化结果。图 4，是不同研究对象汉语元音数据的标准差（SD），图中每个元音的三个柱形图分别代表原始数据的标准差，矫直上颚归一化数据的标准差和采用 TPS 方法归一化数据的标准差。

TPS 方法归一化之后，不同研究对象三个汉语元音/a/i/u/的标准差，x 轴方向减小 3.4mm、y 轴方向减小 0.6mm。此外，从图 4 可以看出，矫直上

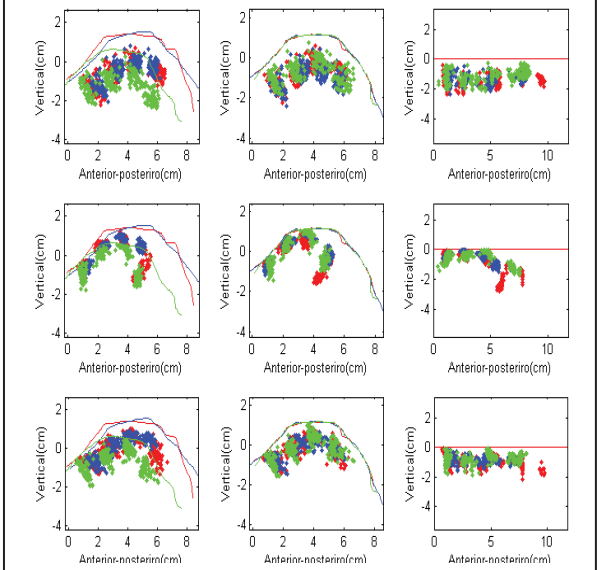

图 3 日语元音的 EMMA 数据。每幅图的含义与图 2 相同

颚壁归一化的方法，使得 X 轴方向的标准差大于原始数据的标准差。原因是矫正上颚壁归一化的方法，令 x 轴数据与上颚之间的相对距离变大。

日语元音采用 TPS 方法归一化的结果，x 轴和 y 轴方向的标准差分别减小 0.2mm、1.5mm。这里省略了日语元音数据的标准差结果图。综合汉语元音和日语元音归一化的结果，表明 TPS 方法的性能优于矫直上颚壁归一化的方法。

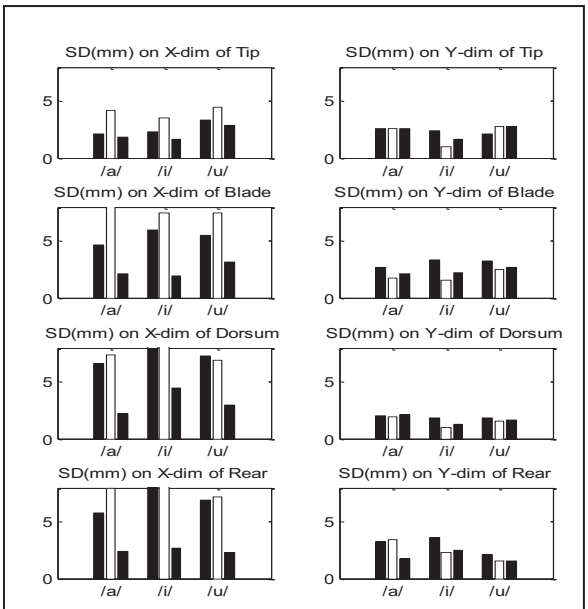

图 4 汉语元音 EMMA 数据，原始数据、TPS 方法归一化数据、拉直声道壁归一化数据的标准差结果

5 小结

本文采用TPS方法对不同研究对象的声道形态归一化处理。归一化之后不同研究对象发音器官的形态差异均减小。汉语元音（x 轴方向减小大约 3.4mm，y 轴方向减小 0.6mm）和日语元音（x 轴方向减小约 0.2mm，y 轴方向减小 1.5mm）的平均标准差均减小。和矫直上颚壁归一化的方法比较，表明 TPS 方法具有更好的归一化效果。虽然本文汉语元音和日语元音的数据对象均只有三个，但是通过归一化处理，两种语言数据分别验证了 TPS 方法的有效性。因此，数据量对实验结果不会产生影响。

6 致谢

这项工作，一部分由中国国家重点基础研究发展计划（973 项目：2013CB329305）资助，一部分由中国国家自然科学基金（61175016，重点项目第 61233009 号）资助。此外，还要感谢微软亚洲研究院（FY11 RESOPP-008）的支持。

参 考 文 献

[1] Beckman, M., Jung, T., Lee, S., Jong, K., Krishnamurthy, A., Ahalt, S., Cohen, K. and Collins, M., "Variability in the production of quantal vowels revisited", J. Acoust. Soc. Am., vol. 97, pp. 471-490, 1995.

[2] Hashi, M., Westbury, J. and Honda, K., "Vowel posture normalization", JASA, vol. 104, pp. 2426–2437, 1998.

[3] FL, B., "Principal warps: Thin plate splines and the decomposition of deformations", IEEE Trans Pattern Anal. Mach. Intell, vol. 11, pp. 567-85, 1989.

[4] Jianguo Wei and Jianwu Dang, "Morphological normalization of vocal tract shape", IEEE Acoustics Speech and Signal Processing, pp. 4186-4189. 2010

[5] Yang, C.-S. and Kasuya, H., “Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child”, IEICE Trans. On Inf. & Syst., Vol.E78-D, No.6, pp.732-737, 1995

[6] Zagorchev, L. and Goshtasby, A., "A comparative study of transformation functions for nonrigid image registration", IEEE Trans. Image Processing, vol. 15, pp. 529-538, 2006.

[7] Beautemps, D., Badin, P., and Laboissière, R. Deriving vocal-tract area function from midsagittal profiles and formants frequencies: A new model for vowels and fricative consosnants based on experimental data. Speech Communication, 16, 27-47.ference on Computer Vision and Pattern Recognition 1995

[8] Yuguang Wang, Jianwu Dang, Xi Chen, Jianguo Wei, Kiyoshi Honda and Hongcui Wang, " An MRI-based Acoustic Study of Mandarin Vowels." Interspeech, August 2013

Morphological Normalization of Vowels for Mandarin and Japanese

Hong Liu*, JianguoWei*, Qiang Fang, Jianwu Dang* , Wenhuan Lu and Liang Ma
* School of Computer Science and Technology, Tianjin University,
#School of Computer Software, Tianjin University,
92 Weijin Road, Nankai District, Tianjin 300072, China
†Chinese Academy of Social Sciences, BeiJing,
5, Jianguomennei Dajie, Beijing 100732, China
School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
Department of Chinese Language and Literature, Fudan University,
220 Handan Road, Shanghai 200433, China
E-mail: wlgc0802@sina.cn; Jianguo.fr@gmail.com

Abstract: Reducing the morphological variances of vocal tract across different subjects would benefit articulatory data analysis and modeling. To further test such a hypothesis by a thin-plate spline warping (TPS) method, this study used articulatory data of /a, i, u/ from 3 Chinese subjects and 3 Japanese subjects, which were collected by Electromagnetic Midsagittal Articulographic (EMMA) system. The templates for the normalization of Chinese and Japanese were obtained by averaging the 3 subjects’ palates and tongue shapes in each group. The 44 landmarks in each template were then defined by a gridline system of the vocal tract. The results show that the variances among subjects were reduced in both horizontal direction and vertical direction. The similar vowel

structures between pre- and post-normalization data indicate that TPS method outperforms the traditional palate-straighten method in that TPS method has reduced mid-sagittal morphological differences among speakers while keeping their individual vowel structures unchanged. The comparison results show that the articulatory differences among the three vowels are consistent with their corresponding acoustic properties.

Key words: vocal tract normalization, articulatory data, thin-plate spline

第十二届全国人机语音通讯
学术会议 中国贵阳 2013
（原载 NCMMSC 2013 中国贵阳 2013年8月）