J. Vis. Commun. Image R. 41 (2016) 352–360

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

# Morphological normalization of vowel images for articulatory speech recognition ☆

Jianguo Wei [a,b], Jingshu Zhang [b], Yan Ji [b], Qiang Fang [c], Wenhuan Lu [a,*]

[a] School of Computer Software, Tianjin University, 135 Yaguan Road, Jin Nan District, Tianjin 300350, China
[b] Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, 135 Yaguan Road, Jin Nan District, Tianjin 300350, China
[c] Chinese Academy of Social Sciences, Beijing, China

ARTICLE INFO

*Article history:*
Received 15 March 2016
Revised 30 June 2016
Accepted 12 October 2016
Available online 17 October 2016

*Keywords:*
Vocal tract normalization
Articulatory data
Acoustic data
Thin-Plate Spline
DNN
Articulatory recognition

ABSTRACT

Minimizing morphological variances of the vocal tract across speakers is a challenge for articulatory analysis and modeling. In order to reduce morphological differences in speech organs among speakers and retain speakers' speech dynamics, our study proposes a method of normalizing the vocal-tract shapes of Mandarin and Japanese speakers by using a Thin-Plate Spline (TPS) method. We apply the properties of TPS in a two-dimensional space in order to normalize vocal-tract shapes. Furthermore, we also use DNN (Deep Neural Networks) based speech recognition for our evaluations. We obtained our template for normalization by measuring three speakers' palates and tongue shapes. Our results show a reduction in variances among subjects. The similar vowel structure of pre/post-normalization data indicates that our framework retains speaker specific characteristics. Our results for the articulatory recognition of isolated phonemes show an improvement of 25%. Moreover, our phone error rate of continuous speech reduced by 5.84%.



## 1. Introduction

In recent years, speech recognition technology has advanced significantly. Speaker adaptive and system robustness factors remain vital to speech recognition systems. Interestingly, much articulatory data used for speech research is also used for acoustic data [1]. However, articulatory data are not widely applied. One reason is that acquiring such data is difficult. Another reason is that variances in vocal tracts prove difficult for usage in multi-subject articulatory data research [2]. Hence, articulatory data are not as popular as acoustic data in spite of its importance in the speech research field. In order to discover the kinematic properties that characterize speaker differences, it is necessary to normalize inter-subject articulatory data so that morphological variances among different speakers are reduced.

As such, it is important to understand that there are differences in vocal tracts among subjects, and that large nonlinear deformations can occur on vocal tracts. Therefore, it is difficult to study vocal tract shape by affine transformation of simple rigid objects. Up to now, researchers have proposed many normalization techniques for articulatory space and acoustic space. For instance, Bechman et al. [3] proposed straightening the walls of vocal tracts in order to transform the coordinates of x-rays into micro beam data. Hashi et al. [4] also proposed a method of normalizing vowel postures for an X-ray micro beam database. The two methods both straighten vocal tract walls in order to normalize vocal tract length; however, this can cause the relative relationship between the palate and tongue surface to change significantly after transformation. Pitz et al., in a study concerning acoustic space, processed the length of vocal tracts by using linear transformation in a frequency domain [5]. Additionally, Saheer et al. normalized the length of the vocal tract by using a linear transformation method [6]. Among these studies, it is evident that they all attempt to normalize vocal length tract length (in either articulatory or acoustic space) without considering the articulatory features of vocal tract shapes.

Because the vocal tract shape usually reflects local and nonlinear deformations, it can be treated as a kind of non-rigid shape deformation. Based on this idea, our study proposes a framework of normalizing speakers' EMA (Electromagnetic Midsagittal Articulographic) data by using a TPS (Thin-Plate Spline warping) method [7] (a non-linear transformation method applied in shape

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
*E-mail addresses:* jianguo@tju.edu.cn (J. Wei), jingshu@tju.edu.cn (J. Zhang), tjujiyan@tju.edu.cn (Y. Ji), fangqiang@cass.org.cn (Q. Fang), wenhuan@tju.edu.cn (W. Lu).

http://dx.doi.org/10.1016/j.jvcir.2016.10.005
1047-3203/

matching and image alignment). Our TPS-based method can realize point-based normalization among subjects' EMA data, in which the fitting shapes of subjects' vocal tracts serve as templates. Furthermore, we use a gridline system to define the landmarks of the vocal tracts. Our proposed framework is able to maintain the vocal tract's relationship between the palate and tongue position, which retains the kinematic properties of the vocal tract throughout the normalization procedure.

Reducing the morphological differences of speakers' vocal tracts helps analyze the properties of speech organs and the rules of pronunciation. Moreover, it enhances the robustness and performance of speech recognition systems. Hence, we propose a new method that combines a vocal tract gridline system with a TPS-based method. Based on this method, morphological variations are reduced in a two-dimensional space while retaining key individual differences. As such, we can study the single kinematical behaviors of a group of speakers, and analyze the variance of kinematical behaviors between different groups of speakers. EMA database was from NTT Communication Science Laboratories [8] for our study.

In this study, we evaluate the overall performance of a TPS method both in an articulatory space and in an acoustic space. We utilized a physiological articulatory model [9] to generate speech (according to normalized articulatory postures). In addition, we compare the vowels of normalized Mandarin and normalized Japanese in an articulatory space. We also study whether there is consistency among normalized vowels in articulatory space and acoustic space when a TPS method is used. And the performance of speech recognition is evaluated in DNN based articulatory recognition system [10–12].

The remainder of this study is organized as follows. In Sections 2 and 3, we introduce the framework of our proposed method. In Section 4, we discuss the experiments we conducted. In Section 5, we evaluate the results we obtained. Finally, we provide a conclusion in Section 6.

## 2. Elastic deformations of vocal tracts and TPS transformations

The movements of the tongue and jaw cause deformations in the shape of the vocal tract. The result is that the shape of vocal tracts cannot be measured properly among speakers. In order to reduce the morphological variations of speakers, some methods [3] straighten the length of the vocal tracts in order to normalize the vocal tract shape. These methods are able to maintain the constriction constant of the vocal tracts. According to the results in Yang et al. [13], the variability of subjects is not only related to vocal tract length, but also to the volumes of the back and front cavities of the vocal tract. Moreover, such methods of straightening vocal tracts fail to take into consideration the nonlinear elastic properties of vocal tracts' deformations. In addition, the relative position information of the four sensors attached to the tongue surface of speakers was lost after the normalization process. These factors may affect the pronunciation kinematics characteristics of the individuals. In the image processing field, a great deal of non-rigid normalization methods has been proposed. Among these approaches, the Thin-Plate Splines method possesses properties suitable for our research [7]. The Thin-Plate Splines interpolation function is a kind of non-rigid mapping function, which is often used in non-rigid shape deformation. In our study, we use this method to smooth the palate and tongue surface. As a condition of the Thin-Plate Splines interpolation function, it is necessary to find the feature points of the vocal tract defined by four sensors attached to the tongue surface and the palate. According to the physical property of the Thin-Plate Splines method, the definition of a reference point is a map from the origin data to the target data.

In Thin-Plate Splines transformations, the basic principle is to create a map from the speakers' articulatory data to the template data. Specifically, it is necessary to have a map between the vocal tract landmarks of the subjects and the landmarks of the template. The specific principle of TPS transformations is as follows. Given a set of n corresponding 2D points, a TPS warp is described by 2(n + 3) parameters, which include 6 global affine motion parameters and 2n coefficients for correspondences of the control points. These parameters are calculated by solving a linear system [14]. We propose that $(\hat{x}_i, \hat{y}_i) \in R^2$, $i = 1,\ldots n$, are the n control points in a 2D planar, and their corresponding function values are $v_i \in R^2$, $i = 1,2,\ldots,n$. Then the Thin-Plate Spline interpolation $f(x,y)$ is defined by a map: $f : R^2 \rightarrow R$. The formula of the Thin-Plate Splines function is shown as follows.

$$f(x,y) = a_1 + a_2 x + a_3 y + \sum_{i=1}^{n} w_i r_i^2 \ln r_i^2 \tag{1}$$

where

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

Eq. (1) expresses that the equation is based on $(\hat{x}_i \hat{y}_i)$ as the center. Moreover, it is a kind of infinite extent deformation under loads. The plate deflects under the imposition of loads to take values $w_i$ [14]. The interpolation Spline function consists of two parts: an affine transformation specified by the first three elements ($a_1$, $a_2$ and $a_3$), and the last warping part. $r_i$ is the distance from the target points to the original points. The goal is to ensure that linear parallelism does not cause changes in the shape of an object. There is also the nonlinear portion of the formula to consider. In our study, according to the theory of the TPS method we must first map the relationship between the template and the speakers. As such, we need to solve the 2(n + 3) parameters in Eq. (1). The Thin-Plate Splines method is an interpolation spline function, which attempts to smooth a surface that can pass through all of the control points with a minimum bending degree, and ensures that the data points can be correctly matched. Hence, we added an energy constraint and three weight coefficient constraints. The minimum bending defined by an energy function is shown below:

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

$E_f$ is the energy constraint for $f(x,y)$. It ensures the minimum second derivative of accumulation of each data point on the surface. Moreover, the three weight constraints are defined by:

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

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

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

Eqs. (3), (4), and (5) are the constraints of the weight coefficients. The equation to the right value controls the degree of the smoothed surface in the three equations. Commonly, the value is zero. Eq. (3) shows that the sum of the weights should be zero for the smoothed palates. Eqs. (4) and (5) restrict the motion of the palate so that it does not rotate after adding weight to the x-axis and the y-axis, respectively.