Temporal Resolution Enhancement of Vocal Tract MRI Sequences Based on Image Registration

Ana L. D. Martins; Nelson D. A. Mascarenhas; Cláudio A. T. Suazo
Universidade Federal de São Carlos, Programa de Pós-Graduação em Biotecnologia
Via Washington Luís, Km 235, CP 676, CEP: 13.565-905, São Carlos, SP, Brazil
{ana_martins;nelson}@dc.ufscar.br; claudio@power.ufscar.br

Abstract—Dynamic magnetic resonance imaging (MRI) is an emerging technique for studying speech production. In general, vocal tract image sequences are acquired during the speech of a word or phoneme. Sequences allow the identification of shapes taken by the vocal tract during speech production. However, there is no prior knowledge about the spatial and temporal resolution requirements, which are expected to vary for different speech tasks. Available approaches try to enhance the resolution of the images by empowering the acquisition devices, which can be very expensive. In this paper, we propose an alternative approach to enhance temporal resolution based solely on the observed image sequences. We use a previous non-rigid image registration method, which provides an intuitive background for temporal resolution enhancement. Based on a motion compensated interpolation (MCI) approach, intermediate images are coherent with the movement present in the whole sequence. Results indicate the effectiveness of our approach.

Keywords-magnetic resonance imaging; speech production; vocal tract image sequences; motion compensated interpolation; temporal resolution enhancement.

I. INTRODUCTION

Detailed knowledge about speech production is of great interest to several research areas (engineering and linguistics, just to name a few). This knowledge allows refined models for speech signal that can be exploited for the design of speech recognition, coding, and synthesis systems. On the other hand, research may be conducted to explore open questions in phonology and phonetics (for instance, variability of speech and language disorders) [1]. For that purpose, knowledge about vocal tract shape and dimensions acquired during the speech of words or phonemes are essential for a full understanding of articulatory and acoustical processes involved in speech production. According to Baer et al. [2], magnetic resonance imaging (MRI) is the only tool that can provide detailed tri-dimensional (3D) data of the entire vocal tract and tongue without any known harmful effects on the subject. However, the temporal and spatial resolution requirements are expected to vary depending on the speech task and there is no information about it in advance. Moreover, according to Narayanan et al. [3], even though MRI advances represent a significant improvement in the quality of information about changes in speech articulators over time, they are still not close to the temporal resolution necessary for capturing the dynamic characteristics of tongue movement.

Available approaches try to enhance the resolution of image sequences by empowering the acquisition devices [1], which can be very expensive. Therefore, it is of great interest to enhance temporal resolution of existing image sequences using only digital image processing techniques. The task of image registration is to find an optimal geometrical transformation between corresponding image data [4]. This transformation may be used for the quantification of changes between images. In this case the primary goal is not only the transformation which maps points in one image into their corresponding counterparts in the second image, but also the motion and deformation characteristics exhibited by this transformation. Thus, we believe that a meaningful transformation estimated by the image registration method could be used for temporal resolution enhancement.

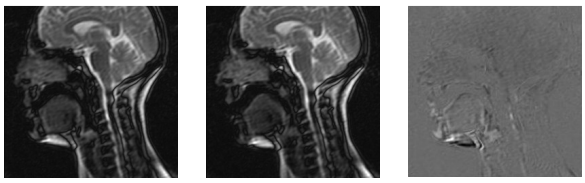
In this context, we propose an approach for temporal resolution enhancement of human vocal tract image sequences using a previous non-rigid image registration method [5]. This method describes the transformation between each pair of images by a free-form deformation (FFD) with B-spline interpolation between uniformly spaced control points. The coordinates of the control points are the only parameters of the transformation. Indeed, we demonstrate that the meshes of control points are a powerful tool for temporal resolution enhancement. Note that, according to the correspondence between two meshes, intermediate images in a sequence can be generated simply by positioning control points in meaningful positions. We compare the cubic spline interpolation considering four neighboring meshes of control points and the linear interpolation in the motion direction considering two adjacent meshes. In both cases, intermediate images are coherent with the movement present in the sequence.

Intermediate images were generated for a visual evaluation of the proposed approach, considering both the cubic spline and linear interpolation methods. In a first moment, one can imagine that the cubic spline interpolation considering four neighboring meshes would perform better because it considers more information in the generation of an intermediate image.

For comparison purposes, the Normalized Mean Square Error (NMSE) criterion was used for the numerical evaluation of both interpolation methods in a simulated situation. The results indicate the equivalence of these methods and the effectiveness of our approach.

II. NON-RIGID IMAGE REGISTRATION

Human vocal tract images used in speech production research present deformations similar to the one in illustrated in Fig. 1. The error image presented in Fig. 1(c) emphasizes the locality of the deformation around the mouth. Therefore, in this context, a non-rigid transformation is necessary to relate a pair of images.



(a) (b) (c)

Figure 1. Two moments of a speech and the error image.

A. FFD based on B-spline Interpolation

Rueckert et al. [5] presented a non-rigid image registration method using FFD based on cubic B-splines. FFD is an approach used in computer graphics applications to model 3D deformable objects. The object is deformed by manipulating a mesh of control points.

Following [5] deformations are modeled by a transformation φ which combines global scene movement with local deformations

$$\varphi(x,y)=\varphi_{\text{global}}(x,y)+\varphi_{\text{local}}(x,y) \tag{1}$$

φ_{global} models an affine transformation applied to the whole image

$$\varphi_{\text{global}}(x,y)=\begin{bmatrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}+\begin{bmatrix}\theta_{13}\\ \theta_{23}\end{bmatrix} \tag{2}$$

where θ_{ij} are the coefficients of this transformation. Local deformations are modeled by FFD based on cubic B-spline basis functions. Considering

$$\Omega=\{(x,y)\,|\,0\le x\le X, 0\le y\le Y\}$$

the image support and Φ the $nx\times ny$ mesh of control points with uniform spacing δ, φ_{local} is given by

$$\varphi_{\text{local}}(x,y)=\sum_{l=0}^{3}\sum_{m=0}^{3}B_l(u)B_m(v)\phi_{i+l,j+m} \tag{3}$$

where $i=\lfloor x/nx\rfloor$, $j=\lfloor y/ny\rfloor$, $u=x/nx-\lfloor x/nx\rfloor$, $v=y/ny-\lfloor y/ny\rfloor$ and B_l is the l-th B-spline basis function

$$\begin{aligned}B_0(u)&=(1-u)^3/6\\ B_1(u)&=(3u^3-6u^2+4)/6\\ B_2(u)&=(-3u^3+3u^2+3u+1)/6\\ B_3(u)&=u^3/6.\end{aligned} \tag{2}$$

Note that the control points $\phi_{i,j}$ are the only parameters of the local transformation. Their identification is automatic and does not depend on expert knowledge or image characteristics. According to Rohr [6], although there are sophisticated ideas for automatically identification of landmarks based on image characteristics, the process is complicated and still not fully automatic. Moreover, cubic B-spline basis functions have finite support. Therefore, each control point affects the transformation only in a local neighborhood.

The mesh resolution determines the kind of non-rigid deformation that can be modeled. Higher resolutions allow deformations localized in small parts of the image. On the other hand, finer grids allow only more global deformations. Lee et al. [7] proposed a multilevel free-form deformation (MFFD) for image metamorphosis. In this approach, FFD based on 2D B-spline approximation is applied to a hierarchy of control lattices to exactly satisfy the feature constraints. Moreover, this approach achieves the best compromise between degree of non-rigid deformation and computational cost. Consider that $\Phi_0,\ldots,\Phi_L$ is a hierarchy of control lattices in which the spacing between control points decreases from Φ_l to Φ_{l+1}. This hierarchy is used to derive a sequence of deformation functions with the FFD manipulation. Each mesh Φ_l and the associated FFD defines a local transformation φ^l_{local} and their sum defines φ_{local}

$$\varphi_{\text{local}}(x,y)=\sum_{l=0}^{L}\varphi^l_{\text{local}}(x,y). \tag{4}$$

The transformation is regularized by imposing the following smoothness constraint

$$C_{\text{smooth}}(\varphi)=\int_0^X\int_0^Y\left(\frac{\partial^2\varphi}{\partial x^2}\right)^2+\left(\frac{\partial^2\varphi}{\partial y^2}\right)^2+2\left(\frac{\partial^2\varphi}{\partial xy}\right)^2 \mathrm{d}x\mathrm{d}y \tag{5}$$

to the spline-based FFD transformation. Moreover, the normalized mutual information (NMI) is used as a similarity criterion to measure the degree of alignment between images. Considering a reference image R and an image to be compared with the reference one T,

$$C_{\text{similarity}}(R,T_\varphi)=\frac{H(R)+H(T_\varphi)}{H(R,T_\varphi)}, \tag{6}$$

is the similarity criterion, where $H(R)$ and $H(T_\varphi)$ denote the marginal entropies of R and the transformed image T_φ and $H(R,T_\varphi)$ denotes the joint histogram of R and T_φ.

The optimal transformation is found by minimizing the cost function

$$C(\Theta,\Phi) = -C_{\text{similarity}}(R,T_\varphi) + \lambda C_{\text{smooth}}(\varphi), \qquad (7)$$

where λ is a weighting parameter that models the compromise between alignment of the two images and smoothness of the transformation. Considering only affine transformations, $C_{\text{smooth}} = 0$. Therefore, in a first step the cost function is optimized for the parameters of the global transformation Θ. One can adopt any registration method that considers affine transformations. The optimization of the local transformation cost is performed by an iterative gradient descent technique which steps in the direction of the gradient vector with a certain step size μ. The algorithm stops if a local optimum of the cost function has been found ($\|\nabla C\| \le \varepsilon$, for a small positive ε). In this way, the method starts with a mesh of uniformly spaced control points that corresponds to the identity transformation in which no deformation is modeled. Control points coordinates are iteratively updated according to the minimization of the cost function $C(\Theta,\Phi)$.

III. Proposed Method

Considering a sequence of vocal tract MR images $I_0,\ldots,I_k$, acquired with a MRI system operating at 1.5 Tesla (quantum gradients; 30mT/m amplitude; 0.24ms rise time; 125T/m/s Slew rate; 50cm FOV), the registration algorithm was applied to each pair of images always selecting the first image as the reference one (Fig. 2). It is important to note that we adopted this procedure because speech articulators movement is restricted to a small area. In another context it would be more interesting to register pairs of consecutive images. Moreover, the images of the meshes of control points are merely illustrative. They are not necessary to the registration method or to the proposed approach.

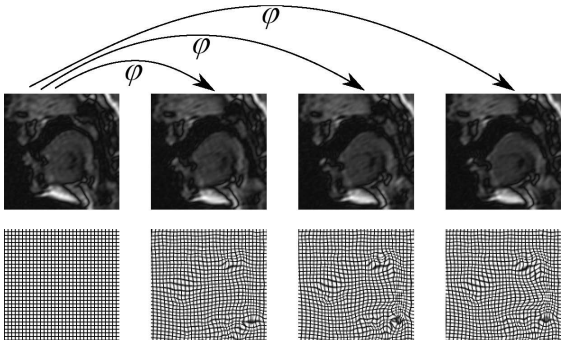


Figure 2. Illustration of the registration of pairs of images always considering the first image as the reference one, and the corresponding meshes of control points.

Speech articulators movement during speech production is a continuous and smooth process. Therefore, intermediate images generated by the temporal resolution enhancement method must respect the movement present in the whole observed sequence. Considering the correspondence between each control point in the sequence of identified meshes, intermediate meshes can be generated simply by positioning control points in intermediate positions. The most intuitive approach is the linear interpolation of the corresponding control points coordinates in adjacent meshes. However, since this approach considers only the motion between adjacent images, we also used the cubic spline interpolation of the coordinates of four corresponding control points in surrounding meshes as illustrated in Fig. 3. In this way, each generated image considers the movement present in practically the whole observed sequence.

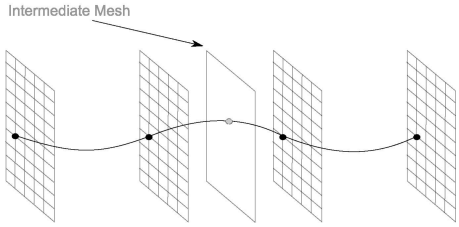


Figure 3. Intermediate mesh generated considering four neighboring meshes.

In both cases, according to the identified intermediate mesh of control points, temporal resolution enhancement is performed first by applying φ_{local} to both images that are next to the new image. After, a weighted sum of the transformed images is performed as illustrated in Fig. 4. The weights ω_1 and ω_2 are defined according to the distance between the new image and each of the neighboring images ($\omega_1 + \omega_2$ is always equal to 1). The transformed image related to the closest image receives a higher weight. Considering the observed images and corresponding meshes of control points, this approach gives an image that is always coherent to the data.

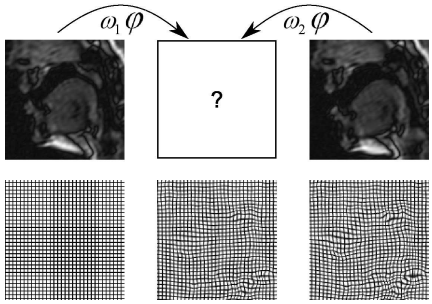


Figure 4. Temporal resolution enhancement by performing a weighted sum of transformed neighboring images.

IV. Results

We generated intermediate images in a sequence of twenty five 256x256 vocal tract MR images for a visual evaluation of the proposed method. The hierarchical non-rigid registration approach resulted in 67x67 meshes of control points. Fig. 5 presents the detail of two images registered by the FFD based approach. The error between the image in Fig. 5(e), which is the image in Fig. 5(a) transformed according to the mesh of control points in Fig. 5(d), and the image in Fig. 5(c) indicates that this non-rigid registration approach accurately detected speech articulators movement in the sequence.

Fig. 6(a) and 6(d) present the first two images in a sequence of MR vocal tract images. Fig. 6(b) and 6(c) are the images generated in between these first two

images using the cubic spline and linear interpolation methods, respectively. Note that there are no artifacts and the images are coherent with speech articulators movement. Since the cubic spline interpolation method considers more information in the generation of each intermediate image, at a first look, it seems to give better results. In order to analyze this, the NMSE was used for the numerical evaluation of these interpolation methods in a simulated situation. In the first experiment, considering a sequence of twenty five images, each of these images was removed one at a time, and each method was used to generate an interpolated version of the removed image. The second experiment was similar, but each pair of consecutive images was removed one at a time. In the third and fourth experiments images were removed to simulate a sub-sampled sequence in time by the factors 2 and 3, respectively.

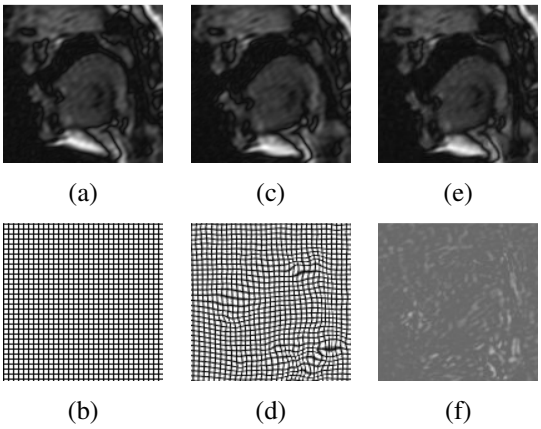


Figure 5. (a) Detail of the reference image and (b) the corresponding mesh of control points. (c) Second observed image and (d) the corresponding mesh of control points. (e) Image (a) transformed according to the mesh (d). (f) Error between (c) and (e).

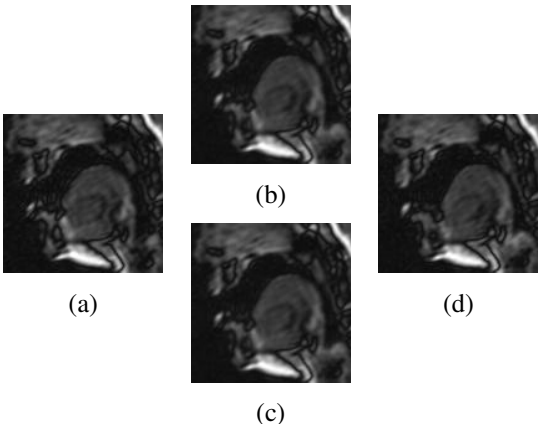


Figure 6. Detail of the intermediate images (b) and (c) using the cubic spline and linear interpolation methods, respectively, in between the first two images of a sequence (a) and (d).

Based on the Wilcoxon matched pairs test, which compares two matched groups without assuming that the distribution of the before-after differences follows a Gaussian distribution, the experiments did not demonstrate statistically significant differences between the NMSE means (with 95 percent confidence, p-values were always higher than 0.5). Based on this statistical evidence, we can conclude that these two interpolation methods give equivalent results in the context of the MR vocal tract images.

V. CONCLUDING REMARKS

We presented an MCI method for temporal resolution enhancement of vocal tract MR image sequences used in speech production research. Our method is based on the a previous non-rigid image registration approach. Indeed, results demonstrate the effectiveness of our approach. Moreover, intermediate frames generated by the proposed method present a coherent movement according to an observed sequence. The adopted non-rigid image registration presents a higher computational cost when compared with block-matching algorithms for motion estimation. However, it is able to precisely detect speech articulators movement and the meshes of control points provide an intuitive background for temporal resolution enhancement. Note that this non-rigid registration method and the proposed MCI approach could be adaptive to the movement among the observed images. In fact, movement among images is localized in parts related to speech articulators. Therefore, we believe that a segmentation method used to identify these parts in a pre-processing step could be used to reduce the computational cost of the registration method.

ACKNOWLEDGMENT

We would like to thank professors Antonio Teixeira and Augusto Silva from Instituto de Engenharia Eletrônica e Telemática de Aveiro (IEETA) of Universidade de Aveiro, Portugal, for the vocal tract images used in this work. These images are part of the HERON Project - A Framework for Portuguese Articulatory Synthesis Research, POSI/PLP/57680/2004. Ana L. D. Martins is supported by FAPESP, Brazil, under grant number 2008/01348-2.

REFERENCES

[1] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan. Seeing speech: Capturing vocal tract shaping using realtime magnetic resonance imaging [Exploratory DSP]. Signal Processing Magazine, IEEE, 25(3):123–132, May 2008.

[2] T. Baer, J. C. Gore, L. C. Gracco, and P.W. Nye. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. J. Acoust. Soc. Am., 90(2):799–828, Aug 1991.

[3] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd. An approach to real-time magnetic resonance imaging for speech production. The Journal of the Acoustical Society of America, 115(4):1771–1776, 2004.

[4] J. Modersitzki. Numerical Methods for Image Registration. Oxford University Press, 2004.

[5] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images. IEEE Transactions on Medical Imaging, 18(8):712– 721, 1999.

[6] K. Rohr. Landmark-Based Image Analysis: Using Geometric and Intensity Models. Kluwer Academic Publishers, Norwell, MA, USA, 2001.

[7] S. Lee, G. Wolberg, K.-Y. Chwa, and S. Y. Shin. Image Metamorphosis with Scattered Feature Constraints. IEEE Transactions on Visualization and Computer Graphics, 2:337–354, 1996.