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 depending on the speech task. Available approaches try to enhance the resolution of the images by empowering the acquisition devices, which can be expensive. In this paper, we propose an alternative approach to enhance temporal resolution based solely on the observed image sequences. Our continuous motion compensated interpolation (MCI) method uses a previous non-rigid image registration method, which provides an intuitive background for temporal resolution enhancement. We consider motion along four neighboring images instead of only the enclosing neighbors. Therefore, the speech articulators movement is more accurately represented. Results indicate the effectiveness of our approach.

1. Introduction

According to Brench et al. [2], 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 phonetics and phonology (for instance, variabilities of speech and language disorders). For that purpose, knowledge about vocal tract shape and dimensions acquired during the speech of words or phonemes are essential to a full understanding of articulatory and acoustical processes involved in speech production.

According to Baer et al. [1], magnetic resonance imaging (MRI) is the only tool that can provide detailed three-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. [11], 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 acquired image sequences by empowering the acquisition devices, 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 [10]. According to Rueckert [14], the 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 [15]. This registration describes the transformation between each pair of images by a free-form deformation (FFD) with B-spline interpolation between uniformly spaced control points. Indeed, we demonstrate that these 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 could be

---

1 Ana L. D. Martins is supported by FAPESP, Brazil, under grant number 2008/01348-2.
generated simply by positioning control points in meaningful positions. We use a tensor product of cubic B-spline functions considering four neighboring images in order to position the control points. In this way, intermediate images are coherent with the movement of the whole sequence.

2. Related Work

To accurately identify the speech articulators contours during a speech task, available approaches try to improve the MR image acquisition process. Brench et al. [2], Narayanan et al. [11] and Kim et al. [7] employed accelerated methods for acquisition, reconstruction and analysis of real time MRI. A sparse sampling of the k-space was used, motivated by the fact that undersampled spiral k-space data can recover images free from aliasing artifacts. Differently from these methods, the focus of our approach is to enhance temporal resolution, and not to improve the identification of the contours of the speech articulators. Thus, the proposed approach can be used as a pre-processing task, before the application of identification methods. In fact, according to Kim et al. [7], high temporal resolution is required for a complete understanding of the production dynamics. To the best of our knowledge, there are no temporal resolution enhancement approaches applied to this context available in the literature.

To increase temporal resolution in a sequence by generating intermediate frames, a simple approach would be to combine pixel values at the same spatial locations. This approach provides acceptable visual quality in the absence of motion, but it usually blurs moving objects and causes motion jerkiness. Motion compensated interpolation (MCI) techniques are applied to reduce these artifacts. These techniques use motion information to enhance temporal resolution. Most MCI algorithms use block-matching algorithms for motion estimation [4] [5]. In general, these algorithms are simple, easy to implement and generate a compactly represented motion field. However, motion vectors from the conventional block matching algorithms are often not faithful to true object motions and present blocking artifacts. An interesting alternative is to use spatial transformation techniques, such as control grid interpolation [6] [18]. Spatial transformation techniques allow superior motion tracking when compared with the translational block models. However, control grid interpolation usually requires the prior identification of control points. Moreover, these methods are often used to enhance temporal resolution of videos, and, therefore, suffer from rigid time restrictions. In our approach, control points are automatically identified and time restrictions are not an important issue (i.e., we are more interested in the accuracy of the motion estimation task).

A free-form registration, similar to the one used in our method, was applied by Rohlfing et al. [12] and von Siebenthal [17] to study respiratory liver motion. However, different from our context, in these approaches the motion analysis is essentially different because it is based on a repetitive movement due to respiratory motion. Moreover, to the best of our knowledge, the proposed approaches do not deal with temporal resolution enhancement.

3. Non-Rigid Image Registration

In many applications a rigid transformation is sufficient to describe the spatial relationship between two images. A transformation is called rigid when only global translations and rotations are allowed. However, there are applications in which the transformation has to accommodate tissue deformations. This kind of transformation is named non-rigid. Human vocal tract images used in speech production research present deformations similar to the one illustrated in Figure 1. The error image presented in Figure 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.

![Figure 1. Two moments of a speech and the error image.](image)

3.1 FFD based on B-spline Interpolation

Rueckert et al. [15] present a non-rigid image registration method using FFD based on cubic B-splines. Introduced by Sederberg and Parry [16], 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. Considering the two-dimensional (2D) case, Figure 2 illustrates an image deformation caused by manipulations of the mesh of uniform spaced control points shown in Figure 2(b).

The approach presented by Rueckert et al. models deformations by a transformation \( \varphi \) which combines global scene movement with local deformations

\[
\varphi(x, y) = \varphi_{\text{global}}(x, y) + \varphi_{\text{local}}(x, y).
\]

(1)

\( \varphi_{\text{global}} \) models an affine transformation applied in the whole image

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

(2)
given by $x \leq n$
deformations localized in small parts of the image. On the 
formation that can be modeled. Higher resolutions allow
finite support. Therefore, each control point affects the trans-
characteristics, the process is complicated and still not fully au-
identification of landmarks based on image char-
Rohr [13], although there are sophisticated ideas for auto-
knowledge or image characteristics. According to
identification is automatic and does not depend on ex-

Figure 2. Lena deformation using FFD based on cubic B-splines.

where $\theta_{ij}$ are the coefficients of this transformation.

Local deformations are modeled by FFD based on cubic B-spline basis functions. Considering $\Omega = \{(x, y)\mid 0 \leq x \leq X, 0 \leq y \leq Y\}$ the image support and $\Phi$ the $n_x \times n_y$ mesh of control points with uniform spacing $\delta$, $\varphi_{\text{local}}$ is given by

$$
\varphi_{\text{local}}(x, y) = \sum_{l=0}^{3} \sum_{m=0}^{3} B_l(u) B_m(v) \phi_{i+1,l+j+m}
$$

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

$$
B_0(u) = (1-u)^3/6, \\
B_1(u) = (3u^3 - 6u^2 + 4)/6, \\
B_2(u) = (-3u^3 + 3u^2 + 3u + 1)/6, \\
B_3(u) = u^3/6.
$$

Control points $\phi_{i,j}$ are parameters of the FFD. Their identification is automatic and does not depend on expert knowledge or image characteristics. According to Rohr [13], 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, finner grids allow only more global deformations. Lee et al. [8] 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 $\Phi_L$ to $\Phi_0$. This hierarchy is used to derive a sequence of deformation functions with the FFD manipulation. Each mesh $\Phi_L$ and the associated FFD defines a local transformation $\varphi_{\text{local}}^l$ and their sum defines $\varphi_{\text{local}}$

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

Rueckert et al. regularized the transformation by imposing the following smoothness constraint

$$
C_{\text{smooth}}(\varphi) = \frac{1}{V} \int_0^X \int_0^Y \left( \frac{\partial^2 \varphi}{\partial x^2} \right)^2 + 2 \left( \frac{\partial^2 \varphi}{\partial x \partial y} \right)^2 \, dx \, dy
$$
to the spline-based FFD transformation. Moreover, they used the normalized mutual information (NMI) 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 $T$, the similarity criterion is given by

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

where $H(R)$ and $H(T_{\varphi})$ denote the marginal entropies of $R$ and the transformed image $T_{\varphi}$ and $H(R, T_{\varphi})$ denotes the joint histogram of $R$ and $T_{\varphi}$.

The optimal transformation is found by minimizing the cost function

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

where $\lambda$ 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 $\Theta$. Rueckert et al. used an iterative multiresolution search strategy. However one can adopt any other registration method that considers affine transformations [3] [19]. We used the algorithm proposed by Lucas and Kanade [9]. 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 $\mu$. The algorithm stops if a local optimum of the cost function has been found ($\|\nabla C\| \leq \epsilon$, for a small positive $\epsilon$).
4. Proposed Method

Considering a sequence of vocal tract images \( I_0, \ldots, I_k \), the registration algorithm was applied to each pair of consecutive images as illustrated in Figure 3. In this way, given the resultant meshes of control points, it is possible to identify the transformation between any pair of images in this sequence. Note that we applied the registration algorithm to pairs of consecutive images because between distant images sometimes deformations were so large that they could not be easily captured.

![Figure 3. Illustration of the registration of each pair of consecutive images in a sequence.](image)

Speech articulators motion during speech production is a continuous process. Therefore, considering the correspondence between each control point in the sequence of image meshes, we used the continuous motion model

\[
\phi^t_{x,y} = \sum_{n=0}^{3} B_n(w) \phi^{k+n}_{x,y}
\]

where \( k = \lfloor t \rfloor \), \( w = t - \lfloor t \rfloor \) and index \( t \) in \( \phi^t_{x,y} \) denotes the control point \( \phi_{x,y} \) of the \( t \)-th image in the sequence. In this way, the deformation at control point \( \phi_{x,y} \) at a new moment \( t \) is computed from the positions of the four corresponding control points of surrounding images in the sequence.

According to the identified mesh of control points, temporal resolution enhancement is performed first by applying \( \varphi_{local} \) to both images that are next to the new image. After, a weighted sum of the transformed images is performed as illustrated in Figure 4. The weights are defined according to the distance between the new image and each of the neighboring images. The transformed image related to the closest image receives a higher weight. Considering observed images and corresponding meshes of control points, this approach gives an image that is coherent to the data.

![Figure 4. Temporal resolution enhancement by performing a weighted mean of transformed neighboring images.](image)

According to the proposed continuous motion compensated interpolation approach presented in Equation (8), it is possible to generate an infinite number of intermediate temporal resolution enhancement method gives a continuous motion model for each pixel of the image. Considering pixels that compose the contours of interest estimated by a segmentation method, our method can be used to identify the continuous movement of this contours during the whole acquisition.

![Figure 5. Speech articulators segmentation [2].](image)

5. Results

We generated intermediate images in a sequence of 256 × 256 vocal tract MR images for visual evaluation. The sequence of four MR images presented in Figure 6 was registered as illustrated in Figure 3. The hierarchical non-rigid registration approach resulted in the 67 × 67 meshes of control points also presented in Figure 6. Each mesh of control points models deformations present between each pair of consecutive images. It is important to that the first mesh is composed by uniform spaced control points, indicating that there is no deformation in this image. Therefore, the first image can be considered the reference image. Note that the non-rigid registration approach accurately detected speech articulators movement in the sequence.

According to the proposed continuous motion compensated interpolation approach presented in Equation (8), it is possible to generate an infinite number of intermediate
Figure 6. Sequence of four 256 × 256 MR vocal tract images and respective meshes of control points.

For comparison purposes, unfortunately we do not have any knowledge about methods of motion compensated interpolation of vocal tract images used for speech production research. Therefore, we compared our method with the image generated by combining pixel values at the same spatial locations in the first two images in the sequence presented in Figure 8(a). Note that this approach did not consider motion between the two observed images. Therefore, it blurs locations that moves the most and decreases resolution in small details of the images. The image generated by the proposed method coherently models the movement without degradation of image details.
6. 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 non-rigid image registration approach presented by Ruckert et al. [15]. 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 method present 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 present 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.

References


