Vocal Tract Model Adaptation Using Magnetic Resonance Imaging

Peter Birkholz[1]*, **Bernd J. Kröger**[2]

[1]Institute for Computer Science, University of Rostock, 18051 Rostock, Germany

[2]Department of Phoniatrics, Pedaudiology, and Communication Disorders
University Hospital Aachen, 52074 Aachen, Germany

piet@informatik.uni-rostock.de, bkroeger@ukaachen.de

Abstract. *We present the adaptation of the anatomy and articulation of a 3D vocal tract model to a new speaker using magnetic resonance imaging. We used two different corpora of the speaker: a corpus of volumetric magnetic resonance (MR) images of sustained phonemes and a corpus with dynamic sequences of midsagittal MR images. Different head-neck angles in these corpora required a normalization of the MRI traces, which was done by warping. The adaptation was based on manual matching of midsagittal vocal tract outlines and automatic parameter optimization. The acoustic similarity between the speaker and the adapted model is tested by means of the natural and synthetic formant frequencies. The adaptation results for vowel-consonant coarticulation are exemplified by the visual comparison of synthetic and natural vocal tract outlines of the voiced plosives articulated in the context of the vowels /a/, /i/ and /u/.*

1. Introduction

In the last few years, we have been developing an articulatory speech synthesizer based on a geometric 3D model of the vocal tract (Birkholz, 2005; Birkholz et al.). Our goals are high quality text-to-speech synthesis as well as the application of the synthesizer in a neural model of speech production (Kröger et al., 2006). Till now, the anatomy and articulation of our vocal tract model were based on x-ray tracings of sustained phonemes of a Russian speaker. However, these data were not sufficient to reproduce the speakers anatomy and articulation very accurately. They neither provided information about the lateral vocal tract dimensions nor on coarticulation of phonemes. These information had to be guessed and impeded a strict evaluation of the synthesizer.

In this study, we started to close this gap by adapting the anatomy and articulation of our vocal tract model to a new speaker using MRI (magnetic resonance imaging). Two MRI corpora were available to us: one corpus of volumetric images of sustained vowels and consonants, and one corpus of dynamic midsagittal MRI sequences with 8 frames/second. Additionally, we had high resolution computer tomography (CT) scans of oral-dental impressions. The CT scans were used to adapt the geometry of the hard palate, the jaw, and the teeth. The articulatory targets for vowels and consonants were

*Supported by the German Research Foundation.

determined by means of the volumetric MRI data. The dynamic MRI corpus were used to determine the influence/dominance of the individual articulators during the production of consonants. This is important for the simulation of vowel-consonant coarticulation in our synthesizer.

Section 2 will discuss the analysis and normalization of the images from both corpora, and Sec. 3 introduces the vocal tract model and describes the adaptation of vowels and consonants. Conclusions are drawn in Sec. 4.

2. Magnetic Resonance Image Processing

2.1. Corpora

We analyzed two MRI corpora of the same native German speaker (JD, ZAS Berlin) that were available to us from other studies (Kröger et al., 2000, 2004). The first corpus contains volumetric images of sustained phonemes including tense and lax vowels, nasals, voiceless fricatives, and the lateral /l/. Each volumetric image consists of 18 sagittal slices with 512 x 512 pixels. The pixel size is 0.59 x 0.59 mm^2 and the slice thickness is 3.5 mm.

The second corpus contains dynamic MRI sequences of midsagittal slices scanned at a rate of 8 frames/second with a resolution of 256 x 256 pixels. The pixel size 1.18 x 1.18 mm^2. The recorded utterances consist of multiple repetitions of the sequences /a:Ca:/, /i:Ci:/ and /u:Cu:/ for nearly all German consonants C.

In addition to these two corpora, we had high resolution CT scans of plaster casts of the upper and lower jaws and teeth of the speaker with a voxel size of $0.226 \times 1 \times 0.226$ mm^3.

2.2. Outline Tracing

The midsagittal airway boundaries of all MR images were hand-traced on the computer for further processing. The manual tracing was facilitated by applying an edge detector (Sobel operator) to the images. Examples of MR images from corpora 1 and 2 are shown in Fig. 1 (a) and (d), respectively. Pictures (b) and (e) show the corresponding results of the Sobel edge detector, and the tracings are depicted in (c) and (f). For corpus 1 phonemes, we additionally traced the tongue outlines approximately 1 cm left from midsagittal plane (dashed line in Fig. 1 (c)).

In corpus 2, we were interested in the articulation of the consonants in the context of the vowels /a:/, /i:/ and /u:/. The analysis of the dynamic MRI sequences revealed, that the sampling rate of 8 frames/second was to low to capture a clear picture of each spoken phoneme. But in the multiple repetitions that we had of each spoken /VCV/-sequence, we identified for each consonant+context at least 2 (usually 4-5) candidate frames, where the consonantal targets were met with sufficient precision. One of these candidates was chosen as template for tracing the outlines. The chosen candidate frame was supposed to be the one that best represented the mean of the candidate set. Therefore, we chose in each candidate set the frame that had the smallest sum of "distances" to all other frames in that set. The distance between two pictures was defined as

$$e = (W \cdot H)^{-1} \sum_{x=1}^{W} \sum_{y=1}^{H} |A(x,y) - B(x,y)|,$$

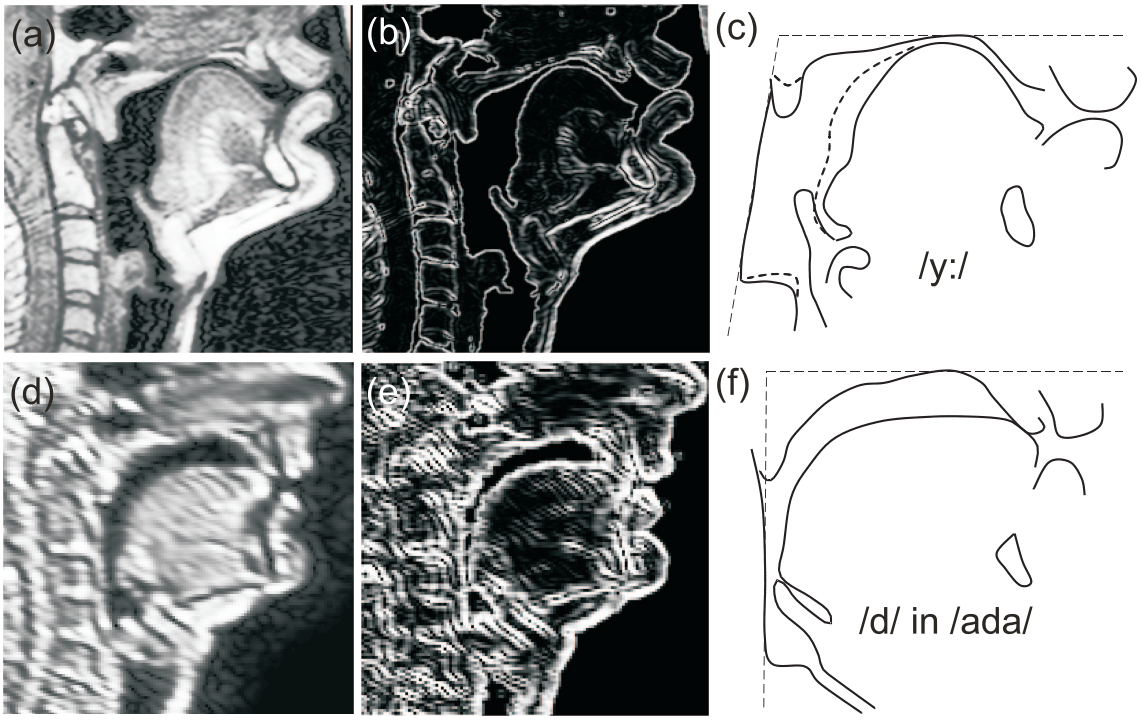

Figure 1. (a) Original image of corpus 1. (b) Edges detected by the Sobel operator for (a). (c) Tracing result for (b). (d)-(f) Same as (a)-(c) for an image of corpus 2.

where $W \times H$ is the resolution of the images, and $A(x, y)$ and $B(x, y)$ are the 8-bit gray values at the position (x, y) in the pixel matrices.

The volumetric CT images of the plaster casts of the upper and lower jaw were exactly measured in the lateral and coronal plane to allow a precise reconstruction of these rigid parts in the vocal tract model.

2.3. Contour Normalization

The comparison of Fig. 1 (c) and (f) shows, that the head was not held in exactly the same way in both corpora. In corpus 1, the neck is usually more ”stretched” than in corpus 1, resulting in a greater angle between the rear pharyngeal wall and the horizontal dashed line on top of the maxilla outline [1]. Smaller variation of this angle also exist within the two corpora. For the vocal tract adaptation it was essential to normalize these differences in head postures.

Our basic assumption for the normalization is, that there exists a fixed point R (with respect to the maxilla) in the region of the soft palate, around which the rear pharyngeal outline rotates when the head is raised or lowered. Given this assumption, the straight lines approximating the rear pharyngeal outlines of all tracings should intersect in R. Therefore, R was determined solving the minimization problem

$$\sum_{i=1}^{N} d^2(R, l_i) \rightarrow min,$$

[1] Both tracings were rotated such that the horizontal dashed line is parallel to the upper teeth.

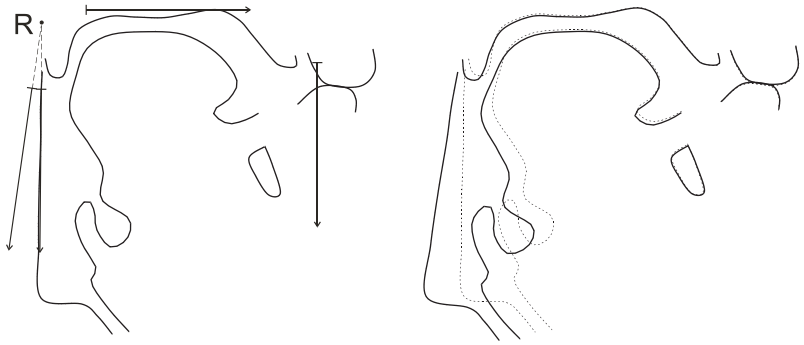

Figure 2. Warping of the MRI-tracing of the consonant /b/ in /ubu/.

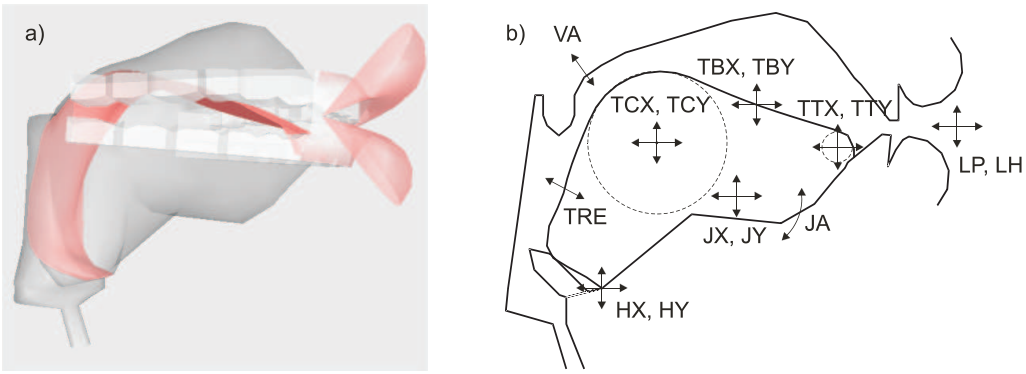

Figure 3. (a) 3D-rendering of the vocal tract model. (b) Vocal tract parameters.

where N is the total number of traced images from both corpora, and $d(R, l_i)$ denotes the shortest distance from R to the straight line l_i that approximates the rear pharyngeal wall of the ith image. Each MRI-tracing was then warped such that its rear pharyngeal outline was oriented at a predefined constant angle. Warping was performed using the method by Beier and Neely (1992) with 3 corresponding pairs of vectors as exemplified in Fig. 2. The horizontal vectors on top of the palate and the vertical vectors at the chin are identical in the original and the warped image, keeping these parts of the vocal tract equal during warping. Only the vectors pointing down the pharyngeal outline make the vocal tract geometry change in the posterior part of the vocal tract. Both of these vectors only differ in the degree of rotation around R. Figure 2 (b) shows the MRI-tracing in (a) before warping (dotted curve) and after warping (solid curve). This method proofed to be very effective and was applied to all MRI-tracings.

3. Adaptation

3.1. Vocal Tract Model

Our vocal tract model consists of different triangle meshes that define the surfaces of the tongue, the lips and the vocal tract walls. A 3D rendering of the model is shown in Fig. 3 (a) for the vowel /a:/. The shape of the surfaces depends on a number of predefined parameters. Most of them are shown in the midsagittal section of the model in Fig. 3 (b). The model has 2 parameters for the position of the hyoid (HX, HY), 1 for

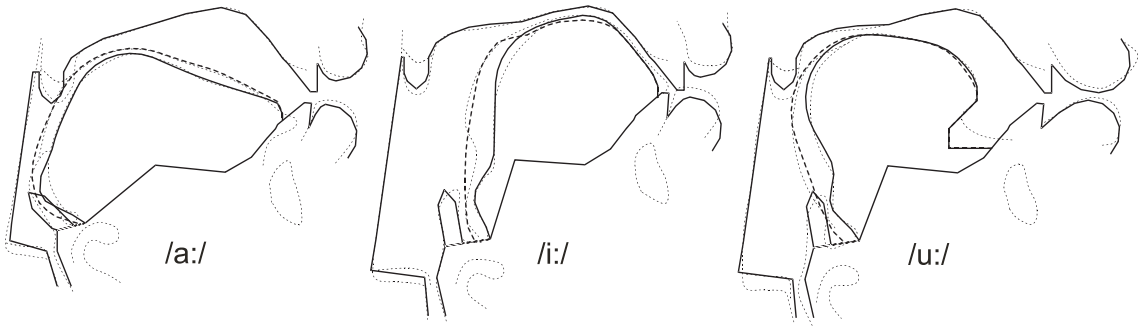

Figure 4. MRI outlines (dotted curves) and the matched model-derived outlines (solid curves) for the vowels /a:/, /i:/ and /u:/.

the velic aperture (VA), 2 for the protrusion and opening of the lips (LP, LH), 3 for the position and rotation of the jaw (JX, JY, JA) and 7 for the midsagittal tongue outline ($TRE, TCX, TCY, TBX, TBY, TTX, TTY$). Four additional parameters define the height of the tongue sides with respect to the midsagittal outline at the tongue root, the tongue tip, and two intermediate positions. A detailed description of the parameters is given in (Birkholz, 2005; Birkholz et al.). The current version of the model is an extension of the model in the cited references. On one hand, we added the epiglottis and the uvula to the model, which were previously omitted. Furthermore, the 3D-shape of the palate, the mandible, the teeth, the pharynx and the larynx were adapted to the (normalized) MR images.

3.2. Vowels

To reproduce the vowels in corpus 1, the vocal tract parameters were manually adjusted aiming for a close match between the normalized MRI tracings and the model-derived outlines. Furthermore, the tongue side parameters were adjusted for a close match of the tongue side outlines. Figure 4 shows our results for the vowels /a:/, /i:/ and /u:/. The model outline is drawn as solid lines and its tongue sides as dashed lines. The corresponding MRI tracings are drawn as dotted lines. In the case of all examined vowels, we achieved a fairly good *visual* match.

The *acoustic* match between the original and synthetic vowels was tested by comparison of the first 3 formant frequencies. The formants of the natural vowels were determined by standard LPC analysis. The audio corpus was recorded independently from the MRI scans with the speaker in a supine position repeating all vowels embedded in a carrier sentence four times. For each formant frequency of each vowel, the mean value was calculated from the 4 repetitions.

The formant frequency of the synthetic vowels were determined by means of a frequency-domain simulation of the vocal tract system based on the transmission-line circuit analogy (Birkholz, 2005). The area functions for these simulations were calculated from the 3D vocal tract model. The nasal port was assumed to be closed for all vowels. In all acoustic simulations, we considered losses due to yielding walls, viscous friction, and radiation. The *piriform fossa* side cavity was included in the simulations and modeled after (Dang and Honda, 1997).

The test results are summarized in Fig. 5 for the first two formants of the tense Ger-

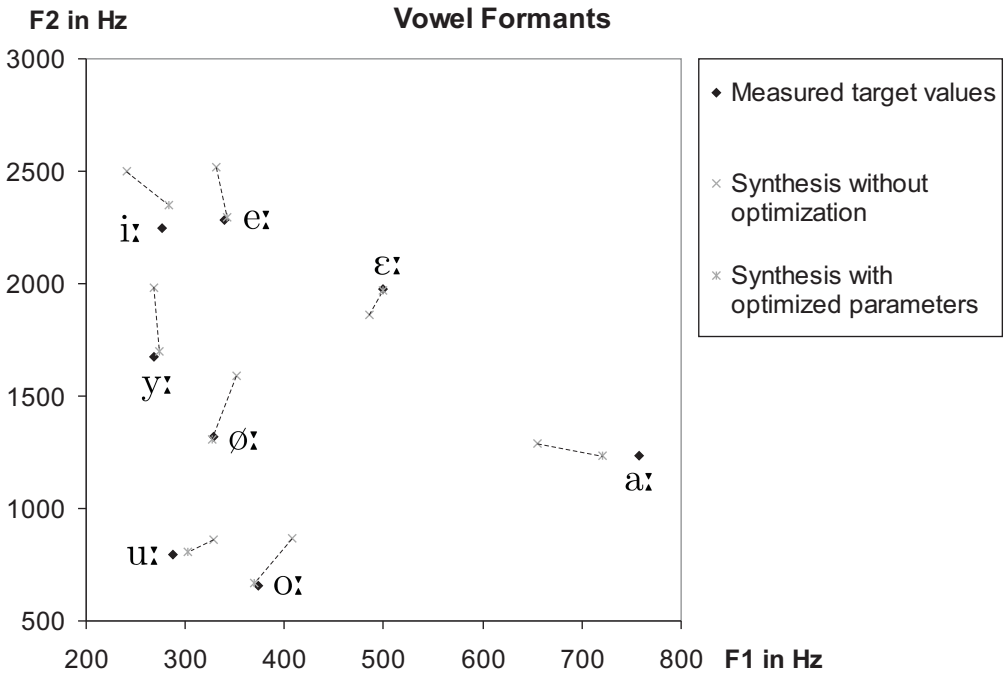

Figure 5. Formant frequencies for the German tense vowels.

man vowels. The error between the natural and synthetic formant frequencies averaged over the first three formants of all vowels shown in Fig. 5 was 12.21%. This error must be mainly attributed to the limited accuracy of the MRI tracings (due to the low image resolution) as well as to the imperfect matching of the outlines. In order to improve the acoustic match, we implemented an algorithm searching the vocal tract parameter space to minimize the formant errors. During the search, each vocal tract parameter was allowed to deviate maximally 5% of its whole range from the value that was determined during the outline matching. Figure 5 shows that the formants were much closer to their ”targets” after this optimization, though the parameters (and so the model geometry) changed only little. The average formant error reduced to 3.41%.

3.3. Consonants

To a certain extend, the articulatory realization of a consonant depends on the vocalic context due to vowel-consonant coarticulation. In our synthesizer, we use a dominance model to simulate this effect Birkholz et al.. The basic idea is, that each consonant has a ”neutral” target shape (just like the vowels), but in addition, each parameter has a weight between 0 and 1, expressing the ”importance” of the corresponding parameter for the realization of the consonantal constriction. For /d/, for example, the tongue tip parameters have a high weight, because the alveolar closure with the tongue tip is essential for /d/. Most of the other parameters/articulators are less important for /d/ and have a lower weight. The other way round, a weight expresses how strong a consonantal parameter is influenced by the context vowels (low weight = strong influencing). Formally, this concept is expressed by

$$x_{c|v}[i] = x_v[i] + w_c[i] \cdot (x_c[i] - x_v[i]), \tag{1}$$

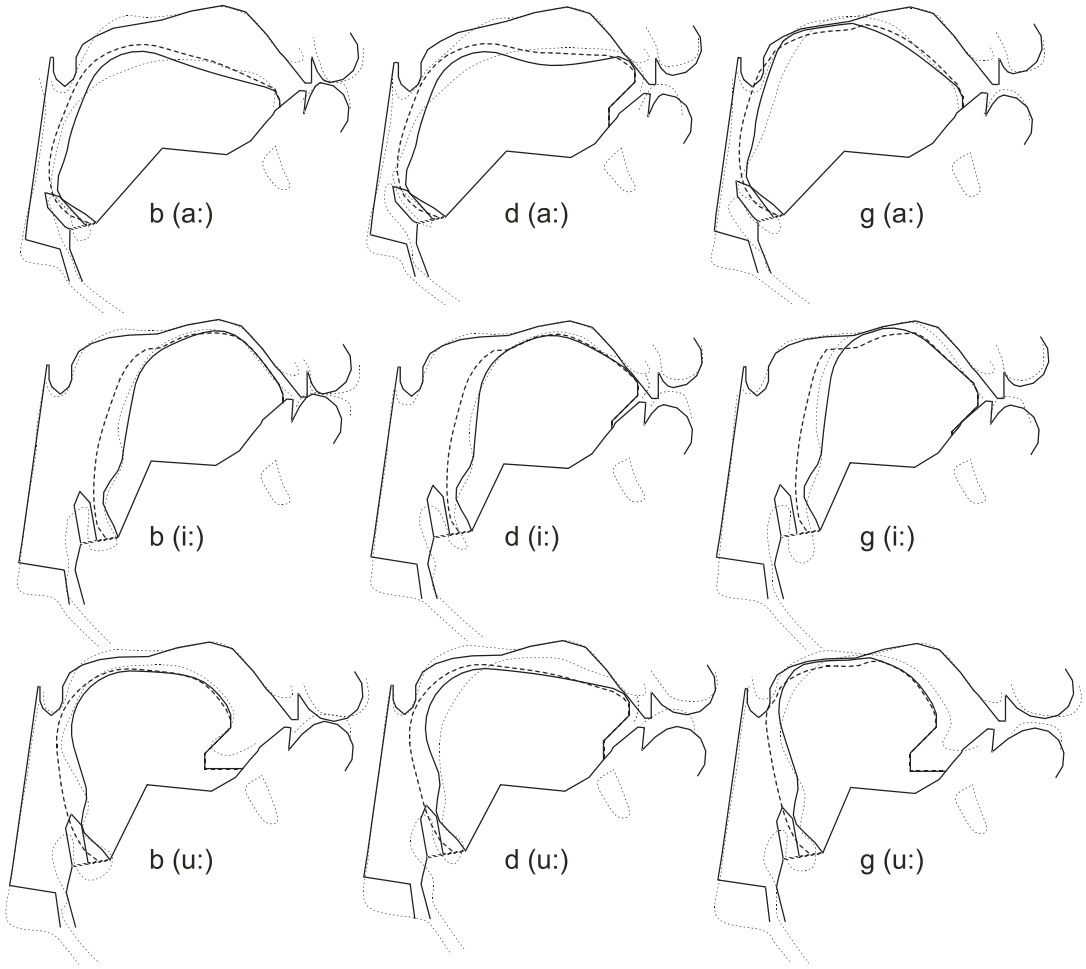

Figure 6. Articulatory realization of the voiced plosives in the context of the vowels /a:/, /i:/ and /u:/. MRI tracings are drawn as dotted curves and model-derived outlines as solid curves.

where i is the parameter index, $x_{c|v}[i]$ is the value of parameter i at the moment of the maximal closure/constriction of the consonant c in the context of the vowel v, $w_c[i]$ is the weight for parameter i, and $x_c[i]$ and $x_v[i]$ are the parameter values of the targets for the consonant and vowel.

Hence, the needed data for the complete articulatory description of a consonant c are $x_c[i]$ and $w_c[i]$. The parameters for the "neutral" consonantal targets were adjusted analogous to the vowel parameters in Sec. 3.2 using the high resolution MRI data from corpus 1. The consonantal weights were determined using the selected MRI tracings from corpus 2, that show the realization of the consonants in symmetric context of the vowels /a:/, /i:/, and /u:/. The vocal tract parameters for these coarticulated consonants were manually adjusted, too. Let us denote these parameters by $x_{c|v_j}$, where $v_j \in \{/\mathrm{a}:/, /\mathrm{i}:/, /\mathrm{u}:/\}$. The optimal weights $w_c[i]$ were determined solving the minimization problem

$$\sum_{j=1}^{N} \left[x_{c|v_j}[i] - x_{v_j}[i] - w_c[i] \cdot (x_c[i] - x_{v_j}[i]) \right]^2 \rightarrow min,$$

where $N = 3$ is the number of context vowels. The solution is

$$w_c[i] = \left[\sum_{j=1}^{N}(x_{c|v_j}[i] - x_{v_j}[i])(x_c[i] - x_{v_j}[i])\right] / \left[\sum_{j=1}^{N}(x_c[i] - x_{v_j}[i])^2\right].$$

Figure 6 contrasts the model-derived outlines of coarticulated consonants using Eq. (1) (solid curves) and the corresponding MRI tracings (dotted curves). Obviously, some of the outlines differ and show the limits of the dominance model. A major (systematic) mismatch can be found in the laryngeal region. We attribute this to the marked differences of the larynx shape in the images of corpus 1 and 2 (cf. Fig. 1 (c) and (f)). Nevertheless, the basic coarticulatory properties are retained in all examples (e. g., the tongue for /b/ is further back in /u:/-context than in /i:/-context).

4. Conclusions

We have presented the anatomic and articulatory adaptation of a vocal tract model to a specific speaker combining data from higher resolution volumetric MRI data and lower resolution dynamic MRI data. We achieved a satisfying visual and acoustic match between the original speaker and the model. The methods proposed in this study can be considered as simple but powerful means for future adaptations to other speakers, provided that the corresponding MRI data are available.

References

Beier, T. and Neely, S. Feature-based image metamorphosis. *Computer Graphics*, 26(5): 35–42, 1992.

Birkholz, P. *3D-Artikulatorische Sprachsynthese*. Logos Verlag Berlin, 2005.

Birkholz, P., Jackèl, D., and Kröger, B. J. Construction and control of a three-dimensional vocal tract model. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, pages 873–876, Toulouse, France.

Dang, J. and Honda, K. Acoustic characteristics of the piriform fossa in models and humans. *Journal of the Acoustical Society of America*, 101(1):456–465, 1997.

Kröger, B. J., Birkholz, P., Kannampuzha, J., and Neuschaefer-Rube, C. Spatial-to-joint coordinate mapping in a neural model of speech production. In *32. Deutsche Jahrestagung für Akustik (DAGA '06)*, Braunschweig, Germany, 2006.

Kröger, B. J., Hoole, P., Sader, R., Geng, C., Pompino-Marschall, B., and Neuschaefer-Rube, C. MRT-Sequenzen als Datenbasis eines visuellen Artikulationsmodells. *HNO*, 52:837–843, 2004.

Kröger, B. J., Winkler, R., Mooshammer, C., and Pompino-Marschall, B. Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. In *5th Seminar on Speech Production: Models and Data*, pages 333–336, Kloster Seeon, Bavaria, 2000.