Centerline articulatory models of the velum and epiglottis for articulatory synthesis of speech

Yves Laprie
Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France
Yves.Laprie@loria.fr

Benjamin Elie
L2S, CentraleSupélec, CNRS, Univ. Paris Sud
Université Paris-Saclay, Gif-sur-Yvette, France
Benjamin.Elie@loria.fr

Anastasiia Tsukanova
Université de Lorraine, CNRS, Inria, LORIA
F-54000 Nancy, France
Anastasiia.Tsukanova@loria.fr

Pierre-André Vuissoz
Université de Lorraine, Inserm, IADI
F-54000 Nancy, France
pa.vuissoz@chru-nancy.fr

***Abstract*—This work concerns the construction of articulatory models for synthesis of speech, and more specifically the velum and epiglottis. The direct application of principal component analysis to the contours of these articulators extracted from MRI images results in unrealistic factors due to delineation errors. The approach described in this paper relies on the application of PCA to the centerline of the articulator and a simple reconstruction algorithm to obtain the global articulator contour.**

The complete articulatory model was constructed from static Magnetic Resonance (MR) images because their quality is much better than that of dynamic MR images. We thus assessed the extent to which the model constructed from static images is capable of approaching the vocal tract shape in MR images recorded at 55 Hz for continuous speech. The analysis of reconstruction errors shows that it is necessary to add dynamic images to the database of static images, in particular to approach the tongue shape for the /l/ sound.

***Index Terms*—Speech, Articulatory models, MRI, Deformable objects.**

I. Introduction

The numerical acoustic simulations of the vocal tract have seen recent progress in the possibility of taking into account the complexity of the vocal tract geometry [1], [2], the specificities of vocal fold vibrations [3], [4] and production of consonants [5] or other complex speech sounds as trills [6]. These advances rely on numerical models that are more faithful to the acoustics of speech, but also on articulatory data that are more precise and accurate thanks to the emergence of Magnetic Resonance Imaging (MRI).

In order to exploit these advances within the framework of articulatory synthesis it is necessary to provide a sufficiently precise geometric description of the vocal tract shape given as an input of aero-acoustical simulations. Precision should be good especially at the place of articulation, i.e. the place where the distance between both walls of the vocal tract is minimal.

Many efforts have been dedicated to the main articulators [7]–[9], i.e. the mandible which corresponds to the jaw opening, the tongue which controls the distribution of the back and front cavities in the vocal tract, and lips which give the vocal tract aperture and the length of the front cavity via their protrusion. However, less voluminous articulators are likely to change the acoustics quite profoundly and should therefore also receive a great deal of attention. This is the case of the velum which controls the velopharyngeal port and fine adjustments of the balance between back and front cavities. Similarly, the epiglottis located in the larynx region may change the resonance at the lower part of the vocal tract, and can also be used to articulate some phonemes.

This work is dedicated to the development of articulatory models for these two articulators and their link with the delineation of the contours of these articulators in MR images. We already developed a 2D articulatory model of the vocal tract from X-ray [9] and then MR images. The choice of a 2D model is motivated by two issues. First, the main geometrical feature from an acoustic point of view is the distribution of cavities and the places where constrictions occur. Taking the third dimension into account does not bring much [10]. Second, the development of a complete 3D model from 3D MR volumes requires efforts of delineation which are often prone to errors. Indeed, the distinction between air and tissues is often problematic and depends on the operator who analyzes the image or supervises semi-automatic tools provided by itk-SNAP [11] or MITK [12].

Our approach uses data analysis methods, i.e. Principal Component Analysis (PCA), to identify articulator deformation modes. The weakness of this approach is that the model is strongly linked to a speaker. To a certain extent the model can be quite easily adapted to another speaker in terms of size and main anatomical parameters including palate shape [13]. Similarly, the up/down or front/rear movements, and the rounding of the tongue are common to all speakers, ensuring that the model can be used without risk of major errors. On the other hand, finer speaker articulatory strategies escape this adaptation.

The underlying assumption for using PCA to build an articulatory model is that the magnitude of actual deformations modes, i.e. PCA factors, is greater than those due to

ISBN 978-90-827970-1-5

delineation errors. From a practical point of view this means that the variance explained by those factors is stronger than those of delineation errors. This assumption is fully verified for the tongue which occupies an important place on the MR mid-sagittal image of the vocal tract, but not for smaller articulators like the epiglottis and the velum. On the mid-sagittal image the epiglottis is a filiform object whose two edges are separated by only a few pixels.

This means that the PCA factors resulting from delineation errors can mix with or even dominate the true deformation factors. The situation is quite similar for the velum which is a small object which does not present a strong contrast on the image and whose edges cannot be delineated easily. A deeper reason is related to the intrinsic form and nature of these organs. The epiglottis appears as a filiform object on the image and it makes little sense to look for transverse deformation factors. Sections II and III are dedicated to the description of these models.

The main advantage of a model derived from vocal tract images is to ensure that the model produces vocal tract shapes similar to those that the speaker realized during the acquisition. X-ray images used to construct articulatory models until the late 1990s were extracted from films and correspond to continuous speech. The main difficulty encountered during the construction of these models was the poor quality of the images on which several contours are superimposed. This is due to the nature of images obtained from X-rays that cross the speaker's head and projects all contours on the same image. In particular, teeth and fillings often hide the tongue contour which was poorly detected.

Models constructed from MR images mainly correspond to static images. Subjects are asked to position their articulators as if they were producing the sound for vowels. They must keep the position without moving during the acquisition which lasts approximately ten seconds. For consonants, they are asked to place their articulators as if they were going to produce the consonant in a particular vocalic context. Each consonant acquisition thus corresponds to a CV and in order to limit the number of acquisitions the vowels are often the cardinal vowels /i,a,u/. The main reason of using static images is that the availability of dynamic MR imaging is recent (in the late 2000s), very limited and of poor quality and resolution (less than 100x100 for the whole vocal tract in most cases). Conversely, static images offer the advantage of covering the complete volume of the vocal tract with a resolution of 1mm or slightly below, and a good contrast between air and tissues, which makes the delineation process easier. The risk of using static images is that the model cannot generate vocal tract shapes that correspond to continuous speech. This issue is addressed in section IV.

In this work we exploited 99 MR static images recorded by a male speaker corresponding to 14 vowels, 72 blocked consonant-vowel articulations for /f,s,ʃ,p,t,k,l,m,n,ʁ/ followed by a cardinal vowel and other vowels in some cases, plus 3 reference images (used to register the dental cast). The data were collected with a GE Signa 3T machine with an 8-channel neurovascular coil array. The protocol consisted in a 3D volume of the vocal tract acquired with a custom modified Enhanced Fast Gradient Echo (EFGRE3D, TR 3.12 ms, TE 1.08 ms, matrix 256x256x76, with spatial resolution 1.02x1.02x1.0 mm3). The mid-sagittal slice was used to construct the models for each of the articulators of the vocal tract.

In addition we also exploited dynamic MR data recorded at Max Planck Institute in Göttingen by using the real-time MRI reconstruction algorithm [14]. This database comprises 200 phonetically balanced sentences and spontaneous speech. The sampling frequency is 55 Hz and each image (189x189 pixels) is a 8 mm slice located at the mid-sagittal plane.

II. Centerline model of the velum

In [15] PCA was applied directly to the velum contours extracted from a X-ray film. Even if this model rendered the velopharyngeal opening correctly, it fails to render complex forms observed in some recently acquired MR images where the velum rolls up on itself as illustrated by Fig. 1.a which corresponds to the static articulation of /ʁa/. As it can be seen on Fig. 1.b, the rolling of the velum on itself is confused with swelling. Coupled with delineation errors this gives unrealistic first PCA factor, which corresponds to the swelling of the velum as shown in Fig. 2. The variance explained by this first factor is 58.2%, while that explained by the second factor, i.e. the actual opening and closing movement of the velopharyngeal port by the velum, is only 32.6%.

Hence the idea of a model construction that reduces the impact of delineation errors, and a model that is capable of representing the rolling movement, and more generally the pendulum movement, when the velum closes the velopharyngeal port. Instead of being applied to the exterior contours of the velum, PCA is applied on the centerline of the velum as illustrated by Fig. 3. The centerline is represented as a regularly sampled curve whose upper extremity is fixed. The PCA input vector is the sequence of segments that define this curve. Since one extremity of the curve is fixed, each segment is entirely determined by the angle it forms with the previous segment. The input vector of PCA is thus formed by the sequence of angles and the total length of the centerline. Actually, to ensure the homogeneity of the input vector each segment is represented by its length and the angle it forms with the previous segment. The integration of the length into the input vector is intended to account for the small elongation of the velum and also the fact that the upper extremity may not be completely fixed.

Since the global articulatory model requires the knowledge of the velum shape and not only the velum centerline a reconstruction algorithm was designed. Its principle consists of reconstructing both edges of the velum from the fixed extremity. The distance of each edge point with respect to the centerline is an affine function of the index of the corresponding center-point. The mobile extremity is represented as an ellipse which connects both edges. The widths of the edges at both extremities were adjusted by hand on a

ISBN 978-90-827970-1-5 © EURASIP 2018

reference image. The algorithm takes into account the high curvature configurations where two segments orthogonal to the centerline, and corresponding to two consecutive center-points, cross.

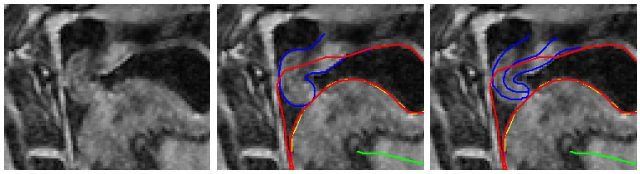

Fig. 1. (a) MRI sagittal image of /ʁa/ (left), (b) Velum (solid blue curve) approximated by the PCA model (middle), (c) Velum and velum centerline (solid blue curves) approximated by the centerline PCA model.

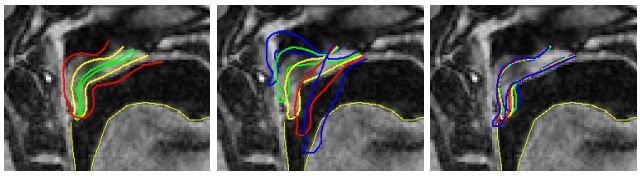

Fig. 2. Nomograms of the first three factors of the standard PCA velum model. The yellow curve is the average configuration. The green and yellow curves are $\pm\sigma$. The blue curves (for the second and third factor) are $\pm 2\sigma$.

Fig. 3. Centerline velum model. The red cross is the anchor point of the centerline.

This time, the first factor explains 43.4 % of the variance (see Fig.4 left) and corresponds to the real lowering/raising movement of the velum. The second factor (15.6% of the variance) corresponds to the rolling deformation, and the third factor captures the slight change in length, whether due to a real lengthening effect (possibly explained by the movement of the anchor point) or delineation errors. Even if the variance explained relates to the overall contour in one case and the central line in the other, we can note that the weight of the first factor, which is now relevant, is less important in this new model. More importantly, the velum is now adequately represented by the PCA factors, and the effect of delineation errors is pushed beyond the fourth factor.

III. Model of the epiglottis

As the velum, the epiglottis can be modeled from its centerline. But unlike the velum, which is largely independent of other articulators, the epiglottis is strongly influenced by the tongue (and therefore the jaw) and larynx.

As a matter of fact, the centerline was determined after delineation of all the epiglottis contours, and the width was set as the average width of all these contours in the upper part where the two epiglottis edges are clearly visible (see Fig. 5). The height of the upper part (where both contours are visible) is adjusted by hand to fit the contours extracted from images. The centerline is approximated as a B-spline and represented by its control points P_l ($0 \leq l < M$ where M is the number of control points) in the form of a two-coordinate vector, and the reconstruction of the epiglottis from the centerline amounts to a line at a distance of half the width from the centerline.

In order to take into account the influences of the jaw, larynx and tongue, we applied multiple linear regression [16] on the control points P_l ($0 \leq l < M$) of the epiglottis contour:

$$P_l = jaw_{0,l}B_0 + \sum_{j=0}^{j=T-1} tg_{j,l}C_j + lx_{0,l}D_0 + E_l$$

where $jaw_{0,l}$ is the control factor of the first linear component of the jaw in the global articulatory model and B_0 the two-coordinate regression vector for the jaw, $tg_{j,l}$ the control factors of the T first linear components of the tongue and C_j the corresponding regression vector, $lx_{0,l}$ the first control factor of the larynx component and D_0 the corresponding regression vector, and E_l is the residue vector.

Actually, since the centerline is represented by the M control points, each centerline occurrence is represented by a $2 \times M$ vector formed by the coordinates of the M points. Each of the N occurrences of the epiglottis is represented by a $2 \times M$ vector named P_i with $0 \leq i < N$. These contour vectors P_i are the observations, and the control factors of the jaw $jaw_{0,i}$, tongue $tg_{j,i}$ and larynx $lx_{0,i}$ are the input explanatory variables. For the jaw and larynx we kept only the first factor (jaw_0, resp. lx_0), which is in both cases the most informative one. For the tongue, which is in front of the epiglottis, we kept the first six deformation factors $tg_{0..5}$. All the variables are centralized and therefore the intercept can be ignored.

By grouping all the explanatory factors jaw_0, tg_j, and lx_0, in a vector of $K = T + 2$ coordinates which are named X_j, the previous equation is

$$P_i = \sum_{j=0}^{j=K-1} b_j X_i + e_i$$

or in a matrix form $P = XB + E$, where P is the $N \times 2M$ matrix of the observations, i.e. the control points of the centerlines, X is the $N \times K$ matrix of the explaining articulatory

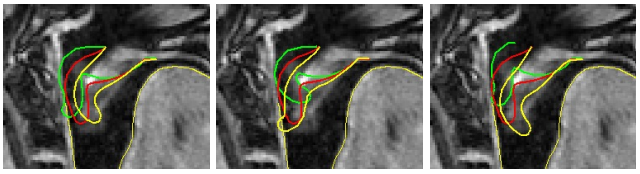

Fig. 4. Nomograms of the first three factors of the centerline velum model. The red curve is the average configuration. The green and yellow curves are $\pm\sigma$.

ISBN 978-90-827970-1-5 © EURASIP 2018

factors, B is the $K \times 2M$ regression matrix and E is the $N \times 2M$ matrix of residue not explained by jaw, tongue and larynx. B is given by $B = (X^tX)^{-1}X^tY$.

Then, the contributions of the jaw, tongue and larynx can be subtracted from the observations of the epiglottis, i.e. $P - XB$, and PCA can be applied to the residue E. Given the nature of epiglottis, i.e. a cartilage, the number of relevant linear components should be small.

We applied this analysis scheme to two sets of data: that of mid-sagittal slice of 3D MRI images presented in the introduction and a corpus of 1021 mid-sagittal images from an X-ray film made of short sentences recorded by a female speaker [17]. Tab. I shows that most of the variance is

σ in mm	male (static)	female (dynamic)
total	8.44	14.32
regression	5.08	6.51
PCA 1st	2.66	3.17
PCA 2nd	1.33	1.73
PCA 3rd	0.78	0.93

TABLE I
STANDARD DEVIATIONS IN MILLIMETER OF THE EPIGLOTTIS: TOTAL, AFTER SUBTRACTING THE INFLUENCE OF OTHER ARTICULATORS, AFTER SUBTRACTING 1^{st}, 2^{nd} AND 3^{rd} LINEAR COMPONENTS.

explained by the articulators influencing the epiglottis and that two linear components are enough to approximate the epiglottis fairly accurately.

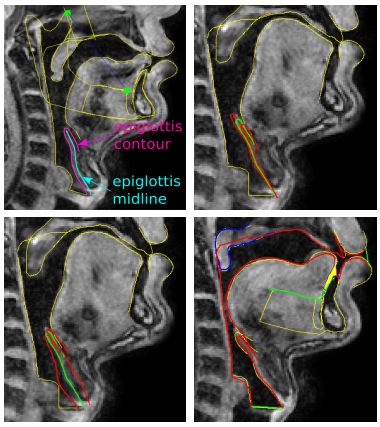

Fig. 5. Epiglottis and its centerline (left), first two deformation modes of the epiglottis for the male speaker. The green line is the neutral position (middle), and the complete articulatory model in red approximating the vocal tract for the articulation /sø/ (right)

The regression stage enables most of the extrinsic influences exerted on the epiglottis to be removed before PCA. It should be noted that these influences explain as much variability as the intrinsic components.

Fig.5 shows that the first factor corresponds to a vertical movement, and the second corresponds to a horizontal movement in the lower and mid part of the epiglottis which contributes to the constriction of the whole laryngeal vestibule. This retraction is not explicitly used in French which is the mother tongue of the two speakers but plays a determining role in the production of glottal consonants in Arabic or Hebrew [18], [19].

IV. Fitting the articulatory model to dynamic images

The main reason of using static MR images of the vocal tract is to provide a much better image quality than that of real-time MR images, and to allow the retrieval of the third dimension. The latter crucial feature can be exploited directly by developing a complete 3D articulatory model, or by utilizing machine learning techniques to automatically restore the area function from the 2D sagittal model. It is this latter avenue that we will use in the future.

However, it is not guaranteed that the articulatory model derived by PCA from static MR images can capture vocal tract shapes of continuous speech. We have included in our database a sufficient number of phonetic contexts to cover most of the articulatory variability of the target language, here French. Despite these precautions it is possible that the blocked articulation of a consonant in a vocalic context, for example /k/ followed by the vowel /a/, does not reflect real dynamic articulations. Of all the articulators the tongue shape is most likely not to be approached correctly. Indeed, it has the greatest variability. Moreover, most consonants require an effort to make and maintain the contact between the tongue and the palate or teeth, during the MR acquisition.

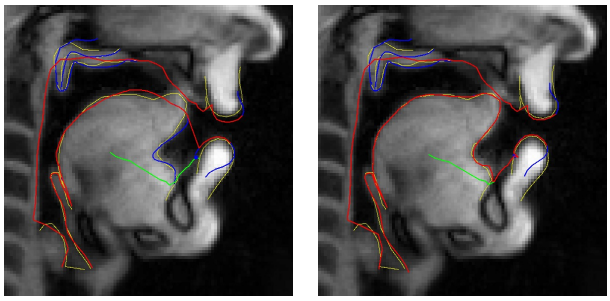

Fig. 6. Vocal tract contours reconstructed by the model constructed from static images (left), and by the model extended with dynamic images (right). The thin yellow lines are the contours delineated by hand. The red line is the reconstructed vocal tract used for acoustic simulations.

To investigate this question we randomly selected a series of 64 real-time MR images. All the articulator contours were carefully delineated by hand and the model constructed from images of static articulatory configurations was fitted to these contours. We found that the model approached most images correctly but failed for some images, especially those corresponding to the /l/ sound (see Fig.6). The average reconstruction error was 1.57 mm ($\sigma = 3.02mm$). Since the errors mainly concerned the tongue shape for /l/ sounds, we added 3 real-time MR images of /l/ for the vowels /i, a, u/ to the images used to build the model. The average reconstruction error then dropped to 1.33 mm ($\sigma = 2.52mm$).

V. Concluding remarks

The construction of an articulatory model is a key stage in the development of articulatory synthesis because it enables

ISBN 978-90-827970-1-5 © EURASIP 2018

the vocal tract shape to be defined with a small number of parameters. These parameters are associated to deformation modes of the speech articulators. The strategy of using centerline models accompanied by simple reconstruction algorithms turned out to guarantee that the deformation modes, corresponding to principal factors, are more relevant for the velum and epiglottis than those directly obtained by applying PCA to global contours. The second lesson of this work is that it is necessary to associate dynamic images with a static MRI image database in order to obtain a model that can correctly approach all vocal tract shapes that a speaker can produce. Even if the dynamic MR images are of a lower quality, they can cover shapes that are difficult to achieve by keeping articulators immobile.

ACKNOWLEDGMENT

This work is supported by an ANR Grant ArtSpeech to the Speech Group LORIA CNRS UMR7503 Nancy France 2015-2019.

REFERENCES

[1] P. Mokhtari, H. Takemoto, and T. Kitamura, "Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches," *Speech Commun.*, vol. 50, no. 3, pp. 179–190, Mar. 2008, ISSN: 0167-6393.

[2] B. Elie and Y. Laprie, "Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink," *Speech Communication*, vol. 82, pp. 85–96, Sep. 2016.

[3] B. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1249–1260, 1995.

[4] N. Lous, G. Hofmans, R. Veldhuis, and A. Hirschberg, "A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prosthesis design," *Acustica*, vol. 42, pp. 1135–1150, 1998.

[5] B. Elie and Y. Laprie, "Acoustic impact of the gradual glottal abduction on the production of fricatives: a numerical study," *Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. 1303–1317, Sep. 2017.

[6] ——, "Simulating alveolar trills using a two-mass model of the tongue tip," *Journal of the Acoustical Society of America*, vol. 142, no. 5, 2017.

[7] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marschal, Eds., Kluwer Academic Publishers, 1990.

[8] D. Beautemps, P. Badin, and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," *Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2165–2180, 2001.

[9] Y. Laprie and J. Busset, "A curvilinear tongue articulatory model," in *The Ninth International Seminar on Speech Production - ISSP'11*, Canada, Montreal, 2011.

[10] C. Ericsdotter, "Detail in vowel area functions," in *Proc of the 16^{th} ICPhS*, Saarbrücken, Germany, 2007, pp. 513–516.

[11] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[12] M. Nolden, S. Zelzer, A. Seitel, D. Wald, M. Müller, A. M. Franz, D. Maleike, M. Fangerau, M. Baumhauer, L. Maier-Hein, K. H. Maier-Hein, H. .-.-P. Meinzer, and I. Wolf, "The medical imaging interaction toolkit: Challenges and advances," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 607–620, Jul. 2013.

[13] Y. Laprie and J. Busset, "Construction and evaluation of an articulatory model of the vocal tract," in *19th European Signal Processing Conference - EUSIPCO-2011*, Barcelona, Spain, Aug. 2011.

[14] A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial FLASH with nonlinear inverse reconstruction," en, *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 477–485, Feb. 2013, ISSN: 07403194.

[15] Y. Laprie, B. Elie, and A. Tsukanova, "2D articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes," in *International Congress of Phonetic Sciences*, 2015.

[16] C. R. Rao, H. S. andC. R. Toutenburg, and C. Heumann, *Linear Models and Generalizations*. Springer, 2008.

[17] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Hecker, L. Ma, J. Busset, and J. Sturm, "DOCVACIM an X-ray database and tools for the study of coarticulation, inversion and evaluation of physical models," in *The Ninth International Seminar on Speech Production - ISSP'11*, Canada, Montreal, 2011.

[18] A. Laufer and I. D. Condax, "The function of the epiglottis in speech," *Language ad Speech*, vol. 21, no. 1, pp. 39–62, 1981.

[19] J. H. Esling, K. E. Fraser, and J. G. Harris, "Glottal stop, glottalized resonants, and pharyngeals: A reinterpretation with evidence from a laryngoscopic study of nuuchahnulth (nootka)," *Journal of Phonetics*, vol. 33, no. 4, pp. 383–410, 2005.