A Segmentation and Analysis Method for MRI Data of the Human Vocal Tract

Johannes Behrends[1], Phil Hoole[2], Gerda L. Leinsinger[1], Hans G. Tillmann[2], Klaus Hahn[3], Maximilian Reiser[1] and Axel Wismüller[1]

[1]Institut für Klinische Radiologie, Klinikum der Universität München, Ziemssenstrasse 1, 80336 München
[2]Institut für Phonetik und Sprachliche Kommunikation, Ludwig-Maximilians-Universität München, Schellingstrasse 3, 80799 München
[3]Klinik und Poliklinik für Nuklearmedizin, Klinikum der Universität München, Ziemssenstrasse 1, 80336 München

Abstract. MRI enables the in vivo analysis of the three-dimensional functional anatomy of the human vocal tract during phonation. For this purpose, MRI examinations are performed during phonation using different slice orientations. Subsequent anatomically correct registration enables a high-precision three-dimensional reconstruction of the vocal tract. Finally, a curvilinear midline is computed from which the three-dimensional functional anatomy of the human vocal tract can be approximated by a cascade of cylindric objects represented by their characteristic location-dependent cross-sectional areas ("area function").

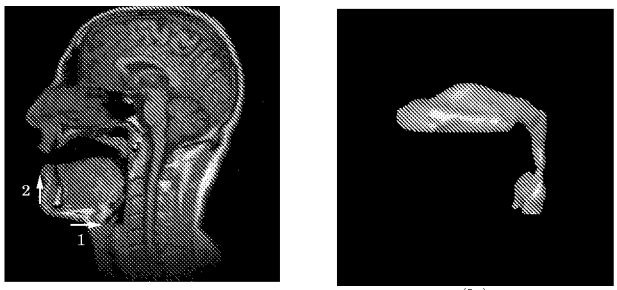
1 Introduction

Obtaining articulatory-acoustic models requires detailed knowledge of the three-dimensional geometry of the human vocal tract. Since most models are based on one-dimensional wave propagation, the vocal tract can be approximated as a tube consisting of a finite number of "stacked" cylindrical area elements from the glottis to the mouth opening. This model can be obtained by determination of intersectional areas of the vocal tract along a *midline* as a function of distance from the glottis. Thus, a particular vocal tract shape can be described by its so-called *area function.*

In early studies of the 1960's and 1970's such models were based on X-ray images and vocal tract impressions [1], [2]. The importance of MRI increased in the last ten years [3], [4], [5], [6] with the aim of achieving more precise articulatory models, i.e. area functions need not be estimated from a midsagittal projection, but can be obtained directly from three-dimensional image data.

This work deals with the segmentation of the human vocal tract and the generation of its area function. The initial segmentation step is performed by three-dimensional region growing (sec. 3). Subsequently, a curved vocal tract midline is computed not only for a midsagittal slice but for the whole three-dimensional data set based on a modified one-dimensional self-organizing feature map approach (sec. 4).

Fig. 1. (a): Midsagittal slice, vowel /a/; (b): Three-dimensional surface-rendered vocal tract shape.



(a) (b)

2 Image Data

Three-dimensional MRI data were acquired from nine healthy professional speakers (eight male, one female), aged 22 to 34 years. A standardized MRI sequence protocol (Siemens^{TM} Vision 1.5 T, T1w FLASH[1], TR=11.5ms, TE=4.9ms) was used. The scans were obtained in three different slice orientations, i.e. in axial, coronal (matrix size $256 \times 256 \times 23$, resolution $1.172 \times 1.172 \times 4\text{mm}^3$), and sagittal (matrix size $256 \times 256 \times 13$, resolution $1.172 \times 1.172 \times 4\text{mm}^3$) planes each in order to improve subsequent software-based three-dimensional analysis of the data sets. The subjects performed prolonged emission of sounds of the German phonetic inventory (vowels /i/, /y/, /u/, /e/, /a/, /o/, /ø/, (post-) alveolar consonants /s/, /sh/, /n/, /l/, and the dental /t/). Audio tape recording two seconds before and during imaging was obtained to control the correctness of the utterances. The dental /t/ could be prolonged during measurement by leaving out the burst.

From each subject, dental impressions were taken. These were scanned by a computer tomography (CT) in order to get three-dimensional data of the teeth with high resolution (matrix size $512 \times 512 \times 200$, resolution $0.156 \times 0.156 \times 0.3\text{mm}^3$) without X-ray exposure of the subjects themselves.

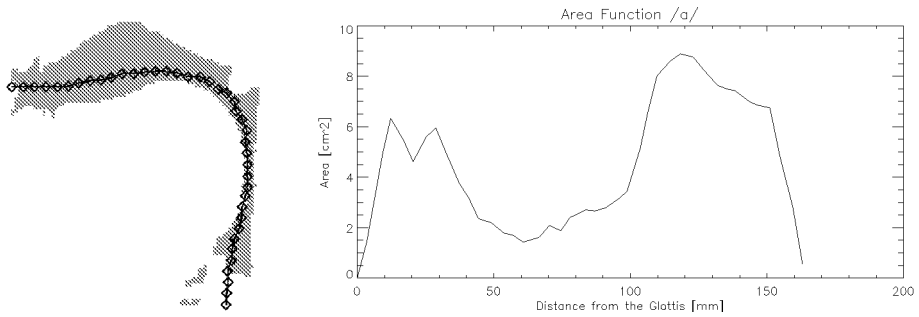
Interactive registration of the teeth phantoms and the MRI data sets was performed on a Picker^{TM} VoxelQ VX workstation, all other computations on a Linux personal computer with an Intel^{TM} Pentium III 900 MHz Processor in Interactive Data Language (IDL) from Research Systems Inc. (RSI^{TM}).

3 Segmentation

The goal of the segmentation process is to generate a vocal tract shape which is completely separated from its surrounding tissue. In other words, we want to obtain binary masks $M \in \{0, 1\}$ of a MRI data set X, representing the vocal

[1] Fast Low Angle SHot.

Fig. 2. (a): Midline through the vocal tract (underlying the midsagittal slice), vowel /a/; (b): resulting area function.



(a) (b)

tract. Fig. 1a shows a midsagittal slice of the human skull. The vocal tract shape is extracted from the glottis (arrow 1) to the mouth opening (arrow 2). Fig. 1b shows the segmentation result as a three-dimensional surface rendered image.

The problem of direction-specific low spatial resolution due to voxel anisotropy was solved by "Automated Image Registration" [7] of the data sets acquired for each phoneme in three slice orientations. Thus, a synthesized high-resolution data set could be obtained which served as the basis for further reconstruction and analysis of the vocal tract.

Since it is desirable to perform segmentation with least possible human intervention and computational expense, we used three-dimensional region growing as in [5] to solve this segmentation problem. However, there are several major problems in vocal tract segmentation: (i) As teeth and the hard palate can hardly be distinguished from air within the vocal tract, region growing would leak into these anatomical structures. (ii) The vocal tract has to be separated from the air outside the body preventing region growing from leaking outside the head region into the extracranial air, which could occur due to the open mouth during phonation. (iii) Region growing can leak through the glottis into the trachea.

(i) was solved by imaging and registration of dental impressions of the subjects as described in sec. 2. The problem of closing the mouth opening can be solved by convolution of each slice of the MRI data set with an I-shaped kernel. After these preprocessing steps, head masks can be generated by three-dimensional region growing starting outside the head. Leakage problems towards the trachea can be prevented by setting a reference point at the bottom of the glottis, thus excluding all caudal voxels from further segmentation procedure.

In a last step, the vocal tract is segmented by three-dimensional region growing leading to the result of fig. 1b.

4 Computation of the vocal tract midline

As conventional two-dimensional midsagittal approaches to midline identification do not account for asymmetries of the vocal tract shape, they cannot provide a realistic evaluation of the functional anatomy during phonation. In order to overcome these problems a three-dimensional curvilinear midline is computed using a modified self-organizing map (SOM) approach [8] based on a one-dimensional topology [9] in which the vocal tract shape is considered as a data distribution in the three-dimensional geometric space.

To avoid over-folding of the codebook the SOM algorithm was modified by keeping the local neighborhood width σ_r of each neuron in the range of its critical value σ_r^c at which the over-folding occurs [8]. If we define $\alpha = |r' - r''|$ as the distance between the closest neuron r' and the second closest neuron r'' to the current data point, we observe topology violation if $\alpha > 1$. In this case, σ_r is increased locally by

$$\sigma_r := \max\left(\sigma_r, \alpha K \exp\left(-\frac{2(r-R)^2}{\alpha^2}\right)\right), \quad \text{where} \quad R = \frac{1}{2}(r' + r''), \qquad (1)$$

and K is an empirical factor.

For the construction of the final midline the resulting one-dimensional SOM chain C is used as an input for subsequent postprocessing steps comprising smoothing and extrapolation: (i) Smoothing of C by convolution with a kernel decreasing exponentially by neighborhood distance. As a result we obtain a smoothed codebook $\tilde{C}$. (ii) Extrapolation of $\tilde{C}$ in the direction of the glottis and the mouth opening, respectively, and resampling $\tilde{N}$ equidistant points $\tilde{P}$ on the resulting curve. (iii) Computation of normal vectors $\tilde{\boldsymbol{n}}_i$ or each point $\tilde{\boldsymbol{p}}_i$ in $\tilde{P}$ by $\tilde{\boldsymbol{n}}_i = \tilde{\boldsymbol{p}}_{i-1} - \tilde{\boldsymbol{p}}_{i+1}$. These normal vectors are perpendicular to an oblique section $\tilde{S}_i$ through the vocal tract. For the edge points $\tilde{\boldsymbol{p}}_1$ and $\tilde{\boldsymbol{p}}_{\tilde{N}}$, $\tilde{P}$ is extrapolated. (iv) Computation of $\tilde{\boldsymbol{q}}_i$ as the center of gravity of the corresponding corss-sectional area $\tilde{S}_i$ through the vocal tract. This results in a curve $\tilde{Q}$ which is again resampled by equidistant points. (v) Convolution of the curve $\tilde{Q}$ by applying step (i) leading to a smoothed midline Q.

With the smoothed normal vectors, we can easily obtain planes which contain vocal tract sections perpendicular to Q corresponding with p_i. A voxel counting algorithm yields the area function shown in fig. 2b.

5 Results

In all the volunteers, we accurately evaluated the structure of the lips, tongue, soft palate, and pharynx. While midsagittal slices were consistent with data acquired by electromagnetic midsagittal articulography (EMMA), the analysis of the posterior and lateral parts of the tongue root revealed quite complex shapes. A sharp groove was found for most phonemes, usually with considerable asymmetry about the midline. The depth of the groove (2–10mm) in relation to the distance between the tongue and the soft palate or the pharyngeal wall

(7–13mm) varied strongly. In several cases, the groove was so deep that it formed an essential part of the cross-sectional area (max 26%). As could be expected, the area functions for different phonemes revealed characteristic reproducible properties with only small inter-speaker variability.

6 Discussion

Fast MRI in different slice orientation followed by subsequent co-registration enables rapid and precise in vivo three-dimensional evaluation of the human vocal tract during phonation. Using this information, acoustic-articulatory models can be obtained by computer-assisted image analysis methods. In this context, the computation of a three-dimensional curvilinear midline through the vocal tract based on a modified self-organizing map approach accounts for asymmetries of the vocal tract shapes, thus improving area function results in comparison to conventional modeling by midsagittal two-dimensional analysis methods. This evaluation may contribute to the construction of speech synthesis methods and, in the long run, may serve as the initial step for the clinical diagnosis of inborn or surgery-caused abnormalities affecting the functional anatomy of the oro-pharyngeal tract during speech production.

References

1. Fant G: Acoustic Theory of Speech Production. Mouton, den Haag 1960.
2. Mermelstein P: Articulatory Model for the Study of Speech Production. Journal of the Acoustical Society of America 53(4):1070–1082, 1973.
3. Baer T, Gore JC, Gracco RC: Analysis of Vocal Tract Shape and Dimension using Magnetic Resonance Imaging: Vowels. JASA 90(2):799–828, 1991.
4. Narayanan SS, Alwan AA, Haker K: Towards Articulatory-Acoustic Models for Liquid Approximants based on MRI and EPG Data. JASA 101(2):1064–1089, 1995.
5. Titze I, Story B: Vocal Tract Area Functions from Magnetic Resonance Imaging. JASA 100(1):537–554, 1996.
6. Soquet A, Lecuit V: Segmentation of the Airway from the Surrounding Tissues on Magnetic Resonance Images: A Comparative Study. ICSLP, 1998.
7. Woods RP, Cherry SR, Mazziotta JC: Rapid automated algorithm for aligning and reslicing PET images. JCAT
16:620–633, 1992.
8. Der R, Herrmann M: Second-Order Learning in Self-Organizing Maps. Kohonen Maps. Oja E (Publisher) 1999.
9. Kohonen T: Self-Organizing Maps. Springer, Heidelberg 2001.