Forschungsberichte des
Instituts für Phonetik und
Sprachliche Kommunikation
der Universität München
(FIPKM) 37 (2001) 179-189

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

Johannes Behrends and Axel Wismüller

Institut für Klinische Radiologie
Ludwig-Maximilians-Universität München
Ziemssenstrasse 1
80336 München, Germany
johannes.behrends@radin.med.uni-muenchen.de
axel@wismueller.de

ABSTRACT

In this paper a method for three-dimensional modeling of the human vocal tract based on T1-weighted MRI data sets is introduced. This includes the reconstruction of the vocal tract shape and the computation of the vocal tract midline. Reconstruction of the vocal tract shape is realized by three-dimensional region growing within the vocal tract after teeth phantoms are matched into the MRI data set. The midline is obtained by using a modified 1d-Kohonen algorithm. Optimization and smoothing of the midline leads to computation of characteristic area functions. With those methods, human time expense could be reduced such that now a study of several subjects is possible.

1 INTRODUCTION

Obtaining articulatory-acoustic models requires detailed knowledge about three-dimensional consistency 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 determining sections 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*.

After obtaining these models based on X-ray images and vocal tract impressions (Fant, 1960; Mermelstein, 1973) in the 60’s and 70’s, the meaning of MRI increased in the last ten years (Baer, Gore,

Table 1: Acquisition parameters for the MRI data sets. Werte fuer TE und TR sind noch nachzutragen!

	Sagittal	Axial	Coronar
Magn. field strength [T]	1.5		
Number of slices	15	23	23
Matrix size	256×256		
Pixel size [mm]	1.172		
Slice Thickness [mm]	3	5	5
Interlice gap [mm]	1	0	0
TR [ms]	11.6		
TE [ms]	4.9		

Gracco, and Nye, 1991; Narayanan, Alwan, and Haker, 1995; Titze and Story, 1996; Soquet and Leciut, 1998) to get articulatory modeling more precise, i.e. area functions need not be estimated from a midsagittal projection but can be directly obtained from three-dimensional image data.

This work deals with the segmentation of the human vocal tract and the generation of its area function. For segmentation a simple algorithm based on three-dimensional region growing is introduced (sec. 3). The vocal tract midline is computed not only for a midsagittal slice but for the whole three-dimensional data set. This is done based on one-dimensional self-organizing feature maps (sec. 4).

2 IMAGE DATA

Three-dimensional MRI data were acquired from 25 healthy professional speakers (13 male, 12 female), aged 22 to 34 years. A standardized MRI sequence protocol (Siemens^{TM} Vision 1.5 T, T1w FLASH[1], see table 2) was used. Scans were made in axial, coronal, and sagittal planes each in order to improve subsequent software-based 3d analysis of the data sets. The subjects prolonged emission of sounds of the German phonetic inventory (vowels /i/, /y/, /u/, /e/, /a/, and (post-) alveolar consonants /s/, /sh/, /n/, /l/, /t/). Simultaneous audio tape recording was obtained for professional control of correct utterance.

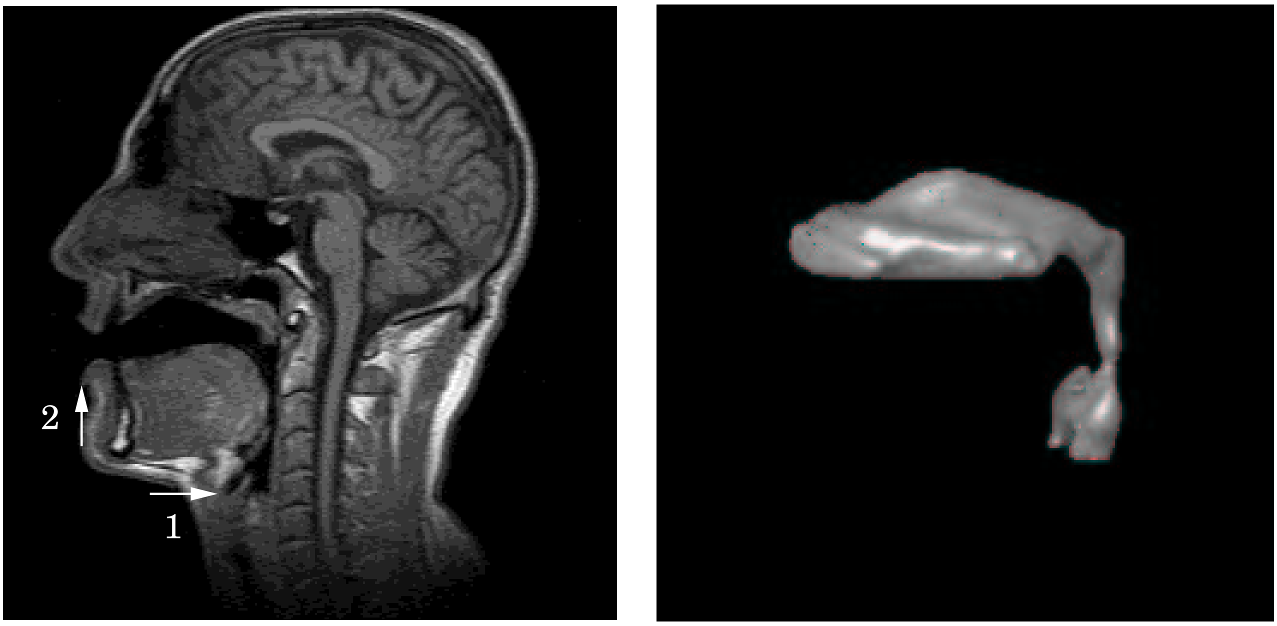
From each subject, teeth impressions were taken. These are scanned in a Computertomograph (CT) in order to get three-dimensional data of the teeth with a high resolution (see table 2) without exposing the subjects themselves to CT radiation (see 3).

The methods explained in the following are tested on one male subject aged 34 years from this group using the sagittal MRI data.

Registration of the teeth phantoms was done on a Picker VoxelQ VX workstation. All other computations were made 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}).

Table 2: Acquisition parameters for the teeth phantom data set.

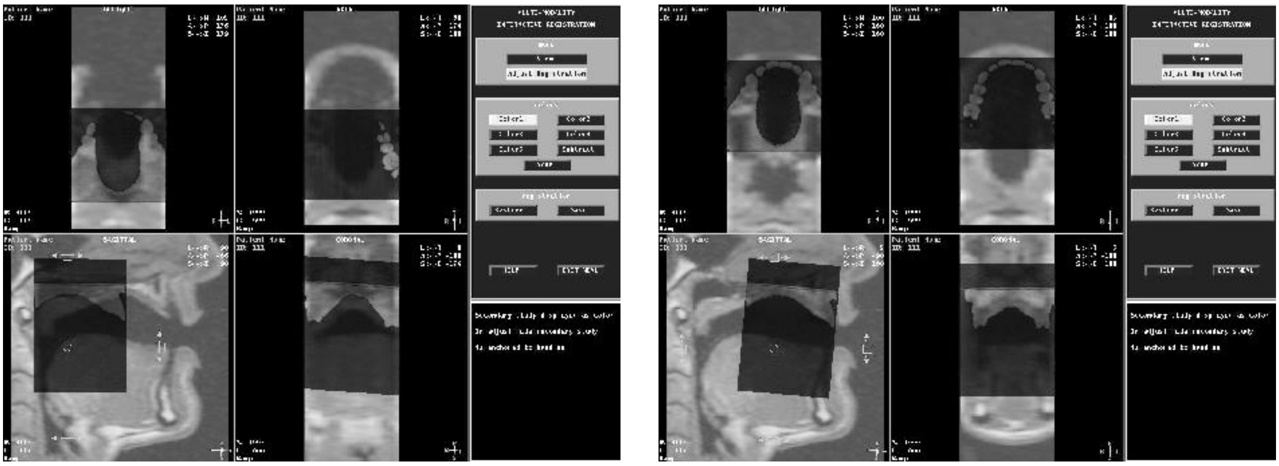
Matrix size	512×512
Pixel size [mm]	0.156
Slice thickness [mm]	0.3



(a)

(b)

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



(a) (b)

Figure 2: Insertion of the teeth phantom data into the anatomic data set; (a) before, (b) after image registration on a Picker VoxelQ VX.

3 Segmentation

Goal of the segmentation process is to generate a vocal tract shape which is completely separated from its surrounding tissue. That means, we want to obtain binary masks $M \in \{0, 1\}$ of a MRI data set X, in which the voxels containing the vocal tract hold the value one, else zero. Fig. 1a shows a midsagittal slice of the human skull. Vocal tract shape is extracted from the glottis (vertex 1) to the mouth opening (vertex 2). The segmentation result can be seen in fig. 1b as a 3d surface rendered image.

It is desirable to perform segmentation with least possible human and computational work. Slice-based segmentation methods like manual contour or edge tracing or even two-dimensional region growing (Soquet and Leciut, 1998) require much human effort and are thus very time-consuming. So, we think that 3d region growing as in (Titze and Story, 1996) is best suited to solve this segmentation problem. However, there are four major problems in vocal tract segmentation:

1. Teeth are only poorly imaged in MRI which causes a gray value like air (black) in the image data. This would cause region growing to leak into the teeth region.

2. The hard palate is also poorly imaged for the same reason as in 1. Thus, three-dimensional region growing would leak into the nasal tract.

3. The vocal tract area has to be non-continuous with air outside the body. Due to the open mouth during phonation, region growing would leak outside the head region and include the whole extracranial air.

4. In some cases region growing can leak through the glottis into the trachea and thus include extracranial air as well.

The first and the second problem were solved by taking teeth impressions from the subjects, taking a separate computer-tomographic (CT) scan of the impressions and inserting these ”teeth phantoms” into the MRI dataset by manual image registration (see fig. 2). This method has two advantages which

[1] Fast Low Angle SHot.

simplify the registration process: (i) teeth data can be extracted by a simple thresholding algorithm and (ii) the impressions of upper and lower jaw can be scanned and inserted, separately. Because the teeth impression of the upper jaw also contains the hard palate, the vocal tract is now separated from the nasal tract.

The problem of closing the mouth opening can be solved by undergoing each slice of the MRI data set a 2d convolution with an I-shaped kernel K frontal from a reference point which is set dorsal to the lips (see vertex 1 in fig. 3a):

$$X_c = K * X, \; K = \begin{pmatrix} 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{n,n}, \; n < x, y, \tag{1}$$

where x and y denote the slice's x- and y-dimension, respectively and n denotes the kernel size.

In a next step, masks can be generated by 3d region growing marking a point within the lips of the convoluted image as an upper threshold and a point in the extracranial air as a lower threshold. In the resulting head masks $H \in \{0, 1\}$ all voxels outside the head within the convolved region obtain the value one.

Leakage problems towards the trachea can be prevented by setting a reference point at the bottom of the glottis and defining all voxels caudal from it to the value one in the mouth opening mask (vertex 2 in fig. 3a). The anatomical image within the resulting masks is shown in fig. 3b.

The zero region inside the head can be filled out with the thresholded image of the corresponding slice of the MRI data set containing the teeth (see fig. 3c). Now, all voxels inside the vocal tract have the value zero and the vocal tract itself is almost completely separated from surrounding zero regions. For checking purposes, all slices are supervised and, if necessary, leakages can be closed.

In the last step, 3d region growing is performed on the vocal tract which leads to the result of fig. 3d.

4 Computation of the vocal tract midline

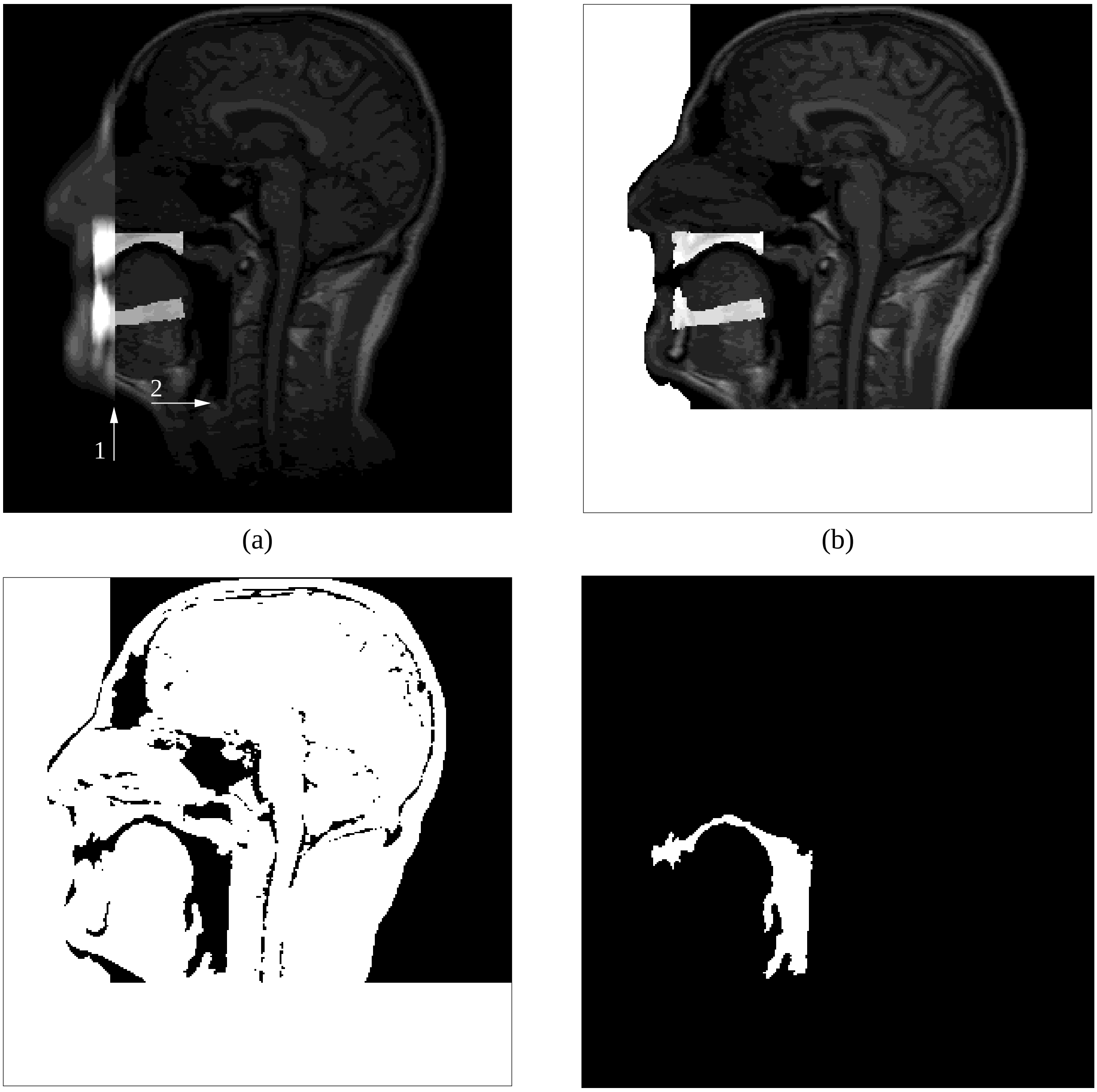
After obtaining a 3d data set of the vocal tract, its midline can be calculated. This was done using a modified Kohonen algorithm like in (Der and Herrmann, 1999) for tracing the midline not only on the midsagittal slice but on the whole 3d data.

Consider the vocal tract data as a distribution in the three-dimensional geometric space. Then we obtain feature vectors $\mathbf{x}$ containing the x-, y- and z-coordinate for each voxel having the value one. On this data distribution, Kohonen's vector quantization algorithm (Kohonen, 1995) can be performed using a one-dimensional network topology (a *Kohonen chain*) consisting of m codebook vectors $\mathbf{w}_r$ with the iteration

$$\Delta \mathbf{w}_r = \epsilon h(r, r^*)(\mathbf{x} - \mathbf{w}_r), \tag{2}$$

where $\mathbf{v} \in \mathbb{R}$ is the input vector and $h(r, r^*)$ denotes the neighbor function depending on the distance from neuron r to the winner neuron r^*:

$$h(r, r^*) = \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{(r - r^*)^2}{2\pi\sigma_r^2}\right), \tag{3}$$



(a)

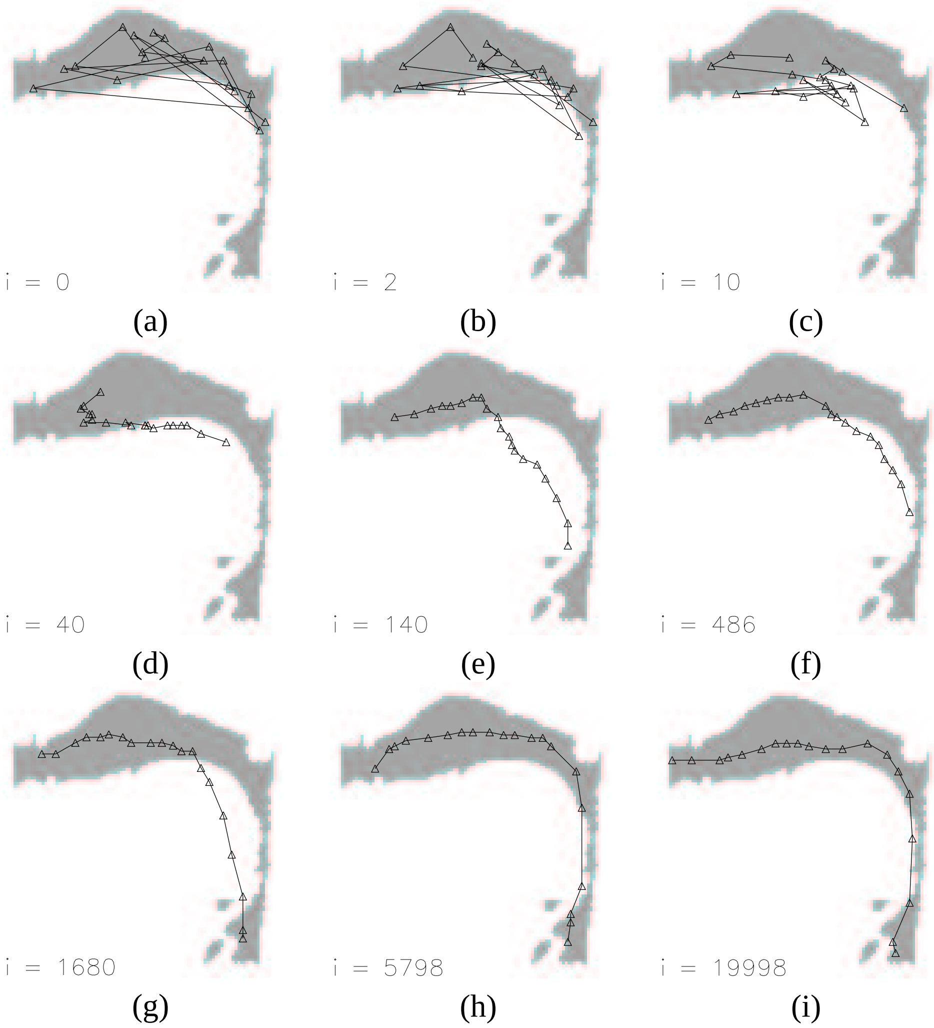
(b)

(c)

(d)

Figure 3: Illustration of the segmentation procedure; (a) convolution ventral from vertex 1 according to (1) and region limitation caudal from vertex 2; (b) resulting mask after region growing; (c) thresholded anatomic image added to the mask from (b); (d) extracted vocal tract after 3d region growing.

where ϵ is an annealing parameter and $\sigma_r(t)$ is the value of the neighborhood width, both decreasing with the time.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 4: Evaluation of the Kohonen chain within the vocal tract mask.

If we consider the Kohonen chain as the principal curve through a data distribution the goal is to adjust σ_r such that the curve fits the distribution in an optimal manner. Usually, σ_r is defined globally for all neurons r. However, below a critical value σ_r^c, the network topology will break down. This results in an over-fitting of the chain towards the data distribution which means

$$\alpha > 1 \quad \text{for} \quad \alpha = |r' - r''|, \tag{4}$$

where r' is the first and r'' is the second closest neuron to the current data point. In other words, if (4) is fulfilled σ_r has fallen below its critical value. This local onset of the phase transition to the over-fitting case makes it necessary to find convenient values of σ_r locally.

Der's approach is to keep σ_r fluctuating around its critical value σ_r^c. For each iteration, in the first step, σ_r is decreased by

$$\Delta\sigma_r = -\frac{1}{NT_\sigma}\sigma_r \quad \forall r, \tag{5}$$

where N is the number of neurons and T_σ denotes the adjustment parameter.

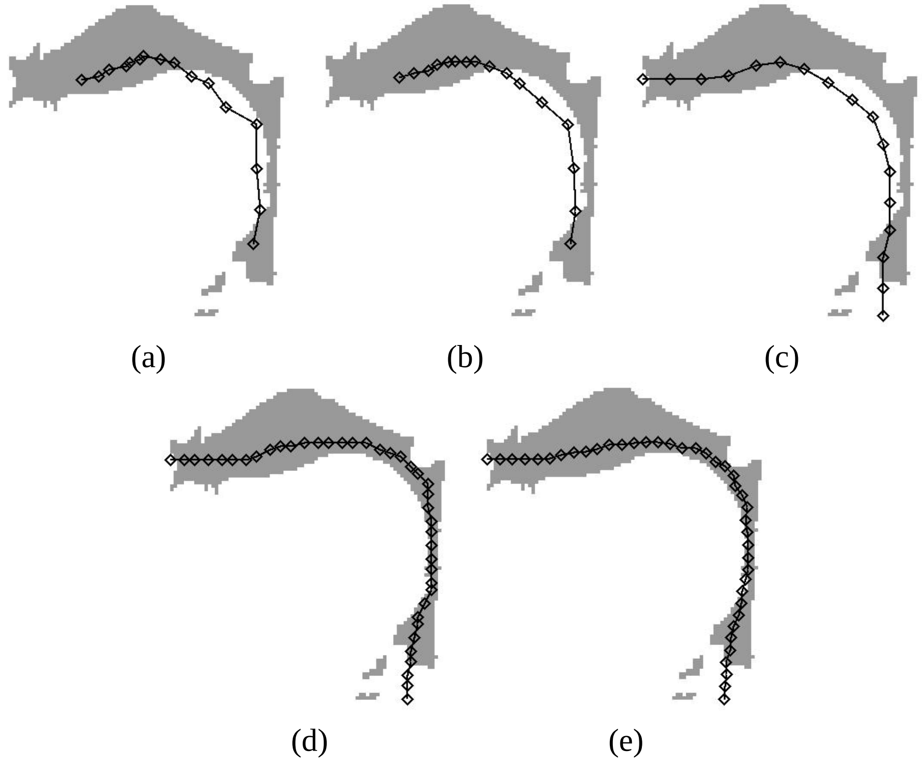


Figure 5: Optimizing the main axis generated by a 1d Kohonen chain: (a) polygon of the native kohonen chain projected on the sagittal plain; (b) after smoothing the initial codebook; (c) after extrapolating the polygon and setting equidistant points onto it; (c) points translated into the center of gravity within its oblique vocal tract section; (d) after translating the points from (c) into the center of gravity within its oblique vocal tract section and resampling to a larger number of equidistant points; (e) final midline after smoothing thr points of (d).

In the second step, σ_r is increased locally if over-fitting is recognized:

$$\sigma_r \;:=\; \max\left(\sigma_r, \alpha K \exp\left(-\frac{2(r-R)^2}{\alpha^2}\right)\right), \quad \text{where} \quad R = \frac{1}{2}(r' + r''), \tag{6}$$

where K is an empirical factor (for details see (Der and Herrmann, 1999)). The result of this method applied on the midline finding problem is shown in fig. 4. After contraction the codebook is iteratively distributed through the vocal tract shape and serves as a basis for generating the area function.

4.1 Determining the Area Function

In our application, we found it the best to perform the modified Kohonen algorithm with only a small number of N codebook vectors. However, the representation of the resulting midline does not suffice for generating an area function because (i) the codebook vectors are not equidistant and (ii) the Kohonen chain does not always lie in areas of thin data distribution density (see fig. 4i). Furthermore, it is desireable to use one parameter set for the image data of all phonemes in order to save time, such that the codebook initially could look like in fig. 5a. Hence, the Kohonen chain C is used as initialization for following optimization algorithm:

1. Smooth C by convolution with a kernel which decreases exponentially by neighborhood distance to the smoothed codebook $\tilde{C}$.

2. Extrapolate $\tilde{C}$ to the point next to the glottis and to the point next to the mouth opening, respectively, and set $\tilde{N}$ equidistant points $\tilde{P}$ in the polygon (see fig. 5b).

3. For each point $\tilde{\mathbf{p}}_i$ in $\tilde{P}$, calculate a normal vector by $\tilde{\mathbf{n}}_p = \tilde{\mathbf{p}}_{i-1} - \tilde{\mathbf{p}}_{i+1}$ which is perpendicular to an oblique section $\tilde{S}_i$ through the vocal tract. For the edge points $\tilde{\mathbf{p}}_1$ and $\tilde{\mathbf{p}}_N$, $\tilde{P}$ is extrapolated.

4. For each oblique section $\tilde{S}_i$, translate $\tilde{\mathbf{p}}_i$ into the center of gravity within its vocal tract area and resample $\tilde{P}$ with $M > \tilde{N}$ equidistant points (see fig. 5c).

5. Convoluting the curve $\tilde{P}$ by applying step 1 leads to a smoothed midline P.

From each point $\mathbf{p}_i \in P$, a normal vector can be calculated. If we perform step 3 with $M > \tilde{N}$ we would get overlap of the oblique sections which could result in undesirable interchange of the codebook vectors . This is avoided by smoothing the normal vector using a neighborhood relation:

$$\mathbf{n}_i = \sum_{j=1}^{k} \mathrm{e}^{-c \cdot j} \frac{p_{i-j} - p_{i+j}}{||p_{i-j} - p_{i+j}||}, \tag{7}$$

where k denotes the neighborhood width and c is a smoothing factor.

With the smoothed normal vectors, we can easily get plains S_i which contain vocal tract sections corresponding with p_i. A voxel counting algorithm leads to the desired area function shown in fig. 6.

5 CONCLUSION AND OUTLOOK

Two methods were introduced which decrease the interactive effort of building an articulatory model of the human vocal tract.

The segmentation method of (Titze and Story, 1996) was improved by inserting the teeth into the MRI data *before* segmentation and by closing the mouth opening using a two-dimensional convolution approach.

The use of a modified Kohonen algorithm to calculate the vocal tract midline has the advantage of obtaining it in all three dimensions without setting reference points, manually.

These methods can serve as a basis for further research of MRI data sets from a larger group (see section 2). Such a wide-area study can lead to interesting conclusions with respect to standardized vocal tract shapes for each phoneme which could eventually improve the model of (Mermelstein, 1973).

Of course, these methods can further be enhanced, e. g. the segmentation method by reference point- and grey value-based feature extraction and voxel-based classification by a neural network using radial basis functions (Wismüller, Vietze, and Dersch, 2000, Behrends, 2000).

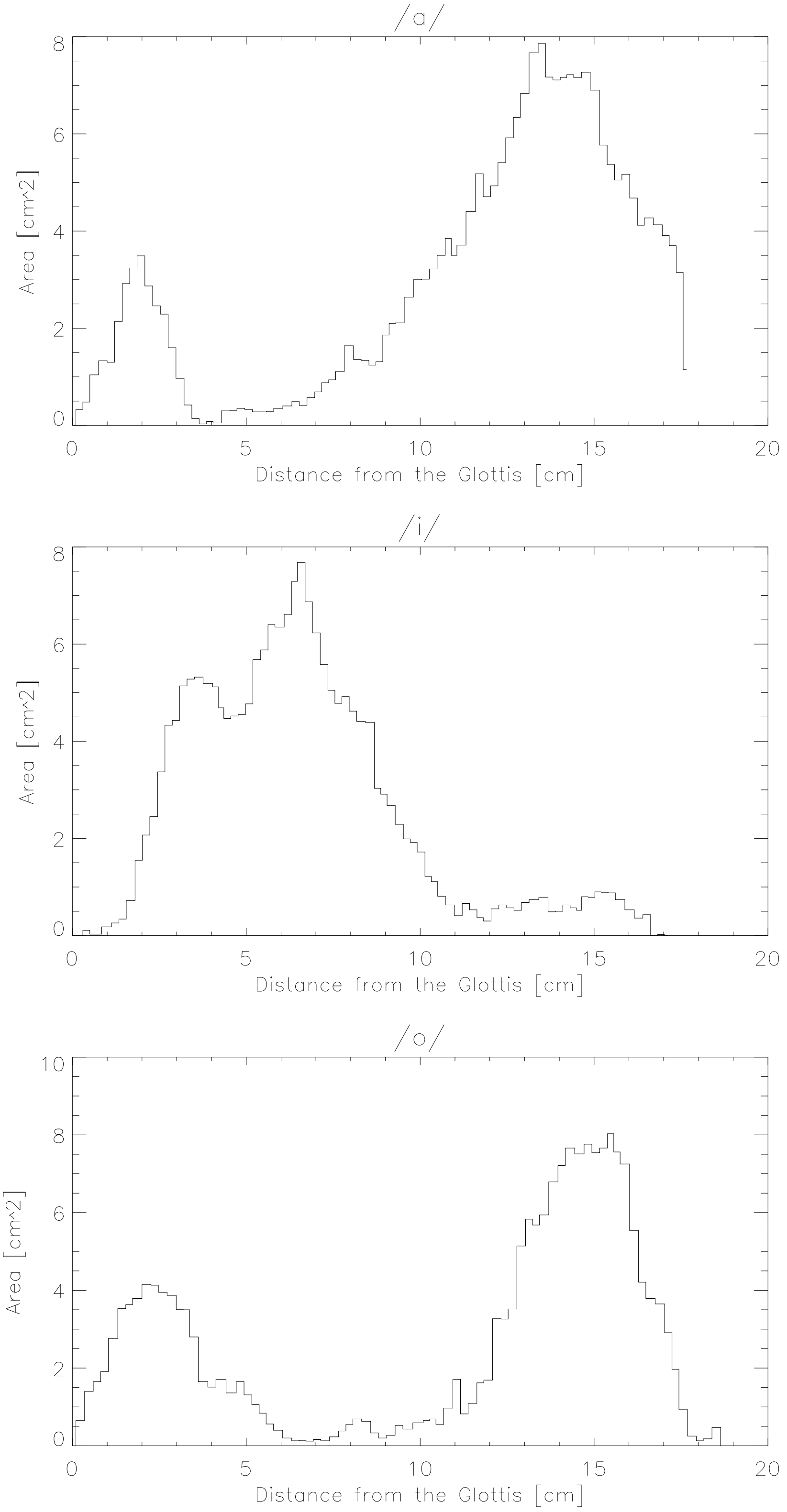


Figure 6: Area functions for the vocals /a/, /i/ and /o/, respectively.

References

T. Baer, J. C. Gore, R. C. Gracco, and P. W. Nye. Analysis of Vocal Tract Shape and Dimension using Magnetic Resonance Imaging: Vowels. *Journal of the Acoustical Society of America*, 90 (2): 799–828, 1991.

J. Behrends. Automatische Randkontur- und Gewebesegmentierung kernspintomographischer Datensätze des menschlichen Gehirns. Master’s thesis, Technische Universität Hamburg Harburg, 2000.

R. Der and M. Herrmann. Second-Order Learning in Self-Organizing Maps. In *Kohonen Maps*. E. Oja, S. Kaski, 1999.

G. Fant. *Acoustic Theory of Speech Production*. Mouton, den Haag, 1960.

T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Heidelberg, 1995.

P. Mermelstein. Articulatory Model for the Study of Speech Production. *Journal of the Acoustical Society of America*, 53 (4):1070–1082, 1973.

S. S. Narayanan, A. A. Alwan, and K. Haker. Towards Articulatory-Acoustic Models for Liquid Approximants Based on MRI and EPG Data. *Journal of the Acoustic Society of America*, 101 (2): 1064–1089, 1995.

A. Soquet and V. Leciut. Segmentation of the Airway from the Surrounding Tissues on Magnetic Resonance Images: A Comparative Study. In *International Conference of Spoken Language Processing*, 1998.

I. Titze and B. Story. Vocal Tract Area Functions from Magnetic Resonance Imaging. *Journal of the Acoustical Society of America*, 100 (1):537–554, 1996.

A. Wismüller, F. Vietze, and D. Dersch. Segmentation with neural networks. In I. Bankman, R. Rangayyan, A. Evans, R. Woods, E. Fishman, and H. Huang, editors, *Handbook of Medical Imaging*, Johns Hopkins University, Baltimore, 2000. Academic Press. ISBN 0120777908.