Human Vocal Tract Analysis by in Vivo 3D MRI during Phonation: A Complete System for Imaging, Quantitative Modeling, and Speech Synthesis

Axel Wismueller[1], Johannes Behrends[1], Phil Hoole[2],
Gerda L. Leinsinger[3], Maximilian F. Reiser[3], and Per-Lennart Westesson[1]

[1] Department of Imaging Sciences and Department of Biomedical Engineering,
University of Rochester, New York,
601 Elmwood Avenue, Box 648, Rochester, NY 14642-8648, U.S.A.
axel_wismueller@urmc.rochester.edu
[2] Department of Phonetics, University of Munich,
Schellingstrasse 3, 80799 Munich, Germany
[3] Department of Radiology, University of Munich,
Ziemssenstrasse 1, 80336 Munich, Germany

Abstract. We present a complete system for image-based 3D vocal tract analysis ranging from MR image acquisition during phonation, semi-automatic image processing, quantitative modeling including model-based speech synthesis, to quantitative model evaluation by comparison between recorded and synthesized phoneme sounds. For this purpose, six professionally trained speakers, age 22-34y, were examined using a standardized MRI protocol (1.5 T, T1w FLASH, ST 4mm, 23 slices, acq. time 21s). The volunteers performed a prolonged ($\geq$21s) emission of sounds of the German phonemic inventory. Simultaneous audio tape recording was obtained to control correct utterance. Scans were made in axial, coronal, and sagittal planes each. Computer-aided quantitative 3D evaluation included (i) automated registration of the phoneme-specific data acquired in different slice orientations, (ii) semi-automated segmentation of oropharyngeal structures, (iii) computation of a curvilinear vocal tract midline in 3D by nonlinear PCA, (iv) computation of cross-sectional areas of the vocal tract perpendicular to this midline. For the vowels /a/,/e/,/i/,/o/,/ø/,/u/,/y/, the extracted area functions were used to synthesize phoneme sounds based on an articulatory-acoustic model. For quantitative analysis, recorded and synthesized phonemes were compared, where area functions extracted from 2D midsagittal slices were used as a reference. All vowels could be identified correctly based on the synthesized phoneme sounds. The comparison between synthesized and recorded vowel phonemes revealed that the quality of phoneme sound synthesis was improved for phonemes /a/, /o/, and /y/, if 3D instead of 2D data were used, as measured by the average relative frequency shift between recorded and synthesized vowel formants ($p<0.05$, one-sided Wilcoxon rank sum test). In summary, the combination of fast MRI followed by subsequent 3D segmentation and analysis is a novel approach to examine human phonation in vivo. It unveils functional anatomical findings that may be essential for realistic modelling of the human vocal tract during speech production.

D. Metaxas et al. (Eds.): MICCAI 2008, Part II, LNCS 5242, pp. 306–312, 2008.

1 Introduction

This work focusses on image-based analysis of the vocal tract analysis from *in vivo* 3D MRI data. Although current limitations of temporal resolution may still impede an immediate clinical usage, it can be expected that MRI of the vocal tract during phonation, in the long run, will play an important role for clinical applications to the diagnosis of speech disorders, such as congenital ones or those acquired due to sugery within the vocal tract region. Here, objective quantitative imaging may be helpful for guiding logopedic training procedures of the speech apparatus. Finally, it can be speculated that the quality of speech synthesis may be improved based on a better understanding of the functional vocal tract geometry during phonation based on quantitative acoustic-articulatory modeling. In this work, we present a complete system for image-based vocal tract analysis ranging from MR image acquisition during phonation, semi-automatic image processing, quantitative modeling including model-based speech synthesis, and quantitative model evaluation by comparison between recorded and synthesized vowel sounds.

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.

A key element of our system is the proper segmentation of the human vocal tract and the generation of an area function derived from MRI data acquired in vivo. The initial segmentation step is performed by three-dimensional region growing. 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. Finally, the result is used to synthesize vowel sounds based on an approximate numeric solution of the wave propagation equation for the calculated area function. For quality evaluation of our 3D versus conventional 2D based speech synthesis, recorded and synthesized sounds are compared quantitatively by computing relative frequency differences of the first three formants between audio-recorded and synthesized speech signals, and differences are tested for statistical significance.

2 Methods

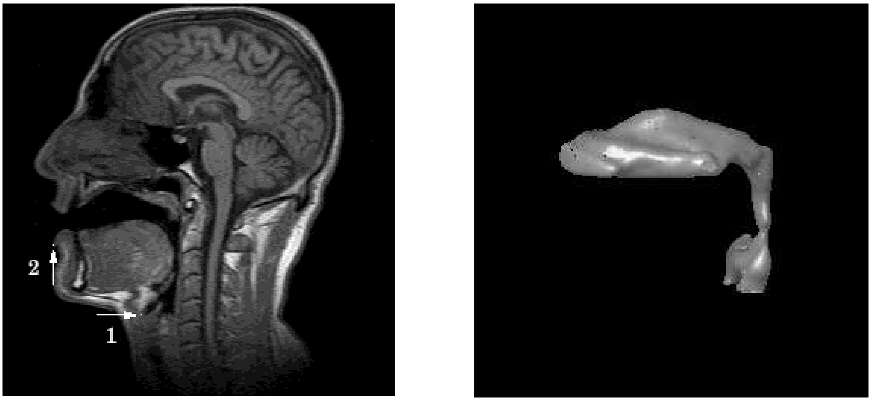
2.1 Image Acquisition

Three-dimensional MRI data were acquired from six healthy professional speakers (four male, two female), aged 22 to 34 years. A standardized MRI sequence protocol

(Siemens VisionTM 1.5 T, T1w Fast Low Angle Shot (FLASH), 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 x 256, 13 slices, resolution 1.172 x 1.172 x 4 mm3), and sagittal (matrix size 256 x 256, 13 slices, resolution 1.172 x 1.172 x 4 mm3) planes each, in order to improve subsequent software-based three-dimensional analysis of the data sets. The acquisition time for each slice orientation was 21 s. 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 computer tomography (CT) in order to get three-dimensional data of the teeth with high resolution (matrix size 512 x 512 x 200, resolution 0.156 x 0.156 x 0.3 mm3) without X-ray exposure of the subjects themselves.

Interactive registration of the teeth phantoms and the MRI data sets was performed on a PickerTM VoxelQ VX workstation, all other computations on a Linux personal computer in *Interactive Data Language (IDL)* from *Research Systems Inc.* (*RSI*^{TM}).



(a) (b)

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

2.2 Segmentation

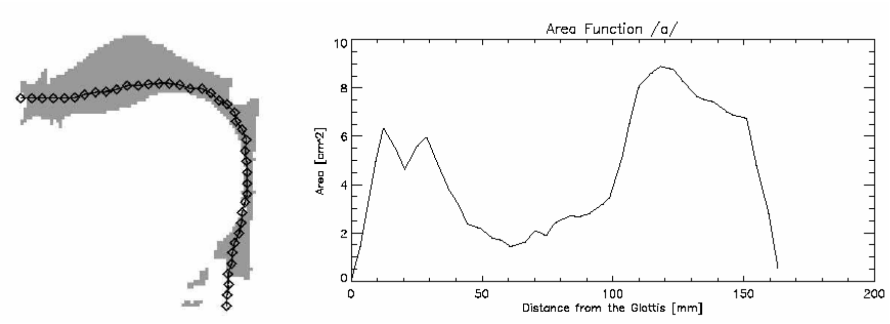
The goal of the segmentation process is to generate a vocal tract shape which is completely separated from its surrounding tissue. In other words, we want to obtain binary masks $M \in \{0,1\}$ of an MRI data set X, representing the vocal tract. Fig. 1a shows a midsagittal slice of the human skull. The vocal tract shape is extracted from the glottis (arrow 1) to the mouth opening (arrow 2).

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.

Problem (i) was solved by imaging and registration of dental impressions of the subjects as described in sec. 1. 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, showing the segmentation result as a three-dimensional surface rendered image.



(a) (b)

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

2.3 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

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

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 including 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 $\tilde{n}_i$ of each point $\tilde{p}_i$ in $\tilde{P}$ by $\tilde{n}_i = \tilde{p}_{i-1} - \tilde{p}_{i+1}$. These normal vectors are perpendicular to an oblique section $\tilde{S}_i$ through the vocal tract. For the edge points $\tilde{p}_1$ and $\tilde{p}_{\tilde{N}}$, $\tilde{P}$ is extrapolated. (iv) Computation of $\tilde{q}_i$ 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 represent vocal tract cross-sections perpendicular to Q corresponding with p_i. A voxel counting algorithm yields the area function shown in fig. 2b.

2.4 Speech Synthesis and Quantitative System Evaluation

The resulting area function was used to synthesize vowel sounds based on an acoustic-articulatory model according to [10]. For quantitative system evaluation, we ompared the quality of speech synthesis from 3D vs. 2D MRI data analysis by computing relative frequency differences of first three formants between audio-recorded and synthesized speech signals. For this purpose, we followed both a (i) *vowel-specific* analysis direction evaluating the statistical significance of the improvement over all vowels, for each formant individually and over all formants, and a (ii) *speaker-specific* direction evaluating the statistical significance of improvement over all speakers, for each formant individually and over all formants. For statistical comparisons, one-sided matched-pairs Wilcoxon rank sum tests with a significance level of $p<0.05$ were used.

3 Results

For quantitative analysis according to section 2.4, the extracted area functions for the vowels /a/,/e/,/i/,/o/,/ø/,/u/,/y/ were used to synthesize phoneme sounds based on an articulatory-acoustic model [10]. The recorded and synthesized phonemes were compared, where area functions extracted from 2D midsagittal slices were used as a reference. All vowels could be identified correctly based on the synthesized phoneme sounds. The comparison between synthesized and recorded vowel phonemes according to the statistical analysis of section 2.4 revealed that the quality of phoneme sound synthesis was improved for (i) *some* phonemes, namely /a/, /o/, and /y/, and for (ii) *some* speakers (2 out of 8), if 3D instead of 2D data were used. For the other vowels and speakers, a statistically significant improvement of using 3D data could not be confirmed.

In addition to this quantitative analysis of speech synthesis quality, we qualitatively evaluated the structure of the lips, tongue, soft palate, and pharynx in all the volunteers. 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 – 10 mm) in relation to the distance between the tongue and the soft palate or the pharyngeal wall (7 – 13 mm) 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.

Despite these notable findings regarding the considerable vocal tract midline asymmetry depicted by our 3D approach, it is interesting that our quantitative analysis of model-based speech synthesis revealed only a moderate, yet statistically significant improvement of modeling quality. It may be concluded that 3D modeling provides some advantage over conventional 2D modeling of the vocal tract, i.e. for *some* phonemes and *some* speakers. On the other hand, 2D modelling may be sufficient in many situations, i.e. 2D models based on fast 2D imaging modalities, such as midsagittal MRI or EMMA, may serve as a first approximation.

Here, our results are in accordance with a hypothesis stated by [5] that small variations of the vocal tract midline may not have a substantial overall effect on speech synthesis quality. Our study, for the first time, supports this hypothesis by empirical data.

However, as shown above, 3D acoustic-articulatory modeling may have advantages for specific phonemes or speakers. To systematically explore the relation between the above mentioned midline asymmetries and their effect on speech synthesis quality is a priority issue for future work.

4 Conclusion

We have developed a system for functional analysis of in vivo 3D MRI data of the human vocal tract during phonation. Here, fast MRI in different slice orientation followed by subsequent co-registration allows for 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, which may improve area function results in comparison to conventional modeling by midsagittal 2D analysis methods. We have quantitatively evaluated this system by comparing audio-recorded and synthesized speech signals, where speech synthesis was based on extracting an area function from the in vivo 3D MRI data, serving as input to an acoustic-articulatory model. We could quantitatively demonstrate from our data that, for specific vowels, the quality of phonetic modelling as tested by speech synthesis, could be improved significantly, by using 3D instead of 2D data. However, this could not be observed as a general result valid for all vowels and all speakers. We conjecture that our results can contribute to the discussion whether 3D data acquisition can improve the quality of acoustic-articulatory modeling.

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, J.C., Gracco, R.C.: Analysis of Vocal Tract Shape and Dimension using Magnetic Resonance Imaging: Vowels. JASA 90(2), 799–828 (1991)
4. Narayanan, S.S., Alwan, A.A., 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. Journal of the Acoustical Society of America 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. In: ICSLP (1998)
7. Woods, R.P., Cherry, S.R., Mazziotta, J.C.: 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. In: Oja, E. (ed.) Kohonen Maps (1999)
9. Kohonen, T.: Self-Organizing Maps. Springer, Heidelberg (2001)
10. Sondhi, M.M., Schroeder, J.: A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer. IEEE Transactions on Acoustics, Speech, and Signal Processing 50, 1070–1082 (1987)