Three-dimensional Vocal Tract Morphology Based on Multiple Magnetic Resonance Images Is Highly Reproducible During Sustained Phonation

*,†Alexander Mainka, ‡Ivan Platzek, *,†Willy Mattheus, §Mario Fleischer, *Ann-Sophie Müller, and *,†Dirk Mürbe, *†‡§*Dresden, Germany*

Summary: Objectives. The extraction of a three-dimensional (3D) morphology of the human vocal tract (VT) from magnetic resonance imaging (MRI) during sustained phonation can be used for various analyses like numerical simulations or creating physical models. The precision of visualizing techniques nowadays allows for very targeted acoustical simulation evaluating the influence of subsections of the VT for the transfer function. The aim of the study was to assess the accuracy of the 3D geometry based on MRI data in repetitive trials.

Study Design. This is a prospective study.

Methods. Four experienced singers underwent an MRI while repeating a specific vocal task 20 times consecutively. Audio recordings were made by means of an optical microphone. Images were restacked and subsections of the VT were segmented on multi-image–based cross sections using a semiautomatic algorithm. Different volume and area measures were evaluated.

Results. A high reproducibility of the morphologic data based on multiple images by means of the applied segmentation method could be shown with an overall variation of around 8%.

Conclusions. 3D modeling of the VT during sustained phonation involves a complex experimental setting and elaborate image processing techniques. Functional comparative analysis or acoustical simulations based on such data should take the found variability into account.

Key Words: Vocal tract–3D–MRI–Reproducibility–Phonation.

INTRODUCTION

During human voice production, the vocal tract (VT) acts as a resonator, which filters the glottal voice signal and modifies the spectrum of the glottal source signal. Thereby, the three-dimensional (3D) VT shape defines both vowel quality and voice timbre characteristics, which are important not only for speech communication, but also for vocal arts.[1]

The *modus operandi* of the VT is the adjustment of its shape or morphology. Magnetic resonance imaging (MRI) has become the state-of-the-art visualization technique for the VT at a functional state.[2] MRI allows for image acquisition during phonation and has helped to deepen the understanding of the physiology of the VT considerably. Thereby, two main approaches can be distinguished: a single-image–based and a multi-image–based analysis. Single-image–based analysis is suitable for so-called dynamic MRI and works on a *pars pro toto* presumption: a two-dimensional cross section of the VT is analyzed (eg, the midsagittal plane) and conclusions are drawn for the behavior of the whole VT. The advantage of this approach lies in a temporal resolution of vocal events. In contrast, multi-image–based analysis or static MRI is based on a series of images that cover the whole 3D anatomy of the VT. Whereas the former approach allows for the analysis of rapid VT adjustments during demanding vocal tasks like singing scales or switching between two different vocal conditions, the latter approach is bound to a single state of vocalization, where pitch, timbre, and loudness are kept constant. This vocal state needs to be maintained throughout the whole MRI recording. The advantage of dynamic MRI lies in a rather naturalistic use of the voice during recordings and an easy access to the data, which comprise distances on single images. Static MRI in contrast allows for 3D modeling of the VT and thereby, numerical simulations or even 3D prints, which can be further analyzed.

A limitation of the dynamic MRI lies in the restricted representation of the whole VT through one chosen plane. The static MRI encounters limitations on the technical level: subjects need to be experienced enough to maintain a stable phonation and articulation during the whole recording time.

Some studies have tried to exploit 3D data of the VT morphology from MRI recordings of subjects sustaining phonation: the main focus of this approach lies on the relationship between the VT geometry and its resonatory properties. Baer et al[3] studied area functions and 3D morphology for vowels. Sulter et al[4] focused on VT dimensions including volume measures and resonatory properties for different vowel configurations. Area functions and vowel formants were analyzed by Story et al,[5] Clement et al,[6] and Kitamura et al.[7] Takemoto et al[8] looked at the transfer function on the basis of area functions, including higher formants like F4 and F5. Aalto et al presented a fully automated method to extract 3D surfaces from MRI VT data.[9] Recently, real 3D geometry calculations including numerical simulations were done by Vampola et al,[10] Takemoto et al,[11] Fleischer et al,[12] and Delvaux and Howard.[13]

Accepted for publication November 10, 2016.

From the *Division of Phoniatrics and Audiology, Department of Otorhinolaryngology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany; †Voice Research Laboratory, Hochschule für Musik Carl Maria von Weber, Dresden, Germany; ‡Department of Radiology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany; and the §Department of Otorhinolaryngology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany.

Address correspondence and reprint requests to Alexander Mainka, Division of Phoniatrics and Audiology, Department of Otorhinolaryngology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany. E-mail: alexander.mainka@uniklinikum-dresden.de

Journal of Voice, Vol. 31, No. 4, pp. 504.e11–504.e20
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.11.009

TABLE 1.
Subject Data

	Subject 1	Subject 2	Subject 3	Subject 4
Voice classification	Lyrical tenor	Baritone	Baritone	Bass
Musical experience	Opera, oratorio international	Concert/oratorio/recital/opera national	Master student	Student
Taxonomy (after Bunch and Chapman[18])	2.1/2.4	4.1/4.4	7.1b	7.1b
Age (y)	36	29	26	22
Height (m)	1.74	1.76	1.82	1.93
Weight (kg)	70	75	86	95
Code	S1	S2	S3	S4

The applied methods for 3D modeling are either a region growing algorithm starting at different seed points as used by some authors[11,13,14] or an image-based segmentation method as applied by Baer et al[3] and Vampola et al.[10] Finally, Aalto et al used a fully automated algorithmic approach.[9] So far, there are no data available comparing those different approaches or further validating one or the other.

In terms of acoustics, vowel characteristics are commonly attributed to differences in the sound spectrum between 0 and 2 kHz and result from VT adjustments within the oral cavity and the tongue base.[1] Application possibilities of 3D data from these regions lie in the field of phonetics including speech simulations. VT adjustments can also induce alterations of the sound spectrum between 2 and 5 kHz, which in terms of perception result in differences of the voice timbre. The regions being held responsible for these modifications are situated in the lower VT.[15] Application possibilities are again linked not only to the field of phonetics, but also to voice pedagogy, voice therapy, and even forensics.

Recently, we have established a semiautomatic segmentation algorithm that works with reoriented re-sliced images and generating highly detailed data even for the lower VT.[16,17] Using this method, the 3D shape of the VT at a functional stage can be precisely quantified. This type of data holds the potential of deepening the understanding of voice physiology by numerous techniques, for example, morphologic analyses, numerical simulations on the basis of finite element models, or physical printed 3D models.

Yet to the best of our knowledge, no evidence derived from MRI data validating the generation of 3D morphology of the human VT during phonation has been reported to date.

Therefore, the questions that gave rise to the present study were the following:

(1) Is the VT configuration for a specific vocal task reproducible when singers are asked to do repetitive recordings?
(2) Which influential factors might impair the reproducibility of the measurements?

MATERIALS AND METHODS

The study was approved by the Ethics Committee of the Faculty of Medicine of the Technische Universität Dresden (No. 402.11). All participants provided written informed consent.

MR image acquisition

Four male subjects, all of them western classically trained (for subject data, see Table 1), were asked to produce a sustained vowel /a/ [like in pyjama] at 220 Hz (A3) in a 3T MRI scanner (Verio, Siemens Medical Solutions, Erlangen, Germany).

The task was conducted in a classical singing style at a medium loudness and well-controlled vibrato and was repeated 19 times, resulting in a total of 20 repetitive trials. The pitch was provided using a pitch pipe *via* the communication system of the MRI facility. The MRI recording started after the onset of phonation, which was maintained throughout the whole sequence. Before the measurements, subjects were given the opportunity to perform several trials to get used to the MRI recording conditions. To monitor the influence of the positioning of subjects within the MRI, subjects were asked to exit the MRI machine after every fifth measurement before being repositioned. Head and neck coils were removed and repositioned for this maneuver.

A 12-element head-neck coil was used for MR data acquisition. The MR examination included 20 3D volume-interpolated breath-hold examination scans in sagittal orientation. Each 3D dataset consisted of 52 adjacent slices with no interslice gap. The following MR sequence parameters were used: slice thickness 1.8 mm (S1, S2, S3) and 1.5 mm (S4), repetition time 4.01 ms, echo time 1.22 ms, matrix 288×288, field of view 300×300 mm, and flip angle 9°. The obtained in-plane resolution of the images was 1.04 mm.

The acquisition time was 12.1 seconds for each recording. Additionally, an audio signal was captured during MRI recordings from an optical microphone (MO 2000, Sennheiser, Wedemark-Wennebostel, Germany) that was positioned at a distance of 8 cm from the subject's mouth.

Image processing

The 52 sagittal images of each dataset were stacked and scaled by a factor of 3.0 with *ImageJ* (National Institutes of Health, Bethesda, MD), resulting in 156 images with a pixel size of 0.35 mm. Then, the images were re-sliced to the coronal view to fit the distance between slices to 0.35 mm and to obtain uniformly sized voxels. The re-slice was repeated a second time with default settings to obtain images in the sagittal plane. A centerline segmentation was used to create segment piles of different parts of the VT. The image stack was transformed to four sets of images. The planes of those images were orthogonal to