Real-time magnetic resonance imaging investigation of resonance tuning in soprano singing

Erik Bresch[a] **and Shrikanth Narayanan**
Department of Electrical Engineering, University of Southern California, 3740 McClintock Avenue,
Los Angeles, California 90089
bresch@usc.edu, shri@sipi.usc.edu

Abstract: This article investigates using real-time magnetic resonance imaging the vocal tract shaping of 5 soprano singers during the production of two-octave scales of sung vowels. A systematic shift of the first vocal tract resonance frequency with respect to the fundamental is shown to exist for high vowels across all subjects. No consistent systematic effect on the vocal tract resonance could be shown across all of the subjects for other vowels or for the second vocal tract resonance.

PACS numbers: 43.75.Rs, 43.75.Bc, 43.75.St, 43.75.Zz [TM]
Date Received: August 5, 2010 **Date Accepted:** September 16, 2010

1. Background

The singing voice has been of considerable interest to the acoustics researcher for a long time, and in particular the concept of resonance tuning has drawn notable attention over the past decades.[1,2] Resonance tuning is a strategy that trained opera singers are hypothesized to employ in order to increase their vocal efficiency and output power. Before the availability of audio power amplification this was an obvious necessity when performing in large concert halls.

During a vocal song production, the artist faces at least three constraints. Besides the need for an adequate intensity, the pitch at any given point in time is dictated by the melodic score of the music. Furthermore, the lyrics of the song have to be rendered with some degree of fidelity, which in turn demands the maintenance of the linguistic identities of the sung sounds (e.g., vowels) to some extent.[3]

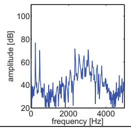
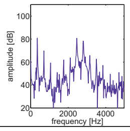
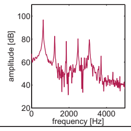
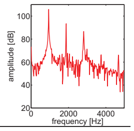
The theory of resonance tuning now contends that the vowel identity requirement is relaxed in practice and that trained singers actively modify their vocal tract shape so as to shift one of the resulting resonance frequencies to a multiple of the current (target) pitch frequency.[4] So, even though the changed formant structure alters the vowel quality, the singer is able to maintain the pitch in accordance with the score of the music while simultaneously maximizing the voice output.

Showing evidence for resonance tuning using audio recordings alone is not straightforward since the estimation of vocal tract resonance frequencies can be difficult, in particular for the case of high-pitched singing, e.g., soprano singing.[5] Here, the glottal source spectrum contains much wider spaced harmonics than in normal speech, so that the estimation of the resonance frequencies from peaks in the spectral envelope of the recorded signal is severely compromised (see, for example, Table 1). Therefore, researchers have resorted to other methods for the investigation of the vocal tract transfer function.

One possibility is the use of an artificial external broad-band noise source to excite the vocal tract while the soprano singer tries to maintain her natural singing vocal tract posture without actually producing any sound.[6] Subsequently, a resonance frequency estimation can be carried out from the reflected sound waves.

[a] Author to whom correspondence should be addressed.

Table 1. 1024-point FFT spectra for /i/ at notes 1, 5, 11, and 15 (subject M1).

F_0 [Hz]	233	349	622	932
spectrum				

Another option is to obtain direct evidence of the vocal tract shaping strategies such as using magnetic resonance imaging (MRI).[7,8] However, to acquire a conventional (static) MRI recording the singer may have to hold the vocal tract posture for an unusually long time, e.g., on the order of a few minutes as would be the case for a high resolution 3-D volumetric scan. To alleviate this issue researchers often restrict themselves to capturing the midsagittal view of the vocal tract and then performing an aperture-to-area function conversion to facilitate a tube model description of the vocal tract. However, even a 2-D static MRI scan can easily take a few seconds.

In contrast to the previous studies, this study employs real-time (RT) MRI technology to obtain midsagittal vocal tract image data from a total of 5 soprano singers. While thus far RT-MRI has been mostly used to study dynamic speech production processes, it also appears well suited for the investigation of scale singing since it allows the subjects to produce vocal sounds in a more natural way, i.e., they are not required to maintain the vocal tract posture for unnaturally long periods of time.[9]

Furthermore, RT-MRI allows the researcher to investigate other aspects of song productions, such as their expressive qualities, rhythm and pausing behavior, etc., which require data from dynamic productions. Though this article focuses on sung vowel scales, it does describe the data acquisition, processing, and analysis steps relevant for general song production (data examples can be found in Ref. 10). In that regards, it can be viewed as providing a proof-of-concept for the use of RT-MRI technology for studies of vocal productions of song.

2. Data collection

The subjects for this study were 5 female sopranos (M1, S2, K3, L4, and H5) trained in Western opera and who were native American English speakers. The subjects sang two-octave vowel scales (/la/, /le/, /li/, /lo/, /lu/) without vibrato, and they were allowed to breathe after the first octave.

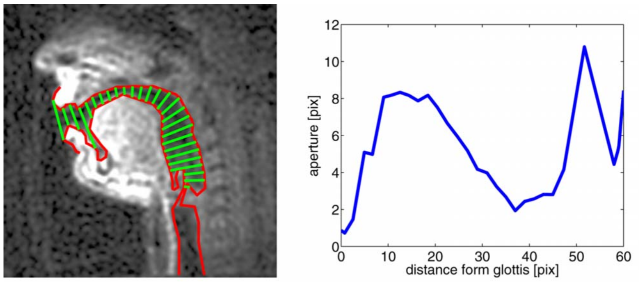
Midsagittal MR images were collected with a GE Signa 1.5T scanner.[11] Synchronized audio recordings were obtained, and the scan noise was subsequently removed.[12] During the data collection the subjects were in a supine position. A sample recording of subject M1 singing the /la/ scale is available in the multimedia file Mm. 1.

Mm.1. Subject M1 singing the /la/ scale.

3. Data analysis

3.1 Audio analysis

Using the noise-cancelled audio recording, a pitch estimation was carried out using the PRAAT software.[13] However, as described above, the estimation of the vocal tract resonances from the audio signal is difficult, especially at high pitch values. This is due to the fact that the harmonics of the source spectrum are widely spaced, and consequently the filter function of the vocal tract gets sampled only at relatively fewer frequency points (see Table 1). Therefore, the vocal tract



(a)Sample midsagittal real-time MR image.

(b)Aperture function from glottis (left) to lips (right).

Fig. 1. (Color online) Subject M1, producing /le/ at note 1.

resonance frequencies were estimated directly using the midsagittal image data. And while these estimates can be noisy, we are mainly interested in statistically significant trends of the resonance frequencies with respect to the fundamental.

3.2 Image analysis

From each of the notes of the scales, one image was extracted corresponding to the midpoint of the vowel segment, i.e., from a relatively stable vocal tract configuration. In these images the vocal tract outline was then automatically detected[14] and then manually corrected if necessary. The glottis position was manually determined in each image. A sample image is shown in Fig. 1(a), showing subject M1 singing /le/ at note 1. Here, the vocal tract outline is shown in red.

Subsequently, the aperture function from the glottis to the lips was derived from the vocal tract contours. This was accomplished by first constructing a vocal tract midline using repeated geometrical bisection, and, second, finding densely spaced perpendiculars along the midline and their intersections with the vocal tract contours.[15] The perpendiculars are the midsagittal aperture lines, and they are shown in green in Fig. 1(a). Figure 1(b) shows the aperture function corresponding to the vocal tract shape of Fig. 1(a). This graph displays the length of the aperture lines as a function of position along the midline. In Fig. 1(b) the left side corresponds to the glottis, while the right side corresponds to the lips. The units used in the graph are pixels.

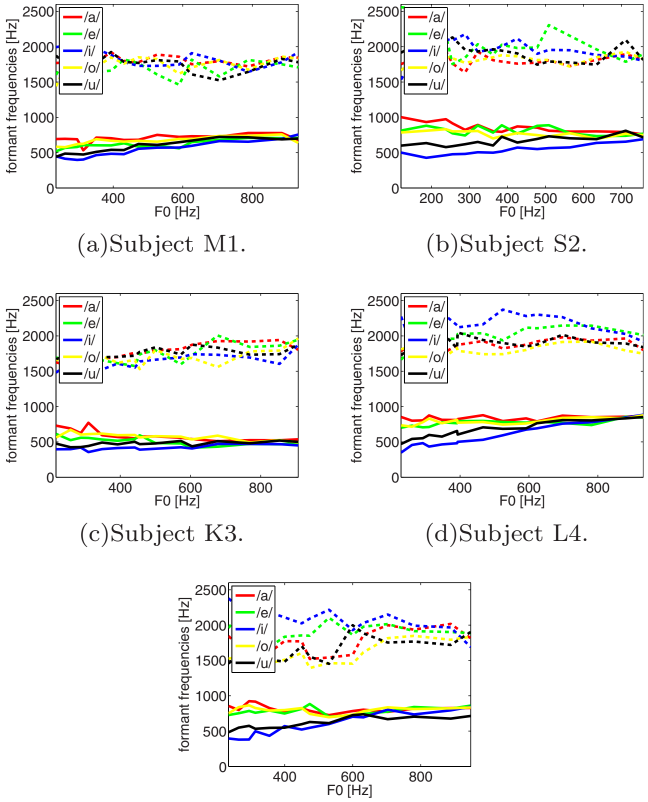
The midsagittal aperture function was then converted to the cross-sectional area function of a tube model whose resonance frequencies were computed using the VTAR (Ref. 16) software. Figure 2 shows the resonances F_1 and F_1 as a function of the fundamental F_0 for all 5 vowels for all 5 subjects. The resonance frequency estimates then form the basis of the statistical analysis in Section 4.

It must be pointed out that numerous methods have been proposed for the aperture-to-area conversion and, in general, their optimum parameters are subject specific.[17] For this study the method described in Ref. 18 and extended in Ref. 19 was employed without adaptation of the parameters. Hence deviations of the computed tube model resonances from the true vocal tract resonances must be expected. However, this study aims at identifying global trends in the formant frequencies with respect to the pitch frequency for a given subject, as opposed to quantifying absolute formant frequency measurements.

4. Results

Table 2 shows the midsagittal images for subject M1 for all 5 vowels at notes 1, 5, 11, and 15 with fundamental frequencies of 233, 349, 622, and 932 Hz, respectively. It can be seen that for the low notes the vocal tract configuration is distinct for the individual vowels, and the distinction decreases as the pitch increases. This behavior was observed for all 5 subjects.

The bottom row in Table 2 shows the aperture functions of subject M1 for the 5 vowels for the notes 1 (blue), 5 (dark purple), 11 (light purple), and 15 (red). It can be seen that at higher notes the individual differences between the vowels decrease, and in particular the shape of the oral cavity converges to a widely open configuration.



(a)Subject M1.

(b)Subject S2.

(c)Subject K3.

(d)Subject L4.

(e)Subject H5.

Fig. 2. (Color online) Resonances F_1 (solid), and F_2 (dashed) versus the fundamental F_0.

Table 2. Sample MR images and midsagittal aperture functions of all 5 vowels at notes 1, 5, 11, and 15 (subject M1).

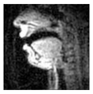
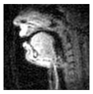
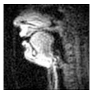
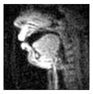
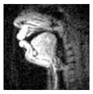
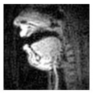
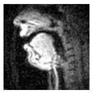
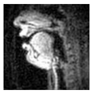
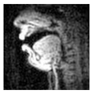
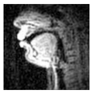
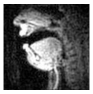
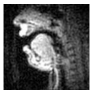
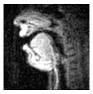
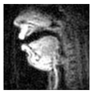
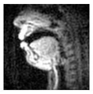
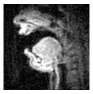
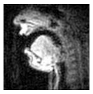
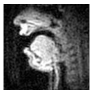
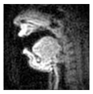
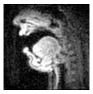
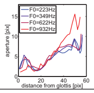
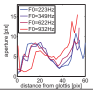
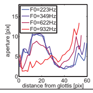
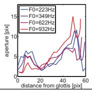
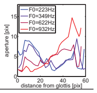
F_0 [Hz]	/a/	/e/	/i/	/o/	/u/
233					
349					
622					
932					
aperture function					

Table 3. Linear regression of the vocal tract resonances versus the fundamental.

		F_1			F_2		
Subject	Vowel	α_1 (Hz)	β_1	p	α_2 (Hz)	β_2	p
M1	/a/	639	0.126	0.061	1769	0.061	0.427
	/e/	506	0.221	3×10^{-5}	1676	0.036	0.783
	/i/	291	0.490	6×10^{-10}	2025	−0.314	0.032
	/o/	580	0.167	0.003	1613	0.230	0.092
	/u/	378	0.405	6×10^{-7}	1884	−0.213	0.088
S2	/a/	975	−0.297	1×10^{-4}	1808	−0.000	0.999
	/e/	851	−0.099	0.21198	2305	−0.598	0.053
	/i/	360	0.425	4×10^{-10}	2133	−0.401	0.115
	/o/	812	−0.108	0.037	1796	0.085	0.412
	/u/	538	0.301	2×10^{-4}	2099	−0.401	0.043
K3	/a/	732	−0.272	6×10^{-4}	1539	0.446	2×10^{-5}
	/e/	603	−0.179	0.003	1516	0.429	0.004
	/i/	357	0.123	4×10^{-4}	1470	0.339	0.002
	/o/	663	−0.178	4×10^{-4}	1582	0.245	0.065
	/u/	431	0.090	0.017	1643	0.217	0.026
L4	/a/	809	0.051	0.186	1782	0.161	0.060
	/e/	692	0.148	0.002	1738	0.465	0.016
	/i/	256	0.671	2×10^{-9}	2269	−0.170	0.352
	/o/	715	0.149	0.002	1784	0.044	0.593
	/u/	418	0.498	2×10^{-8}	1846	0.084	0.464
H5	/a/	853	−0.067	0.282	1579	0.343	0.057
	/e/	729	0.102	0.108	1942	−0.033	0.841
	/i/	237	0.680	2×10^{-8}	2256	−0.393	0.066
	/o/	789	0.016	0.789	1281	0.570	5×10^{-4}
	/u/	460	0.305	5×10^{-5}	1341	0.587	0.002

Corresponding to the /i/-column of Table 2, the 1024-point FFT spectra at notes 1, 5, 11, and 15 are shown in Table 1, which were derived from the noise-cancelled audio recording. These examples illustrate the difficulty of the estimation of the vocal tract resonances at high pitch values. At the low note 1 resonance peaks can be recognized in the spectrum easily, whereas at the high note 15 no resonances are readily observable.

In order to investigate the dependence of the vocal tract resonances F_1 and F_2 on the fundamental F_0, linear models were fit of the form,

$$F_{1,2} = \beta_{1,2} \times F_0 + \alpha_{1,2} + \epsilon \qquad (1)$$

for each vowel. Here, α has the dimension of hertz, and β is the dimensionless slope of the regression line. The value ϵ represents the error. The calculated values are listed in Table 3, and we also list the resulting p-value for the respective β coefficient. In Table 4 we compact this information more, and we list only the sign of the statistically significant trends ($\beta \neq 0$ with significance $\geq 95\%$) for all subjects and all vowels.

These values suggest that for the high vowels /i/ and /u/ for all subjects there is a consistent dependency of the first vocal tract resonance F_1 on the fundamental F_0 in terms of a positive correlation. Other than that, no clear patterns can be readily observed that apply across all subjects.

Table 4. Sign of the statistically significant linear trends of the resonances F_1 and F_2 with respect to the fundamental F_0.

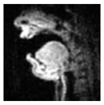
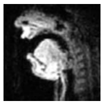
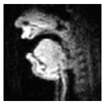
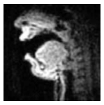
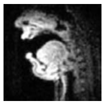
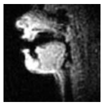
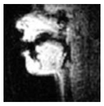
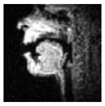
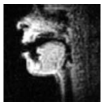
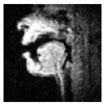
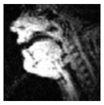
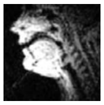
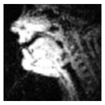
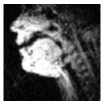
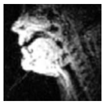
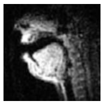
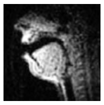
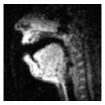
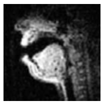
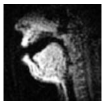
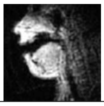
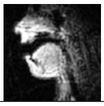
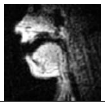
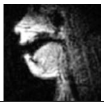
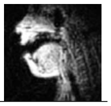
	F_1					F_2				
Subject	/a/	/e/	/i/	/o/	/u/	/a/	/e/	/i/	/o/	/u/
M1		+	+	+	+			−		
S2	−		+	−	+					−
K3	−	−	+	−	+	+	+	+		+
L4		+	+	+	+		+			
H5			+		+				+	+

5. Discussion

The finding that the first resonance of the high vowels rises with the fundamental frequency is consistent with previous findings. Considering the sample images in Table 2, it is easy to see that the front cavity opens more widely as the singer goes to higher fundamental frequencies, and it is well known that F_1 is directly related to the opening degree. The relative opening effect is certainly strongest for the high vowels /i/ and /u/, which are most constricted in their natural oral cavity configuration. Hence the quantitative findings are well in accordance with the expectations, and we conclude that the RT-MRI data and the proposed processing steps offer merit.

However, based on our study, we cannot conclude that all sopranos employ generalizable strategies for resonance tuning the way it has been described in prior literature. To illustrate the qualitative differences in the shaping strategies, we show in Table 5 the MR images for all 5 subjects and all 5 vowels corresponding to note 15 (F_0=932 Hz), which is the highest note in our

Table 5. MR images for all 5 subjects and all 5 vowels at note 15 (F_0 =932 Hz).

Subject	/a/	/e/	/i/	/o/	/u/
M1					
S2					
K3					
L4					
H5					

data set. We observe that in particular subject M1 but also S2 (top 2 rows) show evidence of some of the vowel-specific tongue shaping even at this extreme pitch, whereas the rest of the subjects appear to have converged to a single canonical vocal tract shape for all vowels. Furthermore, the width of the oral cavity varies considerably across subjects, with M1 being on one extreme and K3 on the other.

We speculate that the observed variability in the vocal tract shaping may be due to the individual training that each of the singers had received. In this regard it would be also interesting to see if RT-MRI recordings can be used in the future as a teaching tool for voice teachers to help sopranos acquire consistent tuning strategies. In summary, we find that the interaction between singing and linguistic goals of producing speech sounds is complex and needs further exploration.

Acknowledgment

This work was supported by NIH Grant No. R01 DC007124-01.

References and links

[1] G. Carlsson and J. Sundberg, "Formant frequency tuning in singing," J. Voice **6**, 256–260 (1992).

[2] I. Titze, "A theoretical study of f_0-f_1 interaction with application to resonant speaking and singing voice," J. Voice **18**, 292–298 (2004).

[3] B. Story, "Vowel acoustics for speaking and singing," Acta. Acust. Acust. **90**, 629–640 (2004).

[4] J. Sundberg, "The acoustics of the singing voice," Sci. Am. **236**, 82–91 (1977).

[5] E. Joliveau, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: The soprano voice," J. Acoust. Soc. Am. **116**, 2434–2439 (2004).

[6] E. Joliveau, J. Smith, and J. Wolfe, "Tuning of vocal tract resonance by sopranos," Nature (London) **427**, 116 (2004).

[7] B. H. Story, "Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics," in Proceedings of the Stockholm Music Acoustics Conference SMAC-03 (2003), pp. 435–438.

[8] J. Sundberg, "Research on the singing voice in retrospect," TMH-QPSR Speech, Music and Hearing, KTH, Stockholm, Sweden, **45**, 11–22 (2003).

[9] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," IEEE Signal Process. Mag. **25**, 123–132 (2008).

[10] http://sail.usc.edu/span/ (Last viewed 10/22/2010).

[11] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776 (2004).

[12] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791–1794 (2006).

[13] http://www.fon.hum.uva.nl/praat/ (Last viewed 10/22/2010).

[14] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," IEEE Trans. Med. Imaging **28**, 323–338 (2009).

[15] E. Bresch, J. Adams, A. Pouzet, S. Lee, D. Byrd, and S. Narayanan, "Semi-automatic processing of real-time MR image sequences for speech production studies," in Proceedings of the Seventh International Seminar on Speech Production, Ubatuba, Brazil (2006).

[16] Z. Zhang and C. Y. Espy-Wilson, "A vocal-tract model of American English /l/," J. Acoust. Soc. Am. **115**, 1274–1280 (2004).

[17] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," Speech Commun. **36**, 169–180 (2002).

[18] P. Ladefoged, J. F. K. Anthony, and C. Riley, "Direct measurement of the vocal tract," UCLA Working Papers in Phonetics (WPP) **19**, 4–13 (1971).

[19] S. Lee, "A study of vowel articulation in a perceptual space," Ph.D. thesis, University of Alabama at Birmingham (1991).