Estimates for the measurement and articulatory error in MRI data from sustained vowel production

Daniel Aalto[1,2], Jarmo Malinen[3], Martti Vainio[2], Jani Saunavaara[4], Pertti Palo[3]

[1]Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics, [2]University of Helsinki, Institute of Behavioural Sciences, [3]Aalto University School of Science and Technology, Department of Systems analysis and Mathematics, [4]Medical Imaging Centre of Southwest Finland

daniel@iki.fi

ABSTRACT

Spatial data of the human vocal tract (VT), larynx, and thorax can be obtained by magnetic resonance imaging (MRI) during steady, sustained phonation. Long acquisition time increases the resolution as well as the errors due to involuntary motion of VT.

We discuss two experiments with a single test subject to find a suitable 3D MRI data acquisition procedure in terms of subject's articulatory steadiness.

Keywords: MRI data acquisition, sustained phonation, vowel production, articulatory stability.

1. INTRODUCTION

Magnetic resonance imaging (MRI) is a widely used tool to acquire three dimensional (3D) anatomic data of the vocal tract (VT) for speech production studies, simulation and articulatory synthesis [7], [11]. In comparison to X-ray based computed tomography, MRI is preferred because no known health hazards associated to it.

The sustained vowel production itself is prone to adjustments due to the gravity [4], [10], decrease in the lung volume, and muscle fatigue. The MR images have certain spatial resolution that can be improved by image post-processing and longer acquisition time. However, there is no guarantee that the VT can be kept completely steady. Motionless VT would improve the measurement quality and non-steady productions should be ruled out by a clear rejection criterion.

In this article, we present data from one test subject (the first author) to assess the steadiness of the sustained vowel production and the measurement errors of the MRI data. We record dynamic 2D MRI series which mimic a short (8 s) or a long (18 s) 3D MRI scan. The sustained productions are replicated in an anechoic chamber for acoustic analysis.

The error estimates for the data quality are needed for validating a computational model for sustained vowel production. Also the measurement should be arranged so that the signal is as self-consistent as possible. The control mechanisms of the sustained phonation are well known [12], and there are some estimates for the stability of the acoustic parameters like the fundamental frequency (f_0) [5], or the second formant (F_2) [6]. The sustained vowel production in the supine position, required in MRI, differs from both the up-right and the continuous speech articulations [4], [10]. The quality of 3D MRI data has already been assessed in [3] and [4] qualitatively.

2. EXPERIMENTS AND DATA ACQUISITION

2.1. Hypotheses and experimental settings

In producing a sustained phonation, several muscles are working together while the lungs are emptying gradually and the thorax contracts slowly. The subject tries to keep the articulation constant and uses both tactile and auditory feedback to achieve the goal (see [8] and [9]).

Specifically we hypothesize that 1) longer productions induce greater variability in articulation 2) reducing mobility increases stability of all articulatory and acoustic parameters.

To test the first hypothesis, we acquired mid-sagittal images of the head-neck area during a sustained production of two open, unrounded vowels: [ɑ] and [æ]. We mimicked two different 3D MRI data acquisition procedures corresponding to sequence durations of 8 s and 18 s respectively, as used in [1]. Accordingly, the subject produced each of the two vowels short (ca. 10 s) and long (ca. 20 s). To test the second hypothesis, we used a bite-block and collected the mid-sagittal image sequences for short and long, [ɑ] and [æ], as

previously. The experiment is replicated in an anechoic chamber.

In order to minimize the articulatory artifacts, the subject produced the prolonged vowel with a given fundamental frequency (f_0=110 Hz). The subject heard a cue signal indicating the reference pitch and began the phonation. The MRI sequence started at 1 s after the onset of the phonation.

2.2. Dynamic MRI sequence

We used Siemens Magnetom Avanto 1.5T scanner (Siemens Medical Solutions, Erlangen, Germany). Maximum gradient field strength of the system is 33 mT/m and the maximum slew rate is 125 T/m/s. We used single-shot ultrafast spoiled gradient echo sequence (TurboFLASH) where the time of repetition (TR) and the echo time (TE) are minimized. Single sagittal plane was imaged using parameters TR 178 ms, TE 1.4 ms, flip angle 6, receiver bandwidth 651 Hz/pixel, FOV 230 mm, matrix 120x160, and slice thickness 10 mm. The sequence was run for 45 frames (8 s) or 100 frames (17.8 s) to simulate 3D MRI scan and to gather data.

2.3. Replication in the anechoic chamber

To obtain high quality recordings of the sustained vowel productions, we replicated the experiment in an anechoic chamber and estimated the formant trajectories. In the replicate setting, the same subject produced the same phonetic material in supine position as in the MRI room but the sequencing noise was not simulated, neither the acoustic surroundings inside the MRI machine. The recording was made with high quality microphone and digitized at 44.1 kHz sampling rate.

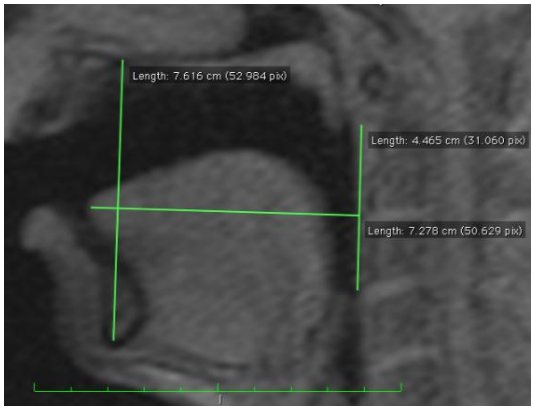
3. RESULTS

Of the collected data we analyzed a subset. The MRI sequences comprised of 580 frames, of which 108 were manually inspected and a total of 126 measurements were made. The sound recordings were analyzed to extract the formant trajectories.

3.1. Image processing

The first and the last frames of each series (total of 16 frames) were inspected manually, and the contrast was adjusted so that the anatomical landmarks were clearly visible. Two distances were measured: mandibular opening (MO: distance from the lower anterior tip of the mandible to the highest point of the hard palate) and tongue tip advancement (TTA: distance of the tongues most distant point from the back wall of pharynx), see Figure 2, and compare with [3].

Figure 1: An example of 2D MR image frame representing the mid-sagittal section of the VT during a sustained phonation of [ɑ]. The anatomic distances, MO and TTA, are given by green lines.



3.2. Sound processing

The data recorded in the anechoic chamber was further analyzed with Matlab 7.10.0.499 (R2010a). The signal was chopped in frames of duration 178 ms to make the comparison to the MRI data easier. Then, three first formants were extracted from each frame by using the linear predictive coding algorithm [2]. The short and long phonations have mean durations of 11.4 s and 21.1 s respectively.

3.3. Image quality

The nominal resolution of the image was 1.4 mm, the pixel size of the 2D image. Since each pixel has a gray-scale value depending on the strength of the signal, the anatomical details can be located more precisely by interpolating. We measured MO from each of the 45 frames taken during a sustained [ɑ] with a bite-block between incisors. The mean distance is 68.2 mm with standard deviation (SD) 0.4 mm. We regard this as the true resolution of the MR image when a manual measurement is made. In other words, the inspector could extract 1.8 bits more information taking the signal strength into account.

3.4. Steadiness of the vocal tract

To assess the steadiness of the VT, we measured both the MO and the TTA. The measurements

were carried out at the beginning of the phonation (first frame), four seconds after the onset of the sequence (23rd frame), and at the end of the phonation (either 45th or 100th frame). In Table 1, all the measured values are listed.

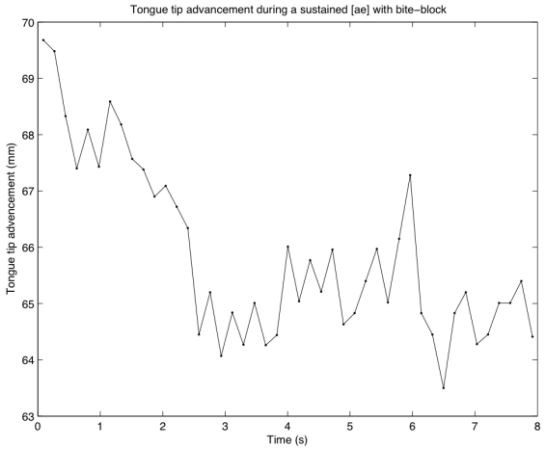
Table 1: The measured distances of the VT based on the mid-sagittal MR images. In the first three columns the MO of four sustained productions without bite-block is given at three distinct time instants.

Distance (mm)	MO beg	MO 4 s	MO end	TT beg	TT 4 s	TT end
[ɑ], 8 s	76.0	74.6	74.7	65.5	64.5	64.9
[ɑ], 18 s	73.9	73.4	72.5	64.5	62.7	61.9
[æ], 8 s	77.0	76.1	76.2	75.3	74.0	73.8
[æ], 18 s	77.6	77.7	77.4	77.1	76.6	73.2

Some observations are clear: tongue tip moves in the posterior direction; MO decreases. Typically, the 4 s and the end measurements are close to each other compared to the measurement in the first frame.

Except for the long [æ] the differences in the MO and the TTA are larger between the first than the second measurements, suggesting that the time development of the distance is not linear. We measured the time development of the TTA during a short [æ] with bite-block, shown in Figure 2.

Figure 2: A measurement series showing the time development of the tongue tip advancement. Each frame represents a time window of 178 ms.



During the beginning of the sequence, there is more variation. The first 22 frames have a SD of 1.7 mm while the last 22 frames have a SD of 0.8 mm. The quality of the image can deteriorate near the sublingual glands, since saliva gives a strong signal in MRI. Hence, the SD of TTA can be partly due to the signal strength, not only fluctuation of the tongue tip position.

3.5. Stability of the formants

The time development of four sustained vowels is shown in the Figure 3. All the shown vowels are produced without a bite-block.

Figure 3: The time development of the first three formants of an [ɑ] (dots) and an [æ] (tilted squares), no bite-block.

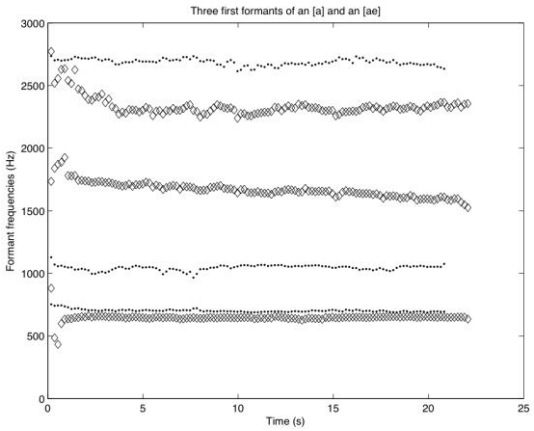


Table 2: Standard deviation of formants over a 4 s period (22 frames) is calculated. The means over the four phonations of the SDs are given as a function of the formant.

Means of standard deviations (Hz)	Frames 1-22	Frames 23-44	Last 22 frames
F_1	21.2	6.55	4.57
F_2	50.5	26.2	24.3
F_3	90.8	32.6	28.9

During first four seconds, formants stabilize. By calculating the SD of formant values during the first four seconds (frames 1-22), during the interval 0-4 s (frames 23-44) and during the end of the sequences (last 4 s), we estimate the stability of the formants.

3.6. The role of the bite-block

A bite-block placed between maxillary and mandibular central incisors was hypothesized to increase stability compared to the non-bite-block condition. The measurements of anatomic distances are shown in Table 3. The MO is fluctuating slightly due to measurement error. However, the TTA is changing more in average with bite-block than without.

The sound recordings in the anechoic chamber do not clearly support the hypothesis either. As shown in the Table 4, the F_1 is rather stable but the fluctuations remain larger towards the end of the phonation compared to the non-bite-block condition. The F_2 stabilizes even stronger with

bite-block. However, the fluctuation of F_3 remains large over all the time intervals. It was observed during the measurements that the lips moved involuntarily to hold the bite-block, which could explain the relatively unstable F_3.

Table 3: The measured distances of the VT based on the mid-sagittal MR images. In the first three columns the MO of four sustained productions with bite-block is given at three distinct time instants.

Distances (mm)	MO beg	MO 4 s	MO end	TT beg	TT 4 s	TT end
[ɑ], 8 s	68.5	68.8	68.2	60.0	57.7	57.2
[ɑ], 18 s	68.5	68.6	68.4	62.4	61.1	57.9
[æ], 8 s	68.6	69.2	68.9	69.7	66.0	64.4
[æ], 18 s	68.5	68.9	69.0	70.7	69.4	68.1

Table 4: Standard deviation of formants over a 4 s period (22 frames) is calculated. The means of the SDs over the four phonations, with bite-block, are given as a function of the formant.

Means of standard deviations (Hz)	Frames 1-22	Frames 23-44	Last 22 frames
F_1	10.4	11.9	8.89
F_2	38.8	12.7	11.2
F_3	116	70.4	109

3.7. The role of the duration

In the Table 5, the mean fluctuations are shown as a function of duration. The phonation is much steadier during the period 4-8 s, if the phonation is to continue longer.

Table 5: Standard deviation of formants over a 4 s period (22 frames) is calculated. The means over the four phonations of the SDs are given as a function of the formant.

Means of standard deviations (Hz)	Frames 1-22	Frames 23-44	Last 22 frames
F_1 short	8.2	10.8	7.6
F_2 short	52.5	23.2	21.3
F_3 short	145	77.8	112
F_1 long	23.3	7.7	5.8
F_2 long	36.8	15.8	14.1
F_3 long	62.0	25.2	25.7

4. CONCLUSIONS

During a sustained vowel phonation, the tongue gradually moves in posterior direction (as in [4]). The change is not uniform in time: the fluctuation is larger during the first seconds both in terms of articulatory distance measures (MO and TTA) and acoustic parameters (F_1, F_2 and F_3). The overall pattern of the TTA trajectory (Fig. 2) and formant trajectories (Fig. 3) suggests that there are two processes responsible for the movements: a gradual movement with diminishing change rate, and a random uniform fluctuation. The strong coupling between the two provides with a way to predict the "stable zone" of the sustained phonation already before MRI measurements are done. More research is needed to find the details of the time development of steady vowel production.

Neither of the initial hypotheses could be supported by the current data. On the contrary, long productions are more stable than short in terms of SD. This surprising evidence should be checked with larger data sets. Finally, the time scales for laryngeal and VT stabilization may be different and vary across subjects.

5. REFERENCES

[1] Aalto, D., Aaltonen, O., Happonen, R.-P., Malinen, J., Palo, P., Parkkola, R., Saunavaara, J., Vainio, M. Recording Speech Sound and Articulation in MRI. *Proc. Biodevices, Rome, 2011*.

[2] Atal, B. S., and Schroeder, M. R. 1967. Predictive coding of speech signals. *Proc. IEEE Conference on Communication and Processing.* 360–361.

[3] Crary, M. A., Kotzur, I. M., Gauger, J., Gorham, M. Burton, S. 1996. Dynamic Magnetic Resonance Imaging in the Study of Vocal Tract Configuration. *Journal of Voice.* 10 No. 4, 378–388.

[4] Engwall, O. 2003. A revisit to the application of MRI to the analysis of speech production—testing our assumptions. *Proc of 6th International Seminar on Speech Production.* 43–48.

[5] Gelfer, M. P. 1995. Fundamental frequency, intensity, and vowel selection: Effects on measures of phonatory stability. *J. Speech and Hearing Research.* 38 (6), 1189–1198.

[6] Gerratt, B. R. 1983. Formant frequency fluctuation as an index of motor steadiness in the vocal tract. *J. Speech and Hearing Research.* 26, 297–304.

[7] Hannukainen, A., Lukkari T., Malinen, J., Palo, P. 2007. Vowel formants from the wave equation. *J. Acoust. Soc. Am.* 122(1), EL1–EL7.

[8] Houde, J., Jordan, M. 1998. Sensorimotor adaptation in speech production. *Science*, 279, 1213.

[9] MacDonald. E. N., Purcell D. W., Munhall, K. G. 2011 Probing the independence of formant control using altered auditory feedback. *J. Acoust. Soc. Am.* 129(2), 955–965.

[10] Yanagihara, N., Koike, Y. 1967. The Regulation of Sustained Phonation. *Folia phoniat.* 19, 1–18.

[11] Stone, M., Stock, G., Bunin, K., Kumar, K., Epstein, M., Kambhamettu, C., Li, M., Parthasarathy, V., Prince, J. 2007. Comparison of speech production in upright and supine position. J. Acoust. Soc. Am. 122 (1), 532–541.

[12] Švancara, P., Horáˇcek, J. 2006. Numerical modelling of effect of tonsillectomy on production of czech vowels. *Acta Acustica united with Acustica.* 92, 681–688.