 doi:10.3176/lu.2012.3.05

PERTTI PALO (Helsinki), *DANIEL AALTO* (Helsinki),
OLLI AALTONEN (Helsinki), *RISTO-PEKKA HAPPONEN* (Turku),
JARMO MALINEN (Helsinki), *JANI SAUNAVAARA* (Turku),
MARTTI VAINIO (Helsinki)

ARTICULATING FINNISH VOWELS: RESULTS FROM MRI AND SOUND DATA

Abstract. We present anatomic and acoustic data from a pilot study on the Finnish vowels [ɑ, e, i, o, u, y, æ, ø].[1] The data were acquired simultaneously with 3D magnetic resonance imaging (MRI) and a custom built sound recording system. The data consist of a single static repetition of each vowel with constant F0. The imaging sequence was 7.6 s long and had an isotropic voxel size of 1.8 mm. We report results of listening tests and acoustic analysis of audio data as well as manual analysis of MR images.

Keywords: Finnish, Formant analysis, spectral LPC, MRI.

1. Introduction

Vowel production has been studied with several imaging methods. The earliest such studies used X-ray imaging (Jones 1929; Sovijärvi 1938; 1963; Chiba, Kajiyama 1941). Nowadays, MRI is preferred because no known health hazards are associated to it (Baer, Gore, Boyce, Nye 1987; Engwall, Badin 1999). Here we report simultaneous MRI and audio data from one test subject pronouncing Finnish vowels. In addition to the images, we assess the quality of the vowels based on a listening experiment of the audio data.

The data examined in this study was acquired for developing a mathematical and computational model of speech production (for a detailed report and further references, see Palo 2011 and references therein). We aim at maximal spatial resolution with minimal movement artifacts. The simultaneous audio recording provides an indirect measure of the stability of the vocal tract and a reference point for model validation.

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 (Hannukainen, Lukkari, Malinen, Palo 2007; Stone, Stock, Bunin, Kumar, Epstein, Kambhamettu, Li, Parthasarathy, Prince 2007; Švancara, Horáček 2006). Bones, teeth and most of the small details under the voxel size (1.8 mm in our case) are not visible in MRI. On the other hand tissues containing water and lipids are clearly visible along with mucus which can be indistinguishable from the actual tissues.

[1] The symbols used in this paper are those of the International Phonetic Alphabet (IPA). Equivalents in the Finno-Ugric transcription system are as follows: ɑ = a, æ = ä, ø = ö, y = ü.

3D MR imaging sequences provide a poor time resolution. We employed a 7.6 s long version of a sequence called VIBE as detailed in Palo 2011 and Aalto, Malinen, Palo, Aaltonen, Vainio, Happonen, Parkkola, Saunavaara 2011. As the conditions are less than ideal for the test subject — requiring a supine position, an extremely long production and being subjected to intense acoustic noise — extra care needs to be taken in evaluating and validating the data.

We use three separate methods to evaluate the same vowel production event. Hence, a single empirical data point is connected to anatomic, acoustic and linguistic contexts.

2. Materials and methods

2.1. The data set

In this study we evaluate a data set consisting of the Finnish vowels [ɑ, e, i, o, u, y, æ, ø]. The set consists of a single production of each of the vowels uttered by a native male speaker. For each production we acquired a simultaneous 3D MRI scan and an audio recording. A detailed report on the data acquisition is available in Palo 2011 and Aalto, Malinen, Palo, Aaltonen, Vainio, Happonen, Parkkola, Saunavaara 2011. For perceptual and acoustic evaluation clear speech samples were extracted from the recording before and after the MRI sequence in the same manner as in Palo 2011.

2.2. Perceptual evaluation of audio data

Two samples of clear speech were extracted manually from the MRI recordings for each of the eight vowels. The first sample — the begin sample — was a 200 ms sample directly before the onset of the MRI noise. The second sample — the end sample — was a 200 ms sample located 100 ms after the end of the MRI noise.

These samples were listened to by 20 female students of phonetics with no known hearing defects and whose ages ranged between 20 and 39 years (mean 26 years, s.d. 5 years). Two listeners were bilingual speakers of Finnish and Swedish and all the rest were native speakers of Finnish. The first three listeners used Sennheiser HD 250 linear II earphones during the test and the rest used Sony MDR-7510 earphones. In both cases the listening experiment was run with Max/MSP software (version 6.0.3) running on a MacBook Pro laptop with Mac OS X (version 10.6.8).

In the experiment, the listeners were asked to categorise the vowels samples they heard and rate the sample's prototypicality and nasality. The test was a forced choice test and the listeners could listen repeatedly to the sample they were rating.

2.3. Acoustic evaluation of audio data

The samples used in the perceptual assessment were analysed with LPC. As the recording system does not have a flat frequency response (Palo 2011), we employed the measured power spectral response of the system in compensating the FFT spectrums of the samples. The spectral linear prediction algorithm (Makhoul 1975) was then used to obtain formant estimates for these samples. The fundamental frequency f_0 of each of the samples was estimated with the autocorrelation method. All of the acoustic analyses were carried out with Matlab release 2010b running on a MacBook Pro laptop with Mac OS X (version 10.6.8).

2.4. Evaluation of MRI data

We measured the cross sectional area of the smallest opening within the vocal tract and the opening distance of the jaw for each vowel articulation. The jaw opening was measured as the distance between the maxilla and the mandible

as shown in Figure 1. Also, we measured the cross sectional area of the lip opening for those articulations where it was possible to define a cutting plane limited by the lips. All articulatory measurements were done with OsiriX (version 3.9) on a MacBook Pro laptop with Mac OS X (version 10.6.8).

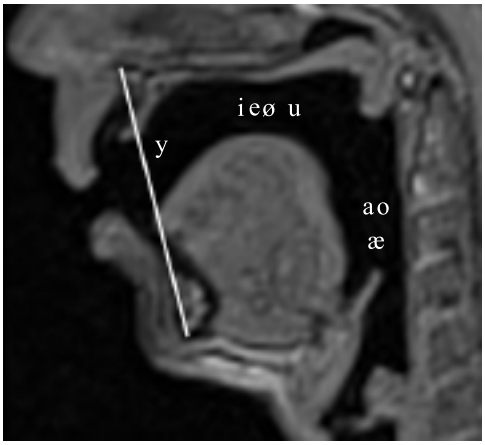

Figure 1. **Position of the narrowest constriction of each vowel shown on the midsagittal cut of** [ø]. **Also shown is a line demonstrating measuring of the jaw opening distance.**

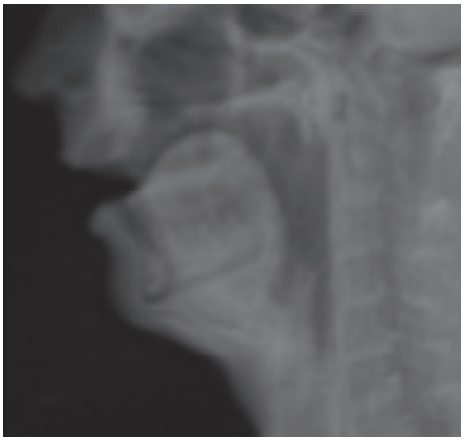

Figure 2. **An average image of the 3D MR image stack of** [e] **produced by averaging in the direction perpendicular to the sagittal plane.**

3. Results

The prototypicality and nasality scoring proved to be inconclusive. In contrast, the categorisation part of the experiment yielded a clear result as seen in Table 1.

Table 1

Confusion matrices for the listening experiments a) beginning samples and b) end samples

a) Categorised as

Target	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
[ɑ]	20	0	0	0	0	0	0	0
[e]	0	20	0	0	0	0	0	0
[i]	0	1	17	0	0	2	0	0
[o]	0	0	0	19	1	0	0	0
[u]	0	0	0	1	19	0	0	0
[y]	0	0	0	0	0	19	0	1
[æ]	0	0	0	0	0	0	20	0
[ø]	0	0	0	0	0	0	0	20

b) Categorised as

Target	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
[ɑ]	20	0	0	0	0	0	0	0
[e]	0	18	0	0	0	0	0	2
[i]	0	1	19	0	0	0	0	0
[o]	0	0	0	8	12	0	0	0
[u]	0	0	0	1	19	0	0	0
[y]	0	0	0	0	0	17	0	3
[æ]	12	0	0	0	0	0	8	0
[ø]	0	0	0	0	0	0	0	20

The confusion matrices displayed there show that [æ] and [u] in this data are not very representative at the end of the productions. It should be noted that many of the listeners reported that the productions in general were not very prototypical, but that they were nonetheless clearly categorisable in most cases. Two other frequently reported observations were the machine like quality of the speech and the fact that some of the listeners felt that some of the samples were shorter than others.

Table 2 lists the results of the acoustic analysis of the samples. As can be seen the subject was able to sustain a fairly stable f_0 and in most cases the formants provided by the analysis show only a small drift. However, there is a relatively large difference in the formants of [e], [i], [u], and [æ]. In the cases of [e] and [i], the formant extraction algorithm has produced one or more artifactual formants. In the cases of [u] and [æ], the articulation has changed considerably. These views are supported by the confusion matrices in Table 1.

The formant data for [e] and [i] was manually adjusted to correct the following artefacts. For [e]: the LPC found a peak at 272 in the begin sample but this was removed as an outlier. For [i]: First, the LPC found peaks at 987 Hz (begin) and 722 Hz (end) but these were removed as outliers. Second, F2 could not be extracted from the beginning sample by the LPC. The value in Table 2 was obtained by visual inspection of the compensated spectrum. Third, the LPC found very strong double peaks about 500 Hz apart corresponding to the actual F3. Accordingly, the F3-values given in Table 2 are defined as their means. Adjusted values are marked in bold face as well as values which have shifted to a lower formant position (e.g. F2 to F1) as the result of removing an outlier.

Table 2

f_0s and formants F1—F4 for vowel samples with target f_0 = 110 Hz

Vowel		[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
f_0 (Hz)	begin	107.6	108.1	108.4	107.3	107.3	108.4	105.3	106.8
	end	110.8	109.2	109.4	110.2	111.4	110.5	107.6	109.7
F1 (Hz)	begin	658	**560**	255	403	269	294	748	419
	end	644	524	238	392	342	303	764	452
F2 (Hz)	begin	1059	**1898**	**2220**	753	636	1577	1532	1488
	end	989	1993	**2183**	717	714	1539	1245	1360
F3 (Hz)	begin	2763	**2625**	**3258**	2298	2186	2057	2278	2008
	end	2530	2436	**3350**	2181	2160	2012	2373	2088
F4 (Hz)	begin	3643	**3402**	**4156**	3487	3381	3281	3511	3321
	end	3715	3504	**4658**	3252	3103	3148	3531	3523

The specialised formant extraction algorithms used for the noisy data in the current context are likely to undergo several improvements in the future. Thus, we provide updated results from our work as they become available at http://math.aalto.fi/en/research/sysnum/formant_extraction.html. The page includes short descriptions of the algorithms used and tables listing the corresponding extraction results.

Figure 1 shows the position of the narrowest constriction for each vowel in the vocal tract (between the lips and the epiglottis). Table 3 lists the articulatory measures: Jaw opening (distance of the maxilla and the mandible), lip opening (inner distance between the lip surfaces), and the smallest area (size of the narrowest constriction in the vocal tract) for each vowel. Lip opening area is also listed for rounded vowels.

Table 3

Articulatory measures from the MR images

	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
Jaw opening (cm)	7.4	7.3	6.8	8.2	8.2	7	8.6	8.2
Lip opening (cm)	1.3	1.7	1.2	0.5	0.6	0.6	3.1	1
Smallest area (cm^2)	1.3	1.6	0.3	0.5	0.9	1.8	2.5	3.9
Lip opening area (cm^2)	na	na	na	0.7	0.3	0.3	na	1.9

In the present data the three dimensional features of the articulations are readily visible. In all of the current vowel productions the tongue is grooved and asymmetric with respect to the mid-sagittal plane. In the vowels [y, i, e, ø, u] the tongue is in contact with the palate and in [u, ɑ, o, æ] with the pharyngeal wall.

4. Discussion

Our observations of the position of the tongue and its groovedness are well in line with the observations of earlier studies (Chiba, Kajiyama 1941; Sovijärvi 1963). It should be noted that this is the first 3D data set on Finnish and as such is potentially richer in detail than previously collected data. An X-ray image produced in the traditional way (rather than with computed tomography) is an average of the tissues in one direction. In contrast, MRI produces images as slices through the tissues. The difference is demonstrated by comparing Figures 1 and 2. However, as can be seen from our results, the MRI data will provide additional detail, while the original understanding of vowel articulation remains well founded.

It is difficult to produce good vowels in the conditions required by MRI. As our data on [u] and [æ] show, the articulatory position is liable to change during the long productions as well as being different from that employed in spontaneous speech (Engwall 2000). This is likely to be due to several different effects acting simultaneously. The supine position is likely to affect the position of the tongue. The noise of the MRI machine will cause a Lombard effect on the subject's speech. The emptying of the lungs will affect the position of the articulatory organs via the movement of the thorax. Furthermore, the long productions are more likely to produce more extreme articulation as can be seen in e.g. the very narrow lip opening of [u] and [y] in Table 3. Taking into account these considerations, this data can be used in modeling speech production not only at the given data points but also by extrapolating from them.

Acknowledgements

We wish to express our gratitude to Mr Olli Santala for helping with the MAX/MSP implementation of the listening experiments. The study has been supported by the Academy of Finland (proj. numbers 128204, 125940, and Lastu 135005).

Addresses

Pertti Palo
Institute of Behavioural Sciences, Faculty of Behavioural Sciences, University of Helsinki and Institute of Mathematics, School of Science, Aalto University
E-mail: pertti.palo@aalto.fi

Daniel Aalto
Institute of Behavioural Sciences, Faculty of Behavioural Sciences, University of Helsinki and Department of Signal Processing and Acoustics, Aalto University

Olli Aaltonen
Institute of Behavioural Sciences, Faculty of Behavioural Sciences, University of Helsinki

Risto-Pekka Happonen
Department of Oral Diseases, Turku University Hospital and Department of Oral and Maxillofacial Surgery, University of Turku

Jarmo Malinen
Institute of Mathematics, School of Science, Aalto University

Jani Saunavaara
Medical Imaging Centre of Southwest Finland, Turku University Hospital

Martti Vainio
Institute of Behavioural Sciences, Faculty of Behavioural Sciences, University of Helsinki

REFERENCES

Aalto, D., Malinen, J., Palo, P., Aaltonen, O., Vainio, M., Happonen, R.-P., Parkkola, R., Saunavaara, J. 2011, Recording Speech Sound and Articulation in MRI. — BioDevices 2011. Proceedings of the International Conference on Biomedical Electronics and Devices, Rome, January 26—29, 2011, Setubal, 168—173.

Baer, T., Gore, J. C., Boyce, S., Nye, P. W. 1987, Application of MRI to the Analysis of Speech Production. — Magnetic Resonance Imaging 5, 1—7.

Chiba, T., Kajiyama, M. 1941, The Vowel, Its Nature and Structure, Tokyo.

Engwall, O. 2000, Are Statical MRI Data Representative of Dynamic Speech? Results from a Comparative Study Using MRI, EMA and EPG. — Proceedings of International Conference on Spoken Language Processing 2000 (ICSLP 2000) I, Beijing, 17—20.

Engwall, O., Badin, P. 1999, Collecting and Analysing Two- and Three-Dimensional MRI Data for Swedish. — TMH-QPSR 3—4/1999, 11—38.

Hannukainen, A., Lukkari, T., Malinen, J., Palo, P. 2007, Vowel Formants from the Wave Equation. — Journal of the Acoustical Society of America Express Letters 122, EL1—EL7.

Jones, S. 1929, Radiography and Pronunciation. — The British Journal of Radiology 2, 149—150.

Makhoul, J. 1975, Spectral Linear Prediction. Properties and Applications. — IEEE Transactions on Acoustics, Speech and Signal Processing 23, 283—296.

Palo, P. 2011, A Wave Equation Model for Vowels. Measurements for Validation (Licentiate Thesis, Aalto University, Institute of Mathematics).

Sovijärvi, A. 1938, Röntgenogrammeja suomen yleiskielen vokaalien ääntymäasennoista. — Vir., 186—191.

—— 1963, Suomen kielen äännekuvasto, Jyväskylä.

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. — Journal of the Acoustical Society of America 122, 532—541.

Švancara, P. Horáček, J. 2006, Numerical Modelling of Effect of Tonsillectomy on Production of Czech Vowels. — Acta Acustica united with Acustica 92, 681—688.

Pertti Palo, Daniel Aalto, Olli Aaltonen, Risto-Pekka Happonen, Jarmo Malinen, Jani Saunavaara, Martti Vainio

ПЕРТТИ ПАЛО (Хельсинки), *ДАНИЭЛЬ ААЛТО* (Хельсинки), *ОЛЛИ ААЛТОНЕН* (Хельсинки), *РИСТО-ПЕККА ХАППОНЕН* (Турку), *ЯРМО МАЛИНЕН* (Хельсинки), *ЯНИ САУНАВААРА* (Турку), *МАРТТИ ВАЙНИО* (Хельсинки)

АРТИКУЛЯЦИЯ ФИНСКИХ ГЛАСНЫХ. РЕЗУЛЬТАТЫ МРТ И ЗВУКОВЫЕ ДАННЫЕ

Мы представляем анатомические и акустические данные пилотного исследования финских гласных [ɑ, e, i, o, u, y, æ, ø]. Данные получены одновременно с помощью 3D-магнитно-резонансной томографии (МРТ) и заказной системы записи звука. Они состоят из одного статического повторения каждого гласного с постоянной частотой основного тона. В статье приведены результаты слухового и акустического анализов аудио-данных, а также ручного анализа изображений МРТ.