RECORDING SPEECH SOUND AND ARTICULATION IN MRI

Daniel Aalto, Jarmo Malinen, Pertti Palo
Dept. of Mathematics and Systems Analysis, Aalto University, P.O. BOX 11100, FI-00076 Aalto, Finland

Olli Aaltonen, Martti Vainio
Institute of Behavioural Sciences, University for Helsinki, P.O. BOX 9, FI-00014 Helsingin yliopisto, Finland

Risto-Pekka Happonen, Riitta Parkkola, Jani Saunavaara
Dept. of Oral Diseases, Dept. of Radiology, Medical Imaging Center of Southwest Finland
University of Turku, FI-20014 Turun yliopisto, Finland

Keywords: MRI, Sound recording, Speech, Vowel, Formant.

Abstract: This article describes an arrangement for simultaneous recording of speech and the geometry of vocal tract. Experimental design is considered from the phonetics point of view. The speech signal is recorded with an acoustic-electrical arrangement and the vocal tract with MRI. Finally, data from pilot measurements on vowels is presented, and its quality is discussed.

1 INTRODUCTION

Helmholtz (1863) put forward the acoustic theory of vowels by showing that the perceived vowel quality depends on the resonance characteristics of the vocal tract. Since then, it has been the main approach to the acoustic theory of speech production (see, e.g., (Fant, 1960)). Based on these ideas, Mrayati, Carr, and Guerin (1988) presented the Distinctive Regions Model (DRM) of speech production suggesting that speech production derives from regions that closely correspond to established vowel and consonant places of articulation.

Based on these earlier ideas, models, and theories, it is hypothesized that the vocal tract configuration can be estimated strictly on the basis of the formant structure. Here we present a data acquisition framework for a mathematical model that not only solves the direct problem of simulating speech sound from a given 3D vocal track configuration, but also allows the prediction of vocal tract shapes on the basis of resonances corresponding to vowel *formants* — the main information bearing parameters in speech.

Such simulators (comprising only of the wave equation in the vocal tract) have been used for studying normal speech production acoustics (Hannukainen, Lukkari, Malinen, & Palo, 2007; Lu, Nakai, & Suzuki, 1993; Švancara, Horáček, & Pešek, 2004). When soft tissue and muscle models are incorporated, we expect that such a simulator is useful for studying speech production from a wider phonetics point of view, and planning and evaluating oral and maxillofacial surgery (Dedouch, Horáček, Vampola, & Černý, 2002; Nishimoto, Akagi, Kitamura, & Suzuki, 2004; Švancara & Horáček, 2006). See also Vahatalo, Laaksonen, Tamminen, Aaltonen, and Happonen (2005) and Niemi, Laaksonen, Peltomaki, Kurimo, Aaltonen, and Happonen (2006) for background.

A computational model of speech production (such as discussed in Hannukainen et al. (2007), Aalto (2009), and Aalto, Alku, and Malinen (2009)) must be validated by comparing simulated sound to measured sound in some metric (such as the resonance structure, i.e., the formants). Since the simulation is based on anatomic data, the validation of the computational model depends on recording a coupled data set: the speech sound and the precise anatomy which produces it. This requires imaging the vocal and nasal tracts from the lips and nostrils to the beginning of the trachea. We chose to use magnetic resonance imaging (MRI) technique because of its safety in contrast to X

Aalto D., Malinen J., Palo P., Aaltonen O., Vainio M., Happonen R., Parkkola R. and Saunavaara J..
RECORDING SPEECH SOUND AND ARTICULATION IN MRI.
DOI: 10.5220/0003137301680173
In *Proceedings of the International Conference on Biomedical Electronics and Devices* (BIODEVICES-2011), pages 168-173
ISBN: 978-989-8425-37-9

ray based CT-imaging.

Using MRI poses many restrictions. When a full 3D scan of the vocal tract is desired, the acquisition time will necessarily be long. During this time, the test subject needs to remain stationary; in particular, all the parts of the speech apparatus have to remain as stable as possible as well as the *fundamental frequency* (f_0). We discuss these matters in Sections 3 and 4. On the other hand, the test subject's voice is recorded simultaneously when his[1] vocal tract is scanned using MRI. This sound recording is carried out using an arrangement that has already been reported by Lukkari, Malinen, and Palo (2007), and Malinen and Palo (2009). For previous work on similar projects, see, for example, Švancara et al. (2004), Ericsdotter (2005), and Bresch, Nielsen, Nayak, and Narayanan (2006).

The purpose of this paper is to outline experimental protocols for acquiring the above mentioned data sets in high quality, discuss these protocols from a wide range of perspectives, and finally present some observations based on pilot data.

2 SPEECH RECORDING

2.1 Phonetic Materials

From an articulatory point of view, the main problem in acquiring simultaneous MRI and audio data is the subject's ability to alter their articulatory settings very effectively in a manner that retains the main acoustic characteristics of the speech sounds. It is well known that human subjects adapt their motor behavior regarding both vision and audition (Houde & Jordan, 1998; Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984). With regard to speech articulation, the adaptation of motor behavior is particularly important since the production goals (i.e., speech sounds) occur in numerous different articulatory contexts. Therefore, it can be expected that speakers are forced to adapt their articulation even during sustained vowel production as their articulatory setup changes due to, e.g., the contraction of the thorax. This causes additional problems when both MRI and audio data are acquired simultaneously and calls for careful design of the speech materials.

There are two main sources for adaptation in sustained vowel production: (1) the changing position of the larynx due to *changes in fundamental frequency*, and (2) the *changing shape of the vocal organs* due to the contracting thorax *during a long exhalation*. The first problem can be circumvented by having the subject produce the sustained vowels with a stationary fundamental frequency (f_0). The second problem cannot be completelety avoided and it calls for other means to control the matching of MRI and audio data. A partial solution is to measure the articulatory movements during the vowel production.

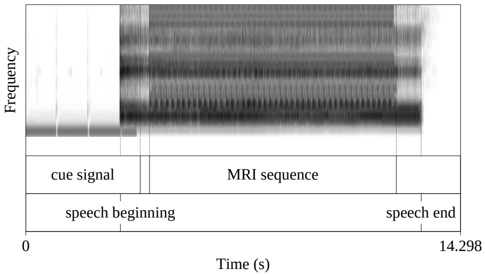

Figure 1: A spectrogram showing a full sound recording. From left to right: cue signal for the subject (≈ 3.5 s from the signal onset, overlapping speech for the last ≈ 500 ms); the clean speech sample (≈ 500 ms); the speech and the imaging noise (≈ 8 s); and the clean speech sample (≈ 500 ms).

With the above problems in mind, we designed a plan according to which the subject was asked to produce the sustained vowels with two different stationary f_0 levels (110 and 137.5 Hz; corresponding to notes A2 and C#3, respectively). Furthermore, two separate MRI imaging techniques were used: (1) a stationary 3D image of the vocal tract was produced; and (2) a dynamic 2D image showing the sagittal section of the vocal tract during the vowel production. The two f_0 levels were used to study the effect of larynx position on the vocal tract shape whereas the dynamic 2D image sequence was used to study the effect of the changing vocal tract shape due to the contracting thorax.

2.2 Experimental Setting

Before the imaging sequence starts, the subject lies inside the machine in supine position in the same way as during a standard MRI procedure of the head and neck area. In addition, the sound collector is placed upon the MRI coil, and it is positioned in front of test subject's mouth. The subject is able to speak to the control room through the sound collector all the time. Moreover, he can hear instructions from the control room as well as his own (de-noised) voice (with a delay of ≈ 20 ms due to acoustic wave guides) through earphones of the MRI machine.

Before the experiment, the subject is given a de-

[1] We shall use a male subject in the pilot stage as analyzing a male voice is generally easier than analyzing a female voice.

scription of what he is asked to do next. When the subject indicates that he is ready the experiment is started. First, the subject hears a sinusoidal cue signal that gives him a count-down for starting the utterance at the right time as well as the desired pitch, i.e., the level of f_0.

A typical sound sample, including the cue signal, is represented in Figure 1. The MRI machine is operated so that a 500 ms "pure sample" of stabilized utterance is obtained immediately before and right after the MRI noise interval.

After each experiment, the image data is inspected visually and the subject gives his comments. During the whole imaging sequence the sound sample is listened by a trained phonetician in the control room, and unsuccessful utterances are usually detected immediately. Particular attention is paid to the phonation type and nasality. When doubt arises, the formants of the sample are extracted from the two "pure samples" using Praat 4.6.15; see Figure 3 below.

2.3 Imaging Sequence

Experiments were performed on a Siemens Magnetom Avanto 1.5T scanner (Siemens Medical Solutions, Erlangen, Germany). Maximum gradient field strength of the system is 33 mT/m (x,y,z directions) and the maximum slew rate is 125 T/m/s.

12-element Head Matrix Coil was combined with the 4-element Neck Matrix Coil in order to completely cover the anatomy of interest. Coil configuration allowed the use of Generalized Auto-calibrating Partially Parallel Acquisition (GRAPPA) technique to accelerate acquisition. Technique was applied in all the scans using acceleration factor 2.

3D VIBE (Volumetric Interpolated Breath-hold Examination) was found out to be the most suitable MRI sequence for the rapid 3D acquisition required in this study. Basically, 3D VIBE is an ultra-fast gradient echo sequence with an isotropic resolution. In addition, the *k*-space scan is typically performed asymmetrically in this sequence, which reduces the number of phase encoding steps in the slice-selection direction leading to faster scan times. As the naming of the sequence suggests, it was originally developed for fast 3D imaging of the abdominal region where breath-hold during the scan is essential. Sequence parameters were optimized in order to minimize the acquisition time. The following parameters allow imaging with 1.8 mm isotropic voxels in just 7.6 s: Time of repetition (TR) was 3.63 ms, echo time (TE) 1.19 ms, flip angle (FA) 6°, receiver bandwidth (BW) 600 Hz/pixel, FOV 230 mm, matrix 128x128, number of slices 44 and the slab thickness of 79.2 mm.

When higher resolution is required, imaging with 1.2 mm isotropic resolution is possible in 17 s when following changes to parameters are applied: TR 3.95 ms, TE 1.34 ms, matrix 192x192, and 64 slices.

Dynamic MRI scans were performed using segmented ultrafast spoiled gradient echo sequence (TurboFLASH) where TR and TE were minimized. This sequence is typically used in cardiac studies but this time magnetization preparation pulse was not applied. Single sagittal plane was imaged with a pace of 5.5 images per second using parameters TR 178 ms, TE 1.4 ms, FA 6°, BW 651 Hz/pixel, FOV 230 mm, matrix 120x160, and slice thickness 10 mm.

2.4 Sound Recording

The MRI room presents a challenging environment for sound recording. Use of metal components and electronics is restricted inside the MRI room, and it is completely excluded near the MRI machine for safety and image quality reasons.

We use the sound recording arrangement detailed in Lukkari et al. (2007), and Malinen and Palo (2009): A two-channel sound collector samples the speech and noise signals in a dipole configuration. The sound collector is an acoustically passive, non-microphonic device which does not cause artifacts in the MR images, and it is also transparent in X-Ray CT. The sound signals are coupled to a RF-shielded microphone array by acoustic waveguides of length 3 m. Again, the waveguides are acoustically passive and linear, but their frequency response is far from flat because of the longitudinal resonances that, however, have been satisfactorily controlled by special impedance terminations at both ends of the wave guides. The microphone array inside its Faraday cage lies at a safe distance from the main coil of the MRI machine. The signals are coupled from the microphone array to custom RF-proof amplifier that is situated outside the MRI room. These analogue electronics are used to optimally substract the noise channel from the speech channel in real time.

The audio signal is digitized using a Digidesign M-Box model 1, 24bit ADC, controlled by a MacBookPro2,2 computer running MacOSX 10.4.9 and Pro Tools LE 7.3.1.

2.5 Acoustic Noise in the MRI Room

Loud acoustic noise is familiar to anyone who has undergone an MRI study. The noise originates from the vibrations of the gradient coil support structure. These vibrations are caused by the interactions between the pulsed magnetic fields created in gradient coils and the main magnetic field.

The dipole sound collector and the analog noise cancellation (as described in Section 2.4) takes care of a good part of the noise at low frequencies, say < 500 Hz. In particular, we are able to obtain sound recordings of vowel utterances in real time, most of which have positive S/N ratios during the MRI noise. However, we are not *always* able to produce a signal that would be directly (without further de-noising) usable for formant extraction by, e.g., the linear prediction algorithm.

For DSP-based post-processing, we record a noise sample from each MRI sequence and configuration used. During these recordings, the test subject lies silently inside the MRI machine so that the acoustic conditions are the same as in actual speech recordings. For comparison, some noise samples are also collected by a directional microphone ≈ 3 m away from the MRI machine. The frequency response of the acoustic wave guides is measured in an anechoic chamber. For further details, and for post-processing and analysis of the data we refer to Aalto, Aaltonen, Happonen, Malinen, Palo, Parkkola, Saunavaara, and Vainio (2011).

The acoustic MRI noise is significantly different for different imaging sequences. Ultrafast sequences — such as 3D VIBE used in this work — require maximal performance of gradient system both in terms of slew rate and amplitude. This results in exceptionally loud acoustic noise. We remark that even smallest changes in parameters of a given MRI sequence may change acoustic noise significantly. It is thereby essential to maintain sequence parameters and patient positioning constant.

3 RESULTS

For successful data acquisition, it is necessary to identify *relevant parameters* concerning the whole experimental setting that must be kept track of at all times. It is not always *a priori* clear what should be regarded as a relevant parameter, or what practical steps must be taken to keep them under control.

Here, the relevant parameters can be divided into two groups: (1) *physiological parameters* involving the human subject; and (2) *physical parameters* of the measurement equipment. As documented in Sections 2.2 – 2.5 above, we have spent much effort to optimize and standardize physical parameters which, indeed, we can control for, measure and compensate to a considerable extent. Some physiological parameters have been considered in Section 2.1, and they have a much more problematic nature. We proceed to discuss them in terms of our pilot experiments.

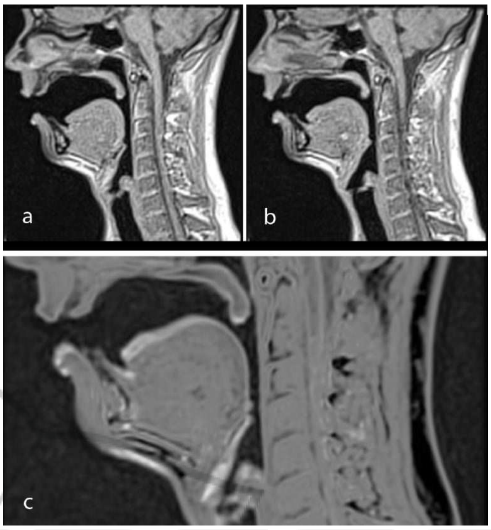

Figure 2: Mid-sagittal sections of an 8 s production of the vowel [ɑ] with a) $f_0 = 110$ Hz; and b) $f_0 = 137.5$ Hz. An overlaid image is shown in c) to indicate their difference that is visible as lighter gray. Notice in particular the difference in position of lower lip, tongue blade and larynx.

Let us start with the *level of* f_0. The subject was instructed to keep f_0 in a given reference value, and he was able to do that with an error of ± 3 Hz in all experiments. Figure 2 indicates that different levels of f_0 result in visible differences of vocal tract configuration while uttering the vowel [ɑ].

We remark that changes in sound pressure may result in a similar change as in Figure 2. To exclude this, the subject tried to keep the sound pressure same in both imaging sessions, but he did not receive any feedback in that respect. However, the (subjective) exhalation time was of the same length in both measurements. After recording and compensating the frequency response of the whole setup, the sound pressure level can be extracted from recorded signals quite precisely. It can be observed that the sound pressure given by the subject always increases towards the end of the sample.

The measured formants of [ɑ] (corresponding to the experiment giving Figure 2) are given in (F_1, F_2) -plane in Figure 3. As explained in Section 2.2, we obtain two "pure samples" of speech, and we therefore have two points (connected by a line) for each experiment in Figure 3. We have used both 7.6 s and 17 s imaging sequences, two f_0 levels, and several phonation types.

Let us next discuss the *changing of shape* of vocal organs. Figure 3 gives us a first indication that the vocal tract geometry somewhat "creeps" for some reasons during the 7.6 s MRI scan. We measured this

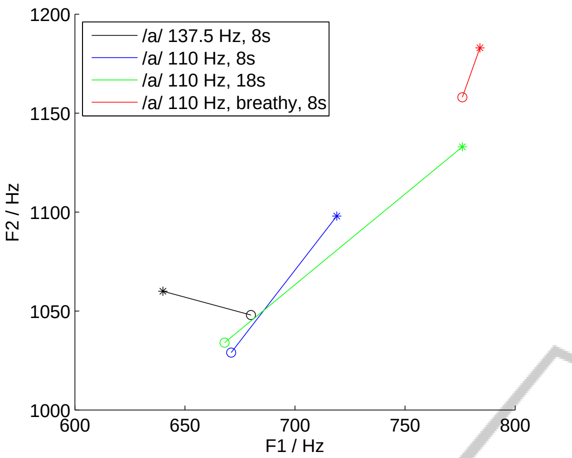

Figure 3: Formants of the vowel [ɑ] produced with $f_0 = 110$ Hz and $f_0 = 137.5$ Hz based on 500 ms samples from the beginning (marked with *) and the end (marked with o) of the sustained phonation. During these samples the MRI sequence was not active.

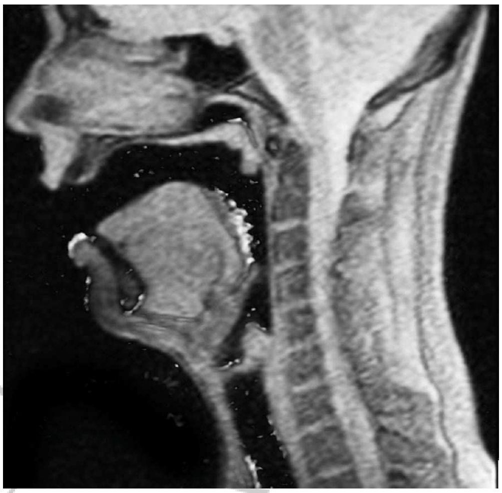

Figure 4: Overlaid image showing the difference between the first and last frame of the 8 s dynamic image sequence of [æ] with $f_0 = 137.5$ Hz. Differences are visible as lighter gray.

phenomenon for vowel [æ] directly using a dynamic MRI sequence in the sagittal plane; see Figure 4.

Figure 4 indicates quite large changes in both the mouth and pharyngeal cavities caused by the different position of the larynx during a long exhalation; in Figure 2(c) there is in addition a (phantom) doubling in the size of the vocal folds. That the larynx has moved, is also supported by the movement of vocal folds that can be observed from Figure 4. As explained in Section 2.1, some movement is to be expected because of the contracting thorax. A part of the observed changes may, however, be due to changing aerodynamic forces in the vocal tract that arise from variable air flow during the MRI scan.

We note that the formants are considered as purely acoustic parameters of the vocal tract geometry, and — as such — they do not depend *directly* on dynamic variables such as the sound pressure and the air flow but, instead, through detectable changes in vocal tract geometry including the time-dependent acoustic termination due to the vocal folds. It is not clear whether one should (for physically motivated reasons) aim at constant sound pressure or at constant air flow in a measurement leading to Figure 3. We remark, however, that observing the flow inside the MRI machine is probably very challenging.

4 CONCLUSIONS

We have described experimental protocols, MRI sequences, and a sound recording system that can be used for simultaneous sound and anatomical data acquisition of human speech. The results and experiences of a pilot experiment on vowel formants and the corresponding vocal tract geometries have been reported. Such data sets are intended for parameter estimation, fine tuning, and validation of a mathematical model for speech production as discussed in Section 1. However, these methods have a wide range of applications in phonetics and medicine.

Phonetic Remarks and Observations

The MR imaging poses severe problems with regard to both speech production by the test subject and sound recording. The MRI requires long sustained vowel production and — at the same time — it produces high levels of acoustic noise which masks the speech sound.

The phonetic and articulatory problems stem from the inability of a subject to maintain a stable vocal tract shape long enough: there is a trade-off between image quality in terms of resolution and speech production “quality” in terms of articulatory stability. The duration of an MRI scan (such as considered in Section 2.3 above) is 7.6–17 s, and the sound recording time is ≈ 2 s longer than that; see Figure 1. The subject should be able to maintain constant position, configuration of the vocal organs, all sound characteristics, and the type of phonation during the whole period. According to our experience, this is a difficult requirement even for a healthy subject.

We designed a set of recording materials to address these problems phonetically as well as possible. Our work indicates that the problems cannot be circumvented altogether. There are, however, simple means to further improve articulatory stability during recordings. We propose, at least, the following:

1. The test subject should be familiar with the MRI noise as well as the cue signal so as to perform optimally in the experimental situation.
2. The intensity of the cue signal should match the MRI noise so that the initial voice production can be maintained, i.e., possible sound intensity fluctuation should be avoided.
3. The MRI noise itself is periodic and interferes with the voice f_0 when they are close. Hence, the cue should be matched with the MRI noise frequency profile.
4. The voice sample f_0 should be standardized but in a way that depends on the test subject.
5. The cue signal should be longer to allow the subject more time to inhale.
6. Externally triggered MRI sequences can be used to introduce noiseless pauses.

All in all, the results from the current experiment are encouraging. They clearly point to directions where the setup refinements and better understanding will iteratively approach a useful solution to the whole problem.

REFERENCES

Aalto, A. (2009). A low-order glottis model with nonturbulent flow and mechanically coupled acoustic load. Master's thesis, TKK, Helsinki. Available at http://math.tkk.fi/research/sysnum/.

Aalto, A., Alku, P., & Malinen, J. (2009). A LF-pulse from a simple glottal flow model. *MAVEBA 2009* (pp. 199–202). Florence, Italy.

Aalto, D., Aaltonen, O., Happonen, R., Malinen, J., Palo, P., Saunavaara J., & Vainio, M. (2011). Recording speech sound and articulation in MRI. Analysis and postprocessing of audio-spatial data. In preparation.

Branderud, P. (2008). Personal communication.

Bresch, E., Nielsen, J., Nayak, K., & Narayanan, S. (2006). Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans (L). *Journal of the Acoustical Society of America*, *120*(4), 1791 – 1794.

Dedouch, K., Horáček, J., Vampola, T., & Černý, L. (2002). Finite element modelling of a male vocal tract with consideration of cleft palate. *Forum Acusticum*. Sevilla, Spain.

Ericsdotter, C. (2005). *Articulatory-Acoustic Relationships in Swedish Vowel Sounds*. PhD thesis, Stockholm University, Stockholm, Sweden.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

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*(1), EL1–EL7.

Helmholtz, H. L. F. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage fr dieTheorie der Musik*. Braunschweig: F. Vieweg.

Houde, J., & Jordan, M. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213.

Kelso, J., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. (1984). Functionally specific articulatory adaptation to jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology*, *10*(6), 812–832.

Lu, C., Nakai, T., & Suzuki, H. (1993). Finite element simulation of sound transmission in vocal tract. *J. Acoust. Soc. Jpn. (E)*, *92*, 2577 – 2585.

Lukkari, T., Malinen, J., & Palo, P. (2007). Recording Speech During Magnetic Resonance Imaging. *MAVEBA 2007* (pp. 163 – 166). Florence, Italy.

Malinen, J., & Palo, P. (2009). Recording speech during MRI: Part II. *MAVEBA 2009* (pp. 211–214). Florence, Italy.

Mrayati, M., Carr, R., & Guerin, B. (1988). Distinctive regions and modes: a new theory of speech production. *Speech Communication*, (7), 257–286.

Niemi, M., Laaksonen, J., Peltomaki, T., Kurimo, J., Aaltonen, O., & Happonen, R. (2006). Acoustic comparison of vowel sounds produced before and after orthognathic surgery for mandibular advancement. *Journal of Oral & Maxillofacial Surgery*, *64*(6), 910–916.

Nishimoto, H., Akagi, M., Kitamura, T., & Suzuki, N. (2004). Estimation of transfer function of vocal tract extracted from MRI data by FEM. *The 18th International Congress on Acoustics*, Vol. II (pp. 1473 –1476). Kyoto, Japan.

Vahatalo, K., Laaksonen, J., Tamminen, H., Aaltonen, O., & Happonen, R. (2005). Effects of genioglossal muscle advancement on speech: an acoustic study of vowel sounds. *Otolaryngology - Head & Neck Surgery*, *132*(4), 636–640.

Švancara, P., & Horáček, J. (2006). Numerical Modelling of Effect of Tonsillectomy on Production of Czech Vowels. *Acta Acustica united with Acustica*, *92*, 681 – 688.

Švancara, P., Horáček, J., & Pešek, L. (2004). Numerical modelling of production of Czech Wovel /a/ based on FE model of the vocal tract. *Proceedings of International Conference on Voice Physiology and Biomechanics*.