Large scale data acquisition of simultaneous MRI and speech

Daniel Aalto[a], Olli Aaltonen[a], Risto-Pekka Happonen[b,c], Päivi Jääsaari[b], Atle Kivelä[d], Juha Kuortti[d], Jean-Marc Luukinen[b,c], Jarmo Malinen[d,*], Tiina Murtola[d], Riitta Parkkola[e], Jani Saunavaara[f], Tero Soukka[b,c], Martti Vainio[a]

[a] *Inst. of Behavioural Sciences (SigMe group), University of Helsinki, Finland*
[b] *Dept. of Oral and Maxillofacial Diseases, Turku University Hospital, Finland*
[c] *Dept. of Oral and Maxillofacial Surgery, University of Turku, Finland*
[d] *Dept. of Mathematics and System Analysis, Aalto University, Finland*
[e] *Dept. of Radiology, University of Tampere, Finland*
[f] *Dept. of Radiology, Medical Imaging Centre of Southwest Finland, University of Turku, Finland*

Abstract

We describe an arrangement for simultaneous recording of speech and vocal tract geometry in patients undergoing surgery involving this area. Experimental design is considered from an articulatory phonetic point of view. The speech signals are recorded with an acoustic-electrical arrangement. The vocal tract is simultaneously imaged with MRI. A MATLAB-based system controls the timing of speech recording and MR image acquisition. The speech signals are cleaned from acoustic MRI noise by an adaptive signal processing algorithm. Finally, a vowel data set from pilot experiments is qualitatively compared both with validation data from the anechoic chamber and with Helmholtz resonances of the vocal tract volume, obtained using FEM.

Keywords: Speech production, speech recording, MRI, noise reduction, formant analysis, vocal tract resonance.

1. Introduction

A. M. Liberman suggested that speech is a special code [1]. Literate people are taught to think that speaking is like writing, and that a speaker produces a distinctive acoustic pattern of energy for every distinct vowel and consonant that we perceive, much as a typewriter produces letters. If human speech were segmented at the acoustic level, the task of speech perception would be simply a matter of identifying sounds one-by-one from the speech signal, chaining them into words, and associating these with meanings stored in memory.

Speech, however, is not perceived, produced, or neurally programmed on a segmental basis. Such spelling out loud is far too slow and tedious for human communication. Instead, utterances are produced and perceived as a whole. We perceive speech by virtue of our tacit knowledge of how speech is produced. Thus, the elements of speech are

[*] Corresponding author

articulatory gestures, not the sounds these phonetic gestures produce. The gestures are the ultimate constituents of language which must be exchanged between a speaker and a listener if communication through language is to occur.

The human articulatory system is the only one anatomically and neurally efficient enough to produce acrobatic manoeuvres of speech organs fast, without errors, and with minimal energy. The main vocal tract elements used in producing phonetic gestures are the lips, tongue tip and tongue dorsum, soft palate, and the larynx. By combining their movements in various ways meaningful linguistic units can be built up and conveyed via sound. Observing as well as modelling the related biophysical features and dynamic phenomena is far from a trivial matter even if state-of-the-art instruments and methods (such as computational modelling based on modern medical imaging technologies) are available. Challenging as they are, these approaches appear quite promising for adding to our current understanding of what happens during normal or pathological speech.

Modelling based on multi-modal data sets

Perhaps the most important reason for modelling and simulations is the inherent difficulty in observing speech biophysics in test subjects directly. Further compelling motivation is provided by many situations where experiments cannot be arranged at will: consider, e.g., the acoustic effect of tonsillectomy [2] or the suitability of the vocal tract structures in *Homo neanderthalensis* for speech [3].

Mathematical models of human speech production have been used for speech analysis, processing, and synthesis as well as studying speech production acoustics for a long time; see, e.g., [4, 5, 6, 7]. Many of the earlier models were based on radical simplifications of the underlying physics and anatomic geometry such as the Kelly–Lochbaum model [8] and many approaches of transmission line type; see, e.g., [9, 10, 11, 12, 13, 14]. Anatomic data for early models used to be rather scarce due to difficulties in data acquisition as has been explained in [11, pp. 799–800]. The vocal tract geometry was often determined approximately by extrapolation, based on the mid-sagittal section as proposed in [15, 16]. Due to modern fast and cheap computing of large scale systems, heavier computational acoustics models [17, 18, 19, 20, 21, 22] and Computational Fluid Dynamics (CFD) models [23, 24] are replacing earlier approaches where higher precision is required. The progress in modelling now depends crucially on getting a large number of three-dimensional (3D) geometries of the whole speech apparatus in high resolution.

This article has background and motivation in anatomic data acquisition for acoustic modelling of a stationary vocal tract by the 3D wave equation (or its resonance version, the Helmholtz equation) and Webster's horn model. Without being numerically as prohibitively heavy as most CFD models are, these acoustics models are well-suited for studying speech for medical purposes as well as for basic research. It is further expected that after incorporating *dynamic* soft tissue response and muscle action into such models, their usability would extend into studying normal and pathological speech production from an articulatory point of view [2, 25, 26, 27]. However, even the high resolution acoustic modelling based on *static* vocal tract geometries in 3D provides novel tools for planning and evaluating oral and maxillofacial surgery and rehabilitation [28, 29].

Before a computational model for speech production (such as those that have been described in our earlier work [17, 30, 32, 33, 31]) can be used for any practical or theoretical purpose, there are always some model parameters that need to be estimated based on measurements from human subjects. Such parameters, of course, include the geometry of the vocal and the nasal tracts from the lips and nostrils to the beginning of the trachea. To have a sufficient degree of confidence in the simulation results, any such model must have been rigorously validated by extensively and methodically comparing simulated speech sounds (or their characteristics) to measurements. One way of doing the validation is by comparing the measured and the simulated formants that are related to the acoustic Helmholtz resonances of the vocal tract; see [33, 34, 35]. In any case, the validation of the computational model depends on recording a coupled, multi-modal data set: speech sound and the precise anatomy which produces it.

Magnetic Resonance Imaging (MRI) has been a popular approach for acquiring geometric data of the vocal tract for a long time, and the current literature is far too extensive to present a full account of. Out of the pioneering work, we should mention at least the seminal papers [10, 11, 12] as well as the more recent works [26, 36, 37, 38, 40] which contain many further references. It is well known that recording speech samples during MRI is challenging due to matters such as the high acoustic noise level [41, 42, 38]. Much attention has been paid to the noise even for reasons that are not related to recording speech [43, 44].

Purpose and outline of the article

We have developed an experimental arrangement to collect a large data set using simultaneous MRI and speech recordings as reported in [45, 46]. The experimental arrangement includes custom hardware, software, and experimental protocols.

In contrast to many recent works that concentrate on dynamic MRI during natural speech (see, e.g., [42, 47] in 2D and [48] in 3D using compressed sensing techniques), our focus is in static 3D MRI during prolonged vowel utterance. This is the kind of data that is most suitable for computational acoustics models utilising Finite Element Method (FEM) for solving the required partial differential equations such as Eqs. (3) below. Compared to earlier similar experiments such as [11, 12, 36, 49, 50], our objective is to create a data set that is sufficiently large for statistically sound modelling and model validation, ultimately comprising several thousands of simultaneous samples of speech and MR images of the vocal tract. The scale, quality, and consistency of such "big data" depends on the optimisation of experimental protocols, and standardisation of measurement arrangements and acoustic details of sound samples during MRI as pointed out in [47]. The requirements of acoustic modelling have been central in designing the post-processing of image and sound data as well. For example, the noise cancellation procedure described in Section 4 is optimised and validated for precise formant extraction from static vowels as opposed to, e.g., the approach in [41] where carefully syncronised samples of spontaneous speech and dynamic MRI are sought.

During a pilot stage in June 2010, a set of measurements were carried out on a healthy 30-year-old male subject (in fact, one of the authors of this article), confirming the feasibility of the arrangement and the high quality of the data obtained [51, 52]. These pilot measurements also revealed a number of issues to be addressed before tackling the main objective: obtaining a clinically relevant data set from a large number of patients. The purpose of this article is to describe the final experimental arrangement including the improvements which take into account these issues. A second pilot experiment was carried out in June 2012 on a healthy 26-year-old male subject (one of the authors of this article as well) in order to validate the final experimental setup. The geometric data in Fig. 3 as well as the recorded vowel data in Section 5 are from these experiments. All patient data is excluded from this article.

This article consists of three parts that document the main aspects of MRI experiments during speech. In Section 2 the experimental design is discussed from the phonetic and physiologic points of view. Technical questions related to MRI and the simultaneous speech recording are discussed in Section 3. The acoustic instrumentation is treated only briefly, and we refer to earlier work [45, 46, 51, 52] for details. Instead, we concentrate on the software and digital parts of the measurement system, optimisation of the MRI sequences, and the automated control and timing of the experiments. Sections 4 and 5 are devoted to digital signal processing of the recorded signals: removing acoustic MRI noise and artefacts, extracting formants, and validating the results.

The patient group

Since large data sets are notoriously expensive to create, the acquired data should have multiple uses in addition to modelling of speech. For this reason, the experimental procedures have been designed to assess acoustic and anatomic changes in patients undergoing oral or maxillofacial surgery which causes changes in the vocal tract. Patients of orthognathic surgery are a particularly attractive study group for mathematical modelling of the speech production. Not only are these patients mostly young adults without any significant underlying diseases, but there is a strong medical motivation for a comparative study of their pre- and post-operative speech as well.

Orthognathic surgery deals with the correction of abnormalities of the facial tissues. The underlying cause for abnormality may be present at birth, or it may be acquired during the life as the result of distorted growth. Orthodontic treatment alone is not adequate in many cases due to severity of the deformities. In a typical operation, the position of either one jaw (mandible or maxilla) or both jaws is surgically changed in relation to the skull base. The movement of the jaws in orthognathic surgical treatment can cause noticeable changes in the relative position of the jaw in anteroposterior, vertical, and lateral direction.

Change in anterior or posterior direction varies usually in the range of 5 to 12 mm. This movement has a profound effect on the shape and volume of vocal tract, resulting in detectable changes in acoustics [28] that can be measured from speech samples recorded in optimal conditions. Although the surgery involves mandibular and maxillary bone, changes occur also in the position and shape of the soft tissues defining the vocal tract. This change is easily quantifiable using 3D MRI, and the typical change in the anatomy is 5–20 times as large as the accuracy of the resulting computational geometry for numerical acoustics; see [53]. Any linear partial differential equation describing the vocal tract acoustics (of which the Helmholtz model in Eqs. (3) is one example) can be solved using FEM practically without introducing any numerical error. Thus, the acoustic change of a typical orthognathic operation can be replicated or even predicted computationally by using modified vocal tract geometries. However, the change may not be detectable in the speech signals that have been recorded in pre- and post-operative MRI examinations[1].

[1] The full set of simultaneous MRI and speech recordings is to be used for improving and validating a computational acoustics model of speech, and not for detecting an acoustic change in a single patient. To observe the effect of surgery in speech directly, one should carry out pre- and post-operative speech experiments in an anechoic chamber under optimal conditions.

At the time of writing of this article, ten orthognathic surgery patients (out of which six are female) have undergone their pre-operative MRI examinations following the methods and protocols described here. We expect to enroll the total of 20 patients (10 adults of both sexes) in this research. The study design has been approved by the Ethics Committee of the Hospital District of Southwest Finland.

2. Experimental arrangement

Generally speaking, the experimental setting is similar to the setting in which the pilot arrangement was tested [51, 52] but with numerous improvements. They are related to instructing and cueing the patient, the role of the experimenter, and the automated control and timing of MR imaging.

The creation of the original pilot data reported in [51] required 3 – 4 people working simultaneously in the MRI control room. The improved arrangement described in this article requires only two people: one for MR imaging and the other for running the integrated experimental control system and sound recording. Moreover, it is now possible to produce as many as 90 takes during a session of 1 h which is about four times as fast a data collection rate as can be attained using a non-automated system. The streamlining of all procedures is vital because laboratory downtime and cost must be minimised when gathering a large data set. Overly long MRI sessions also compromise patient comfort and performance.

2.1. Phonetic material

The speech materials have been chosen to provide a phonetically rich data set of Finnish speech sounds. The chosen MRI sequences require up to 11.6 s of continuous articulation in a stationary position. We use the Finnish speech sounds for which this is possible: vowels [ɑ, e, i, o, u, y, æ, œ], nasals [m, n], and the approximant [l]. A long phonation is possible also for, e.g., [j, s, ŋ] but these have been excluded because of unpleasantness in supine production ([ŋ]) and turbulences in the vocal tract ([j, s]).

Patients are instructed to produce each of the sounds at a sustained fundamental frequency (f_0). We use two different f_0 levels (104 and 130 Hz for men, 168 and 210 Hz for women) for the sounds [ɑ] and [i] to obtain the vocal tract geometry with different larynx positions. The rest of the sounds are produced at the lower f_0 only. The f_0 levels have been matched with the acoustic MRI noise frequency profile to avoid interference.

In a sustained phonation, the long exhalation causes contraction in the thorax and hence a change in the shape of vocal organs. The stationary 3D imaging sequence which is used to obtain the vocal tract geometry provides no information on this adaptation process, so additional dynamic 2D imaging on the mid-sagittal section for the sounds [ɑ, i, u, n, l] is used to monitor articulatory stability.

Speech context data is also acquired by asking the patient to repeat 12 phonetically rich sentences containing all Finnish monophones [54]. In addition, the cardinal vowels [ɑ] and [i] are produced in delexical nasal stop context (i.e., syllable repetition). These continuous speech samples are imaged using the same dynamic 2D sequence which is used for checking articulatory stability.

An instruction and cue signal is used to guide the patient through each measurement. The signal consists of three parts as shown in Fig. 1: (i) recorded instructions specifying

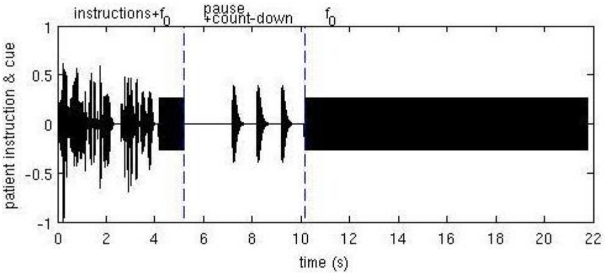


Figure 1: *Patient instruction and cue signal structure.*

the task with a sample of the desired f_0, (ii) a 2 s pause and three count-down beeps one second apart, and (iii) continuous f_0 for 11.6 s. In case of speech context experiments, the recorded instructions specify the sentence to be repeated and f_0 is left empty in both parts (i) and (iii). Audibility of the f_0 cues over MR imaging noise is achieved by using a sawtooth waveform.

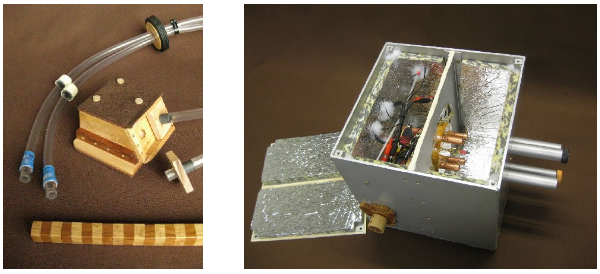
2.2. Setting for experiments

The patient lies supine inside the MRI machine with a sound collector placed on the Head Coil in front of the patient's mouth. The patient can communicate with the control room through the sound collector and the headphones of the MRI machine. The patient can also hear his or her own (denoised) voice through the standard Siemens MRI headphones (having 26 dB nominal damping of external noise) with a delay of approximately 90 ms.

The patients familiarise themselves with the tasks and the phonetic materials before the beginning of a measurement session. They also practice the tasks under the supervision and are given feedback on their performance. At the start of a measurement, the experimenter selects the phonetic task following a pre-defined random order. The patient then hears the recorded instructions. The instructions, the following pause, and count-down beeps give the patient time to prepare for the speech production task. The phonation is started immediately after the count-down beeps. The patient hears the target f_0 in the headphones added to his or her own (denoised) voice throughout the phonation.

MR imaging for static 3D and dynamic stability check sequences is started 2 s after the start of phonation and finishes approximately 500 ms before the end of phonation. Thus “pure samples” of stabilised utterance are available before and after the imaging sequence. Two 200 ms breaks are inserted into the MRI sequences to give two more pure samples. The duration of these breaks has been determined based on the half-time of the imaging noise in the MRI room, which was measured to be approximately 20 ms. Sentence and syllable imaging sequences start simultaneously with phonation and end after 3.2 s.

The experimenter listens to the speech sound throughout the experiment, allowing unsuccessful utterances to be detected immediately. At the end of the experiment, the experimenter writes comments and observations into a meta-data file. The recorded sound pressure levels are also inspected. Unsuccessful measurements are repeated, at the experimenter's discretion, immediately.



(a) (b)

Figure 2: (a) The sound collector with one of the two audio wave guides attached. (b) The microphone array inside the Faraday cage.

3. Simultaneous MRI and speech recording

The MRI room presents a challenging environment for sound recording due to acoustic noise and interference to electronics from the MRI machine. For safety and image quality reasons, use of metal is restricted inside the MRI room and prohibited near the MRI machine. Our approach is to use non-microphonic, passive acoustic instruments (without moving or vibrating parts) for collecting the sound samples and transmitting them to a safe distance from the MRI machine. Alternative solutions would be (i) using an optical microphone inside the MRI machine [38, 41, 42, 55], (ii) recording by conventional directional microphones sufficiently far away from the MRI machine that has an open construction [56, 57], (iii) taking an electret microphone inside a low-field MRI machine [11], and even (iv) using the internal microphone of the MRI machine itself [2].

3.1. Speech recording

We use instrumentation specially developed for speech recording during MRI [45, 46]: A two-channel sound collector (Fig. 2a) samples the speech and primary noise signals in a dipole configuration. The separation of these two channels is excellent because the acoustic space inside the MRI head and neck coil is well separated from the exterior acoustic space by the construction and placement of the sound collector; for experimentation with optical microphones in dipole configuration, see [41, p. 1791]. The sound signals are coupled to a microphone array inside a sound-proof Faraday cage (Fig. 2b) by acoustic waveguides of length 3.00 m (the ends of which can be seen in Fig. 2a). The microphone array contains four electret microphones of type Panasonic WM-62 (sensitivity -45 ± 4 dB re 1 V/Pa at 1 kHz) provided with 5.0 V bias, of which two are used for the signals coming from the sound collector. Two additional “ambient noise” samples are collected: one from the microphone array inside the Faraday cage (by one of the two reserve microphones in the array) and another from inside the MRI room using a custom directional microphone (containing another Panasonic WM-62 unit) near the patient’s feet, pointing towards the patient’s head and the MRI coil.

The four signals are coupled from the microphones to a custom RF-proof amplifier that is situated in the measurement server rack (shown in Fig. 4a) outside the MRI room. The amplifier contains additional circuitry (i.e., a long-tailed pair with a constant emitter current source) for optimal, real time analogue subtraction of the primary noise

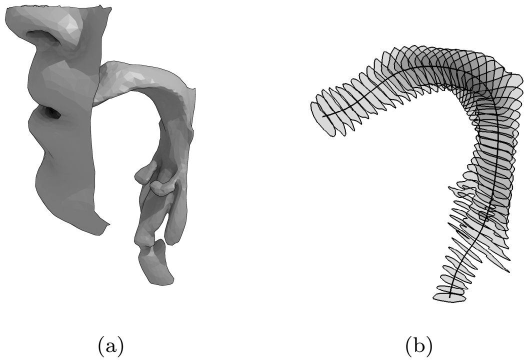


Figure 3: (a) The surface model of the tissue-air interface of a male vocal tract while pronouncing [œ]. (b) The centreline and intersection areas extracted from the same geometry.

channel from the speech channel. This is intended to produce the denoised signal played back to the patient, and it is used for audio signal quality observation in the MRI control room as well. The final, high-quality denoised signal is not produced this way but from digitised component signals by the algorithm discussed in Section 4. We remark that the hardware appears to be able to transmit good signal at least up to 10 kHz but we use only the phonetically relevant frequency range below 4.5 kHz where the measured frequency response is given in Fig. 4b.

Audio signals are converted between analogue and digital forms using a M-Audio Delta 1010 PCI Audio Interface. A measurement server is used which has an Intel Core i7-860 processor clocked at 2.80GHz, and is equipped with 4Gb RAM and a SSD drive for fast booting. For immediate internal data backup, three additional 1.5TB discs are set up in RAID1 configuration by a HighPoint RocketRaid 2302 controller. The whole setup is powered by an APC Smart-UPS SC 450VA, and it is installed to a portable 10U rack as shown in Fig. 4a. All user access to the server is done with laptops (in fact, MacBooks) running X11 servers, either via 1GB LAN or a wireless access point.

3.2. Magnetic resonance imaging

Measurements are 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. A 12-element Head Matrix Coil and a 4-element Neck Matrix Coil are used to cover the anatomy of interest. The coil configuration allows the use of Generalized Auto-calibrating Partially Parallel Acquisition (GRAPPA) technique to accelerate acquisition. This technique is applied in all the scans using acceleration factor 2.

3D VIBE (Volumetric Interpolated Breath-hold Examination) MRI sequence [58] is used as it allows for the rapid static 3D acquisition required for the experiments. Sequence parameters have been optimized in order to minimize the acquisition time. The following parameters allow imaging with 1.8 mm isotropic voxels in 7.8 s: Time of repetition (TR) is 3.63 ms, echo time (TE) 1.19 ms, flip angle (FA) 6°, receiver bandwidth (BW)

	3D static	2D stability	2D sentence/syllable
Pulse separation	240 ms	140 ms	150 ms
Number of sequence parts	35	69	20
Pause after sequence part	12 and 24	23 and 43	no pause

Table 1: External triggering parameters used in MRI scans. In 3D scans, the sequence parts are slice encoding segments. In 2D scans, the parts refer to the number of measurements.

600 Hz/pixel, FOV 230 mm, matrix 128x128, number of slices 44 and the slab thickness of 79.2 mm. Running the sequence with these settings creates broadband acoustic sound with pitch at 275.5 Hz.

Dynamic MRI scans are performed using segmented ultrafast spoiled gradient echo sequence (TurboFLASH) where TR and TE have been minimized. Single sagittal plane is imaged 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. For this sequence, the acoustic MRI noise has a more variable sound profile with less clear pitch compared to the 3D VIBE sequence discussed above.

In many earlier experiments, the MRI sequence and other parts of the experiment have been syncronised by special arrangements for different reasons: see, e.g., [41] where speech recording was precisely syncronised with dynamic MRI using the 10 MHz clock signal from the MRI machine. We use external triggering of the MRI machine in all three different types of experiments not only for syncronisation of the experiment but also for introducing pauses during which the acoustic noise is greatly reduced. Siemens Magnetom Avanto 1.5T units have inputs that accept external syncronisation signals for timing the MRI sequences. The triggering signal is always a train of 12 ms TTL level 1 pulses separated by TTL level 0 of variable duration. The pulse train is generated with a custom-made device which converts 1 kHz analogue sine signal from the sound system to the logic pulses in the same time base as the cue signals. External triggering with the additional pauses increases the 3D imaging time to 9.1 s. The details of triggering are given in Table 1.

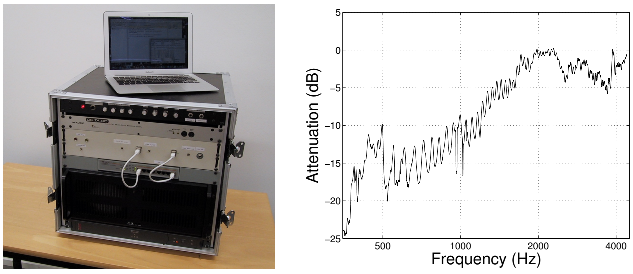
Post-processing of the MR images and the resolution of the obtained vocal tract geometries have been discussed in [53, 59].

Visibility of teeth

Teeth are not visible in MRI but they are an important acoustic element of the vocal tract. Hence, it is necessary to add teeth geometry into the soft tissue geometry obtained from the MR images during post-processing. Optical scans of teeth or digitalised dental casts can be readily obtained from the patients but automatic alignment of the two geometries is a non-trivial problem. Markers containing vegetable oil attached to the surface of the teeth appear to be a practical approach that produces sufficient MRI visibility; see also [38] where Gd-based markers were used. Further work is still required to get a solution for alignment that does not require extensive manual work.

3.3. Control of measurements

Measurements are controlled with a custom code in MATLAB [39] 7.11.0.584 (R2010b) running on the portable server with operating system Ubuntu 10.04 LTS on Linux



(a) (b)

Figure 4: (a) The measurement server with the RF-proof four-channel amplifier, M-Audio Delta 1010 Audio Interface, and networking facilities. The laptop is used for remote access. (b) The measured frequency response of the whole audio signal path from the sound collector to the amplifier input. The non-flat frequency response is compensated by DSP in post-processing stage.

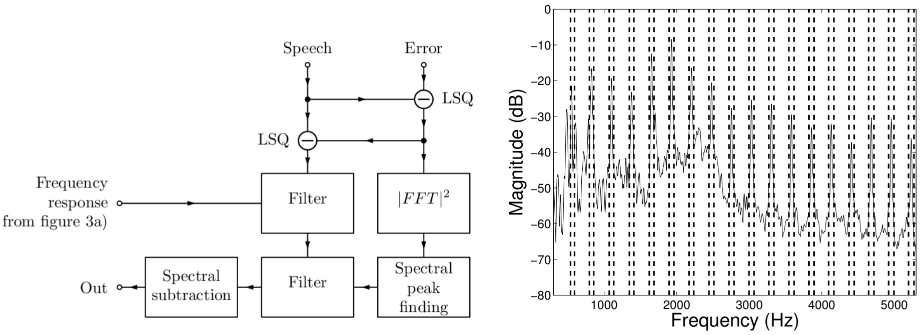
2.6.32.38 kernel (Fig. 4a). Access from MATLAB to the Audio Interface is arranged through Playrec (a MATLAB utility, [60]), QjackCtl JACK Audio Connection Kit (v. 0.3.4), and JackEQ 0.4.1.

The custom code computes the input signal to the MRI triggering device, reads the patient instruction and cue audio file, and assembles the two signals into a playback matrix. Recording is started simultaneously with playback and carried on for an equal number of samples. In addition to the speech and the three noise signals, recording also includes the analogically denoised signal and the patient instruction signal.

Both digital and acoustic parts of the audio configuration cause delays. Speech and noise signals are transmitted acoustically with identical delays from near the patient to the microphone assembly, and the instruction and cue signal is also transmitted acoustically to the patient headphones that are part of the Siemens Avanto 1.5T unit. The noise produced by MR imaging is first recorded approximately 60 ms (MRI machine delays excluded) after the onset of a trigger pulse in MATLAB. This is accounted for by the method of locating the “pure samples”. The corresponding delay in the cue-patient-record -loop is approximately 90 ms (patient reaction time excluded). The difference in the delays causes the MRI machine to start and finish the tasks 30 ms ahead of the patient. However, as the patient is asked to begin phonation 2 s before and carry on 500 ms after imaging, the impact is negligible in practice. The patients also hear their own voices with the delay of 90 ms which may cause an echo effect. If this disturbs the patient, particularly during sentence repetition tasks, speech feedback may be turned off or its volume reduced independent of the cue signal.

The control code automatically saves the recorded sounds as a six-channel Waveform Audio File. A separate file containing meta-data is also saved automatically. The meta-data file contains all experimental parameters, including task specification, and the locations of the pure samples in the sound file.

The control system requires user input for three tasks. First, the experimenter selects the next phonetic task (target sound or sentence and f_0) and MR imaging sequence. Second, comments and observations may, if necessary, be written about each measurement separately. They are saved automatically in the meta-data file in JSON format. And third, patient headphone volume and recorded sound pressure levels may be adjusted



(a) (b)

Figure 5: (a) Block diagram of the noise reduction algorithm. (b) The spectral noise peaks detected during the phonation of [ɑ] . Note the regular harmonic structure defining the stop bands.

manually based on feedback from the patient and rudimentary post-experimental sound data checks. The sound data checks consist of histograms of recorded signal levels, and they are displayed to the experimenter automatically at the end of each measurement. This allows detection and correction of settings for which the recorded signal levels, which vary for different speech sounds, are outside the optimal range.

A single measurement takes on average 30-40 s, including task selection by the experimenter and writing additional information and observations in the meta-data file. At the time of writing of this article, ten patients of orthognathic surgery have taken part in the experiments, and the times spent inside the MRI machine were between 50–95 min. When running the experiment at a comfortable pace for the patient, between 93 and 107 MRI scans have been produced in a single session.

4. Post-processing of speech signals

As explained in Section 3.1, two sound channels are acoustically transmitted from near the test subject inside the MRI machine. One of the channels provides the speech sample $s(t)$ (which is contaminated by acoustic MRI noise), and the other is reserved for the acoustic MRI noise sample $n(t)$ (which, in turn, is contaminated by speech). The analogically produced weighted difference of these signals is fed back to the subject's headphones during the experiment in almost real time. Both the signals $s(t)$ and $n(t)$ are also recorded separately, so that more refined numerical post-processing can be performed later.

Because of the multi-path propagation of the noise in particular around the MRI coil surfaces, the recorded noise sample is a weighted sum of more simple signals with distributed delays. As a further complication, the chassis of the MRI machine acts as a spatially distributed acoustic source, and its dimensions are large compared to wavelengths in air at frequencies of interest. Hence, some residual higher frequency noise will remain after an optimised subtraction of the noise $n(t)$ from the contaminated speech signal $s(t)$. To reduce this residual noise, *adaptive spectral filtering* is used. The approach is based on the observation that the typical noise spectrum of a MRI machine consists of narrow and high peaks with significant harmonic overtones; see [43] for a good treatment of such noise. Adaptivity is desirable because the peak positions depend on the MRI sequence used, and they are not invariant of time even within a single MRI sequence.

The noise reduction algorithm is outlined in Fig. 5a, and it consists of the following Steps 1–7 that have been realised as MATLAB code:

1. **LSQ:** Noise is removed from speech using linear, least squares optimal subtraction as detailed below. This reproduces roughly the same quality of speech signal that was produced analogically during the experiment for patient's headphones in real time.
2. **Frequency response compensation:** The measured non-flat frequency response of the measurement system, shown in Fig. 4b, is compensated.
3. **Noise peak detection:** The noise power spectrum is computed by FFT, and the most prominent spectral peaks of noise are detected.
4. **Harmonic structure completion:** The set of noise peaks is completed by its expected harmonic structure to ensure that most of the noise peaks have been found; see Fig. 5b.
5. **Chebyshev peak filtering:** Each of the noise peaks defines a centre frequency of a corresponding stop band. The width of the stop band is a function of the centre frequency given by Eq. (2). The corresponding frequencies are attenuated from the denoised speech signal (that has been produced in Step 2 above) by Chebyshev filters of order 20 at these stop bands.
6. **Low pass filtering:** The resulting signal is low pass filtered by a Chebyshev filter of order 20 and cut-off frequency 10 kHz.
7. **Spectral subtraction:** A sample of the acoustic background of the MRI room (without patient speech and the noise during the MRI sequence) is extracted from the beginning of the speech recording. Finally, the averaged spectrum of this is subtracted from the speech signal in frequency plane using FFT and inverse FFT; see [61].

The optimal linear subtraction in Step 1 is carried out by producing denoised signals $\tilde{s}(t)$, $\tilde{n}(t)$ from the original signals $s(t)$ and $n(t)$ according to

$$\tilde{n} = n - \frac{\langle n, s \rangle}{||n|| \cdot ||s||} s \quad \text{and} \quad \tilde{s} = s - \frac{\langle s, \tilde{n} \rangle}{||s|| \cdot ||\tilde{n}||} \tilde{n} \tag{1}$$

where $\langle n, s \rangle = \int n(t)s(t)\, dt$ and $||s||^2 = \langle s, s \rangle$. The bandwidths $w(\cdot)$ in Step 5 are defined as a function of the centre frequency f by the empirical formula

$$w(f) = C \ln f \quad \text{satisfying} \quad w(550 \text{ Hz}) = 50 \text{ Hz}. \tag{2}$$

The numerical parameters values (i.e., the bandwidth parameter C in Eq. (2), the filter order, and the cut-off frequency) have been determined by trial and error to get audibly good separation of speech and noise in prolonged vowel samples. In particular, choosing the bandwidth parameter C for Step 5 is crucial for the outcome. The cut-off frequency of 10 kHz in Step 6 is chosen well above the phonetically relevant part of the frequency range that extends up to 4.5 kHz corresponding to Fig. 4b.

The algorithm produces denoised speech signals where the S/N ratio is audibly much improved compared to the mere optimal subtraction as defined in Eq. (1). Each speech sample contains 2 s of undisturbed speech before acoustic MRI noise starts, and comparing the amplitude of the speech channel signal just before and right after the noise onset, we can get an estimate for the S/N ratio (assuming that the speech amplitude remains reasonably constant at the MRI noise onset, and that speech and noise are uncorrelated). As a rule of thumb, we obtain cleaned-up vowel signals whose S/N ratio lies between 1.9 dB and 3.6 dB, the average being 2.8 dB. The S/N ratio depends on the vowel because the emitted acoustic power tends to be larger for vowels with larger mouth opening area. The Chebyshev filtering in Step 5 creates an audible "musical noise" artefact to speech signals but we have not carried out perceptual evaluation of the denoised signals as was done in [62].

The subtraction of noise with a "spiky" power spectrum from, e.g., speech is a classical problem in audio signal processing. In [42], sufficient noise reduction was achieved by time-domain subtraction of carefully syncronised speech and noise samples. The nonlinear *cepstral transform* is a popular procedure, and it has been used successfully in [56] for MRI noise cancellation. This algorithm is based on computing the logarithm of the power spectrum (in order to compress all high spectral peaks "softly" and non-adaptively), returning to time domain by FFT, and reconstructing the phase information from the original signal. The cepstral transform does not take into account the harmonic structure of noise at all. The multi-path propagation of noise would seem to invite an approach based on *deconvolution*. However, an accurate estimation of the convolution kernel (i.e., the delays and the weights in multi-path propagation) does not seem to be feasible even though the autocorrelation of the noise signal is easy to compute. The multi-path propagation of noise was treated by using a continuously adjusted, adaptive, non-causal FIR filter in [41] where the noise sample was collected from outside the MRI machine or, alternatively, generated artificially by a MRI noise model. In contrast to [41], our approach does not tune the causal Chebyshev filter "on the run" but once the filter has been optimised for a given sample, it is then applied to the entire sample as such. This more simple approach makes the noise cancellation procedure tractable: if an unexpected artefact appears in the cleaned-up signal, it is always possible to exclude the post-processing algorithm as its source.

5. Evaluation of the audio measurement system using pilot data

The purpose of the audio measurements is to obtain precise estimates for formant frequency values from the speech signal gathered during a noisy MRI recording. The acoustic MRI noise induces two sources of uncertainty in measured values. Firstly, the speaker receives somewhat noisy auditory feedback. Secondly, the ambient noise and possible post-processing artefacts may increase the formant error. To analyse the impact of confounding factors (such as the ambient noise) to vocal production and general measurement precision, two comparison data sets are analysed: a set of similar recordings in an anechoic chamber and computational analysis of MR images obtained during the vowel production.

5.1. Extracting power spectra and spectral envelopes

Formants are the main information bearing component of vowel sounds. They can be understood as acoustic energy concentrations around discrete frequencies in the power spectrum of the speech signal. The measured formant frequencies $F_1, F_2, \ldots$ are related to the acoustic resonance frequencies $R_1, R_2, \ldots$ of the vocal tract. In contrast to harmonic overtones of the fundamental frequency f_0 of the glottal excitation, the formants have a much wider bandwidth. Thus, the extraction of formants from speech can be carried out by a frequency domain smoothing process that downplays the narrow bandwidth harmonics of f_0.

Perhaps the most popular formant extraction tool is Linear Predictive Coding (LPC); see, e.g., [63, 64]. LPC is mathematically equivalent to fitting a low-order rational function $R(s)$ to the power spectrum function defined on the imaginary axis, and the pole positions of $R(s)$ give the estimated formant values. Hence, plotting the values of $|R(i\omega)|$ for real ω yields *LPC envelopes* whose peaks indicate the formant frequencies $F_1, F_2, \ldots$.

All data for this article has been recorded from a healthy 26-year-old male in supine position, and it includes sound samples during an MRI scan as well as comparison samples that have been recorded in an anechoic chamber. Formants and LPC envelopes from these sound samples have been produced by the MATLAB function `lpc` for each of the eight Finnish vowels [ɑ, e, i, o, u, y, æ, œ]. The formant values from comparison measurements are given in Table 2, and the spectral envelopes of all signals are given in Figs. 7–8. Formants for Fig. 6a have been extracted using Burg's method [65] (MATLAB function `arburg`) which was observed to give better resolution for sound data having some residual MRI noise. Acoustic resonances under 5 kHz have been computed by FEM from Eqs. (3) using the vocal tract geometries obtained by MRI, and they are shown in Figs. 7–8 as vertical lines. Formant frequencies F_1, F_2, F_3, and the corresponding Helmholtz resonance frequencies R_1, R_2, R_3, are compared in Table 3 by their discrepancies in logarithmic semitone scale.

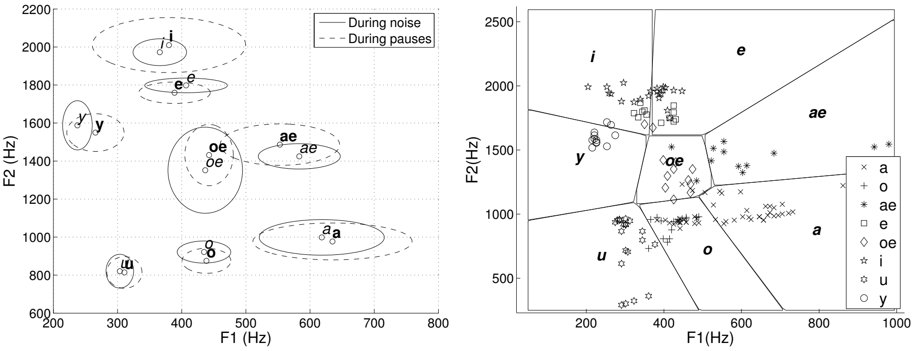
Further observations and details concerning the data are explained below.

Sound data during MRI

As pointed out in Section 1, a second set of pilot MRI experiments was carried out in 2012. The test subject was able to produce 107 speech samples during a single MRI session of 1.5 h according to the experimental specifications given in Section 2. Out of these speech and MRI samples, 69 are vowels imaged by static 3D MRI, out of which 40 with $f_0 \approx 104$ Hz were chosen as the data for this article. The vowel samples were processed by the noise reduction algorithm detailed in Section 4, and their formants F_1, F_2, F_3 as well as their LPC envelopes (shown in Figs. 7–8) were produced by MATLAB using `lpc` with filter order 40, applied on a 3 s interval taken from the middle of each sample.

The noise reduction algorithm does not spoil the extraction of first formants as can be seen in Fig. 6a. For this figure, F_1 and F_2 have been estimated from the noisy parts of the speech signal as well as from those parts where the MRI sequence was paused. We note that in many but not in all cases, the lowest formants F_1, F_2, and F_3 could be correctly revealed by Praat [66] (using default settings) from the denoised signals. Reflections from the MRI coil walls produce spurious "external formants" to measured speech signals, which is likely the cause of the extra peaks (such as the one appearing at ≈ 1 kHz) that can be seen in many of the upper curves in Figs. 7–8.

The test subject had occasional difficulties in producing the prolonged [ɑ] during the MRI which results in a large internal variation of F_1[ɑ] and F_2[ɑ]. However, there are many "good" samples of [ɑ] during MRI whose spectral envelopes resemble those that



(a) (b)

Figure 6: (a) Sample means and standard deviations of F_1 and F_2 that were extracted using Burg's method from speech samples recorded (i) during the MRI noise after denoising, and (ii) during the intermediate pauses in MRI as explained in Section 3.2. The means are depicted as the center points of ellipses, and standard deviations as their semi-axes. (b) F_1 and F_2 extracted from denoised samples recorded during MRI. They have been classified using the Linear Discriminants Analysis and formant data from the anechoic chamber as training data.

were recorded in the anechoic chamber. None of the samples were rejected since even the less successful vowel productions reflect correctly the physical relation of speech and the anatomy that produces it.

Comparison sound data from anechoic chamber

To obtain high-quality comparison data, vowel samples were recorded in the anechoic chamber from the same test subject in supine position. The recordings were carried out about one year after the MRI experiments, and the subject was now more skilled in producing prolonged vowels and following a pitch reference.

The Brüel & Kjæll 2238 Mediator integrating sound level meter was used as a microphone, coupled to RME Babyface digitiser with software TotalMix FX v.0.989 and Audacity v.1.3.14 running on MacBook Air OSX 10.7.5. The microphone was placed 0.5 m from the mouth of the test subject, at 45° angle on the right-hand side. The test subject heard from headphones (Bose QuietComfort 15) his own, algorithmically denoised vowel signal from pilot MRI experiments as a pitch reference (one sample for each vowel). The vowels were given in a randomised order and also shown on a computer screen. The statistics from these experiments are reported in Table 2.

	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[œ]
#	10	10	10	9	10	10	6	14
F_1	580±23	465±14	276±38	494±23	323±93	394±51	563±23	465±21
F_2	1018±63	1608±52	1849±57	861±35	822±148	1527±53	1442±37	1400±50

Table 2: Number of vowel samples as well as their sample means and standard deviations of F_1 and F_2, extracted (by LPC, as explained in the text) from the comparison data recorded in the anechoic chamber.

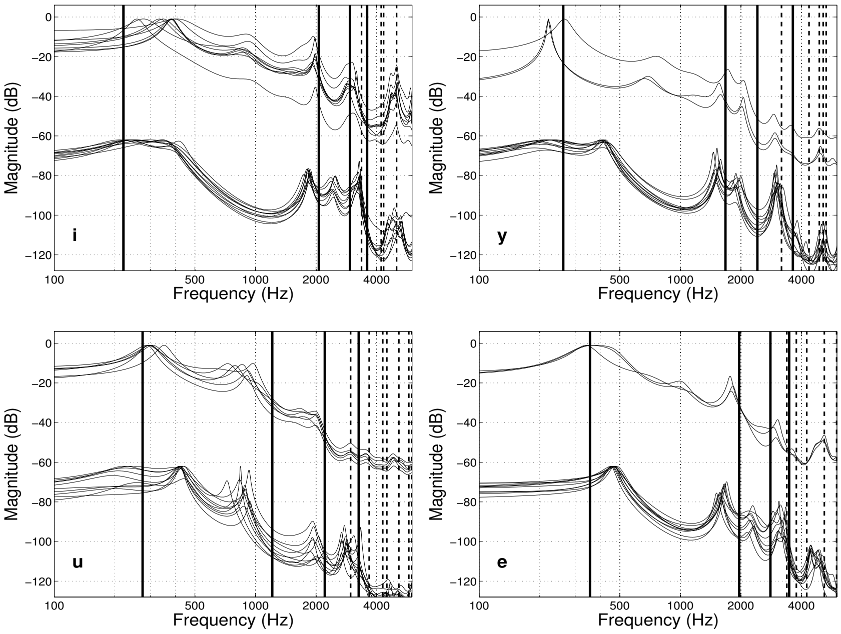


Figure 7: Spaghetti plots of LPC envelopes from Finnish vowels [i, y, u, e]. In each panel, the upper graphs have been produced from recordings during the MRI using the noise reduction detailed in Section 4. The lower graphs have been produced from recordings in the anechoic chamber from the same test subject. The vertical lines indicate the resonance frequencies $R_1, R_2, \ldots$ computed by FEM from the Helmholtz model Eqs. (3). Some of the resonances were identified as purely longitudinal, and they have been marked using a solid line.

The spectral envelopes of the anechoic chamber data were produced in the same way as described above for vowel samples during MRI. The results have been included in Figs. 7–8 where they appear as the lower and more regular family of curves. In Fig. 6b, the vowel samples during MRI were classified in (F_1, F_2) space by MATLAB function `classify` using the anechoic chamber data (whose statistics are given in Table 2) as the training set. Most of the data gets classified correctly but there is some mixing of [i] with [e], and [ɑ] with [o]. That [i] is not correctly separated from [e] is due to the systematic error in the extraction of F_1[i] from speech during MRI as discussed in Section 6. A perceptual classification experiment on similarly recorded data (but without using the noise reduction algorithm of Section 4) was reported in [62].

As expected, the recordings during the MRI have more formant variation than the recordings in the anechoic chamber. It can be seen in Figs. 7–8 that, for example, productions of [ɑ] are much more consistent in the anechoic chamber: the variance of F_1[ɑ] is significantly smaller there (F-test, $F = 8.2$ with $p = 0.0031$). However, the test subject had similar problems controlling F_2[u] during both types of the experiments whereas all spectral envelopes of [œ] are remarkably similar.

Even though experiments in the anechoic chamber were designed to resemble the conditions during the MRI scan in many respects, there are significant differences. Firstly, the acoustic noise of the MRI machine was not replicated in the anechoic chamber. Secondly, the test subject fatigue played lesser role in the anechoic chamber since the

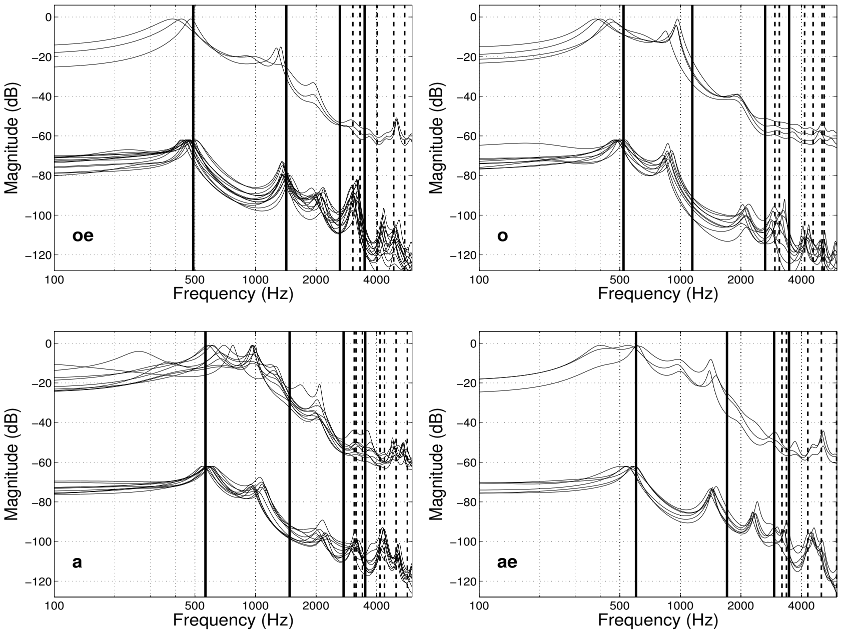


Figure 8: Spaghetti plots of LPC envelopes of Finnish vowels [œ, o, ɑ, æ]. The presentation is similar to Fig. 7.

total duration of a single experimental session was only about 10 minutes. Thirdly, the head and neck MRI coil is a rather closed acoustic environment whereas there was no similar acoustic load present in the anechoic chamber.

5.2. Computation of the Helmholtz resonances

For each vowel presented in Figs. 7–8, the corresponding MRI scans were processed as described in [53] to produce the corresponding air-tissue interface surface Γ_2 as shown for [œ] in Fig. 3. However, three geometries of both [ɑ] and [i] could not be processed without manual intervention, and they were discarded from computational resonance analysis (leaving 33 that are presented in Table 3). Completing the air-tissue interface Γ_2 with the mouth opening Γ_1 and the (virtual) control surface Γ_3 right above the glottis, we obtain the air column volume $\Omega \subset \mathbb{R}^3$ of the vocal tract whose boundary satisfies $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$. The models did not include teeth geometries.

As detailed in [17], eigenvalues λ and the corresponding (velocity potential) eigenfunctions Φ_λ were computed from the Helmholtz resonance problem

$$\begin{cases} \lambda^2 \Phi_\lambda = c^2 \Delta \Phi_\lambda \quad \text{on} \quad \Omega, \quad \Phi_\lambda = 0 \quad \text{on} \quad \Gamma_1, \\ \frac{\partial \Phi_\lambda}{\partial \nu} = 0 \quad \text{on} \quad \Gamma_2, \quad \text{and} \quad \lambda \Phi_\lambda + c \frac{\partial \Phi_\lambda}{\partial \nu} = 0 \quad \text{on} \quad \Gamma_3 \end{cases} \tag{3}$$

where $c = 350$ m/s and $\frac{\partial \Phi_\lambda}{\partial \nu}$ denotes the exterior normal derivative. As explained in [33], the numerical solution of Eqs. (3) was carried out by Finite Element Method (FEM) with piecewise linear shape functions and tetrahedral meshes with approximately 10^5 elements, using a MATLAB-based FEM solver and the eigenvalue function `eigs`. The imaginary parts of the smallest eigenvalues λ give resonance frequencies $R_1, R_2, \ldots$ of the vocal tract air column in increasing order of frequency. The chosen number of elements was observed to be large enough so that all computed resonances can be regarded as accurate, given Eqs. (3).

One of the surface models was randomly selected to represent each vowel [ɑ, e, i, o, u, y, æ, œ], and some of the lowest Helmholtz resonances $R_1, R_2, \ldots$ are represented by vertical lines in Figs. 7–8. Some of the resonances were identified as purely longitudinal by using a MATLAB-based FEM solver for Webster's model [33, Eq. (3)] on the area function from the same vocal tract geometry. These have been marked using solid vertical lines in Figs. 7–8. As expected, the FEM computation reveals a cloud of higher Helmholtz resonances $R_4, R_5, \ldots$ near the expected fourth formant position as reported in [33, 34, 35].

The rather trivial *Dirichlet boundary condition* $\Phi_\lambda = 0$ was used at the mouth opening Γ_1 in Eqs. (3), leading to an overestimation of F_2 and F_3 by the respective R_2 and R_3 as explained[2] in [17]. There is a particularly striking discrepancy between F_2[ɑ] and R_2[ɑ] even when they are extracted from the same MRI and speech data pair. The Helmholtz model in Eqs. (3) gives consistently the result that R_2[ɑ] $\approx 1.5 \cdot F_2$[ɑ], and the same observation holds in the data reported in [51] obtained from a different test subject. We conclude that the Dirichlet boundary condition is at its worst for vowels such as [ɑ] that have largest mouth cavity volumes or opening areas.

The discrepancy between measured and computed formants has been reported in many works: the digital simulator of S. Maeda [67] was used in [50], and the Kelly–Lochbaum model or its generalisation was used in [11, 12, 36]. Various model improvements have been used to explain or to reduce the discrepancy: tuning the area function [12]; inclusion and exclusion of piriform sinuses and valleculae [11, 50]; energy dissipation to tissues [36, 50]; and tuning the termination impedance at mouth and glottis [12]. For higher resolution acoustics models such as Eqs. (3), the exterior space should be modelled as faithfully as the vocal tract volume. Hence, the Dirichlet boundary condition in Eqs. (3) should be regarded only as a minimal assumption, giving a baseline for the discrepancy and a starting point for proper exterior space acoustics modelling. Instead of tuning the boundary conditions in Eqs. (3), a preferred approach is to model the exterior space by FEM as in [34] but the high computational cost makes alternative approaches attractive [68, 69].

The discrepancy between measured formants and computed resonances has been estimated from the set of 33 vowel geometries and sound samples in Table 3. Given in semitones, the average discrepancy is 2.2 for F_1 (excluding [i] where a crude error in algorithmic formant extraction remains), 3.1 for F_2, and 2.1 for F_3. This is in line with the discrepancy of 3.5 semitones reported in [17] (based on a single geometry of Swedish [ø] using the data from [70]), and the discrepancy of 2.1 semitones given in [71] (based on four vowel geometries of Finnish [ɑ, i, u, œ] from experiments reported in [62]).

[2] Considering the resonances of a tube with one end $x = 0$ closed and the other $x = 1$ having a mixed boundary condition, we may solve the modes from $v_{xx} + k^2 v = 0$, $v_x(0) = 0$, and $v_x(1) + bv(1) = 0$ where $b, k \geq 0$. We get $b/k = \tan k$ whose positive solutions $k = k_j(b)$, $j = 1, 2, \ldots$, (proportional to resonance frequencies) increase with increasing b. When $b \to \infty$, the boundary condition becomes the Dirichlet condition $v(1) = 0$.

	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[œ]
#	5	3	7	4	5	3	3	3
D_1	-2.8	-1.5	(-10.1)	3.4	-3.3	1.5	1.4	1.4
D_2	7.1	1.8	1.2	4.4	5.2	1.5	1.8	2.0
D_3	2.6	-0.8	-1.0	4.2	1.3	2.6	-1.7	2.2

Table 3: Discrepancies D_1, D_2, D_3 (in semitones) of measured vowel formants F_1, F_2, F_3, and computed Helmholtz resonances R_1, R_2, R_3, estimated from 33 simultaneously recorded MRI/speech data pairs at $f_0 = 104$ Hz. The discrepancies are given by $D = 12 \ln{(R/F)} / \ln 2$, and a positive discrepancy implies that the Helmholtz resonance is higher. The number of data pairs is given on the topmost row.

6. Conclusions

We have described experimental protocols, MRI sequences, a sound recording system, and a customised post-processing algorithm for contaminated speech that, in conjunction with previously reported arrangements [45, 46, 51, 52], can be used for simultaneous speech sound and anatomical data acquisition not only on healthy test subjects but also on a large number of oral and maxillofacial surgery patients.

The data set obtained from such measurements are primarily intended for parameter estimation, fine tuning, and validation of a computational acoustics model for speech production. However, these methods and procedures may be used in a wider range of applications related to anatomy and physiology of the vocal tract, including medical research and clinical use.

Collecting such multi-modal data from numerous patients is far from a trivial task even when suitable instrumentation is available. Several phonetic aspects must be taken into account to ensure that the task is within the ability of the patients, regardless of background and skills. It must be possible to monitor the quality of articulation and phonation despite the acoustic noise in the MRI room, and data collection procedures must be reliable to minimise the number of repetitions and the amount of useless data obtained. All this must be achieved in as short a time as possible to minimise cost and maintain patient interest in the project.

The experimental setting and phonetic tasks require the patients to have abilities in concentration, remaining still, and sustaining prolonged phonation not significantly reduced from young adults in good health. At the time of writing of this article, ten patients (out of which six are female) have already undergone such MRI examinations preceding their orthognathic procedures, and they are expected to take part in a similar examination after their post-operative treatment will have been completed. All of these examinations have succeeded without major troubles, and the resulting MRI image and the speech sound data quality is very satisfactory as well. Applications to other patient groups are under consideration but may require adaptations to the required time of phonation and the total number of measurements.

Some questions and problems in the measurement arrangements remain open, in particular, involving acoustic noise and its impact on articulation. Acoustic noise during measurements remains a problem from two points of view. Firstly, formant extraction from denoised, prolonged vowel samples is sometimes problematic as observed in Section 5.1. Note that reliable formant extraction may be difficult for reasons unrelated to noise contamination: consider, e.g., vowels with low F_1 in high pitch speech samples such as [i] pronounced by female subjects. Secondly, the onset of MRI noise may cause a significant adaptation in the patients' articulation. It may be possible to reduce this problem by running the 3D MRI sequence once while the patient receives the task instructions to adapt the patient to the noise, and a second time during phonation to obtain the vocal tract geometry. For the 2D sequences, the sequence may be started before phonation. It is likely that there is adaptation in patient's speech during MRI noise because of the Lombard effect. A possible counter measure is to use MRI-proof acoustic earmuffs, and lead the cue and instruction signal to the patient using an arrangement described in [44].

Automatic formant extraction from denoised vowel data requires further refinement. Perhaps, the computed resonances of the vocal tract could be used as *a priori* data when sorting out the peaks in the spectral envelopes of vowel signals. Low formants such as F_1[i] are particularly difficult to extract from the recordings during MRI, leading to a quite high statistical variance irrespective of the method chosen. In Figs. 6a, 6b, and 7, the positions for F_1[i] are systematically too high, and an artificial pre-emphasis filter for frequencies under 300 Hz should be determined by trial and error to make the formant extraction algorithm place a pole in the right position. Computing the Helmholtz resonances from all vocal tract geometries of [i] from the test subject gives the following sample means and standard deviations: R_1[i] = 180 ± 56 Hz and R_2[i] = 2064 ± 40 Hz. These standard deviations are of the same magnitude as those given in Table 2 for recordings in the anechoic chamber, and they may reflect the underlying natural variation in vowels produced by this test subject.

7. Acknowledgements

The authors were supported by the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n^0 287678, the Academy of Finland (projects 135005, 128204, 125940), Aalto Starting Grant, a grant from Magnus Ehrnrooth Foundation, and Research Grant of the Hospital District of Southwest Finland (EVO-project 13939).

The authors wish to thank professor P. Alku (Department of Signal Processing and Acoustics, Aalto University) for providing resources and consultation for the acquisition of the comparison data. The comments of two anonymous reviewers resulted in significant improvements of the article. Åbo Akademi Institute of Mathematics has provided resources for the research taking place in Turku, Finland. The patient data collection procedure was accepted by the Ethics Committee of the Hospital District of Southwest Finland. The raw MRI data, the speech sound data, and the surface models used in Section 5 are freely downloadable for non-commercial academic and educational use from `http://speech.math.aalto.fi`.

[1] A. M. Liberman, Speech: A special code, The MIT press, Cambridge, Massachussets, 1996.

[2] P. Švancara, J. Horáček, Numerical modelling of effect of tonsillectomy on production of Czech vowels, Acta Acustica united with Acustica 92 (2006) 681 – 688.

[3] L.-J. Boë, J.-L. Heim, K. Honda, S. Maeda, The potential Neandertal vowel space was as large as that of modern humans, Journal of Phonetics 30 (3) (2002) 465–484.

[4] T. Chiba, M. Kajiyama, The vowel, its nature and structure, Phonetic Society of Japan, 1958.

[5] G. Fant, Acoustic Theory of Speech Production, Mouton, The Hague, 1960.

[6] H. L. F. Helmholtz, Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, Braunschweig: F. Vieweg, 1863.

[7] K. Ishizaka, J. L. Flanagan, Synthesis of voiced sounds from a two-mass model of the vocal cords, Bell System Technical Journal 51 (1972) 1233–1268.

[8] J. Kelly, C. Lochbaum, Speech synthesis, in: Proceedings of the 4th International Congress on Acoustics, 1–4, 1962.

[9] H. Dunn, The calculation of vowel resonances, and an electrical vocal tract, J. Acoust. Soc. Am. 22 (1950) 740 – 753.

[10] T. Baer, J. Gore, S. Boyce, P. Nye, Application of MRI to the analysis of speech production, J Magn Reson Imaging 5 (1987) 1–7.

[11] T. Baer, J. Gore, L. Gracco, P. Nye, Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, J. Acoust. Soc. Am. 90 (2) (1991) 799–828.

[12] A. Greenwood, C. Goodyear, P. Martin, Measurement of vocal tract shapes using magnetic resonance imaging, Communications, speech and vision, IEE Proceedings-I 139 (6) (1992) 553–560.

[13] S. El Masri, X. Pelorson, P. Saguet, P. Badin, Development of the transmission line matrix method in acoustics. Applications to higher modes in the vocal tract and other complex ducts, Int. J. Numerical Modelling 11 (1998) 133 – 151.

[14] J. Mullen, D. Howard, D. Murphy, Waveguide physical modeling of vocal tract acoustics: Flexible formant bandwith control from increased model dimensionality, IEEE Transactions on Audio, Speech and Language Processing 14 (3) (2006) 964 – 971.

[15] J. Heinz, K. Stevens, On the derivation of area functions and acoustic spectra from cineradiographic films of speech, J. Acoust. Soc. Am. 36 (1968) 1037.

[16] J. Sundberg, On the problem of obtaining area functions from lateral X-ray pictures of the vocal tract, Royal Inst. Technol. STL-QPSP (1969) 43–45.

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

[18] C. Lu, T. Nakai, H. Suzuki, Finite element simulation of sound transmission in vocal tract, J. Acoust. Soc. Jpn. 14 (1993) 63 – 72.

[19] H. Suzuki, T. Nakai, N. Takahashi, A. Ishida, Simulation of vocal tract with three-dimensional Finite Element Method, Tech. Rep. IEICE EA93-8 (1993) 17 – 24.

[20] T. Vampola, J. Horáček, J. Švec, FE modeling of human vocal tract acoustics. Part I: Production of Czech vowels, Acta Acustica united with Acustica 94 (2008) 433–447.

[21] T. Vampola, J. Horáček, J. Vokřál, L. Černý, FE modeling of human vocal tract acoustics. Part II: Influence of velopharyngeal insufficiency on phonation of vowels, Acta Acustica united with Acustica 94 (2008) 448–460.

[22] P. Švancara, J. Horáček, L. Pešek, Numerical modelling of production of Czech vowel /a/ based on FE model of the vocal tract, in: Proceedings of International Conference on Voice Physiology and Biomechanics, 2004.

[23] J. Horáček, V. Uruba, V. Radolf, J. Veselý, V. Bula, Airflow visualization in a model of human glottis near the self-oscillating vocal folds model, Applied and Computational Mechanics 5 (2011) 21–28.

[24] P. Šidlof, J. Horáček, V. Řidký, Parallel CFD simulation of flow in a 3D model of vibrating human vocal folds, Computers and Fluids 80 (2013) 290–300.

[25] K. Dedouch, J. Horáček, T. Vampola, L. Černý, Finite element modelling of a male vocal tract with consideration of cleft palate, in: Forum Acusticum, Sevilla, Spain, 2002.

[26] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, C. Savariaux, Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images, Journal of Phonetics 30 (2002) 533–553.

[27] H. Nishimoto, M. Akagi, T. Kitamura, N. Suzuki, Estimation of transfer function of vocal tract extracted from MRI data by FEM, in: The 18th International Congress on Acoustics, vol. II, Kyoto, Japan, 1473 –1476, 2004.

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

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

[30] A. Aalto, A low-order glottis model with nonturbulent flow and mechanically coupled acoustic load, Master's thesis, Aalto University, Department of Mathematics and Systems Analysis, 2009.

[31] T. Murtola, Modelling vowel production, Master's thesis, Aalto University, Department of Mathematics and Systems Analysis, 2014.

[32] A. Aalto, P. Alku, J. Malinen, A LF-pulse from a simple glottal flow model, in: Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications

(MAVEBA2009), Florence, Italy, 199–202, 2009.

[33] A. Kivelä, J. Kuortti, J. Malinen, Resonances and mode shapes of the human vocal tract during vowel production, in: Proceedings of 26th Nordic Seminar on Computational Mechanics, 112–115, 2013.

[34] T. Vampola, A.-M. Laukkanen, J. Horáček, J. Švec, Finite element modelling of vocal tract changes after voice therapy, Applied and Computational Mechanics 5 (1), (2011).

[35] T. Vampola, J. Horáček, A.-M. Laukkanen, J. Švec, Human vocal tract resonances and the corresponding mode shapes investigated by three-dimensional finite-element modelling based on CT measurement, Logopedics Phoniatrics Vocology (2013) 1–10.

[36] B. Story, I. Titze, E. Hoffman, Vocal tract area functions from magnetic resonance imaging, J. Acoust. Soc. Am. 100 (1) (1996) 537–554.

[37] S. Narayanan, A. Alwan, An articulation study of fricative consonants using magnetic resonance imaging, J. Acoust. Soc. Am. 98 (3) (1995) 1325–1347.

[38] C. Ericsdotter, Articulatory-acoustic relationships in Swedish vowel sounds, Ph.D. thesis, Stockholm University, Sweden, 2005.

[39] MATLAB, 7.11.0.584 (R2010b), Mathworks Inc., 2010.

[40] A. Soquet, V. Lecuit, T. Metens, D. Demolin, Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI, Speech Communication 36 (3) (2002) 169 – 180.

[41] E. Bresch, K. Nielsen, K. Nayak, S. Narayanan, Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans, J. Acoust. Soc. Am. 120 (4) (2006) 1791–1794.

[42] M. NessAiver, M. Stone, V. Parthasarathy, Y. Kahana, A. Kots, A. Paritsky, Recording high quality speech during tagged cine-MRI studies using a fiber optic microphone, J Magn Reson Imaging 23 (2006) 92–97.

[43] X. Shou, X. Chen, J. Derakhsan, T. Eagan, T. Baig, S. Shvartsman, J. Duerk, R. Brown, The suppression of selected acoustic frequencies in MRI, Applied Acoustics 71 (2010) 191–200.

[44] Y. Nota, T. Kitamura, K. Honda, H. Takemoto, H. Hirata, Y. Shimada, I. Fujimoto, Y. Shakudo, S. Masaki, A bone-conduction system for auditory stimulation in MRI, Acoust. Sci. & Tech. 28 (1) (2007) 33–38.

[45] T. Lukkari, J. Malinen, P. Palo, Recording speech during magnetic resonance imaging, in: Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2007), Florence, Italy, 163–166, 2007.

[46] J. Malinen, P. Palo, Recording speech during MRI: Part II, in: Proceedings of the 6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2009), Florence, Italy, 211–214, 2009.

[47] S. Narayanan, K. Nayak, S. Lee, A. Sethy, D. Byrd, An approach to real-time magnetic resonance imaging for speech production, J. Acoust. Soc. Am. 115 (4) (2004) 1771–1776.

[48] Y.-C. Kim, S. Narayanan, K. Nayak, Accelerated three-dimensional upper airway MRI using compressed sensing, Magn. Reson. Med. 61 (2009) 1434–1440.

[49] S. Narayanan, D. Byrd, A. Kaun, Geometry, kinematics, and acoustics of Tamil liquid consonants, J. Acoust. Soc. Am. 106 (4) (1999) 1993–2007.

[50] P. Clément, S. Hans, D. Hartl, S. Maeda, J. Vaissiére, D. Brasnu, Vocal tract area function for vowels using three-dimensional magnetic resonance imaging. A preliminary study, Journal of Voice 21 (5) (2007) 522–530.

[51] D. Aalto, O. Aaltonen, R.-P. Happonen, J. Malinen, P. Palo, R. Parkkola, J. Saunavaara, M. Vainio, Recording speech sound and articulation in MRI, in: Proceedings of BIODEVICES 2011, Rome, 168–173, 2011.

[52] P. Palo, A wave equation model for vowels: Measurements for validation, Licentiate thesis, Aalto University, Department of Mathematics and Systems Analysis, 2011.

[53] D. Aalto, J. Helle, A. Huhtala, A. Kivelä, J. Malinen, J. Saunavaara, T. Ronkka, Algorithmic surface extraction from MRI data: modelling the human vocal tract, in: Proceedings of BIODEVICES 2013, Barcelona, 257–260, 2013.

[54] M. Vainio, A. Suni, H. Järveläinen, J. Järvikivi, V. Mattila, Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish, J. Acoust. Soc. Am. 118 (2005) 1742–1750.

[55] Optoacoustics, Ltd., http://www.optoacoustics.com/, accessed Aug. 25th, 2009, 2009.

[56] J. Přibil, J. Horáček, P. Horák, Two methods of mechanical noise reduction of recorded speech during phonation in an MRI device, Measurement Science Review 11 (3) (2011) 92–99.

[57] J. Přibil, A. Přibilová, I. Frollo, Analysis of spectral properties of acoustic noise produced during

magnetic resonance imaging, Applied Acoustics 73 (8) (2012) 687–697.

[58] N. Rofsky, V. Lee, G. Laub, M. Pollack, G. Krinsky, D. Thomasson, M. Ambrosino, J. Weinreb, Abdominal MR imaging with a volumetric interpolated breath-hold examination, Radiology 212 (3) (1999) 876 – 884.

[59] D. Aalto, J. Malinen, M. Vainio, J. Saunavaara, J. Palo, Estimates for the measurement and articulatory error in MRI data from sustained vowel phonation, in: Proceedings of the International Congress of Phonetic Sciences, 180–183, 2011.

[60] R. Humphrey, Playrec, http://www.playrec.co.uk/, accessed Jul. 15th, 2012.

[61] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, Acoustics, Speech and Signal Processing, IEEE Transactions 27 (2) (1979) 113 – 120.

[62] J. Palo, D. Aalto, O. Aaltonen, R.-P. Happonen, J. Malinen, J. Saunavaara, M. Vainio, Articulating Finnish Vowels: Results from MRI and sound data, Linguistica Uralica 48 (3) (2012) 194–199.

[63] J. Makhoul, Linear prediction: A tutorial review, Proceedings of the IEEE 63 (4) (1975) 561–580.

[64] J. Makhoul, Spectral linear prediction: Properties and applications, Acoustics, Speech and Signal Processing, IEEE Transactions 23 (3) (1975) 283 – 296.

[65] J. Burg, Maximum entropy spectral analysis, Ph.D. thesis, Stanford University, 1975.

[66] P. Boersma, D Weenink, Praat v.4.6.15, http://www.fon.hum.uva.nl/praat/, accessed July. 5th, 2010.

[67] S. Maeda, A digital simulation method of the vocal-tract system, Speech Communication 1 (1982) 199 – 229.

[68] R. Udawalpola, E. Wadbro, M. Berggren, Optimization of a variable mouth acoustic horn, Internat. J. Numer. Methods Engrg. 85 (2011) 591–606.

[69] M. Arnela, O. Guasch, F. Alías, Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations, J. Acoust. Soc. Am. 134 (4) (2013) 2946–2954.

[70] O. Engwall, P. Badin, Collecting and analysing two- and three-dimensional MRI data for Swedish, TMH-QPSR, (1999)

[71] D. Aalto, A. Huhtala, A. Kivelä, J. Malinen, P. Palo, J. Saunavaara, M. Vainio, How far are vowel formants from computed vocal tract resonances? arXiv:1208.5963, 2012.