Biomedical Signal Processing and Control 39 (2018) 11–22

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Research Paper

Post-processing speech recordings during MRI

Juha Kuortti[a], Jarmo Malinen[a,b,*], Antti Ojalammi[a]

[a] Department of Mathematics and Systems Analysis, Aalto University, Finland
[b] Department of Signal Processing and Acoustics, Aalto University, Finland

ARTICLE INFO

Article history:
Received 24 August 2016
Received in revised form 5 April 2017
Accepted 20 July 2017

Keywords:
Speech
MRI
Noise reduction
DSP
Helmholtz

ABSTRACT

We discuss post-processing of speech samples that have been recorded simultaneously during Magnetic Resonance Imaging (MRI) of the upper airways. Speech recordings contain acoustic noise from the MRI scanner. The required noise reduction is based on adaptive comb filtering designed for accurate formant extraction.

Two kinds of speech materials were used to validate the post-processing algorithm. The primary material consists of samples of prolonged vowel productions during MRI. The comparison data was obtained from the same test subject, and it was recorded in anechoic chamber in a similar configuration as used during the MRI. Spectral envelopes and vowel formants were computed from the post-processed speech and from the comparison data. Vowel samples (with a known formant structure) were artificially contaminated using MRI scanner noise to determine performance of the post-processing algorithm. Resonances computed from a numerical acoustic model and spectra measured from 3D printed vocal tract physical models were used as comparison data.

The properties of the recording instrumentation or the post-processing algorithm do not explain the observed frequency dependent discrepancy between the vowel formant data from two kinds of experiments: recordings during MRI and comparison data. It is shown that the discrepancy is statistically significant, in particular, where it is largest at ca. 1 kHz and 2 kHz. Numerical and experimental evidence suggests that the surfaces of the MRI head coil change the acoustics of speech which results in "exterior formants" at these frequencies. The discrepancy is too large to be neglected if the recordings during MRI are to be used for parameter estimation or validation of a numerical speech model, based on the MR images. However, the role of test subject adaptation to noise and constrained space acoustics during an MRI examination cannot be ruled out.

1. Introduction

Modern medical imaging technologies such as Ultrasonography (USG), X-ray Computer Tomography (CT), and Magnetic Resonance Imaging (MRI) have revolutionised studies of speech and articulation. There are, however, significant differences in applicability and image quality between these technologies. Considering the imaging of the whole speech apparatus, the use of inherently low-resolution USG is often impractical, and the high-resolution CT exposes the test subject to potentially significant doses of ionising radiation. MRI remains an attractive approach for large scale articulation studies but there are, unfortunately, many other restrictions on what can be done during an MRI scan as discussed in [1,2].

Since the intra-subject variability of speech may often be of the same magnitude as the inter-subject variability within the same gender and language background, it is desirable to sample speech simultaneously with the MRI experiment in order to obtain *paired data*. Such paired data is a particularly valuable asset in developing and validating a computational model for speech such as proposed in [3]. Unfortunately, speech signal recorded during MRI contains many artefacts that are mainly due to high acoustic noise level inside the MRI scanner. There are additional artefacts due to the nonflat frequency response of the MRI-proof audio measurement system and further challenges related to the constrained space acoustics inside the MRI head and neck coils.

Noise cancellation is a classical subject matter in signal processing that in the context of speech enhancement can be divided into two main classes: *adaptive noise cancellation* techniques and the *blind source separation* methods such as FastICA introduced in [4]. The purpose of this article is to introduce, analyse, and validate a

* Corresponding author.

http://dx.doi.org/10.1016/j.bspc.2017.07.017
1746-8094/

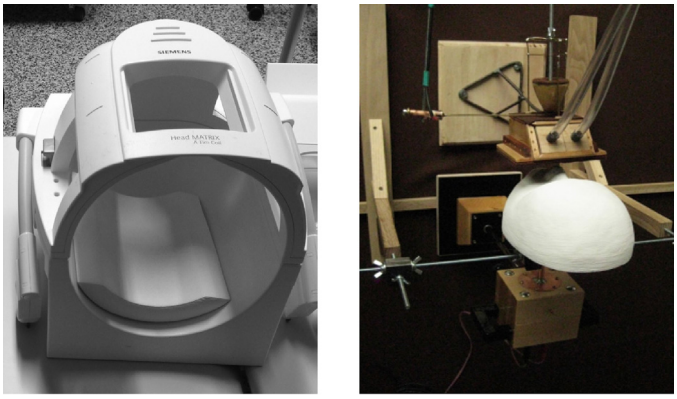
post-processing algorithm of the former type for treating speech that has been recorded during MRI.[1] Compared to blind source separation, the tractability of the processing algorithm favours adaptive noise cancellation that may take place in time domain, in frequency domain, or partly in both. The algorithm discussed in this article is designed based on lessons learned from an earlier algorithm introduced in [2,Section 4]. For different approaches for dealing with the MRI noise, see also [5–8] that will be discussed at the end of the article.

When designing a practical solution, one should consider, at least, these three aspects of the noise cancellation problem: (i) what kind of noise should be rejected, (ii) what kind of signal or signal characteristic should be preserved, and (iii) how the resulting de-noised signal is to be used. In this work, the noise is generated by an MRI scanner, the preserved signal consists of prolonged, static vowel utterances, and the de-noised signals should be usable for high-resolution spectral analysis of speech formants. The noise spectrum of the MRI scanner (in these experiments, Siemens Magnetom Avanto 1.5T) has a lot of harmonic structure on few discrete frequencies as shown in Fig. 2b, and it changes during the course of the MRI scan. The proposed algorithm estimates the harmonics of the noise, and removes their contribution by tight notch filters as explained in Fig. 2. There are additional heuristics to prevent the removal of multiples of the fundamental glottal frequency (f_0) of the speech that, unfortunately, somewhat resemble the noise spectrum of the MRI scanner. One of the caveats is not to have the algorithm "bake" noise energy into spurious spectral energy concentrations that would skew the true formant content – this may be a serious cause of worry in nonlinear signal processing that is able to move energy from one frequency band to another.

Since the de-noised vowel data is used in, e.g., [2,9] for parameter estimation and validation of a computational model, it is imperative that the extracted formant positions, indeed, reflect precisely the acoustic resonances of the corresponding MRI geometries of the vocal tract. For model validation, the proposed post-processing algorithm is applied to noisy speech data consisting of prolonged vowel samples from which vowel formants should be extracted without bias. In a typical speech sample, the noise component is of a comparable level as the speech component, but there is great variance between different test subjects and even between different vowels from the same test subject: a smaller mouth opening area results in lower emission of sound power.

The outline of this article is as follows: after the data acquisition has been described in Section 2, the post-processing algorithm is described in Section 3. The validation of the algorithm is carried out in Section 4 through four different approaches: (i) accuracy of the formant extraction using a synthetic test signal with known formant structure, (ii) comparison of spectral tilts (i.e., the roll-off) of de-noised speech recorded during the MRI to similar data recorded in the anechoic chamber, (iii) comparison of the formants from de-noised speech to computationally obtained resonances (see [9]) as well as to spectral peaks measured from 3D printed physical models from the simultaneously obtained MRI geometries, and finally (iv) a perceptual vowel classification experiment (see [10]) based on de-noised speech recorded during the MRI. These four validation experiments support the conclusion that the proposed noise cancellation algorithm can be used with good confidence for, at least, obtaining formants from speech contaminated by MRI noise. In Section 5, we apply the post-processing algorithm to speech that has been recorded during MRI scans as detailed in [2]. The objective is no longer to validate the algorithm rather than to draw conclusions about the speech data itself. We again use comparison

[1] Some experiments on the same speech data have been carried out using FastICA as well but adaptive methods seem to give better results.



(a) (b)

Fig. 1. Panel (a): The MRI head coil of Siemens Magnetom Avanto 1.5T scanner. The two-channel acoustic sound collector fits exactly the opening on the top. Panel (b): The sound collector positioned above a head model similarly as in the MRI experiments. The noise sample is acquired using a horn on the top surface of the collector and the speech sample from another similar horn pointing downwards.

samples that have been recorded in the anechoic chamber. There is a statistically significant ($p < 0.05$) discrepancy between some of the vowel formants extracted from these two kinds of data. It is further observed that the formant discrepancy has a consistent frequency dependent behaviour shown in Fig. 6 with steps at around 1 kHz and 2 kHz. In Section 6, a computational study is carried out based on the Helmholtz equation and the exterior space model shown in Figs. 7–8. It is observed that the acoustic space between the test subject's head and the MRI head coil produces a family of spectral energy concentrations. They appear as a common feature (i.e., as "external formants") in vowel recordings during MRI but not in similar recordings carried out in the anechoic chamber. In particular, the frequencies 1 kHz and 2 kHz get identified as external formants near some of the true vowel formants, explaining the increased formant discrepancy observed in Fig. 6.

2. Speech recording during MR imaging

2.1. Arrangements

The experimental arrangement has been detailed in [11,1,2]. Briefly, a two-channel acoustic sound collector samples speech and MRI noise in a configuration shown in Fig. 1. The signals are acoustically transmitted to a microphone array inside a sound-proof Faraday cage by waveguides of length 3.00 m. The microphone array contains electret microphones of type Panasonic WM-62. The preamplification and A/D conversion of the signals is carried out by conventional means, see [2, Section 3.1]. The experiments were carried out using Siemens Magnetom Avanto 1.5T using 3D VIBE (Volumetric Interpolated Breath-hold Examination) MRI sequence [12] as it allows for sufficiently rapid static 3D acquisition. Imaging parameters, etc., have been described in [2, Section 3.2].

2.2. Phonetic and geometric materials

The speech materials consist of Finnish vowels [ɑ, e, i, o, u, y, æ, œ] that were pronounced by a 26-year-old healthy male (in fact, the first author) in supine position during the MRI. The number of samples varies between 3 and 9 depending on the vowel. The MRI sequence requires up to 11.6 s of continuous articulation in a stationary supine position. The test subject produced the vowels at a fairly constant fundamental frequency f_0, given by the cue signal to the earphones. Two different pitches $f_0 = 104$ Hz and $f_0 = 130$ Hz

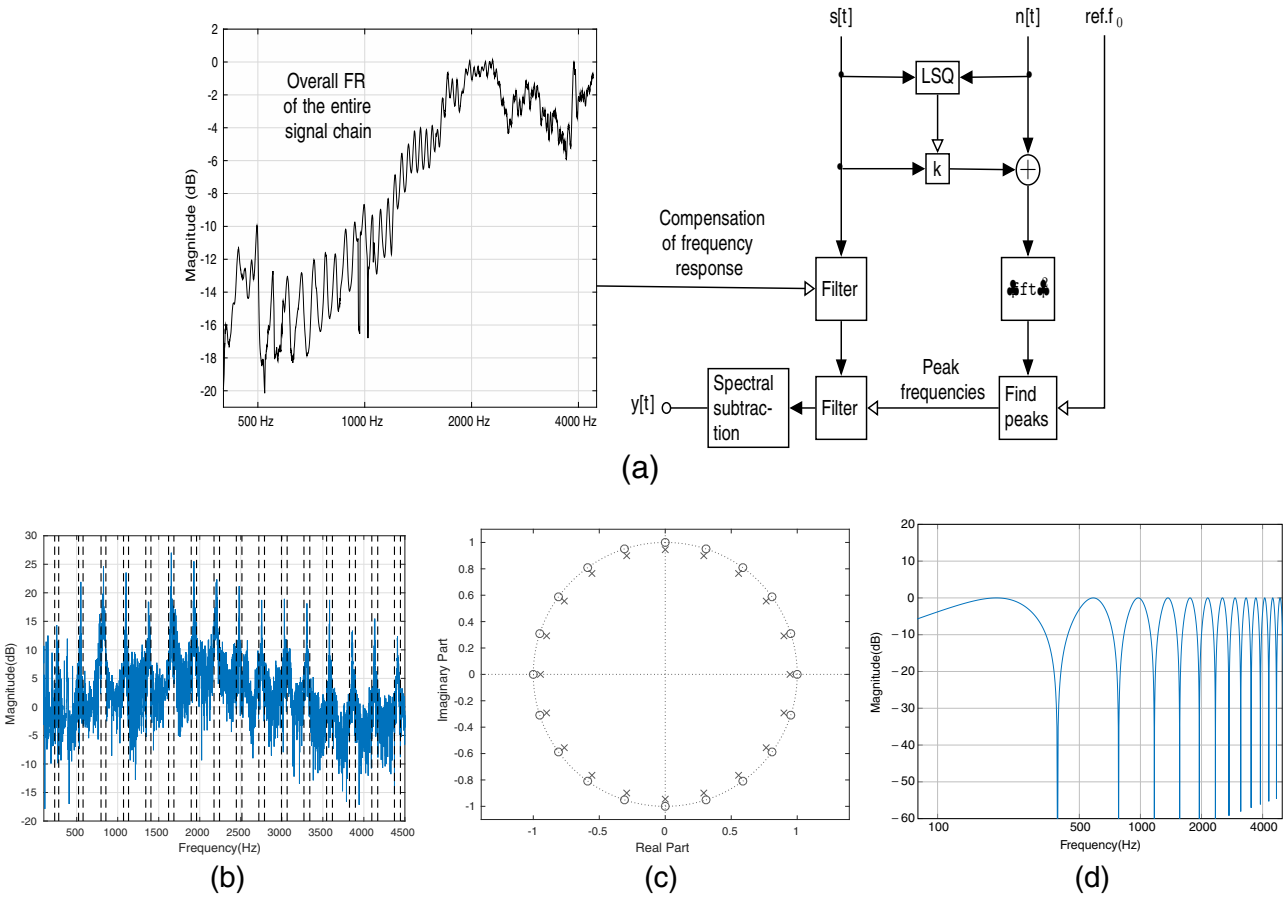


Fig. 2. Panel (a): A block diagram of the post-processing algorithm. Here $s[t]$ and $n[t]$ denote the discretised speech and noise samples at $f_s = 44\,100$ Hz, respectively. The signal $y[t]$ is de-noised speech. Panel (b): Harmonic structure of the MRI noise and stop bands estimated from it. Panel (c): The zero/pole placement in z-plane of the notch filter of degree 20 for removing the frequency $f_s/20$ and its harmonics below the Nyquist frequency $f_s/2$. Panel (d): The frequency response of a 112 order notch filter on the frequency band 80 Hz . . . 5 kHz of the kind that typically appears in the algorithm.

were used, and they had been chosen so as to avoid spectral peaks of the MRI noise.

The paired MRI/speech data for this article was acquired during a single session of 82 min in the MRI laboratory using the protocols reported in [1,2]. We obtained 107 MRI scans which is only possible using well-optimised experimental arrangements. Of the 107 scans, no more than 36 were prolonged vowels at $f_0 \approx 104$ Hz (with sample lengths ≈ 11.2 s) deemed usable for this study. To obtain comparison data, same kind of speech recordings were carried out in the anechoic chamber but neither the MRI coil reflections nor the ambient noise were replicated. Compared to MRI experiments, there are no similar restrictions in the anechoic chamber, apart from test subject fatigue. Thus, each vowel was now produced 10 times since the larger sample number was possible as a benefit of less demanding experimental arrangement.

3. MRI noise cancellation

We treat the measurement signals from speech and acoustic MRI noise $s[t]$ and $n[t]$ for $t \in \{h, 2h, 3h, \dots\}$ in their digitised form where $h = 1/f_s$, and the sampling frequency $f_s = 44{,}100$ Hz. The post-processing algorithm for these discrete time signals is outlined in Fig. 2a, and it consists of the following Steps 1–6 that have been realised as MATLAB code:

1. **LSQ:** Speech channel crosstalk is optimally removed from noise signal using coefficient k from least squares minimisation.
2. **Frequency response compensation:** The frequency response of the whole measurement system, shown in Fig. 2a, is compensated. The peaks in the frequency response are due to the longitudinal resonances of the waveguides, used to convey the sound from inside the MRI scanner to the microphone array placed in a sound-proof Faraday cage.
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 as shown in Fig. 2b. There are heuristics involved so that the harmonics of the reference value of f_0 do not get accidentally removed. Details are described below in pseudocode.
5. **Notch filtering:** The noise peaks are removed by using notch filters provided by the MATLAB function `iircomb` with parameters `n` equal to the number of different harmonic overtone structures detected, and the -3 dB bandwidth `bw` set at $6 \cdot 10^{-3}$.
6. **Spectral subtraction:** A sample of the acoustic background (including, e.g., noise from the helium pump) of the MRI laboratory (without patient speech and scanner noise) is extracted from the beginning of the speech recording. Finally, the averaged spectrum of this "silent sample" is subtracted from the speech signal using FFT and inverse FFT; see [13].

Algorithm 1. Adaptation to spectral structure.

We associate with each spectral peak p its location in spectrum $loc(p)$ in Hz, and its height $mag(p)$ in dB.
1: $P \longleftarrow$ set of all peaks found in the spectrum.
2: **procedure** FINDHARMONICS(P)
3: **while** $P \neq \emptyset$ **do**
4: $p \longleftarrow \max_{mag} P$
5: $P \longleftarrow P \setminus p$
6: **for** $q \longleftarrow P$ sorted by $|loc(p) - loc(P)|$ **do**
7: $d \longleftarrow |loc(p) - loc(q)|$
8: **if** $d < cf_0$ **then**
9: **continue**
10: **if** ∃ harmonics with fundamental d **then**
11: $F \longleftarrow F \cup$ `iircomb`(f_s/d)
12: $P \longleftarrow P \setminus \{r \in P : r = nd, n \in \mathbb{Z}\}$
13: **return** F

Harmonics are considered successfully found at step 10, if P contains four consecutive peaks with distance d. The value 1.5 has been used for the parameter c.

The proposed approach differs essentially from the earlier approach proposed in [2, Section 4]. Firstly, now there is no direct time-domain subtraction of the measured noise component from speech which makes the present approach more similar to [5]. For that reason, the low frequency components of speech are not attenuated as a result of the proximity of recording sound effect in dipole configurations. Secondly, using notch filters instead of high-order Chebyshev produces sharper removal of unwanted spectral components with much reduced musical noise artefact compared to what was reported in [2]. The comb filter is a more efficient way of removing higher harmonics of spectral peaks in the entire spectrum. In the current approach, the filter degree is determined by the Nyquist frequency $f_s/2$ =22 050 Hz and the number of notches required, making the computations much less intensive. However, using Chebyshev filters made it possible to vary the bandwidth of the stop bands as a function of frequency which possibility is now lost.

In [2], the post-processed speech recordings during MRI were classified with linear discriminant classifier, using the speech recorded in the anechoic chamber as a learning set. This experiment yielded 62% correct classifications. Repeating the experiment using the same speech data, the improved post-processing algorithm, and better accounting for the strong exterior resonance at ≈1 kHz as discussed in Section 6 below, the proportion of correctly classified vowels increases to 72%. Further significant improvement in classification accuracy does not seem possible since a strong systematic component is present in classification errors of both classification experiments, reflecting the properties of the speech data. More precisely, many [æ] get classified as [e], and many [e] get classified as [i]. Looking at the spectral envelopes (i.e., low resolution power spectral densities) of [æ] in Fig. 10, two different kinds of behaviour can be seen in the upper curves. Based on only F_1 and F_2, samples with the lower first peak location (i.e., F_1[æ]) are almost indistinguishable from [e] recorded in the anechoic chamber. This results in the first kind of systematic error. The second type of error is due to the systematic overestimation of F_2[e] ≈ 2 kHz in speech recorded during MRI as can be seen in Fig. 6. This artefact is connected to the acoustics inside the MRI head coil in Sections 5 and 6.

4. Performance analysis

4.1. Validation through synthetic signals

The formant extraction from noisy speech can be validated using artificially noise contaminated speech where the original formant positions are known precisely. Before going to details, let us first discuss how formants usually are extracted from speech signals.

Vowel formants can be understood as centre frequencies of relatively wide peaks appearing in the power spectral envelope of the signal. Such spectral envelope is a low resolution version of the power spectral density where the narrow peaks from the harmonic structure of the glottal fundamental frequency have been sufficiently downplayed to reveal the resonance structure due to the vocal tract. There are several methods available for estimating the power spectral density (and, hence, the spectral envelope) of a signal. The approaches can be divided into parametric and nonparametric methods. The nonparametric methods assume nothing of the signal apart from its stationarity, but they require long signals for accurate results. This also precludes using them as an online application. Parametric methods assume that the signal is produced by a parametric model whose structure is chosen *a priori*, and an attempt is made to find its optimal model parameter values for matching the signal. For details on both kinds of methods, see, e.g., [14]. The underlying parametric model is usually autoregressive (AR) in speech sciences (see [15]), and the most common variants include the Yule-Walker method, the Burg method, and the covariance method, all of which are examples of Linear Predictive Coding (LPC). A comprehensive overview can be found in [16,17]. All low-order rational spectral envelopes in this article have been produced using Burg's method which is the traditional approach and found in the speech analyser Praat [18] as well. More precisely, the MATLAB function `arburg` is used for producing rational spectral envelopes from which the formants are extracted by locating the poles.

Table 1
Original formants (left) and formants extracted after the artificial addition of MRI noise and subsequent noise cancellation (right).

Vowel	F_1	F_2	F_3
[ɑ]	598	1094	1918
[e]	453	1691	2255
[i]	318	1900	2097
[o]	465	815	2233
[u]	410	898	1934
[y]	379	1535	2034
[æ]	562	1452	2375
[œ]	436	1400	2076

Pure vowel signals were taken from comparison data for each vowel in [ɑ, e, i, o, u, y, æ, œ], and their formants F_1, F_2, and F_3 were computed. A sample of MRI noise (without any speech content) was recorded using the experimental arrangement detailed in [2, Section 3], and it was mixed with each vowel sample so that the speech and noise components have equal energy contents (SNR ≈ 0 dB). The post-processing algorithm described in Section 3 was then applied to these signals, of which an example is shown in Fig. 3.

It was first observed that the post-processing increases the SNR of the artificially noise-contaminated signals by 9 . . . 14 dB depending on the vowel. The three formants F_1, F_2, and F_3 were extracted from artificially noise contaminated vowels after they had been post-processed. The resulting formant frequencies are within −0.5 . . . 0.3 semitones from those measured from the original pure vowels, except for the outlier F_2[o] where the discrepancy is 1.1 semitones (Table 1).

The average formant discrepancies of under 2.8 semitones were reported in [2, Table 3] between speech formants and Helmholtz resonances computed from vocal tract geometries (without any model for the surrounding space) that were obtained by simultaneous MRI. Also, the observations in [19] provide magnitudes for formant error that results from inherent variation in long vowel productions due to test subject adaptation and fatigue. Comparing these values with the results on artificially contaminated speech, we conclude that formant extraction from algorithmically post-processed signals can be regarded as a relatively small error source.

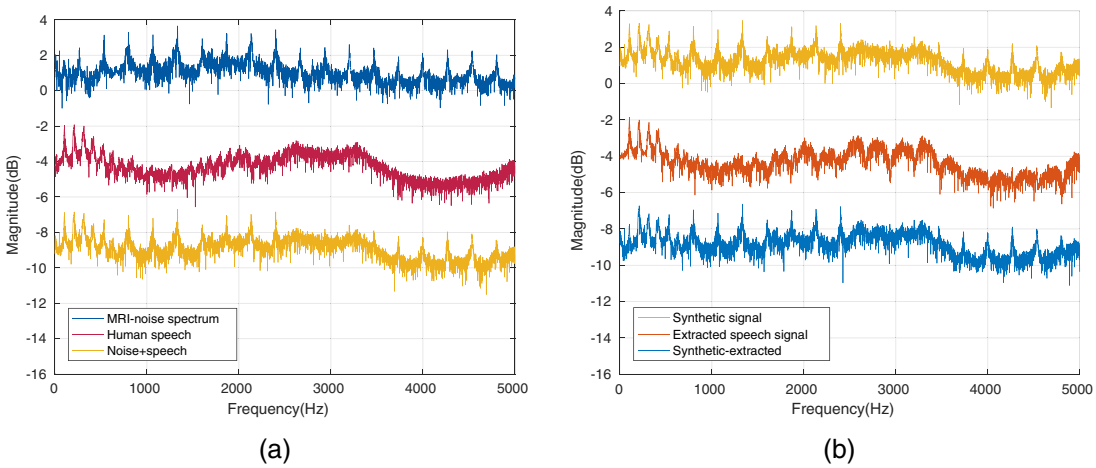


Fig. 3. Illustration of the artificially noise-contaminated vowel signal. Panel (a): MRI noise (upmost), pure vowel signal (middle), and the synthetic signal as their sum (lowest). Panel (b): Synthetic signal (upmost), signal after post-processing using the proposed algorithm (middle), and the reconstructed noise (lowest).

Table 2
Spectral tilts (in dB/octave) from recordings in the anechoic chamber and from samples recorded during the MRI noise after post-processing.

	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[œ]
Anech	12.2	11.9	9.0	14.5	15.6	12.6	11.3	12.7
MRI	15.7	13.9	9.2	17.9	15.3	13.5	14.0	15.2

4.2. Comparison of spectral tilts

In addition to formants, another important spectral characteristic of speech signals is the *spectral tilt* or *roll-off*. It is a measure of attenuation at higher frequencies that are still relevant to speech. We quantify the spectral tilt by first fitting a low-order rational spectral envelope on the frequency range of speech, and then finding the LSQ regression line to the envelope on the logarithmic frequency range between 465 Hz and 5 kHz. The bound 465 Hz is the mean of all F_1's present in the dataset.

The spectral tilt data is given in Table 2 where the unit is dB/octave, and the positive value indicates attenuation at higher frequencies. The roll-off in post-processed speech during the MRI is systematically larger than in comparison data (in average by 1.9 dB/octave), the only exception being the vowel [u]. We point out that the two kinds of spectral tilt data in Table 2 correlate strongly ($R = 0.78$). As can be seen from Fig. 5d, the difference of the average spectral tilts is quite small. The difference is, at least, partly explained by the fact that there was a lot of more attenuating material (such as cushions to keep the test subject's head in place, cloth to keep instruments sterile, etc.) around the test subject in the MRI scanner, compared to experiments in the anechoic chamber. The filtering operations performed by the proposed algorithm do not change the spectral tilt significantly (see Fig. 2d).

The nominal value of 12 dB/octave for the spectral tilt is given in [20] but the authors point out that the value is too low for falsetto and breathy phonations. Vowel productions extending over 10 s are often breathy because the test subject needs to save air but, to our knowledge, this has not been further studied. However, there exists literature on the spectral tilt in consonant–vowel syllables at the phonation onset. The value of 13.5 dB/octave can be concluded from [21, Fig. 1], describing the first four glottal waveforms after the onset of phonation.

4.3. Comparison to sweeps in physical models

Three of the MR images corresponding to Finnish quantal vowels [ɑ, i, u] were processed into 3D surface models (i.e., STL files) and intersectional area functions for Webster's equation as explained in [22]. Fast prototyping was used to produce physical models from the STL files in ABS plastic with wall thickness 2 mm. The printed models extend from the glottal position to the lips, and they were coupled to a custom acoustic source (see Fig. 4) whose design resembles the loudspeaker-horn construction shown in [23, Fig. 1]; see also [24].

The acoustic source contains an electret (reference) microphone (ø 9 mm, biased at 5 V) at the glottal position, and another similar (signal) microphone was placed near the lips. A sinusoidal logarithmic sweep was preweighted by the iteratively measured inverse response of the acoustic source in order to obtain a uniform sound pressure level at the reference microphone for all frequencies of interest. The frequency responses of the physical models (and reference resonators with known resonant frequencies were measured using this arrangement between 80 Hz . . . 7 kHz.

As can be seen from Fig. 5, there is good correspondence between the spectra of de-noised speech from MRI experiments and the spectra from physical models of the simultaneously imaged vocal tract geometry. There are some extra peaks in both kinds of spectra that correspond to spurious resonances not due to the vocal tract geometry. We point out that the physical models did not contain the face, and the sweep measurements were carried out in an open acoustic environment in the anechoic chamber. This is in contract to the speech recordings that were carried out within MRI head and neck coils [1,2].

It is worth observing from Fig. 5 that the spectral tilt (as defined in Section 4.2) of the frequency response from physical models is practically 0 dB/octave. This is due to two reasons: (i) A 3D printed vocal tract is a virtually lossless acoustic system apart from the radiation losses through mouth opening, and (ii) the glottal excitation in natural speech has its characteristic roll-off estimated to be within 12 . . . 18 dB/octave in [25, p. 991] whereas the measurements from the physical models were carried out keeping the sinusoidal sound pressure constant at the glottal position. Another way to quantify the spectral tilt in terms of the Harmonic Richness Factor (HRF) is given in [20], where values 9.1 . . . 19.1 dB (HRF) were given, not directly comparable with the interval given in [25].

4.4. Perceptual evaluation

A listening experiment was carried out to evaluate the effect of post-processing on vowel recognition. In the experiment, 12 subjects listened to 48 samples of vowel phonation. There were six samples of each Finnish vowel in [ɑ, e, i, o, u, æ, œ]; three unpro-

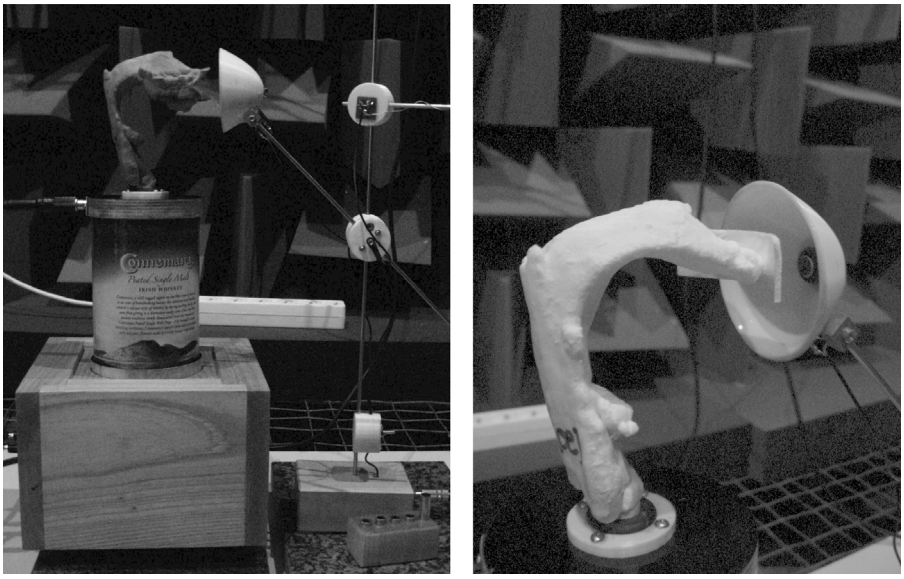


Fig. 4. A detail of the sweep measurement arrangement for 3D printed vocal tract configurations of [ɑ, œ].

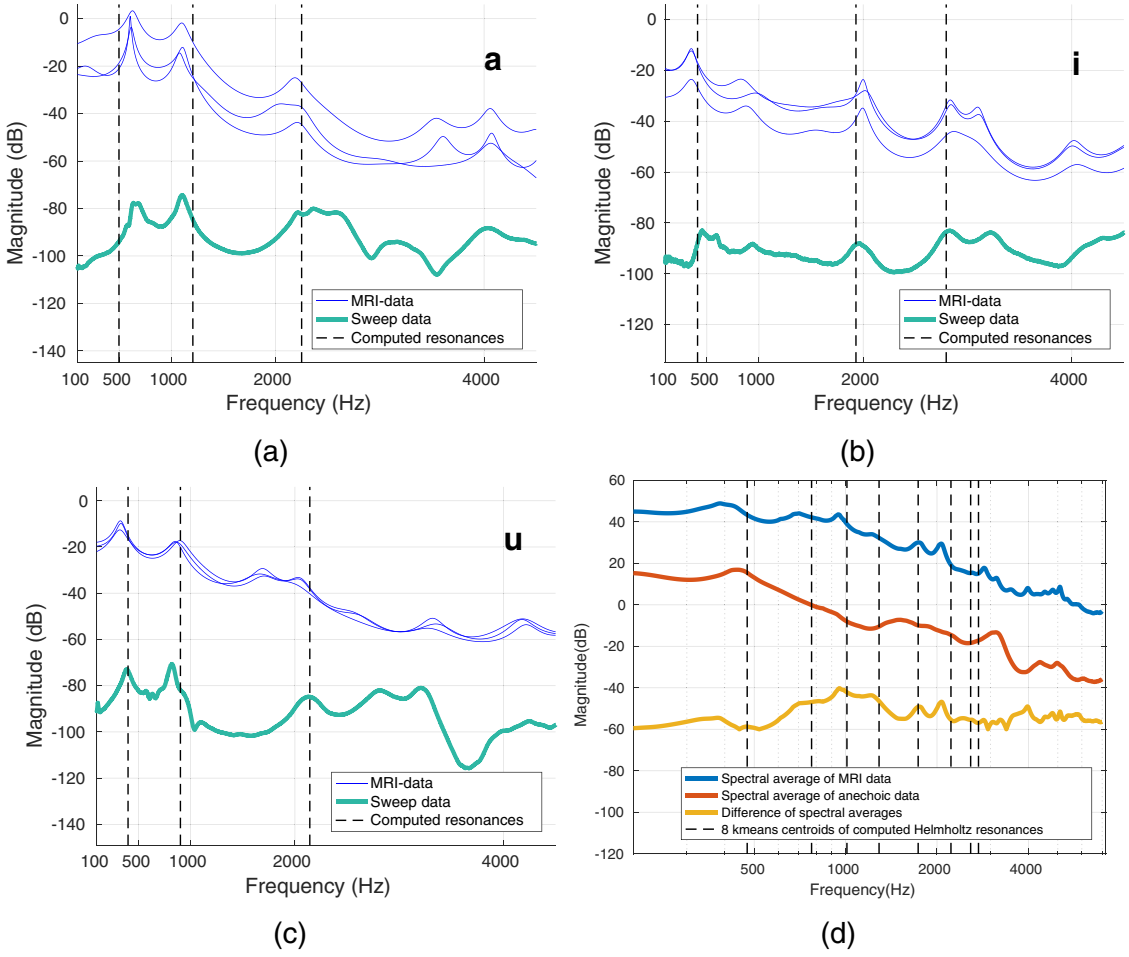


Fig. 5. Panels (a–c): Spectral envelopes and computationally obtained resonances of [ɑ, i, u]. The upper curves are power spectral densities of speech recorded during an MRI scan. The lower curves are frequency responses measured from the physical models that have been produced from the MR images. The vertical lines indicate the three lowest resonances computed by Webster's model from the same vocal tract geometry using the mouth impedance optimisation process introduced in [9]. Panel (d): Averages of spectral envelopes of Finnish vowels [ɑ, e, i, o, u, y, æ, œ] from two different kind of recordings. Each vowel appears in the averages with the same weight. The topmost curve describes speech recorded during the MRI scan, the middle curve recordings in the anechoic chamber, and the lowest curve is their difference. The averaging highlights the common features (partly due to the exterior acoustics) within both kinds of vowel recordings. The vertical dashed lines represent *k*-means cluster centroids of the Helmholtz resonant frequencies computed using a 3D model of the MRI head coil.

Table 3
Results of the perceptual comparison experiment on vowels, some of which were artificially contaminated by MRI noise and then de-noised. Quite many target samples of [u] were classified as [o] in both kinds of samples.

Vowel samples from anechoic chamber								
	Categorised as							
Target	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[œ]
[ɑ]	36	0	0	0	0	0	0	0
[e]	0	33	0	0	0	0	0	3
[i]	0	0	36	0	0	0	0	0
[o]	6	0	0	30	0	0	0	0
[u]	0	0	0	13	23	0	0	0
[y]	0	0	0	0	0	32	0	4
[æ]	0	1	0	0	0	0	32	1
[œ]	0	3	0	0	0	0	0	33

Table 4
The p-values computed with Smith-Satterwaith procedure for distributions with unequal variances. Formant samples that reject the null hypothesis H_0 at $p \leq 0.05$ are written in bold.

	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[œ]
F_1	**0.01**	**0.02**	0.16	0.86	0.30	**0.05**	0.75	0.93
F_2	0.79	**0.01**	**0.01**	**0.01**	**0.02**	**0.01**	0.19	**0.02**
F_3	0.18	**0.01**	**0.01**	0.40	0.83	**0.01**	0.39	0.25

cessed samples from the anechoic chamber while the remaining three had undergone the MRI noise contamination and de-noising process described in Section 4.1. The duration of each sample was 10 s. The experiment yielded thus 36 datapoints for both unprocessed and processed sound. The test subjects' ages ranged from 20 to 55; two of the subjects were female. The experiment was conducted in an empty office room using Bose AE2 headphones. The test subjects were allowed to listen each sample as many times as they wanted. Using a custom made computer interface, they reported the vowel that the phonation resembled the most in their opinion.

The results of the perceptual experiment are given in Table 3. As a conclusion, there is a slight increase in classification mistakes induced by the proposed algorithm, but the increase is a fraction of the classification mistakes due to natural speech variation in the samples used. To draw statistically significant conclusions on such small effects would require a considerably larger data set.

5. Formant extraction from noisy speech

After four validation experiments on the post-processing algorithm described in Section 3, it is time to apply it on true speech data, recorded during an MRI scan. Our purpose is to show by comparative studies that the acoustic environment in the MRI scanner introduces resonant artefacts to speech signals that are large enough to be clearly quantifiable using the proposed algorithm.

To increase the number of vowel sound samples from MRI experiments, six partial samples of 1 s were taken from each recording. These partial samples are separated from each other by at least 1 s of time to enhance the independence of the samples. This six-fold increase of the original sample number improves the statistical analysis given in Table 4. Spectral envelopes of all speech samples are shown in Figs. 9 and 10 where variance between same vowel productions in different MRI scans (or different parts of the same scan) can be observed.

We proceed to show that some of the extracted formant means of samples from the anechoic chamber and the MRI laboratory are significantly nonequal. The estimated formant means μ_{ac} and μ_{mri} are compared using Student's t-distribution where the degrees-of-freedom is determined by the Smith-Satterwaithe procedure; see the unequal variance test statistics in, e.g., [26, Section 10.4]. In case of the vowel formant F_j [ɑ] for $j = 1, 2, 3$, our null hypothesis is that

$$H_0 : \mu_{ac}\left(F_j[\text{ɑ}]\right) = \mu_{mri}\left(F_j[\text{ɑ}]\right) \tag{1}$$

We try to reject H_0 by showing that its converse H_1 is true so the probability p of type 1 error (i.e. rejecting the null hypothesis when it is true) is low, say $p < 0.05$. In this case the experiment indicates that the formants extracted from the two data sources are not consistent but, instead, a feature of the underlying data requiring further investigations.

The results of the experiments are given in Table 4 where the p-values are given. We conclude that H_0 gets typically rejected for F_2 in all vowels except [ɑ, æ] and for all formants in vowels [e, y]. The formant means from post-processed speech during the MRI are plotted in Fig. 6 against their counterparts recorded in the anechoic chamber from the same test subject. If these two datasets were perfectly consistent, all data points would be expected to appear between the two diagonal dashed lines, representing the maximum error of formant extraction from noisy speech as discussed in Section 4.1. We conclude that (at least) 12 of the discrepancies shown in Fig. 6 reflect actual differences of the speech data recorded in MRI laboratory, compared to similar data from the anechoic chamber.

It is worth observing that the formant discrepancy in Fig. 6 shows a peculiar staircase pattern where two plateaus appear near 1 kHz and 2 kHz. More precisely, we observe that in samples recorded during the MRI, we have F_2[y], F_2[œ] $\rightarrow$ 1 kHz from above and F_2[e], F_2[i] $\rightarrow$ 2 kHz from below. The vertical level at 1 kHz coincides with an extra peak appearing in Figs. 9 and 10 in most of spectral envelopes of signals recorded during the MRI; notable exceptions are the vowels [ɑ,u,o] where $F_2 \approx 1$ kHz would conceal any extra peak. These extra peaks can also be seen in Fig. 5d where the spectral envelopes of all vowel recordings in the MRI laboratory (in the anechoic chamber, respectively) have been averaged to downplay the vowel specific formant peaks. It has been excluded by frequency response measurements and ensuing equalisation that these peaks could be an artefact of the speech recording instrumentation.

A similar staircase pattern to Fig. 6 near frequencies 1 kHz and 2 kHz has been observed in [27, Chapter 5, Fig. 5.4] where measured formant and computed resonance pairs have been plotted against each other. The vocal tract resonances in [27] have been computed by the Helmholtz equation from MRI data without exterior space modelling, and the formants have extracted from recordings during the MRI as explained in [2, Section 5].

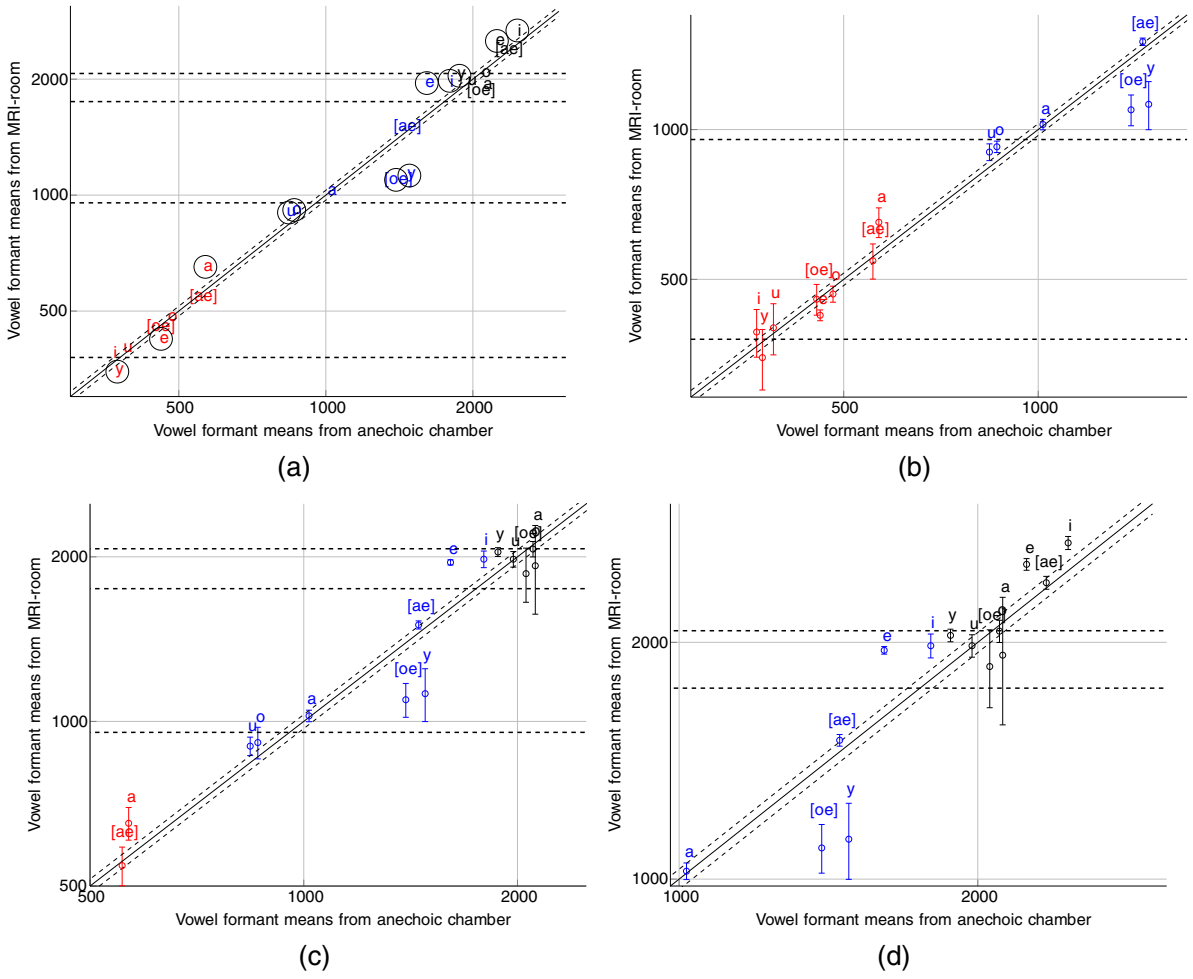


Fig. 6. Estimates of formants F_1, F_2, and F_3 that have been extracted from the vowel samples of [ɑ, e, i, o, u, y, æ, œ] recorded during the MRI. They are plotted against the comparison data recorded in the anechoic chamber from the same test subject. Panel (a) presents the overall situation on the entire frequency range of interest. Panels (b), (c), (d) are close ups where the error bars represent standard deviations of the estimated formants from recordings during MRI. Similar standard deviations of the comparison data are much smaller. The diagonal dashed lines describe the error bounds of ±0.5 semitones as obtained in Section 4.1. Where the estimated formant discrepancy is statistically significant at $p \leq 0.05$, the vowel has been encircled in panel (a); see also Table 4. The horizontal dashed lines show peaks of the spectral envelopes in Fig. 5d that were identified as resonance clusters external to the vocal tract.

6. Identification of exterior resonances

The statistically significant discrepancy in Fig. 6 is expected to be a combination of three different sources: (i) Perturbation[2] of the vocal tract resonances by the adjacent exterior space resonances, caused by reflections from test subject's face and MRI head coil surfaces; (ii) Lombard speech due to the acoustic noise during the MRI (see [28,29]); and (iii) active adaptation of the test subject to the constrained space acoustics inside the MRI head coil. Of these three possible partial explanations, only the first can be studied without carrying out extensive experiments with test subjects. Instead, we can use the simultaneously obtained MR image of the vocal tract for numerical resonance computations in order to investigate the acoustic artefacts in speech caused by the MRI coil.

We extract the vocal tract geometries from the MR images by custom software as explained in [27]. The vocal tract geometries are joined with an idealised geometric model of the head coil as well as a head geometry as shown in Fig. 7. The head geometry was purchased from TurboSquid [30]. The computational domain Ω is split into the interior part Ω_1, the exterior part Ω_2, and the spherical interface $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ as shown in Fig. 7. Both Ω_2 and Γ are same in all computations but Ω_1 (containing the vowel dependent vocal tract) changes.

We use the finite element method (FEM with piecewise linear elements on a tetrahedral mesh with discretisation parameter $h > 0$) to solve the Helmholtz equation $\Delta u = \kappa^2 u$ in Ω and identify those resonances that have strong excitations in Ω_2. Here $\kappa = \omega/c$ where c is the speed of sound, and ω is the complex angular velocity. Using FEM and Nitsche's method (see [31]) on the interface Γ, the Helmholtz equation takes the variational form

$$a(u, v) = \kappa^2 b(u, v) \text{ for all } v \in V \tag{2}$$

where the bilinear form $a(\cdot, \cdot)$ is defined as

$$\begin{aligned} a(u, v) = \sum_{i=1}^{2} (\nabla u, \nabla v)_{\Omega_i} - \langle \left\{ \frac{\partial u}{\partial n} \right\}, [\![v]\!] \rangle_\Gamma \\ - \langle [\![u]\!], \left\{ \frac{\partial v}{\partial n} \right\} \rangle_\Gamma + \nu_h \langle [\![u]\!], [\![v]\!] \rangle_\Gamma. \end{aligned} \tag{3}$$

[2] The discrepancy in vowel formants extracted from speech may be due to misidentification of exterior formants as adjacent vocal tract formants, or there may be "frequency pulling" of a correctly identified vowel formant by an adjacent exterior formant. In Helmholtz computations, we can always tell the true formants by looking at the corresponding pressure eigenmodes. Only spectrogram data is available from measured speech.

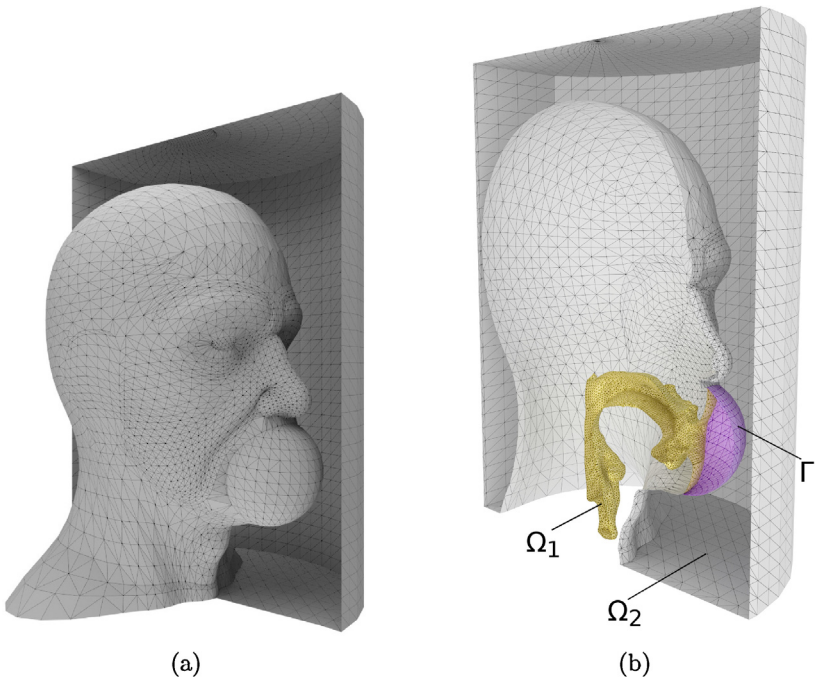


Fig. 7. Panel (a): An illustration of the computational domains used for identifying the acoustic resonances within MRI head coil. Panel (b): The computational domains Ω_1, Ω_2, and the interface Γ.

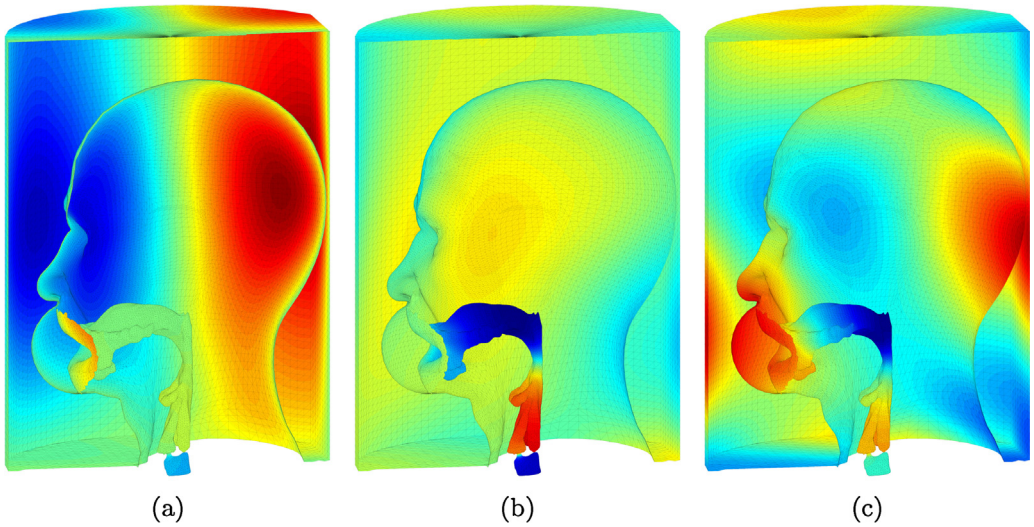


Fig. 8. The modal pressure distributions at the domain boundary of the vocal tract geometry of [ɑ] at the resonant frequencies 1062 Hz, 1120 Hz and 1460 Hz. These represent the three archetypal variants encountered. Panel (a): The exterior space is active but the vocal tract is not. Panel (b): The vocal tract is active but exterior space is not. Panel (c): A mixed mode is shown where both the domain components are active.

Here $\{u\}$ ($[\![u]\!]$) is the average (respectively, the jump) of u over the interface Γ, and ν_h is a mesh size dependent parameter. The bilinear form $b(\cdot,\cdot)$ in (2) is the inner product of $L^2(\Omega)$. Using Nitsche's method on interface Γ makes it possible to use the same discretisation of Ω_2 for all vowel geometries. For a similar kind of numerical experiment, see [32].

The resonance structures of each of the 51 vowel geometries in the data set were computed on Ω by FEM as explained above. The resulting 3060 complex angular velocities ω were processed as follows:

(i) Depending on the vowel, three or four ω's, corresponding obviously to the lowest formants of the vocal tract volume Ω_1, were excluded. This was based on comparing the energy densities in Ω_1 and Ω_2 of the respective eigenfunctions u. A total of 2866 ω's remain that indicate significant acoustic excitation in the exterior domain Ω_2.

(ii) Next, 1075 of the 2866 eigenfunctions u having largest $Re\omega$ (i.e., being least attenuated) were identified, with frequencies between 300 Hz . . . 3 kHz.

(iii) Eight frequency clusters were formed by the k-means algorithm (see [33]) from the remaining 1075 complex wavenumbers ω based on the resonant frequencies $f = Im\omega/2\pi$.

The cluster centroids indicate concentrations of acoustic energy around the eight frequencies, shown by vertical dashed lines in Fig. 5. The energy concentrations coincide quite well with the peaks of the topmost average-of-envelopes curve in Fig. 5d, produced

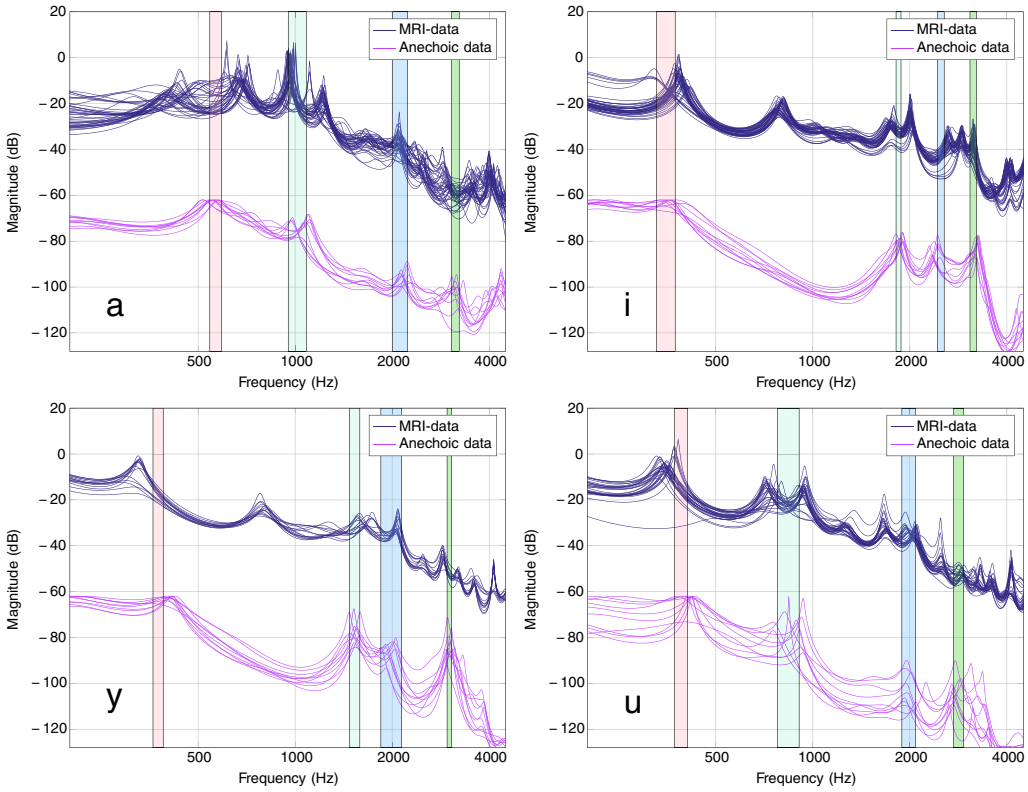


Fig. 9. Spectral envelopes of all vowel [ɑ, i, y, u] samples in the dataset. In each panel, the upper curves represent post-processed signals recorded during the MRI experiments. The lower curves are similar envelopes without any post-processing of signals, obtained from the same test subject in the anechoic chamber. These two families of curves are comparable to curves given in [2, Figs. 7 and 8]. The vertical bars are error intervals for formants $F_1, \ldots, F_4$ extracted from the recordings in the anechoic chamber.

from speech during the MRI. There is much less match with the middle curve in the same figure, produced similarly from speech in the anechoic chamber. We conclude that some effects of the MRI coil reflections are, indeed, present in speech recorded during the MRI. The corresponding artefact peaks in speech spectrograms (see Fig. 5d) occur at the frequencies 380 Hz, 955 Hz, 1750 Hz, 2070 Hz, 3230 Hz, and 3970 Hz of which the four lowest are displayed as horizontal lines in Fig. 6.

7. Discussion

When trying to match a computational model of speech to true speech biophysics, some sort of paired data is necessary. For example, if the acoustic modelling is based on vocal tract geometries acquired by MRI, then the most suitable accompanying data consists of speech samples recorded during the same MRI scan. Unfortunately, these samples are always contaminated by high levels of scanner noise and other acoustic artefacts that must be eliminated before a reliable extraction of desired features (such as the formant positions and the spectral tilt) is possible. Applications related to, e.g., modelling of oral and maxillofacial surgery require extreme precision that is feasible in model computations only by careful parameter estimation and validation of model components. Such models can only be as reliable as their validation data.

In our measurement setting, speech and noise samples are collected essentially at the same point (see Fig. 1 and [2]) although from opposite directions. Issues related to delays and multiway propagation are less serious compared to settings where the sound is collected further away as was done in [5,6]. Hence, it is not necessary to develop a high-order noise model as in [5], but a computationally less intensive and a more tractable post-processing of speech can be used. Additionally, the goal of the work presented above is not to reconstruct speech process in its entirety, but rather allow the extraction of its certain characteristic features, i.e., formants with high degree of accuracy.

The proposed algorithm operates almost entirely in frequency domain which is unavoidable, regardless of all other aspects, for compensating the frequency response of the recording system. We point out that also a real-time, time-domain, analogue subtraction of MRI noise from recorded speech is used during the experiment to provide instant feedback to patient's earphones. The analogue circuit removes low frequency noise very effectively but is useless at higher frequencies where noise arrives to the sound collector channels in different phase.

The notch filtering adds a large number of transmission zeros to processed signals which causes the phase response of the algorithm to be nonlinear. This may be a showstopper if the post-processed signal is to be used as an input for another speech processing algorithm such as the Glottal Inverse Filtering (GIF) for glottal pulse extraction, see [34,35]. To produce signals with linear phase response, one should use, e.g., noncausal spectral filtering (see [36]) instead of notch filters.

Scanners with lower magnetic field intensity (such as used in [6,7]) typically have an open construction where speech may be recorded rather successfully by directional microphones, located at a safe distance from the scanner. Low-field scanners unfortunately produce worse image resolution, and they require longer scanning durations which are undesirable features in speech studies. Here, the recording setup is built around a Siemens Magnetom Avanto 1.5T MRI scanner having higher magnetic field intensity but a closed construction. Using the arrangement detailed in Fig. 1, we are able to obtain an accurate estimate of the scanner noise near the test subject's mouth since the MRI coil surfaces act as an additional acoustic shield between the speech and the noise channels. Thus, the spectral peaks of noise can be extracted quite accurately, and a set of comb filters can be designed to precisely and econom-

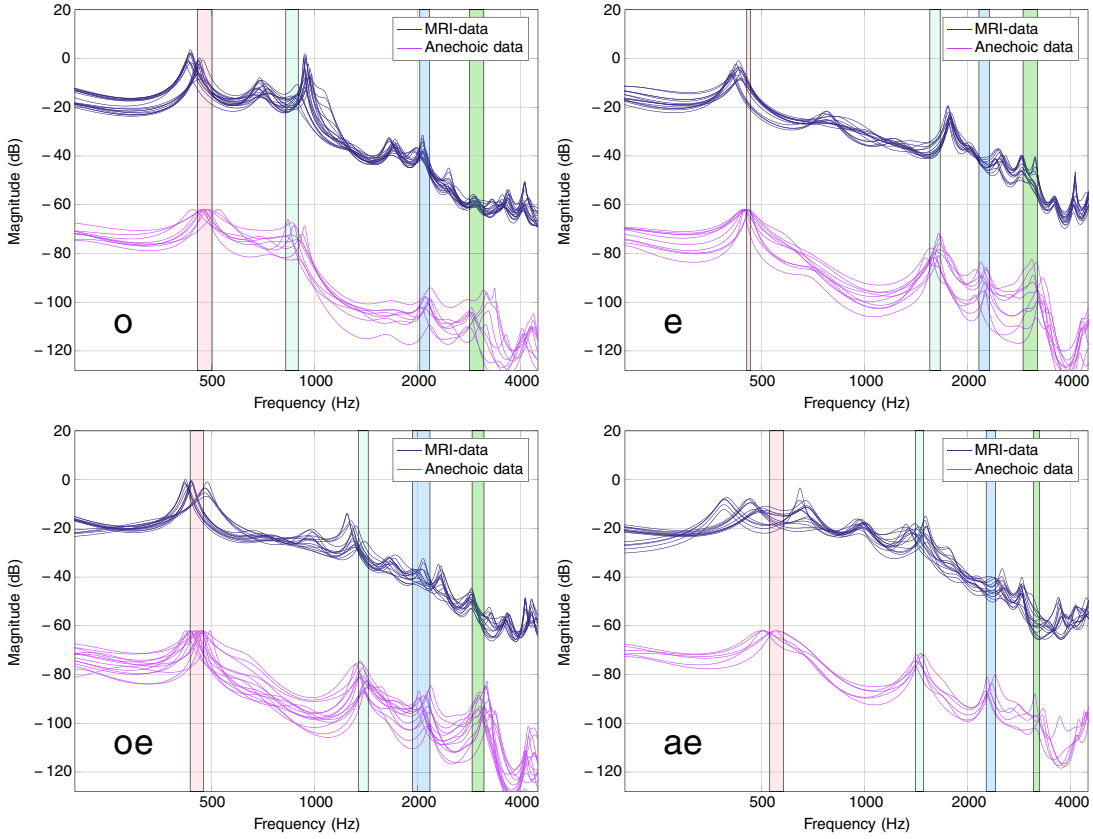


Fig. 10. Spectral envelopes of all vowel [o, e, œ, æ] samples in the dataset. The presentation is similar to Fig. 9.

ically remove these frequency bands from speech recordings. This makes it unnecessary to resort to methods such as the spectral noise gating [8] or the cepstral transformation [6] that affect the entire frequency range. Moreover, the proposed algorithm can make good use of the fact that our main interest lies in long vowel utterances at a fixed f_0, chosen not to coincide with the dominant spectral peaks of the scanner noise. The zeroes of the comb filters are chosen adaptively for each recording which makes it possible to apply the proposed algorithm to different MRI sequences.

8. Conclusions

A post-processing algorithm was proposed for removing acoustic noise from speech that has been recorded during the MRI using special MRI-proof instrumentation. It is one of the salient features of MRI scanner noise that it mainly consists of few strong fundamental frequencies accompanied by their harmonic overtones. The algorithm outlined in Section 3 first identifies such harmonic structure and then adapts a collection of notch filters to the detected frequencies. The algorithm is realised as MATLAB code.

The proposed algorithm is significantly different from the approaches presented in [5–8]. Many of these differences are motivated by dissimilarities in experimental arrangements for data acquisition. The post-processing algorithm was validated by using artificially noise-contaminated vowels where the noise has been recorded from the MRI scanner running the same MRI sequence as in the prolonged vowel experiments. Such artificially MRI noise contaminated vowels have known formant positions and predetermined SNR's which makes it possible to assess the achievable noise reduction in post-processing. In the proposed approach, we observe that 9 . . . 14 dB reduction of MRI scanner noise is attainable for prolonged vowel signals, and the formant extraction error due to post-processing is less than half a semitone. Considering the natural formant variation in speech signals recorded in optimal conditions (see the lower curve families in Figs. 9 and 10), this is an adequate level of performance for the validation and the parameter estimation of a computational speech model such as proposed in [3].

The algorithm was applied on real speech data. A set of prolonged vowels was recorded during the MRI, and this data was post-processed. Comparison measurements were recorded in optimal conditions from the same test subject. Vowel formants were extracted from both types of data, and it was observed that the formant discrepancy between the two kinds of data has a strongly frequency dependent behaviour. Particularly large deviations were observed near 1 kHz and 2 kHz. At these frequencies, the formant discrepancy is several times as large as the formant estimation error due to the post-processing algorithm, and the deviations are statistically significant (Student's t-test with $p < 0.05$). We presented computational evidence that the deviant frequencies are related to the acoustic resonances of the space between test subject's face and MRI coils. However, some of the formant error may also be due to test subject's adaptation to his acoustic environment during the MRI scan.

Even though the algorithm has been designed for the main purpose of formant extraction, it gives audibly quite satisfactory results from natural speech that has been recorded during dynamic MRI of midsagittal sections.

Acknowledgements

The authors wish to thank many collegues for consultation and facilities: Dept. Signal Processing and Acoustics, Aalto University (Prof. P. Alku), PUMA research group at Dept. Oral and Maxillofacial Surgery, University of Turku (Prof. R.-P. Happonen and Dr. D. Aalto), Medical Imaging Centre of Southwest Finland (Prof. R. Parkkola and

Dr. J. Saunavaara), and Aalto University Digital Design Laboratory (M. Arch. A. Mohite). The authors wish to express their gratitude to the five anonymous reviewers for their comments and proposed improvements.

The authors have received financial support from Instrumentarium Science Foundation, Magnus Ehrnrooth Foundation, Niilo Helander Foundation, and Vilho, Yrjö and Kalle Väisälä Foundation.

References

[1] 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, 2011, pp. 168–173.
[2] D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.M. Luukinen, J. Malinen, T. Murtola, R. Parkkola, J. Saunavaara, M. Vainio, Large scale data acquisition of simultaneous MRI and speech, Appl. Acoust. 83 (1) (2014) 64–75.
[3] A. Aalto, T. Murtola, J. Malinen, D. Aalto, M. Vainio, Modal Locking Between Vocal Fold and Vocal Tract Oscillations: Simulations in Time Domain, 2017, arXiv:1506.01395.
[4] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Netw. 13 (4–5) (2000) 411–430.
[5] 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.
[6] J. Přibil, J. Horáček, P. Horák, Two methods of mechanical noise reduction of recorded speech during phonation in an MRI device, Meas. Sci. Rev. 11 (3) (2011) 92–99.
[7] J. Přibil, A. Přibilová, I. Frollo, Analysis of spectral properties of acoustic noise produced during magnetic resonance imaging, Appl. Acoust. 73 (8) (2012) 687–697.
[8] J. Inouye, S. Blemker, D. Inouye, Towards undistorted and noise-free speech in an MRI scanner: correlation subtraction followed by spectral noise gating, J. Acoust. Soc. Am. 135 (3) (2014) 1019–1022.
[9] J. Kuortti, J. Kivi, J. Malinen, A. Ojalammi, Mouth impedance optimisation for vocal tract resonances of vowels, in: Proceedings of 27th Nordic Seminar on Computational Mechanics, 2015, pp. 93–96.
[10] 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.
[11] J. Palo, A Wave Equation Model for Vowels: Measurements for Validation, Aalto University, Department of Mathematics and Systems Analysis, 2011 (Licentiate thesis).
[12] 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.
[13] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. 27 (2) (1979) 113–120.
[14] P. Stoica, R.L. Moses, Spectral Analysis of Signals, Prentice-Hall, 2005.
[15] L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.
[16] J. Makhoul, Linear prediction: a tutorial review, IEEE Proc. 63 (4) (1975) 561–580.
[17] S.M. Kay, Modern Spectral Estimation: Theory and Application, Prentice-Hall, 1988.
[18] P. Boersma, Praat, a system for doing phonetics by computer, Glot Int. 5 (9/10) (2001) 341–345.
[19] 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, 2011, pp. 180–183.
[20] D. Childers, C. Lee, Vocal quality factors: analysis, synthesis, and perception, J. Acoust. Soc. Am. 90 (5) (1991) 2394–2410.
[21] J. Alexander, K. Kluender, Spectral tilt change in stop consonant perception, J. Acoust. Soc. Am. 123 (1) (2008) 386–396.
[22] 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, 2013, pp. 257–260.
[23] D. Tze, W. Chu, K. Li, J. Epps, J. Smith, J. Wolfe, Experimental evaluation of inverse filtering using physical systems with known glottal flow and tract characteristics, J. Acoust. Soc. Am. 133 (5) (2013).
[24] H. Takemoto, P. Mokhtari, T. Kitamura, Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method, J. Acoust. Soc. Am. 128 (6) (2010) 3724–3738.
[25] R. Monsen, A. Engebretson, Study of variations in the male and female glottal wave, J. Acoust. Soc. Am. 62 (4) (1977) 981–993.
[26] J. Milton, J. Arnold, Introduction to Probability and Statistics, 4th ed., McGraw-Hill, 2003.
[27] A. Kivelä, Acoustics of the Vocal Tract: MR Image Segmentation for Modelling, Aalto University School of Science, Department of Mathematics and Systems Analysis, 2015 (Master's thesis).
[28] V. Hazan, J. Grynpas, R. Baker, Is clear speech tailored to counter the effect of specific adverse listening conditions? J. Acoust. Soc. Am. 132 (5) (2012) EL371–EL377.
[29] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, P. Alku, Effect of noise type and level on focus related fundamental frequency changes, in: INTERSPEECH, 2012, pp. 1–4.
[30] Turbosquid, Head, 2005, Turbosquid, New Orleans, LA, http://www.turbosquid.com/3d-models/3d-model-male-head-morph-targets/261694 (last viewed 09.06.16).
[31] R. Becker, P. Hansbo, R. Stenberg, A finite element method for domain decomposition with non-matching grids, ESAIM: Math. Model. Numer. Anal. 37 (2) (2003) 209–225.
[32] 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.
[33] J.B. MacQueen, Some methods for classification and analysis of multivariate observation, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.
[34] P. Alku, Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications, Sadhana 36 (5) (2011) 623–650, http://dx.doi.org/10.1007/s12046-011-0041-5, ISSN 0256-2499.
[35] P. Alku, Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, Speech Commun. 11 (2–3) (1992) 109–118.
[36] W.R. Gardner, B. Rao, Noncausal all-pole modeling of voiced speech, IEEE Trans. Acoust. Speech Signal Process. 5 (1) (1997) 1–10.