Post-processing speech recordings during MRI

Juha Kuortti* and Jarmo Malinen*†
*School of Science, Department of Mathematics and Systems Analysis, Aalto University
†School of Electrical Engineering, Department of Signal Processing and Acoustics, Aalto University

***Abstract*—We discuss post-processing of speech that has been recorded during Magnetic Resonance Imaging (MRI) of the vocal tract area. These speech recordings are contaminated by high levels of acoustic noise from the MRI scanner. Also, the frequency response of the sound signal path is not flat as a result of restrictions on recording instrumentation and arrangements due to MRI technology. The post-processing algorithm for noise reduction is based on adaptive spectral filtering, and it has been designed keeping in mind the requirements of subsequent formant extraction.**

Speech material was used for validation of the post-processing algorithm, consisting of samples of prolonged vowel productions during MRI. The comparison data was recorded in anechoic chamber from the same test subject. Spectral envelopes and formants were computed for the post-processed speech and the comparison data. Artificially noise-contaminated vowel samples (with a known formant structure) were used for validation experiments to determine performance of the algorithm where using true data would be difficult. Resonances computed by an acoustic model and, similarly, those measured from 3D printed vocal tract physical models were used as comparison data as well.

The properties of recording instrumentation or the post-processing algorithm do not explain the observed frequency dependent discrepancy between formant data from experiments during MRI and in anechoic chamber. It is shown that the discrepancy is statistically significant, in particular, where it is largest at 1 kHz and 2 kHz. There is evidence that the reflecting surfaces of the MRI head and neck coil change the speech acoustics which results in "exterior formants" at these frequencies. However, the role of test subject adaptation to noise and constrained space acoustics during an MRI examination cannot be ruled out.

***Index Terms*—Speech, MRI, noise reduction, DSP**

I. 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, e.g., 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. The 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 often appears to be of the same magnitude as the inter-subject variability, 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 non-flat frequency response of the MRI-proof audio measurement system and further challenges related to the constrained space acoustics inside the MRI head and neck coil. In this article, we deal only with the artefacts of the first and the second kind here, leaving the treatment of the constrained space acoustics to future work.

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 introduce, analyse, and validate a 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], [6], [7], [8].

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 a 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. 1 (lower panel), 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. 1. 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 peaks that would skew the true formant

Manuscript received XX.XX.XX Corresponding author: J. XX (email:.

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

content – this may be a serious cause of worry in non-linear 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 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 II, the post-processing algorithm is described in Section III. The validation of the algorithm is carried out in Section IV 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 MRI to similar data recorded in 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 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 V, we apply the post-processing algorithm to speech that has been recorded during an MRI scan as detailed in [2]. The objective is no longer to validate the algorithm rather than draw conclusions about the speech data itself. We again use comparison samples that have been recorded in anechoic chamber. There is a statistically significant ($p > 0.95$) 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. 5 with steps at 1kHz and 2kHz. We hypothesise that it is due to the constrained space acoustics inside the MRI head and neck coils, resulting in exterior formants not related to vocal tract configurations.

II. Speech recording during MR imaging

A. Arrangements

The experimental arrangement has been detailed in [11], [1], [2]. Briefly, a two-channel acoustic sound collector samples the speech and the MRI noise. 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 [58] as it allows for sufficiently rapid static 3D acquisition. Imaging parameters, etc., have been described in [2, Section 3.2].

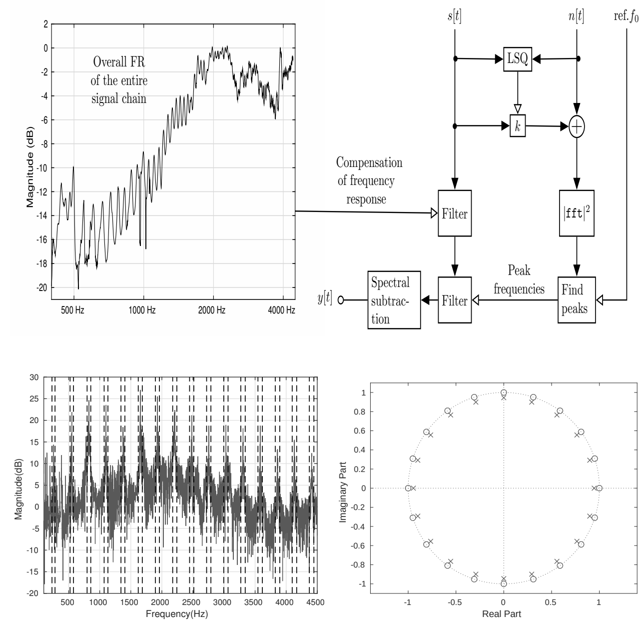

Fig. 1: Upper panel: A block diagram of the post-processing algorithm. Here $s[t]$ and $n[t]$ denote the discretised speech and noise samples at $f_s = 44\,100$ Hz, respectively. The signal $y[t]$ is de-noised speech. Lower panel on the left: Harmonic structure of the MRI noise and stop bands estimated from it. Lower panel on the right: The zero/pole placement in z-plane of the notch filter of degree 20 for removing the frequency $f_s/10$ and its harmonics below the Nyquist frequency $f_s/2$.

B. 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 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 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 anechoic chamber but neither the MRI coil reflections nor the ambient noise were replicated. Compared to MRI experiments, there are no similar restrictions in 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.

III. MRI NOISE CANCELLATION

We treat the measurement signals from speech and acoustic MRI noise $s[t]$ and $n[t]$ for $t \in \{h, 2h, 3h, \ldots\}$ in their digitised form where $h = 1/f_s$, and the sampling frequency $f_s = 44\,100\,\mathrm{Hz}$. The post-processing algorithm for these discrete time signals is outlined in Fig. 1 (upper panel), 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. 1 (upper panel), 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. 1 (lower panel on the left). 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 $-3\,\mathrm{dB}$ 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 [12].

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 similar to [5]. For that reason, the low frequency components of speech are not attenuated as a result of the proximity effect in dipole configurations. Secondly, using notch filters instead of high-order Chebyshev produces sharper removal of unwanted spectral components; see also [13]. These changes improve the audible outcome considerably.

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 ← set of all peaks found in the spectrum.
2:  procedure FINDHARMONICS(P)
3:      while P ≠ ∅ do
4:          p ← max_mag P
5:          P ← P \ p
6:          for q ← P sorted by |loc(p) − loc(P)| do
7:              d ← |loc(p) − loc(q)|
8:              if d < c f_0 then
9:                  continue
10:             if ∃ harmonics with fundamental d then
11:                 F ← F ∪ iircomb(f_s/d)
12:                 P ← P \ {r ∈ P : r = nd, n ∈ 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.

IV. PERFORMANCE ANALYSIS

A. Validation through synthetic signals

The formant extraction from noisy speech can validated using artificially noise contaminated speech where the original formant positions are known. 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[2]. 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 $\approx 0\,\mathrm{dB}$). The post-processing algorithm was then applied to these signals, of which an example is shown in Fig. 2.

The three formants F_1, F_2, and F_3 were extracted from artificially noise contaminated vowels after they had been post-processed as described in Section III. The resulting formant frequencies are within $-0.5 \ldots 0.3$ semitones from those measured from the original pure vowels, except for the outlier F_2[o] where the discrepancy is 1.1 semitones.

Vowel	F_1	F_2	F_3	Vowel	F_1	F_2	F_3
[ɑ]	598	1094	1918	[ɑ]	615	1129	2021
[e]	453	1691	2255	[e]	443	1714	2299
[i]	318	1900	2097	[i]	327	1909	2293
[o]	465	815	2233	[o]	451	858	2088
[u]	410	898	1934	[u]	416	921	2041
[y]	379	1535	2034	[y]	390	1533	2015
[æ]	562	1452	2375	[æ]	559	1476	2319
[œ]	436	1400	2076	[œ]	428	1421	2099

TABLE I: Original formants (left) and formants extracted after the artificial addition of MRI noise and subsequent noise cancellation (right).

The average formant discrepancies of under 2.8 semitones were reported in [2, Table 3] between speech formants and

[2]Throughout this article, the MATLAB function `arburg` is used for producing low-order rational spectral envelopes from which the formants are extracted by locating poles.

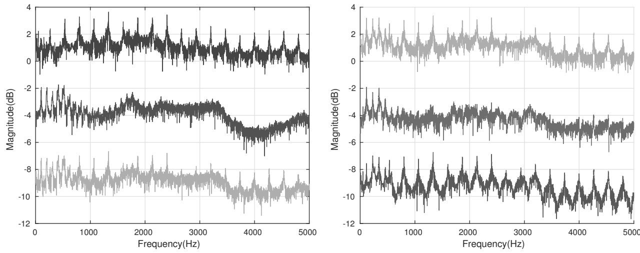

Fig. 2: Illustration of the artificially noise-contaminated vowel signal. On the left, MRI noise (upmost), pure vowel signal (middle), and the synthetic signal as their sum (lowest). On the right, synthetic signal (upmost), signal after post-processing using the proposed algorithm (middle), and the reconstructed noise (lowest).

Helmholtz resonances computed from vocal tract geometries that were obtained by simultaneous MRI. Also, the observations in [14] 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 small error source.

We further observe that the post-processing algorithm described in Section III increases the SNR of the artificially noise-contaminated signals by 9 . . . 14 dB depending on the vowel.

B. 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 of the signal 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.

	[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

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

The spectral tilt data is given in Table II. The roll-off in post-processed speech during MRI is systematically larger than in comparison data (in average by 1.9 dB), the only exception being the vowel [y]. We point out that the two kinds of spectral tilt data in Table II correlate strongly ($R = 0.78$). As can be seen from Fig. 4 (last panel), the difference of the average spectral tilts is quite small. The difference is partly explained by the fact that there was a lot of more attenuating material around the test subject in the MRI scanner, compared to experiments in anechoic chamber.

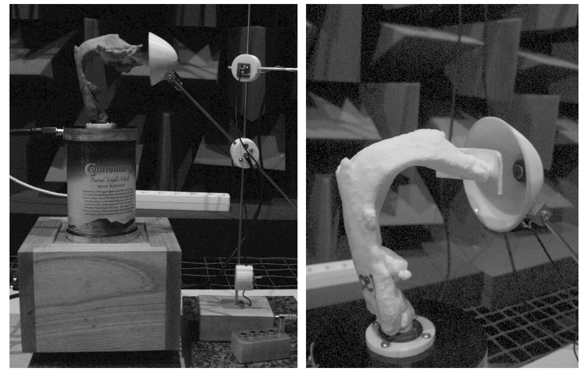

Fig. 3: A detail of the sweep measurement arrangement for 3D printed vocal tract configurations of [ɑ, œ].

C. 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 [15]. Fast prototyping was used to produce physical models in ABS plastic (wall thickness 2 mm) from the STL files. The printed models extend from the glottal position to the lips, and they were coupled to a custom acoustic source (see Fig. 3) whose design resembles the loudspeaker-horn construction shown in [16, Fig. 1]; see also [17].

The acoustic source contains an electret (reference) microphone (⌀ 9 mm, biased at 5 V) at the glottal position, and another similar (signal) microphone was placed near the lips. These two units were picked from a set of 10 units to ensure that their frequency responses between 80 Hz and 10 kHz are practically identical. 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 resonant frequencies between 80 Hz and 7 kHz of the physical models (and reference resonators with known resonant frequencies) were measured using this arrangement.

As can be seen from Fig. 4, 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 an 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. 4 that the spectral tilt (as defined in Section IV-B) 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 of 11 . . . 16 dB/octave whereas the measurements from the physical models were carried out

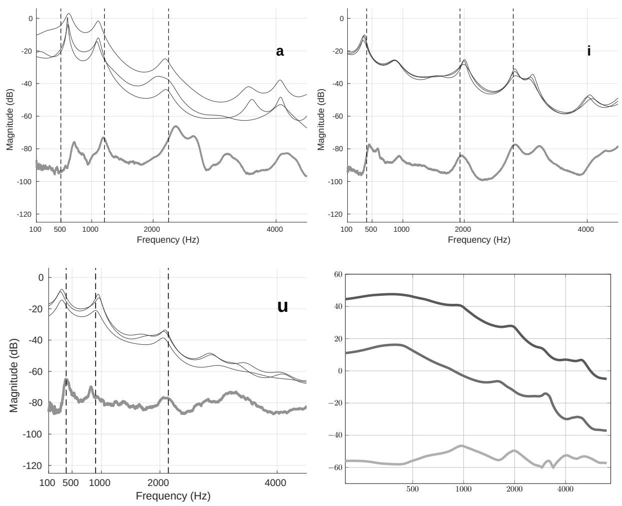

Fig. 4: The first three panels: Spectral envelopes and computationally obtained resonances of [ɑ, i, u]. The upper curves are power spectral densities of speech recorded during an MRI scan. The lower curves are frequency responses measured from the physical models that have been produced from the MR images. The vertical lines indicate the three lowest resonances computed by Webster's model from the same VT geometry using the mouth impedance optimisation process introduced in [9]. The last panel: Averages of spectral envelopes of all Finnish vowels [ɑ, e, i, o, u, y, æ, œ]. Vowels appear in the averages with the same weight. The topmost curve describes speech recorded during the MRI scan, and the center curve recordings in anechoic chamber. The lowest curve is their difference.

keeping the sinusoidal sound pressure constant at the glottal position.

D. Perceptual evaluation

A listening experiment was carried out to evaluate the effect of post-processing on vowel recognition. In the experiment, 12 subjects (of which two were female) listened to 48 recordings of vowel phonation. The recordings consisted of 6 samples of each Finnish vowel in [ɑ, e, i, o, u, æ, œ]; half of the samples were unprocessed recordings from anechoic chamber (24 in total, three for each vowel), while the rest had undergone the MRI noise contamination and de-noising process described in Section IV-A. The duration of each sample was 10 s.

The test subjects were allowed to listen each sample as many times as they wanted. Using a 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 III. 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.

a) 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
[oe]	0	3	0	0	0	0	0	33

b) Artificially MRI noise contaminated samples

	categorised as							
target	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[œ]
[ɑ]	36	0	0	0	0	0	0	0
[e]	0	30	0	0	0	0	0	6
[i]	0	0	36	0	0	0	0	0
[o]	8	0	0	28	0	0	0	0
[u]	0	0	0	15	21	0	0	0
[y]	0	0	0	0	0	27	0	9
[æ]	0	0	0	0	0	0	36	0
[oe]	0	0	0	1	0	0	0	35

TABLE III: 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.

V. Formant extraction from noisy speech

After four validation experiments on the post-processing algorithm described in Section III, 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 sixfold increase of the original sample number improves the statistical analysis given in Table IV. Spectral envelopes of all speech samples are shown in Fig. 6 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 anechoic chamber and 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., [18, Section 10.4]. In case of the vowel formant $F_j[\text{ɑ}]$ for $j = 1, 2, 3$, our null hypothesis is that

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

We try to reject H_0 by showing that its converse H_1 is true with high probability, say $p > 0.95$, in which case the

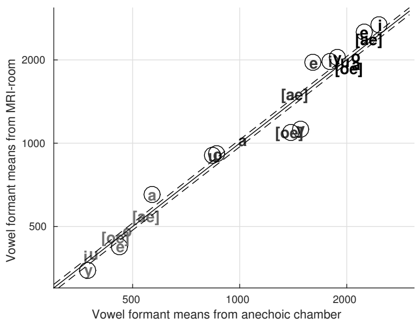

Fig. 5: The means of formants F_1, F_2, F_3 have been extracted from the vowel samples of [ɑ, e, i, o, u, y, æ, œ] recorded during MRI. They are plotted against the comparable data recorded in anechoic chamber from the same test subject. The dashed lines describe the error bounds of ±0.5 semitones due to formant extraction from post-processed noisy speech; see Section IV-A. Where the formant discrepancy is statistically significant at $p \geq 0.95$, the vowel has been encircled; see Table IV.

experiment indicates that the formant extraction from the two data sources is not consistent. The results of the experiments are given in Table IV where the p-values are given. We conclude that H_0 gets typically rejected for F_2 in all vowels except [ɑ, o, æ] and for all formants in vowels [e, i].

The formant means from post-processed speech during MRI are plotted in Fig. 5 against their counterparts recorded in 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 IV-A. We conclude that (at least) 12 of the discrepancies shown in Fig. 5 reflect actual differences of the speech data recorded in MRI laboratory, compared to similar data from anechoic chamber.

It is worth observing that the formant discrepancy in Fig. 5 shows a peculiar staircase pattern where two plateaus appear near 1 kHz and 2 kHz. More precisely, we observe that in samples recorded during 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 Fig. 6 in most of spectral envelopes of signals recorded during MRI; notable exceptions are the vowels [ɑ,u,o] where $F_2 \approx 1$ kHz would conceal any extra peak. The extra peaks can be seen in Fig. 4 (last panel) where the spectral envelopes of all vowel recordings in MRI laboratory (in anechoic chamber, respectively) have been averaged. The acoustic resonances of the exterior space are expected to show up in the averaged envelopes, and there are, indeed, spectral peaks near 1 kHz and 2 kHz in recordings within MRI head and neck coils that do not appear in the corresponding averages of the comparison data. It has been excluded by frequency response measurements and ensuing compensation that these peaks could be an artefact of the speech recording instrumentation. A similar staircase pattern to Fig. 5 appears in [19, Chapter 5] where formant and resonance pairs have been plotted against each other. The vocal tract resonances in [19] have been computed by Helmholtz equation from MRI data without exterior space modelling, and the formants have extracted from recordings during MRI as explained in [2, Section 5].

	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[œ]
F_1	**0.99**	**0.98**	0.84	0.14	0.70	**0.95**	0.25	0.07
F_2	0.21	**0.99**	**0.99**	**0.99**	**0.98**	**0.99**	0.81	**0.98**
F_3	0.82	**0.99**	**0.99**	0.60	0.17	**0.99**	0.61	0.75

TABLE IV: The p-values computed with Smith-Satterwaith procedure for distributions with unequal variances. Formant samples that reject the null hypothesis H_0 at $p > 0.95$ are written in bold.

The statistically significant discrepancy in Fig. 5 is expected to be a combination of three different sources: (i) "frequency pulling" of the vocal tract resonances by the adjacent exterior space resonances, caused by reflections from test subject's face and MRI head and neck coil surfaces; (ii) Lombard speech due to the acoustic noise during MRI (see [20], [21]); and (iii) active adaptation of the test subject to the constrained space acoustics inside the MRI head and neck coil.

VI. Conclusions

When trying to match a computational model of speech to true speech biophysics, some sort of paired data is necessarily required. 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 the 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. Thus, the model can only be as reliable as its validation data.

A post-processing algorithm was proposed for removing acoustic noise from speech that has been recorded during 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 III 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 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. This is an adequate level of performance for

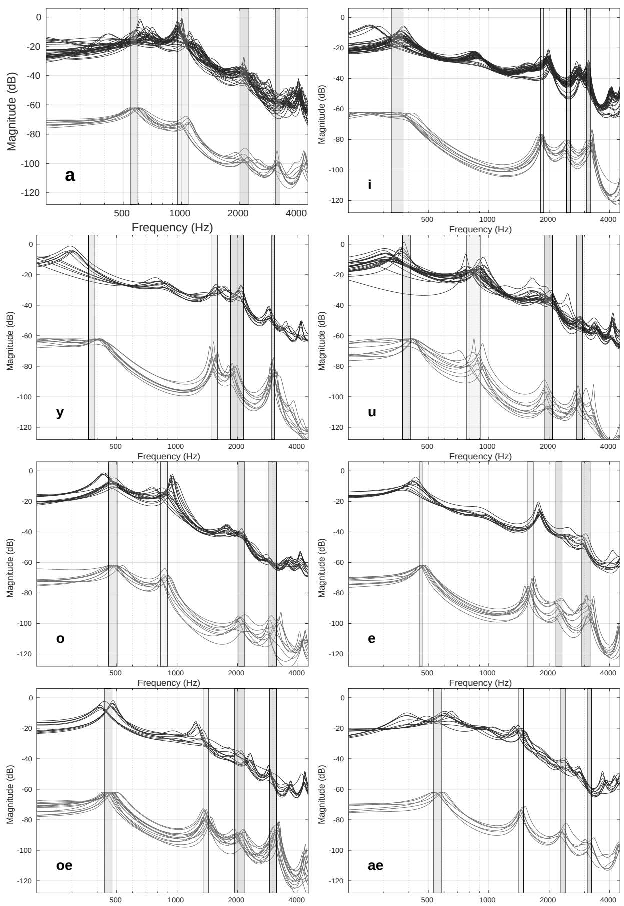

Fig. 6: Spectral envelopes of all vowel samples in the dataset. In each panel, the upper curves represent post-processed signals recorded during MRI experiments. The lower curves are similar envelopes without any post-processing of signals, obtained from the same test subject in the anechoic chamber. These two families of curves are comparable to curves given in [2, Figs. 7–8]. The vertical bars are error intervals for formants $F_1, \ldots, F_4$ extracted from the recordings in the anechoic chamber.

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 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 at 1 kHz and 2 kHz. At these frequencies, the formant discrepancy is several magnitudes larger than the formant estimation error from post-processed speech, and the deviations are statistically significant (Student's t-test with $p > 0.95$). There is 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.

The notch filtering adds a large number of transmission zeros to processed signals which causes the phase response of the algorithm to be non-linear. 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 [23], [24]. To produce signals with linear phase response, one should use, e.g., non-causal spectral filtering (see [22]) instead of notch filters.

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

Acknowledgements

The authors wish to thank many collegues for consultation and facilities required in data acquisition: 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 (Mr. A. Mohite).

The authors have received financial support from Instrumentarium Science Foundation and Magnus Ehrnrooth Foundation.

References

[1] D. Aalto, O. Aaltonen, R.-P. Happonen, J. Malinen, P. Palo, R. Parkkola, J. Saunavaara, and M. Vainio, "Recording speech sound and articulation in MRI," in *Proceedings of BIODEVICES*, 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, and M. Vainio, "Large scale data acquisition of simultaneous MRI and speech," *Applied Acoustics*, vol. 83, no. 1, pp. 64–75, 2014.

[3] A. Aalto, T. Murtola, J. Malinen, D. Aalto, and M. Vainio, "Modal locking between vocal fold and vocal tract oscillations: Simulations in time domain," arXiv:1506.01395, 2015, submitted.

[4] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.

[5] E. Bresch, K. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.

[6] J. Přibil, J. Horáček, and P. Horák, "Two methods of mechanical noise reduction of recorded speech during phonation in an MRI device," *Measurement science review*, vol. 11, no. 3, pp. 92–99, 2011.

[7] J. Přibil, A. Přibilová, and I. Frollo, "Analysis of spectral properties of acoustic noise produced during magnetic resonance imaging," *Applied Acoustics*, vol. 73, no. 8, pp. 687–697, 2012.

[8] J. Inouye, S. Blemker, and D. Inouye, "Towards undistorted and noise-free speech in an MRI scanner: correlation subtraction followed by spectral noise gating," *Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1019–1022, 2014.

[9] J. Kuortti, J. Kivi, J. Malinen, and 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, and M. Vainio, "Articulating Finnish vowels: Results from MRI and sound data," *Linguistica Uralica*, vol. 48, no. 3, pp. 194–199, 2012.

[11] J. Palo, "A wave equation model for vowels: Measurements for validation." Licentiate Thesis, Aalto University School of Science, Department of Mathematics and Systems Analysis, 2011.

[12] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113 – 120, 1979.

[13] X. Shou, X. Chen, J. Derakhsan, T. Eagan, T. Baig, S. Shvartsman, J. Duerk, and R. Brown, “The suppression of selected acoustic frequencies in MRI,” *Applied Acoustics*, vol. 71, pp. 191–200, 2010.

[14] D. Aalto, J. Malinen, M. Vainio, J. Saunavaara, and 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.

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

[16] D. Tze Wei Chu, K. Li, J. Epps, J. Smith, and J. Wolfe, “Experimental evaluation of inverse filtering using physical systems with known glottal flow and tract characteristics,” *Journal of the Acoustical Society of America*, vol. 133, no. 5, 2013.

[17] H. Takemoto, P. Mokhtari, and T. Kitamura, “Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method,” *Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3724–3738, 2010.

[18] J. Milton and J. Arnold, *Introduction to probability and statistics*, 4th ed. McGraw-Hill, 2003.

[19] A. Kivelä, “Acoustics of the vocal tract: MR image segmentation for modelling,” Master’s thesis, Aalto University School of Science, Department of Mathematics and Systems Analysis, 2015.

[20] V. Hazan, J. Grynpas, and R. Baker, “Is clear speech tailored to counter the effect of specific adverse listening conditions?” *Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. EL371–EL377, 2012.

[21] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, and P. Alku, “Effect of noise type and level on focus related fundamental frequency changes,” in *INTERSPEECH*, 2012, pp. 1–4.

[22] W. R. Gardner and B. Rao, “Noncausal all-pole modeling of voiced speech,” *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 1, pp. 1–10, 1997.

[23] 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*, vol. 36, no. 5, pp. 623–650, 2011.

[24] ——, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992.