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
Available online 4 August 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/© 2017 Elsevier Ltd. All rights reserved.

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 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.

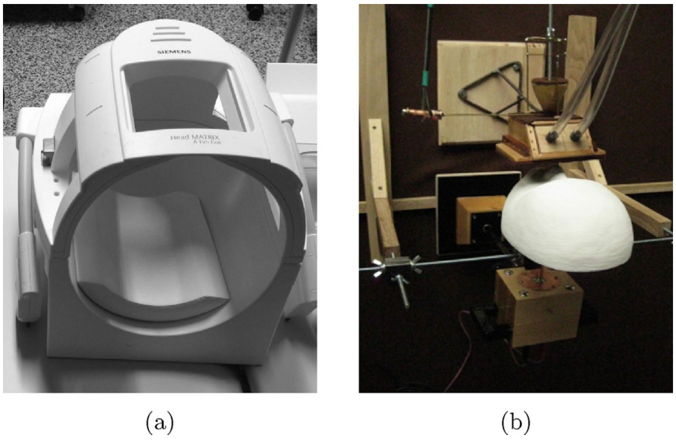

Fig. 1. Panel (a): The MRI head coil of Siemens Magnetom Avanto 1.5T scanner. The two-channel acoustic sound collector fits exactly the opening on the top. Panel (b): The sound collector positioned above a head model similarly as in the MRI experiments. The noise sample is acquired using a horn on the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

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

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