LETTERS TO THE EDITOR

This Letters section is for publishing (a) brief acoustical research or applied acoustical reports, (b) comments on articles or letters previously published in this Journal, and (c) a reply by the article author to criticism by the Letter author in (b). Extensive reports should be submitted as articles, not in a letter series. Letters are peer-reviewed on the same basis as articles, but usually require less review time before acceptance. Letters cannot exceed four printed pages (approximately 3000–4000 words) including figures, tables, references, and a required abstract of about 100 words.

Towards undistorted and noise-free speech in an MRI scanner: Correlation subtraction followed by spectral noise gating (L)

Joshua M. Inouye and Silvia S. Blemker[a)]
Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia 22902

David I. Inouye
Department of Computer Science, University of Texas at Austin, Austin, Texas 78712

(Received 17 July 2013; revised 22 January 2014; accepted 24 January 2014)

Noise cancellation in an MRI environment is difficult due to the high noise levels that are in the spectral range of human speech. This paper describes a two-step method to cancel MRI noise that combines operations in both the time domain (correlation subtraction) and the frequency domain (spectral noise gating). The resulting filtered recording has a noise power suppression of over 100 dB, a significant improvement over previously described techniques on MRI noise cancellation. The distortion is lower and the noise suppression higher than using spectral noise gating in isolation. Implementation of this method will aid in detailed studies of speech in relation to vocal tract and velopharyngeal function.
 [http://dx.doi.org/10.1121/1.4864482]

PACS number(s): 43.70.Jt, 43.70.Fq, 43.70.Aj [SSN] Pages: 1019–1022

I. INTRODUCTION

Speech production studies have recently combined dynamic MRI sequences with speech recordings during the scan to study vocal tract shaping and velopharyngeal (i.e., soft palate) function.[1–6] These studies are designed to provide basic information on the mechanics of speech production,[2–6] investigate the acoustic characteristics of speech,[1] elucidate strategies for speech therapy,[2] and optimize surgeries associated with speech production such as cleft palate repair surgeries.[3]

The high levels of MRI noise (over 110 dB during most scans[7]) and the fact that the majority of the MRI noise power lies in the frequency range of human speech (Fig. 1), makes simple bandpass filters ineffective. However, the periodicity (if present) and predictability of the gradient noise—produced by the high rate of on-off switching of the gradient magnetic coils—can be exploited for effective noise suppression. Because the signal to noise ratio is so low, precise acoustical analyses are difficult if not impossible without powerful noise suppression.

II. RELATED WORK

The principal technique currently used for noise cancellation in MRI scanner environments is adaptive filtering, first published by Bresch *et al.*[6] This method requires two synchronized signals: one signal with speech and one noise signal to use as a reference for the adaptive filter. The noise signal can be recorded by a second microphone outside the MRI or generated by a scanner-specific model[8] of the noise. Adaptive filtering has been used successfully in both academic research studies[4,5,9] and advanced commercial noise cancellation systems for MRI environments.[10] However, one main disadvantage of this technique is that it introduces an echoing artifact into the recordings.

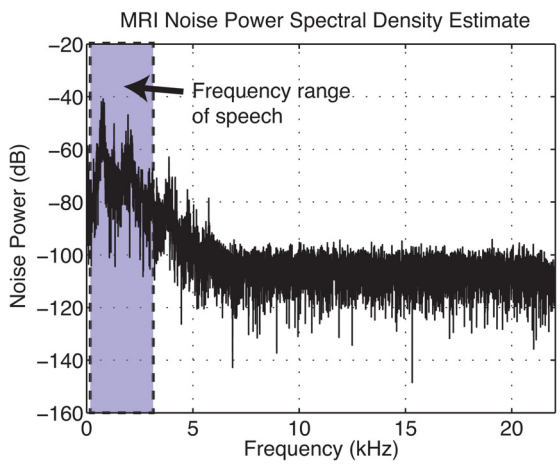

FIG. 1. (Color online) MRI noise power lies in the frequency range of speech, making simple bandpass filters ineffective.

[a)]Author to whom correspondence should be addressed. Electronic mail: ssblemker@virginia.edu

Another technique that has been used for MRI gradient noise cancellation is correlation subtraction,[1] which only requires one microphone. A recording of the noise is taken without speech and then another is taken with speech. Since the noise is periodic, cross-correlation of the audio signals and subsequent subtraction can remove a significant portion of the noise. No echoing artifact is present in this technique, but the noise suppression in NessAiver *et al.*[1] is less than that of adaptive filtering in Bresch *et al.*[6]

The general technique of spectral subtraction has been developed for noise reduction in speech recordings.[11] It cancels noise by operating in the frequency domain. Spectral noise gating[12] is similar to spectral subtraction. In spectral noise gating, the frequencies of the speech recordings pass through a spectral "gate" if their power is higher than the power of the corresponding frequencies from a noise-only period. The other frequencies are suppressed by a specific level. To the best of our knowledge, spectral noise gating has never been applied to noise cancellation in an MRI environment.

This paper combines the method of correlation subtraction with spectral noise gating—operating in the time and frequency domains, respectively. We show that using either method alone produces inferior results to the combination of the methods. Unlike adaptive filtering, our method only requires one microphone and does not produce the echoing artifact associated with adaptive filtering. In addition, our two-step method produces significantly better noise suppression than any previous method (over 100 dB vs 28 dB for adaptive filtering).

III. METHODS

A. Audio acquisition and signal synthesis

A fiber optic microphone (FOM1-MR-30 m, Micro Optics Technologies, Middleton, WI) was used to record utterances from subjects while they were undergoing dynamic scanning. The scanning sequence[13] is a spiral steady-state free precession sequence that produced a temporal resolution of 21.4 frames per second and spatial resolution of 1.2×1.2 mm^2 via a combined spatial and temporal parallel reconstruction method. The frames were acquired at this resolution for each of two image slices: one mid-sagittal and one oblique-coronal, each with a 150 mm field of view. 120 frames from both slices were acquired during our recordings on a Siemens Avanto 1.5 T scanner, and the total scan time was approximately 6.6 s.

We recorded six phrases (e.g., "multiscale muscle mechanics lab") from a subject during one scan and noise samples without speech at a sampling rate of 44.1 kHz with a resolution of 24 bits.

In order to systematically investigate the temporal and spectral effects of the procedures on noise power suppression and signal distortion, we synthesized a "ground-truth" signal composed of sine waves equally spaced by 50 Hz) in the voice frequency range (i.e., 300, 350, 400,…, 3000 Hz) of equal amplitude and random phases. The power of the synthesized signal was normalized to be equal to the power of the noise signal. We added a pure noise signal to this synthesized signal to quantify noise suppression and distortion.

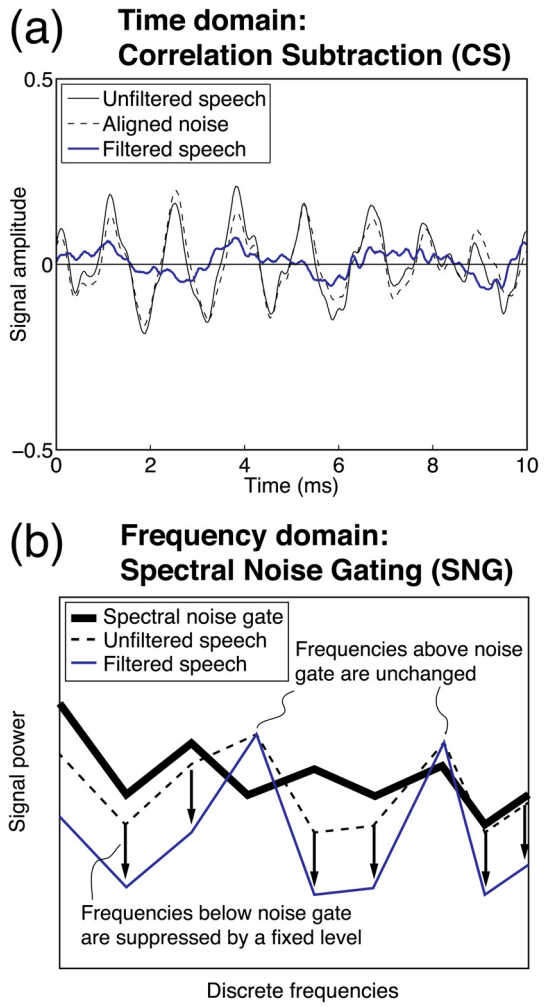

FIG. 2. (Color online) CS is performed in the time domain, while SNG is performed in the frequency domain. (a) CS filters speech by subtracting an aligned noise signal from an unfiltered speech signal. Actual data are shown. (b) SNG creates a gate based on a noise-only sample of the unfiltered speech and suppresses frequencies that are below the gate level. Data are hypothetical.

B. Noise cancellation

The basic concepts behind the two main methods—correlation subtraction (CS) and spectral noise gating (SNG)—are depicted in Fig. 2.

The CS technique is performed by finding the delay d^* that gives the maximum discrete cross-correlation r defined as

$$d^* = \arg\max_d r[d] = \arg\max_d \sum_{i=-\infty}^{\infty} s[d]n[d+i], \qquad (1)$$

where s is the speech signal and n is the noise signal. The time delay d^* is the delay at which the entire speech and noise recordings are most aligned in the time domain. After shifting the noise recording by d^* samples, the noise is subtracted from the speech recording—leaving a signal filtered in the time domain.

The SNG technique is conceptually composed of several steps. The first step is to create a spectral "fingerprint" of a noise-only portion of a speech recording using a Fourier transformation (FT). This fingerprint is used as a "gate" for

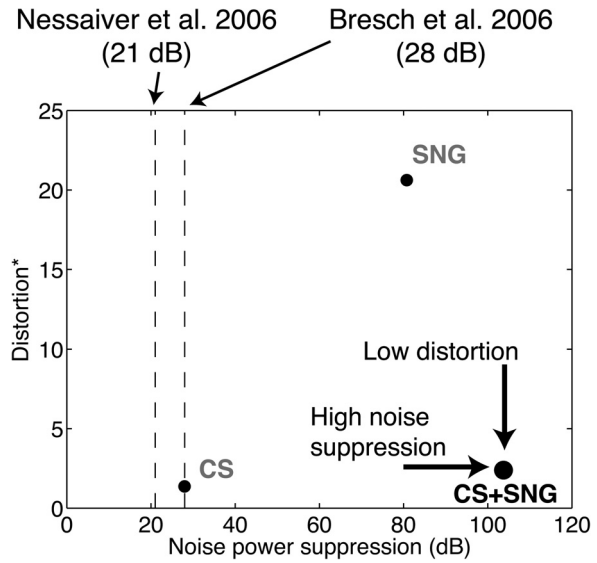

FIG. 3. CS + SNG has the highest noise power suppression and much less distortion than SNG alone. The two previous studies did not report distortion measures. Data are from synthesized signal. *Distortion is measured as the L2 norm of the difference between the clean signal and filtered signals in the time domain and is not calculated in these earlier studies.

the rest of the speech recording. The frequencies in the speech segments that are above the gate are "let through" while the frequencies that are below the gate are "gated" and suppressed by a fixed level. An inverse FT then produces the filtered signal. In our tests, we used the software package SOX,[12] the command line version of AUDACITY,[14] to implement SNG. In this implementation, the FTs of both the noise sample and the speech signal employ windowed time frames, and frequency smoothing is applied followed by time smoothing after the gating process and before the inverse FT. See the software documentation[14] for further details. The inputs to the SOX algorithm are the speech recording, the noise-only recording, and a noise reduction coefficient between 0 and 1 (we step through the coefficients in increments of 0.05 for each recording and choose the coefficient with the maximal noise suppression). The optimal coefficient ranged from approximately 0.3 to 0.5 in our experiments.

We quantified noise power suppression and distortion for the synthesized signal for CS, SNG, and CS followed by SNG (CS + SNG). Noise power suppression was determined by estimating the SNR increase using the method in Bresch *et al.*[6] We quantify distortion by calculating the L^2 norm of the difference between the "ground-truth" signal and the filtered version of the signal after adding noise.

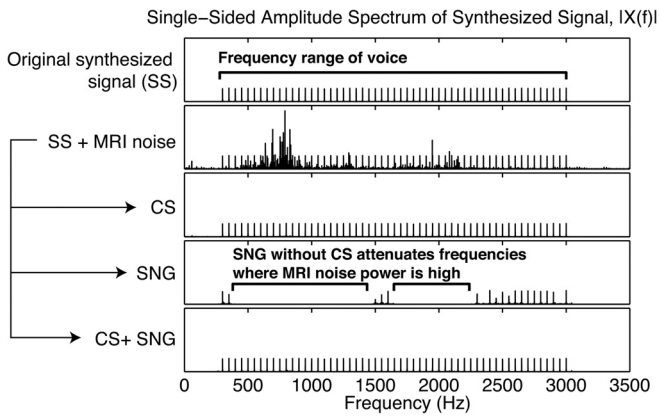

FIG. 4. Using CS as a preprocessing step before SNG retains frequency components in the voice range better than SNG alone, reducing distortion.

IV. RESULTS

For the synthetic data, both SNG and CS + SNG have significantly higher noise suppression than previous studies (Fig. 3).[1,6] Furthermore, distortion is much lower when using CS as a preprocessing step for SNG. CS followed by SNG outperforms either method on its own.

The distortion with SNG alone mainly occurs at the frequencies in which MRI noise power is high (Fig. 4). CS removes most of the MRI noise power in these frequencies and then SNG removes more noise by itself.

The waveforms and spectrograms of the original recording and the various filtering techniques for the phrase ("multiscale muscle mechanics lab") are shown in Fig. 5. The residual noise floor levels of SNG and CS + SNG are similar, but there is increased signal retention—and therefore less distortion—with CS + SNG. The noise power

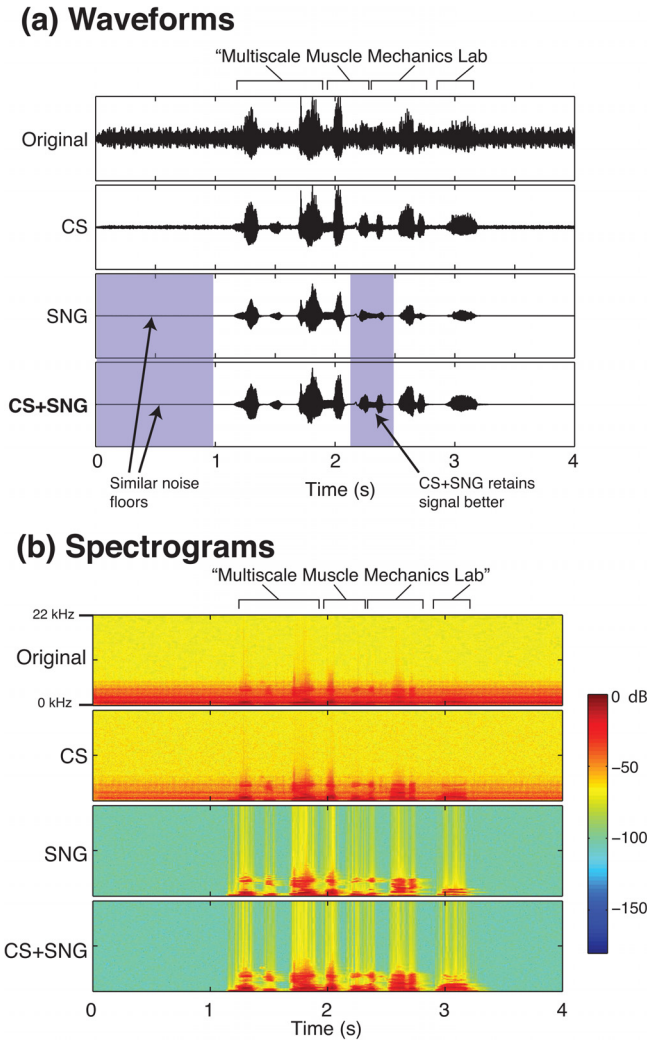

FIG. 5. (Color online) Using CS as a preprocessing step before SNG retains the speech signal better while maintaining a similar noise floor.

suppression for this phrase was 100 dB. The noise power suppression for the other five recorded phrases ranged between 100 and 104 dB.

Audio clips and noise cancellation code are archived along with this publication and available online.[15]

V. DISCUSSION

This two-step method of CS followed by SNG enables significantly greater suppression of noise from audio recordings acquired during an MRI scan than previously published methods while avoiding any echo artifacts. Because the CS method is a good preprocessing step for SNG, distortion is much less and noise cancellation higher than with SNG alone.

While this method works well for the tested periodic sequence, future research could test whether general periodic sequences can be filtered effectively with this method. The noise from aperiodic sequences that are repeatable should be able to be reasonably cancelled with the CS method although the SNG step may not work as well. As this method is tested on other MRI pulse sequences and developed for other noise cancellation purposes, phonetic classification and listening tests could quantify the effectiveness of this method for voice communication and speech recognition applications.

VI. CONCLUSION

This method provides a significant improvement for research on speech recordings produced during MRI scans and future research can now utilize and extend this technique for other advanced noise cancellation applications.

ACKNOWLEDGMENTS

This work was supported by a grant from The Hartwell Foundation. D.I.I. was supported by the NSF Graduate Research Fellowship under Grant No. DGE-1110007. The authors would like to acknowledge useful discussions with Ali Dhanaliwala and Geoffrey Handsfield and the assistance of John Christopher, Katie Pelland, Lucas Thomeer, Craig Meyer, and Xue Feng with MRI data acquisition.

[1]M. S. NessAiver, M. Stone, V. Parthasarathy, Y. Kahana, and A. Paritsky, "Recording high quality speech during tagged cine-MRI studies using a fiber optic microphone," J. Magn. Reson. Imag. **23**, 92–97 (2006).
[2]J. M. Inouye, X. Feng, C. H. Meyer, K. Y. Lin, K. Borowitz, T. Altes, T. Kovach, W. El-Nahal, K. Pelland, and S. S. Blemker, "New technique to assess velopharyngeal function with multi-planar real-time MRI," in *Proceedings of the 12th International Congress on Cleft Lip/Palate and Related Craniofacial Anomalies* (2013).
[3]J. M. Inouye, C. M. Pelland, N. M. Fiorentino, W. G. El-Nahal, K. Y. Lin, K. C. Borowitz, and S. S. Blemker, "A 3D model of the human soft palate muscle function during speech: Implications for surgical repair of cleft palates," in *Proceedings of the 11th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering* (2013).
[4]A. Niebergall, S. Zhang, E. Kunay, G. Keydana, M. Job, M. Uecker, and J. Frahm, "Real-time MRI of speaking at a resolution of 33 ms: Undersampled radial flash with nonlinear inverse reconstruction," Magn. Reson. Med. **69**(2), 477–485 (2013).
[5]A. D. Scott, R. Boubertakh, M. J. Birch, and M. E. Miquel, "Adaptive averaging applied to dynamic imaging of the soft palate," Magn. Reson. Med. **70**(3), 865–874 (2013).
[6]E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791 (2006).
[7]S. A. Counter, A. Olofsson, H. Grahn, and E. Borg, "MRI acoustic noise: sound pressure and frequency analysis," J. Magn. Reson. Imag. **7**, 606–611 (1997).
[8]Z. Wu, Y. Kim, M. C. Khoo, and K. S. Nayak, "Evaluation of an independent linear model for acoustic noise on a conventional MRI scanner and implications for acoustic noise reduction," Magn. Reson. Med. (published online).
[9]E. Bresch, Y. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," IEEE Sign. Process. Mag. **25**, 123–132 (2008).
[10]FOMRI-III Noise Cancellation System, http://www.optoacoustics.com/medical/fomri-iii (Last viewed February 5, 2014).
[11]S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Sign. Process. **27**, 113–120 (1979).
[12]sox-Sound eXchange, http://sox.sourceforge.net (Last viewed February 5, 2014).
[13]X. Feng, J. M. Inouye, S. S. Blemker, K. Y. Lin, K. C. Borowitz, T. Altes, T. Kovach, W. El-Nahal, C. M. Pelland, and C. H. Meyer, "Assessment of velopharyngeal function with multi-planar high-resolution real-time spiral dynamic MRI," in *Proceedings of the 21st Annual Meeting of the International Society for Magnetic Resonance in Medicine, Program 1228* (2013).
[14]audacity software, http://audacity.sourceforge.net (Last viewed February 5, 2014).
[15]http://bme.virginia.edu/muscle/cleftpalate/noisecancellation.html (Last viewed February 5, 2014).