A two-step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis

Colin Vaz, Vikram Ramanarayanan, and Shrikanth Narayanan

Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA – 90089
<cvaz,vramanar>@usc.edu, shri@sipi.usc.edu

Abstract

We present a method for speech enhancement of data collected in extremely noisy environments, such as those found during magnetic resonance imaging (MRI) scans. We propose a two-step algorithm to perform this noise suppression. First, we use probabilistic latent component analysis to learn dictionaries of the noise and speech+noise portions of the data and use these to factor the noisy spectrum into estimated speech and noise components. Second, we apply a wavelet packet analysis in conjunction with a wavelet threshold that minimizes the KL divergence between the estimated speech and noise to achieve further noise suppression. Based on both objective and subjective assessments, we find that our algorithm significantly outperforms traditional techniques such as nLMS, while not requiring prior knowledge or periodicity of the noise waveforms that current state-of-the-art algorithms require.

Index Terms: rtMRI, noise suppression, wavelets, pLCA, dictionary learning.

1. Introduction

Speech science researchers use a variety of methods to study articulation and the associated acoustic details of speech production. These include Electromagnetic Articulography [1] and x-ray microbeam [2] methods that track the movement of articulators while subjects speak into a microphone. Data from these methods offer excellent temporal details of speech production. Such methods, however, are invasive and do not offer a full view of the vocal tract. On the other hand, methods using real-time MRI (rtMRI) offer a non-invasive method for imaging the vocal tract, affording access to more structural details [3]. Unfortunately, MRI scanners produce high-energy broadband noise that corrupts the speech recording. This affects the ability to analyze the speech acoustics resulting from the articulation and requires additional schemes to improve the audio quality.

The Least Mean Squares (LMS) algorithm is a popular technique for signal denoising. The algorithm estimates the filter weights of an unknown system by minimizing the mean square error between the denoised signal and a reference signal. This approach removes noise from the noisy signal very well, but it severely degrades the quality of the recovered speech. Bresch et al. proposed a variant to the LMS algorithm in [4] to remove MRI noise from noisy recordings. This method uses knowledge of the MRI pulse sequence to design an artificial reference "noise" signal that can be used in place of a recorded noise reference. We found that this method outperforms LMS in denoising speech corrupted with noise from certain types of pulse sequences. Unfortunately, it performs rather poorly when the noise frequencies are spaced closely together in the frequency domain. Furthermore, the algorithm creates a reverberant artifact in the denoised signal, which makes speech analysis challenging. The LMS formulation assumes additive noise, so these algorithms may not perform well in the presence of convolutive noise in the signal, which we encounter during MRI scans.

Source separation techniques provide a way to separate the speech and noise. Duan et al. proposed a probabilistic component analysis (PLCA) algorithm in [5]. This algorithm learns the dictionaries and their associated time activation weights for the speech and noise, thus separating the speech from noise. In recent decades, wavelets have been used for denoising speech and images [6]. Discrete wavelet transforms, wavelet packet analysis, and lifting have been developed to aid signal denoising. Both PLCA and wavelet analysis are useful for removing convolutive noise from signals because there is no underlying assumption of additive noise. We propose an algorithm that takes advantage of source separation and wavelet analysis to denoise speech recorded in an MRI scanner.

This paper is organized as follows. Section 2 discusses properties of MRI noise. In Section 3, we describe the method we used to perform denoising. Section 4 discusses the results of our method on data acquired from MRI scans and artificially-created noisy speech. Finally, we state our conclusions and future work in Section 5.

The authors would like to acknowledge the support of NIH Grant DC007124.

2. MRI Noise

A primary source of MRI noise arises from Lorentz forces acting on receiver coils in the body of an MRI scanner. These forces cause vibrations of the coils, which impact against their mountings. The result is a high-energy broadband noise that can reach as high as 115 dBA [7]. The noise corrupts the speech recording, making it hard to listen to the speaker, and can obscure important details in speech.

MRI pulse sequences typically used in rtMRI produce periodic noise. The fundamental frequency of this noise, i.e., the closest spacing between two adjacent noise frequencies in the frequency spectrum, is given by:

$$f_0 = \frac{1}{\text{repetition time} \times \text{number of interleaves}}\ \text{Hz} \qquad (1)$$

The repetition time and number of interleaves are scanning parameters set by the MRI operator. Choice of these parameters inform the spatial and temporal resolution of the reconstructed image sequence, as well as the spectral characteristics of the generated noise. Importantly, the periodicity of the noise allows us to design effective denoising algorithms for time-synchronized audio collected during rtMRI scans. For instance, the algorithm proposed by Bresch et al. [4] relies on knowing f_0 to create an artificial “noise” signal which can then be used as a reference signal by standard adaptive noise cancellation algorithms.

However, a few pulse sequences do not exhibit this exact periodic structure. In addition, there are other useful sequences that are either periodic with an extremely large period, resulting in very closely-spaced noise frequencies in the spectrum (i.e., f_0 is very small), or are periodic with discontinuities that can introduce artifacts in the spectrum. To handle these cases, it is essential that denoising algorithms do not rely on periodicity. One example of such sequences which we will consider in this article is the Golden Ratio (GR) sequence [8], which allows for retrospective and flexible selection of temporal resolution of the reconstructed image sequences (typical rtMRI protocols do not allow this desirable property).

3. Denoising Algorithm

We propose a denoising algorithm that uses PLCA and wavelet packet analysis. A noisy recording is given to PLCA, which separates the signal into estimated speech and noise components. Then, the estimated speech is passed to a wavelet packet algorithm for further noise removal. The result of the wavelet packet algorithm is a denoised speech recording. Figure 1 shows the spectrograms of the signal at each stage of the algorithm. The following subsections describe PLCA and wavelet packet analysis in greater detail.

3.1. Step 1: PLCA

PLCA uses non-negative matrix factorization (NMF) to factor a spectrogram of the noisy speech into noise and speech dictionaries and their corresponding time activation weights. We first train the algorithm with the MRI noise to learn a noise dictionary and its time activation weights. Once learned, the noise dictionary stays fixed for the duration of the PLCA algorithm. We obtain the noise-only recording from the beginning 1 second of the noisy speech recording before the speaker speaks (it is usually the case that the speaker speaks at least 1 second after the start of the recording).

After training on the noise, we give PLCA the noisy speech spectrogram for source separation. The algorithm takes each frame of the spectrogram and computes the KL divergence between the spectrogram frame and the current estimate of the noise spectrum. If the KL divergence is low, then it updates the time activation weights of the noise. If the KL divergence is high, then it updates the speech dictionary and the time activation weights for the speech and noise. PLCA uses the EM algorithm to update the speech dictionary.

After the algorithm processes all the spectrogram frames, it returns an estimate of the speech and noise. The algorithm performs well at removing noise in silence regions and suppressing some of the noise in speech regions. To remove the residual noise in the speech estimate, we turn to wavelet packet analysis. Nonetheless, PLCA removes enough noise to make wavelet packet analysis a viable option for denoising; performing wavelet packet analysis on the original noisy speech recording does not work well because the energy of the MRI noise is too high compared to the energy of the speech.

3.2. Step 2: Wavelet Packet Analysis

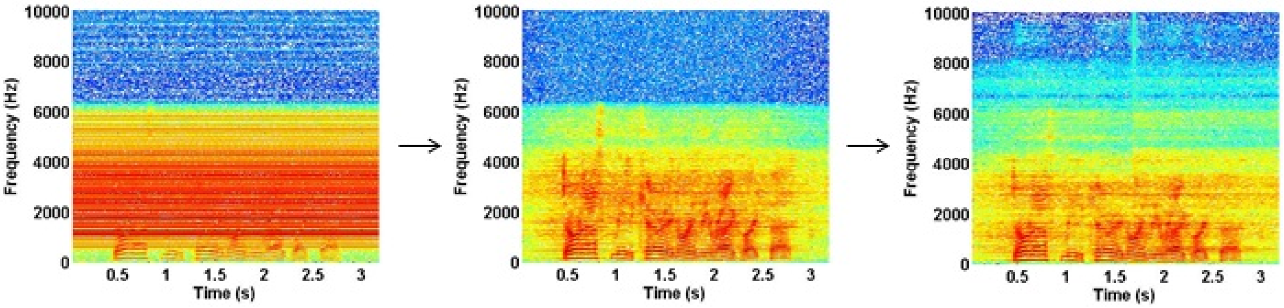
Wavelet packet analysis iteratively decomposes a signal into lowpass and highpass bands using a quadrature mirror filter (QMF) to produce different levels of frequency resolution. We pass the estimated speech from PLCA into a D-level wavelet packet, which yields wavelet coefficients in 2^D subbands.

We threshold the wavelet coefficients to remove the noise. Tabibian et al. proposed a threshold in [9] that minimizes the symmetric KL divergence between the noisy speech coefficients and noise coefficients in the range of $-\lambda$ to λ, where λ is the threshold value. With this formulation, they solved for the threshold to get:

$$\lambda = \frac{\hat{\sigma}^2_{N_k}}{\xi_k}\sqrt{2\left(\xi_k + \xi_k^2\right)\ln\left(\sqrt{1+\frac{1}{\xi_k}}\right)} \qquad (2)$$

where

$$\xi_k = \frac{\hat{\sigma}^2_{X_k}}{\hat{\sigma}^2_{N_k}} \qquad (3)$$



(a) Noisy speech (b) Estimated speech (c) Denoised speech

Figure 1: Spectrograms of the TIMIT sentence "Don't ask me to carry an oily rag like that" spoken by a male. PLCA processes the recording from the MRI scanner (a) to produce a speech estimate (b). Wavelet analysis subsequently removes residual noise in the estimated speech to produce the denoised speech (c).

Here, $\hat{\sigma}^2_{N_k}$ is the estimated variance of the noise coefficients in subband k of level D and $\hat{\sigma}^2_{X_k}$ is the estimated variance of the noisy signal coefficients in subband k of level D, $k = 1, 2, \ldots, 2^D$. To compute the threshold, we need an estimate of the noise. If the MRI noise is periodic, we can estimate the noise with

$$v[n] = \sum_k \alpha_k \cos(2\pi f_0 k n) \tag{4}$$

where f_0 is calculated using Equation 1 and α_k is a scalar that shapes the spectrum of $v[n]$ to match the spectral shape of the MRI noise. For non-periodic MRI noise, we can estimate the noise from the beginning 1 second of the estimated noise calculated by PLCA. This gives us the flexibility to denoise speech corrupted by non-periodic MRI noise. Since the noise in our experiments is periodic, we use $v[n]$ for the noise estimate because it performs marginally better than estimating the noise from PLCA's noise estimate. Once we calculate the threshold, we soft threshold the wavelet coefficients in each subband and reconstruct the denoised signal from the thresholded coefficients.

Soon et al. reported very little difference in the SNR of the denoised signal when using different wavelets, even accounting for varying SNR of the noisy signal and male/female speakers [10]. They evaluated denoising performance using biorthogonal, Daubechies, Coiflet, and Symmlet wavelets with different wavelet orders. Our experiments corroborated their findings; we found very little difference in the quality of the denoised signal, both quantitatively and perceptually, when using different wavelets. Thus, we empirically found the Beylkin wavelet to give the maximum noise suppression, and we used this wavelet for the wavelet analysis and synthesis.

4. Experimental Evaluation

We tested our algorithm on a set of 6 TIMIT utterances recorded in an MRI scanner, with two different scanner settings that produce two different periodic noises we will call seq1 and GR. The drawback with using these recordings for evaluation is the lack of a clean reference signal. Consequently, we supplemented our evaluation with clean speech recordings from the Aurora 5 digits database. We added the two MRI noises to the clean speech with an SNR of -6 dB, which is similar to the SNR in the TIMIT utterances.

We compared the performance of our proposed algorithm to the normalized LMS algorithm (denoted LMS-1) and the LMS variant proposed in [4] (denoted LMS-2). For LMS-1, we used a filter length of 3000 and a step size of 1. The LMS-2 algorithm did not need any parameter tuning; these are set by the algorithm and vary based on the MRI pulse sequence used to acquire the recording. LMS-2 is known to perform well with seq1 noise and is currently used to remove seq1 noise from speech recordings. However, its performance degrades with GR noise, preventing speech researchers from collecting better MRI images using GR pulse sequences.

4.1. Quantitative Performance Metrics

To quantify the performance of our denoising algorithm, we calculated the noise suppression, which is given by:

$$\text{noise suppression} = 10 \log \left(\frac{P_{\text{noise}}}{\hat{P}_{\text{noise}}} \right) \tag{5}$$

where P_{noise} is the power of the noise in the noisy signal and $\hat{P}_{\text{noise}}$ is the power of the noise in the denoised signal. We use a voice activity detector (VAD) to find the noise-only regions in the denoised and noisy signals. We calculate the noise suppression measure instead of SNR because we do not have a clean reference signal for the TIMIT utterances.

Ramachandran et al. proposed the log-likelihood ratio (LLR) and distortion variance measures in [11] for evaluating denoising algorithms. The LLR calculates the mismatch between the spectral envelopes of the clean signal and the denoised signal. It is calculated using:

$$\text{LLR} = \log \frac{\boldsymbol{a}_{\hat{s}}^T R_s \boldsymbol{a}_{\hat{s}}}{\boldsymbol{a}_s^T R_s \boldsymbol{a}_s} \tag{6}$$

where $\boldsymbol{a}_s$ and $\boldsymbol{a}_{\hat{s}}$ are p-order LPC coefficients of the clean and denoised signals respectively, and R_s is a

$(p+1) \times (p+1)$ autocorrelation matrix of the clean signal. An LLR of 0 indicates no spectral distortion between the clean and denoised signals, while a high LLR indicates the presence of noise and/or distortion in the denoised signal. The distortion variance is given by:

$$\sigma_d^2 = \frac{1}{L} \|s[n] - \hat{s}[n]\|^2 \tag{7}$$

where $s[n]$ and $\hat{s}[n]$ are the clean and denoised signals respectively, and L is the length of the signal. A low distortion variance is more desirable than a high distortion variance.

4.2. Qualitative Performance Metrics

To supplement the quantitative results, we created a listening test to compare the denoised signals from our proposed algorithm, as well as LMS-1 and LMS-2. We created 12 sets of audio clips in 4 different environments: TIMIT utterances with seq1 noise, TIMIT utterances with GR noise, Aurora digits with seq1 noise, and Aurora digits with GR noise. Each environment contained 3 sets of audio clips. Each set contained a noisy signal and denoised versions of the signal from the proposed algorithm, LMS-1, and LMS-2. For the sets with Aurora digits, we also included the clean signal. Thus, each set with TIMIT utterances had 4 clips and each set with Aurora digits had 5 clips. The sets and the clips within each set were randomized and presented in an online survey. 25 volunteers ranked each clip within a set from 1 to 4 or 5, with 1 meaning best quality and intelligibility.

4.3. Results

Objective measures: Table 1 lists the noise suppression for the TIMIT utterances. Table 2 shows the noise suppression, LLR, and distortion variance results for the Aurora digits. For TIMIT utterances corrupted by seq1 and GR noises, our proposed algorithm suppresses noise better than LMS-1 and LMS-2. Our algorithm performs slightly worse than LMS-1 for Aurora digits corrupted by seq1 and GR noises. This is because the noise in the Aurora recordings is purely additive, while the noise in the direct MRI TIMIT recordings is more convolutive in nature. Our experiments confirmed that LMS-2 performs better on seq1 noise than GR noise, both for the TIMIT utterances and Aurora digits. Importantly, our proposed algorithm performs comparably to LMS-2 in seq1 noise. The LLR and distortion variance results show that our algorithm reconstructed the spectral characteristics of the clean signal more faithfully than LMS-1 and LMS-2. Preserving spectral characteristics of the signal is a key result when considering denoising speech for subsequent speech analysis and modeling.

Subjective measures: Table 3 shows the median rankings obtained from the listening test for the audio clips in the 4 environments. A nonparametric Kruskal-Wallis Test showed that the medians of rankings obtained for each denoising algorithm were significantly different at the $\alpha = 99\%$ level. We then used the post-hoc Wilcoxon rank-sum test to check for pairwise differences in the median ranks. The Wilcoxon test results show that the median ranks for each pair of clips are significantly different at the $\alpha = 99\%$ level, except for the case of the LMS-1/noisy pair for the TIMIT utterances with seq1 noise environment. Hence, we can say with some certainty that listeners ranked our algorithm as the best for removing GR noise and second best for removing seq1 noise.

Table 1: Noise suppression results for TIMIT sentences.

	Proposed	LMS-1	LMS-2
seq1	19.27	18.01	18.79
GR	24.1	18.37	9.17

Table 2: Noise suppression (NS), LLR, and distortion variance (DV) results for the Aurora 5 digits.

Metric	Sequence	Proposed	LMS-1	LMS-2
NS (dB)	seq1	30.23	32.55	26.53
	GR	24.14	27.88	10.91
LLR	seq1	0.17	0.4	0.42
	GR	0.11	0.41	0.33
DV ($\times 10^{-5}$)	seq1	7.52	34.8	21.4
	GR	9.56	35.8	37.7

5. Conclusions

We have proposed a denoising algorithm to remove noise from speech recorded in an MRI scanner. The two-step algorithm uses PLCA to separate the noise and speech, and wavelet packet analysis to further remove noise left by the PLCA algorithm. Objective measures show that our proposed algorithm achieves better noise suppression and less spectral distortion than LMS methods. A listening test shows that our algorithm yields higher quality and more intelligible speech than LMS methods.

To further extend our work, we will compare our proposed algorithm to other denoising methods, such as signal subspace and model-based approaches. Additionally, we need to evaluate how well our algorithm aids speech analysis, such as formant extraction. Finally, we will evaluate the performance of our algorithm in other low-SNR speech enhancement scenarios, such as those involving Gaussian, Cauchy, babble, and traffic noises.

Table 3: Median rankings of the audio clips for the four environments

ENVIRONMENT	ALGORITHM				
	Clean	Proposed	LMS-1	LMS-2	Noisy
TIMIT, seq1 noise		2	3	1	4
TIMIT, GR noise		1	2	3	4
Aurora, seq1 noise	1	3	4	2	5
Aurora, GR noise	1	2	3	4	5

6. References

[1] W. F. Katz, S. V. Bharadwaj, and B. Carstens, "Electromagnetic Articulography Treatment for an Adult With Broca's Aphasia and Apraxia of Speech," *J. Speech, Language, and Hearing Research*, vol. 42, no. 6, pp. 1355–1366, Dec. 1999.

[2] M. Itoh, S. Sasanuma, H. Hirose, H. Yoshioka, and T. Ushijima, "Abnormal articulatory dynamics in a patient with apraxia of speech: X-ray microbeam observation," *Brain and Language*, vol. 11, no. 1, pp. 66–75, Sep. 1980.

[3] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan, "Timing effects of syllable structure and stress on nasals: A real-time MRI examination," *J. Phonetics*, vol. 37, no. 1, pp. 97–110, Jan. 2009.

[4] E. Bresch, J. Nielsen, K. S. Nayak, and S. Narayanan, "Synchronized and Noise-Robust Audio Recordings During Realtime Magnetic Resonance Imaging Scans," *J. Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, Oct. 2006.

[5] Z. Duan, G. J. Mysore, and P. Smaragdis, "Online PLCA for Real-time Semi-supervised Source Separation," in *Proc. Int. Conf. Latent Variable Analysis/Independent Component Analysis*, Tel-Aviv, Israel, 2012, pp. 34–41.

[6] Y. Ghanbari and M. R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Commun.*, vol. 48, no. 8, pp. 927–940, Aug. 2006.

[7] M. McJury and F. G. Shellock, "Auditory Noise Associated with MR Procedures," *J. Magnetic Resonance Imaging*, vol. 12, no. 1, pp. 37–45, Jul. 2001.

[8] Y. Kim, S. S. Narayanan, and K. S. Nayak, "Flexible retrospective selection of temporal resolution in real-time speech MRI using a golden-ratio spiral view order," *Magnetic Resonance in Medicine*, vol. 65, no. 5, pp. 1365–1371, 2011.

[9] S. Tabibian, A. Akbari, and B. Nasersharif, "A New Wavelet Thresholding Method for Speech Enhancement Based on Symmetric Kullback-Leibler Divergence," in *14th Int. Computer Society of Iran Computer Conf.*, Tehran, Iran, 2009, pp. 495–500.

[10] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Wavelet for Speech Denoising," in *Proc. IEEE Region 10 Annu. Conf. Speech and Image Technologies Computing and Telecommunications*, Brisbane, Australia, 1997, pp. 479–482.

[11] V. R. Ramachandran, I. M. S. Panahi, and A. A. Milani, "Objective and Subjective Evaluation of Adaptive Speech Enhancement Methods for Functional MRI," *J. Magnetic Resonance Imaging*, vol. 31, no. 1, pp. 46–55, Dec. 2009.