31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Integrated Speech Enhancement For Functional MRI Environment

Nishank Pathak , Ali A. Milani *Member, IEEE*,
Issa Panahi, *Senior Member, IEEE* and Richard Briggs, *Member, IEEE*

Abstract— **This paper presents an integrated speech enhancement (SE) method for the noisy MRI environment. We show that the performance of SE system improves considerably when the speech signal dominated by MRI acoustic noise at very low SNR is enhanced in two successive stages using two-channel SE methods followed by a single-channel post processing SE algorithm. Actual MRI noisy speech data are used in our experiments showing the improved performance of the proposed SE method.**

I. INTRODUCTION

FUNCTIONAL MRI (fMRI) is an important tool for investigating the human brain function. During FMRI experiments, researchers (in a control room) usually communicate verbally with the patient (in the scanner room) to give instructions and monitor the brain activity in response to the various questions and auditory stimuli. However, strong acoustic noise (noise levels greater than 120 dB SPL) generated by the scanner and the ancillary equipments in the MRI room overwhelms the subject's speech and interferes with diagnosis and imaging process [1]. The enhancement of this corrupted speech signal and improvement of the auditory communication is an existing challenge in FMRI research [1],[2],[3]. The background noise component must be considerably reduced to ensure reliable speech perception. At the same time care should be taken to minimize the distortion hence to preserve the non verbal cues of the speech signal. FMRI acoustic noise has a dominant periodic component in it. In such type of noise signals the most trivial solution would be to use a comb filter to suppress all the noise harmonics. This is not a pragmatic solution for suppressing the FMRI acoustic noise due to the fact that there are many peaks present under 8 kHz (Fig.1) which are almost equally distributed over the frequency range.

As far as the patient head movement is concerned, our experiments in the MRI machine confirm that head movement inside the head matrix of the machine is very limited due to the padding provided around the head.

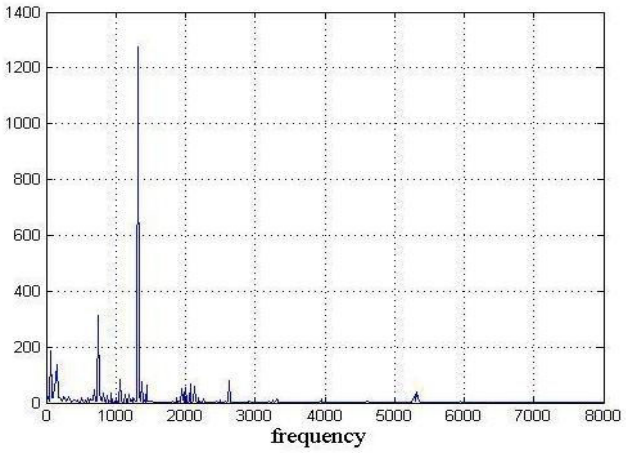

Fig. 1. Spectrum of FMRI noise band limited to 8 kHz

In this paper the performance of the two channel speech enhancement [2] is improved by using an integrated speech enhancement method. The proposed algorithm consists of an adaptive filtering stage followed by a conventional single channel speech enhancement algorithm. The organization of the paper is as follows. In Section II we introduce the general two-stage speech enhancement architecture and describe various algorithms used in the two stages. Section III describes the data acquisition setup used. Section IV reports the experimental results and compares the different methods used in our experiments. Section V concludes the paper.

II. INTEGRATED SPEECH ENHANCEMENT ALGORITHM

In this section we propose a new integrated speech enhancement system which is designed for low SNR conditions. The idea is to use a two channel speech enhancement method as the first stage to increase SNR prior to the second stage. The second stage is a single channel post processing algorithm.

A. Stage 1: Two channel speech enhancement algorithm

In the proposed structure, the first stage is a two channel speech enhancement algorithm using adaptive filtering [2]. In this method (Fig.2) a microphone is placed near the noise source to record the noise signal as the reference noise. The second channel is another microphone placed near the

This study was supported by the VA IDIQ contract number VA549-P-0027 awarded and administered by the Dallas, TX VA Medical Center. The content of this paper does not necessarily reflect the position or the policy of the U.S. government, and no official endorsement should be inferred

Nishank Pathak is a Graduate Teaching Assistant at the University of Texas at Dallas, Richardson, TX 75080 USA (phone: 214-830-0799; fax: 972-883-2710) nishank.pathak@student.utdallas.edu

Ali A Milani is a Graduate Research Assistant at the University of Texas at Dallas, Richardson, TX 75080 USA ali.a.milani@student.utdallas.edu

I. M. S. Panahi is an Assistant Professor with the Department of Electrical Engineering, University of Texas, Dallas, TX 75080 USA : issa.panahi@ieee.org

R. Briggs works with the Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX 75390 : richard.briggs@utsouthwestern.edu

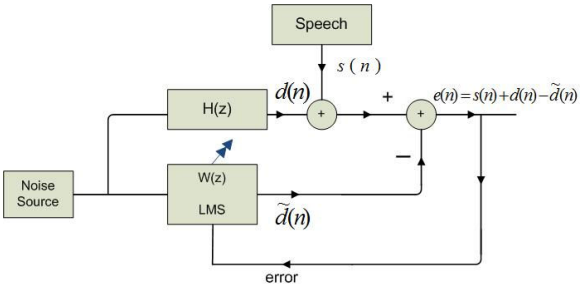

Fig. 2. 2-Channel Speech Enhancement.

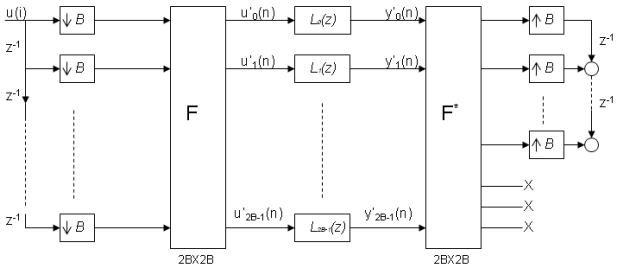

Fig. 3. The block adaptive processing algorithm proposed by [4]

speaker that records the speech which is contaminated with the background acoustic noise. An adaptive filter is used to estimate and suppress the noise in the second channel from the reference noise. The performance and stability of adaptive filters is highly related to the eigenvalue spread of the reference noise. Since FMRI acoustic noise has a very high eigenvalue spread (order of 10^3 compared to 1 ,for the white noise), in order to improve the system performance, the subband adaptive filtering algorithm given in [4] has been used which not only results in improved stability and performance but also reduces the computational complexity. The filterbank structure is shown in Fig.3.

B. Stage 2: Single channel post processing algorithm

As mentioned before, a conventional single-channel speech enhancement algorithm is used as the second stage. In this paper we compare the performances of three categories of single-channel speech enhancement methods which are spectral subtraction, sub-space, and MMSE [5]. The sub-space algorithm used here is the perceptual KL transform [6] which incorporates human hearing properties in speech enhancement. This method is chosen to reduce the distortion in speech which is generally caused by the subspace class of algorithms [5]. Among the MMSE methods, the widely used LOG-MMSE [4] algorithm and among the available spectral subtraction methods, the over subtraction algorithm [7] are chosen. In the following subsections the three algorithms are briefly explained.

1) Perceptual K-L Transform (PKLT): PKLT is an eigenvalue decomposition method which uses the masking property of the human ears in eigenvalue domain. The filter generated by this algorithm is given by [6] $H = U_1 G U_1^H$. Where U_1 is the unitary eigenvector matrix spanning the signal subspace and G is the gain matrix given by $G = diag(g_1,\ldots,g_q)$ and $g_i = e^{-\nu\xi_i/min(\lambda_{s,i},\theta_i)}$ where ν is the parameter which controls the tradeoff between residual noise level and the signal distortion. $\lambda_{i,s}$ are the eigenvalues in the signal sub-space, ξ_i is the noise energy in the spectral direction of the eigenvectors of speech covariance matrix and θ_i are the masking energy in the spectral direction of the eigenvectors of speech covariance matrix.

2) Log Minimum Mean Square Estimate (LogMMSE): As described in [8] this estimator minimizes the mean-square error of the log-magnitude spectra $E\{(\log(x_k) - \log(\hat{x}_k))^2\}$. The optimal LogMMSE estimator can be obtained by evaluating the conditional mean of $\log(\hat{x}_k)$, i.e. $\log(\hat{x}_k) = E\{\log(x_k)|y_k\}$and assuming a Gaussian model of the noise and speech we get

$$\hat{x}_k = \frac{\xi_k}{\xi_k + 1} \exp\{\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\} y_k \tag{1}$$

$$y_k = x_k + d_k \tag{2}$$

$\hat{x}_k$ is the estimated magnitude of the clean speech. x_k is the actual magnitude of the clean speech. y_k is the noisy speech and d_k is the nosie at frequency ω_k. ξ_k is the apriori SNR (Signal to Noise Ratio).

3) Spectral Subtraction: Spectral subtraction subtracts the estimated noise energy from the noisy speech file. Over subtraction algorithm [7] estimates the clean speech magnitude using

$$|\hat{x}(\omega)|^2 = \begin{cases} |\hat{y}(\omega)|^2 - \alpha|\hat{d}(\omega)|^2 & \text{for} \\ & |\hat{y}(\omega)|^2 > (\alpha+\beta)|\hat{d}(\omega)|^2 \\ \beta|\hat{d}(\omega)|^2 & \text{else} \end{cases} \tag{3}$$

α and β are the over subtraction factor and spectral floor parameter, respectively

III. DATA ACQUISITION SETUP

The fMRI noise was recorded from a 3 Tesla Siemens Magnetron Trio. The acoustic signal was recorded using a diffuse-field microphone that had an Omni-directional response and a good dynamic range. 30 second segment of the analog data was digitized at 64 kHz sampling rate using NI PCI 4472 A/D board. LabVIEW 8.0 was used to control the data acquisition. Three microphones were used. One was held by the subject speaking the sentences from the NOIZEUS database [5] and the other two microphones were placed on the outer side of the head matrix to collect the reference noise simultaneously (Fig.4).

In addition to this, we simulated the actual setup in the Laboratory using a manikin and a test bed mimicking the MRI bore to get a close estimate of the amount of speech enhancement, as the clean speech was not available from the UTSW data. We played pre-recorded MRI noise using NI 6733DAC through loudspeaker-A. The frequency response

Fig. 4. Data collection experiment in a 3-T MRI machine

of loudspeaker-A ranged from 55 Hz to 20 kHz with a maximum power output of 125 W. The clean speech signals were played through another loudspeaker (referred to as loudspeaker-B) and the SPL of the speech at loudspeaker B was kept 10dB higher than the noise SPL at loudspeaker B because humans tend to talk 10dB louder than the background noise. The loudspeaker-B was placed close to the mouth of the manikin and clean speech was played from NOIZEUS database using NI6733DAC. The reference noise and the corrupted speech signals were acquired using NI4472 DAC.

IV. SIMULATION AND RESULTS

To evaluate the performance of the algorithm, the actual data recorded from the UTSW was used. The original data sampled at 64 KHz was decimated by factor 4 and down-sampled to 16 KHz because, most of the energy of speech was concentrated below 8 kHz. The time waveform of the one of the recorded sentence is shown in Fig.5(a). The uniform subband adaptive filtering (filter bank) was used. 32 subbands were selected and a noise canceling filter of length 1024 was trained by the first 2 seconds of data. This assumption is valid because there is no speech during the initial few seconds of the data. Moreover in any MRI scan experiment the patient does not start speaking unless instructed to do so by the operator. This can give ample amount of time to the operator for training the filter. The step size of 0.1190 was used to train the filter. In the second stage, we used a 20ms hamming window for spectral subtraction and PKLT, and a 32ms hamming window for LogMMSE. The residual noise was measured from the 100 ms of silence (i.e. no speech signal) frame after the first 20000 samples since the speech starts only after 30000 samples. The residual noise was measured as $20\log_{10}(||x_e||/||x_n||)$ where $||x_e||$ is the second norm of 100ms of the silence segment of the enhanced signal, and $||x_n||$ is the second norm of the 100ms of the silence segment of the noisy signal.

TABLE I
PERFORMANCE OF ALGORITHMS ON RECORDING FROM UTSW

	Segmental SNR
Two channel enhanced (stage1)	-5.0852
PKLT (stage 2)	-18.824
Spectral Subtraction (stage 2)	-15.3058
Log MMSE (stage 2)	-22.3219

TABLE II
PERFORMANCE OF ALGORITHMS ON RECORDING FROM LAB

	Residual noise (dB)	PESQ
Original noisy speech		1.53
Two channel enhanced (stage1)	-16.083	2.5195
PKLT (stage 2)	-34.3551	2.5240
Spectral Subtraction (stage 2)	-27.077	3.0458
Log MMSE (stage 2)	-37.2404	3.0345

Noise Suppression by itself is not an accurate measure of speech enhancement. As a standard objective measure we also used perceptual evaluation of speech quality (PESQ) [9] as an enhancement measure because it was proved in [5] that PESQ gives the best estimate of speech enhancement. The problem in using PESQ is that we need access to the clean speech for the enhancement estimation. Since this was not possible with the actual MRI experiment data, we used the simulation recordings done in the lab as described in the last section for the measurement of PESQ. Table I and II show the results for both data sets.

V. CONCLUSION

The paper demonstrated that the noise suppression is improved in a dual stage enhancement when compared to a single stage two channel enhancement and hence the fatigue of the listener is reduced.

After adding a second stage we found that though PKLT is very aggressive in noise suppression it is very high on computation and has a lower PESQ measure than LogMMSE. Also, subjective tests done on human subjects have shown that LogMMSE has the least distortion on the speech signal [5]. Spectral subtraction is found to be highly prone to musical noise and the noise suppression is not as good as in the other two algorithms. Hence there is a tradeoff between complexity and performance.

REFERENCES

[1] M. Ravicz, J. Melcher, , and N. Kiang, "Acoustic noise during functional magnetic resonance imaging," *Journal of Acoustic Society of America*, vol. 108, no. 4, pp. 1683–1696, 2000.

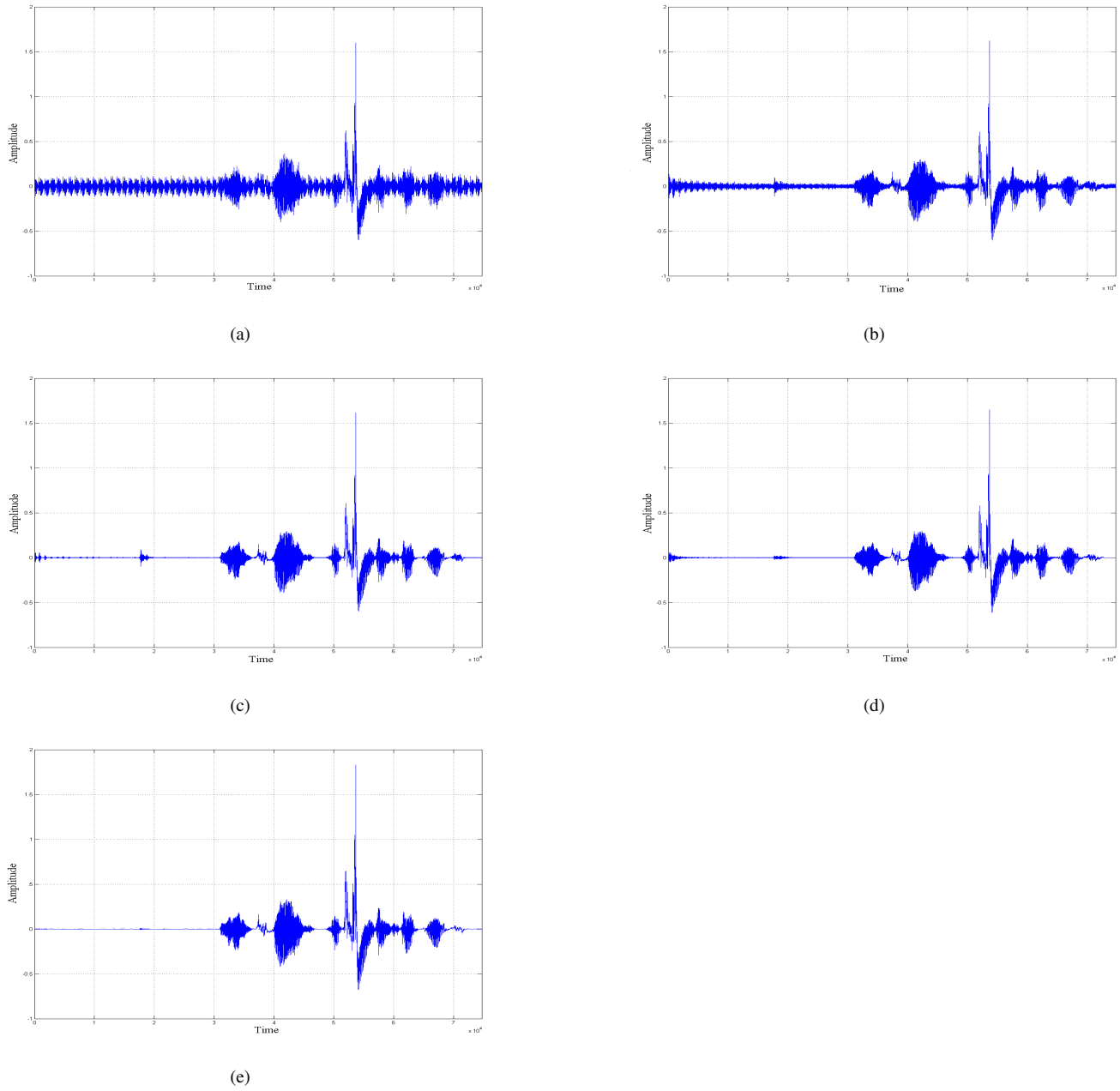

Fig. 5. Time domain plot for (a) noisy speech (b) speech after enhancement by the first stage(c) speech after enhancement by spectral subtraction (d) speech after enhancement by LogMMSE (e)speech after enhancement by PKLT . All the plots are for 40 EPI sequences per second.

[2] V. R. Ramachandran, G. Kannan, A. A.Milani, and I. M. Panahi, "Speech enhancement in functional mri environment using adaptive sub-band algorithms," in *International Conference on Acoustics, Speech and Signal Processing,ICASSP 2007*, vol. 2. IEEE, 2007, pp. :II–341 – II–344.

[3] Mayer.J., "Prosody processing in speech production: Pre-evaluation of a fmri study," in *Proceedings XIVth International Congress of Phonetic Sciences*, San Fransisco, 1999, pp. 2339–2342.

[4] R. Merched and A. H. Sayed, "An embedding approach to frequency domain and sub band adaptive filtering," *IEEE Transactions On Signal Processing*, vol. 48, no. 9, pp. 2607–2619, 2000.

[5] P. C.Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[6] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 700–708, 2003.

[7] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing,ICASSP '79.*, vol. 4, pp. 208–211, 1979.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[9] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii–psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, 2002.