# A SPATIO-TEMPORAL-SPECTRAL PROCESSOR FOR DUPLEX HANDS-FREE COMMUNICATION SYSTEMS

*Siow Yong Low, Sven Nordholm and Hai Quang Dam*

{siowyong,sven,damhai}@watri.org.au
Western Australian Telecommunications Research Institute (WATRI)[†],
Crawley, WA 6009, Australia

## ABSTRACT

This paper introduces a new scheme, which combines a spatio-temporal decorrelator and a non-coherent based post processor to jointly suppress noise and acoustic echo. The new $L$ element structure extends upon the spatio-temporal decorrelator proposed in [1] by incorporating a non-coherent based post-processor. The non-coherent processor compensates for the presence of non-linearity and relaxes the demand on the temporal decorrelator. Real room evaluations demonstrate the efficacy of the scheme in both noisy double-talk and non double-talk situations with an average gain of $5$ dB in noise and echo suppression compared to using only the spatio-temporal decorrelator.

## 1. INTRODUCTION

Hands-free communication systems have revolutionized the way humans communicate with each other. Without a doubt, hand-held communication devices will become outdated with everyday essentials such as personal digital assistants (PDA) and mobile phones becoming hands-free compliant. Nonetheless, when it comes to hands-free system, there are several disadvantages. Since the user is at a distant from the microphone, the microphone will also capture the background noise (such as babble) as well as the interference due to the hands-free loudspeaker. Therefore, a scheme with both noise and acoustic echo cancellation capability is instrumental in this application.

This paper extends upon the idea in [1] to jointly suppress the background noise and acoustic echoes through a spatio-temporal decorrelator and a spectral processor. In [1], the temporal decorrelator is proposed to compensate for the separation quality of the blind signal separation (BSS). However, in the presence of the non-linearity of audio devices, the channel non-linearity or non-converging solution, the linear temporal decorrelator fails to perform satisfactorily. Here, the spectral processor is suggested to combat the aforementioned potential problems. The extension to the spatio-temporal decorrelator yields a spatio-temporal-spectral processor, which spatially and temporally decorrelates the target signal from the noise and echoes, and spectrally suppresses the residue noise and echoes.

[†]WATRI is a joint institute between Curtin University of Technology and The University of Western Australia. Research (partially) supported by National ICT Australia (NICTA). NICTA is funded by the Australian Research Council (ARC).

## 2. OVERVIEW

Figure 1 shows the block diagram of the three main processors in the proposed spatio-temporal-spectral processor. The proposed structure closely resembles the spatio-temporal decorrelator introduced in [1] except for the addition of a spectral based post-processor to remove residue noise and echoes. The first block, which is the BSS acts as a front-end spatial processor to separate the target signal from the interference (e.g., acoustic echo, ambient noise or babble) using the $L$ observations. As noted in [1], there is a fundamental limitation in the separation quality of the BSS. To compensate for that, the desired output is compensated temporally through the adaptive noise canceller (ANC) to remove any dependencies on noise and echoes. Finally, the spectral post-processor jointly estimates the residue noise and echo power to further boost the suppression capabilities.

## 3. THE SPECTRAL PROCESSOR

### 3.1. Preliminary

A non-coherent spectral processor is proposed to remove the residue noise and echoes in the overall output. As explained previously, the purpose of the non-coherent processor is primarily to compensate for the limitations given that the channel possesses non-linear characteristics and for the case of non-converging solutions. Thus, assuming non-perfect noise and echo cancellation coupled with the fact that the target signal, noise and acoustic echo are statistically independent, the output from the spatio-temporal processor can be expressed as

$$z(\omega,k) = z_{\text{target}}(\omega,k) + \underbrace{z_{\text{noise}}(\omega,k) + z_{\text{echo}}(\omega,k)}_{\text{residue noise and echo}}, \quad (1)$$

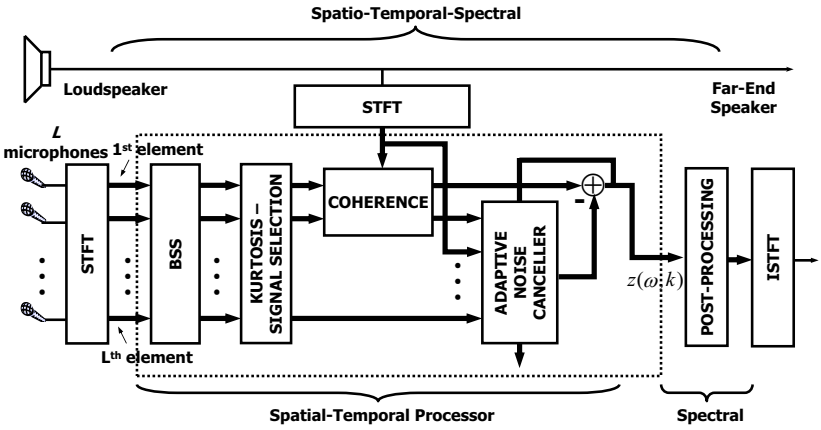


Figure 1: *The proposed spatio-temporal-spectral processor with $L$ microphones.*

where $z_{\text{target}}(\omega,k)$, $z_{\text{noise}}(\omega,k)$ and $z_{\text{echo}}(\omega,k)$ are the short-time Fourier transforms of the target signal, residue noise and residue echo, respectively. The task at hand is to suppress both the residue noise and echo through a post-filter [2]. Note that with the inclusion of a post-filter, the demand on the temporal decorrelator can be relaxed, which in turn speeds up convergence, more robust to perturbation and reduces computational burden [3].

Essentially, the efficiency of the post-filter lies in the estimation of both the residue noise and echo power. In this paper, it is proposed that the background noise be estimated by using the minimum statistics [4] whilst the residual echo be estimated through the coherence method [5]. It is also worth mentioning that as the SNR of the spatio-temporal processor output improves, the noise floor (minimum statistics) becomes more discernible and hence a better estimation of the noise statistics can be achieved. This in turn reduces artifacts caused by randomly distributed spectral peaks due to "over and under" estimation of the noise spectrum.

### 3.2. Residue Noise - Minimum Statistics

The minimum statistics noise tracking method is based on the observation that even during speech activity, a short term power spectral density estimate of the noisy signal frequently decays to values, which are representatives of the noise power levels [6]. Thus, by tracking the minimum power within a finite window (large enough to bridge high power speech segments), the noise floor can be estimated. The short-time subband power estimate, $P_z(\omega,k)$ can be estimated using the recursion as follows

$$P_z(\omega,k) = \alpha P_z(\omega,k-1) + (1-\alpha)|z(\omega,k)|^2. \quad (2)$$

A smaller value of $\alpha$, the smoothing constant, allows better tracking capabilities, whereas a larger value provides a smoother estimate. Following the recursion, the minimum noise power estimate, $P_{z_{\text{noise}}}(\omega,k)$ is obtained by sample-wise comparison of the current smoothed signal with the $\Delta$ previous estimated minimum power (stored in a buffer) according to

$$P_{z_{\text{noise}}}(\omega,k) = \mathsf{min}\left[P_z(\omega,k-\Delta) : P_z(\omega,k)\right], \quad (3)$$

where $\mathsf{min}[\cdot]$ denotes the minimum value of operator. After a designated window size, the $\Delta$ buffer is shifted and its last stored value is updated with the most recent power estimate. The motivation for removing the last stored value after a certain time period is to track the changes in the noise level.

### 3.3. Residue Acoustic Echo - Coherence Method

A popular method for estimating the residual echo power is the coherence method [5]. This estimation method rests on the assumption that the difference between the residual echo and the line echo can be described by the following transfer function

$$z_{\text{echo}}(\omega,k) = H(\omega)y_{\text{line}}(\omega,k), \quad (4)$$

where $H(\omega)$ is the residual echo transfer function and $y_{\text{line}}(\omega,k)$ is the far-end line echo. Since the target signal, residual noise and echo are statistically independent, the cross correlation between the far-end line echo $y_{\text{line}}(\omega,k)$ and the output signal $z(\omega,k)$ is

$$\mathsf{E}[y_{\text{line}}(\omega,k)z(\omega,k)] = \mathsf{E}[y_{\text{line}}(\omega,k)z_{\text{echo}}(\omega,k)], \quad (5)$$

where $\mathsf{E}[\cdot]$ denotes the expectation operator. Substituting (4) into (5) gives

$$\mathsf{E}[y_{\text{line}}(\omega,k)z(\omega,k)] = H(\omega)\mathsf{E}[y^2_{\text{line}}(\omega,k)]. \quad (6)$$

The residual echo transfer function in (6) can then be rewritten in terms of power spectra as

$$H(\omega) = \frac{P_{y_{\text{line}},z}(\omega,k)}{P_{y_{\text{line}}}(\omega,k)}, \tag{7}$$

where the cross power and auto power spectra are $P_{y_{\text{line}},z}(\omega,k) = \mathsf{E}[y_{\text{line}}(\omega,k)z(\omega,k)]$ and $P_{y_{\text{line}}}(\omega,k) = \mathsf{E}[y^2_{\text{line}}(\omega,k)]$, respectively. From (4), the residual echo power is given as

$$P_{z_{\text{echo}}}(\omega,k) = |H(\omega)|^2 P_{y_{\text{line}}}(\omega,k). \tag{8}$$

Thus by inserting (7) into (8), the residual echo power becomes

$$P_{z_{\text{echo}}}(\omega,k) = \frac{P^2_{y_{\text{line}},z}(\omega,k)}{P^2_{y_{\text{line}}}(\omega,k)} P_{y_{\text{line}}}(\omega,k) = \frac{P^2_{y_{\text{line}},z}(\omega,k)}{P_{y_{\text{line}}}(\omega,k)}. \tag{9}$$

Alternatively, the residual echo power can be expressed in terms of the coherence function as

$$\begin{aligned} P_{z_{\text{echo}}}(\omega,k) &= \frac{P^2_{y_{\text{line}},z}(\omega,k)}{P_{y_{\text{line}}}(\omega,k)} \\ &= \frac{P^2_{y_{\text{line}},z}(\omega,k)}{P_{y_{\text{line}}}(\omega,k)P_z(\omega,k)} P_z(\omega,k) \\ &= \mathsf{Coh}\left[y_{\text{line}}(\omega,k), z(\omega,k)\right] P_z(\omega,k), \end{aligned} \tag{10}$$

where $\mathsf{Coh}\left[y_{\text{line}}(\omega,k), z(\omega,k)\right] = \frac{P^2_{y_{\text{line}},z}(\omega,k)}{P_{y_{\text{line}}}(\omega,k)P_z(\omega,k)}$. Equation (10) shows that the residual echo power can be computed in terms of the received signal (output of the spatio-temporal decorrelator) and the far-end line echo. As such, the coherence method is especially appealing since both signals are readily available.

### 3.4. Joint Residue Noise and Echo Postfilter

Taking the magnitude of (1) and rewriting it in terms of the target signal, $z_{\text{target}}(\omega,k)$ gives

$$|z_{\text{target}}(\omega,k)| = |z(\omega,k)| \cdot \underbrace{\left[1 - \frac{|z_{\text{noise}}(\omega,k)| + |z_{\text{echo}}(\omega,k)|}{|z(\omega,k)|}\right]}_{\text{gain function}}. \tag{11}$$

Based on (11), the gain function for the postfilter can be found as

$$G(\omega,k) = \left[1 - \varphi(\omega,k)\frac{|z_{\text{noise}}(\omega,k)|^a + |z_{\text{echo}}(\omega,k)|^a}{|z(\omega,k)|^a}\right]^{\frac{1}{a}} \tag{12}$$

where $a$ is the parameter that determines the type of subtraction used (e.g., $a = 1$ is the magnitude subtraction whereas $a = 2$ is the power subtraction) and $\varphi(\omega,k)$ is the subtraction factor. By inspecting (12), it is observed that the subtraction of the (smoothed) estimates of $|z_{\text{noise}}(\omega,k)|$ and $|z_{\text{echo}}(\omega,k)|$ from the actual spectrum may result in spectral peaks due to spectral mismatch. As the estimates are time-varying, the spectral mismatch is inherently time-varying and will lead to the infamous "musical noise". One straightforward solution to reduce the musical noise is to "oversubtract" ($\varphi(\omega,k) > 1$) [7]. By doing so, the residue spectral peaks will be made lower compared to the case of $\varphi(\omega,k) = 1$. Nevertheless, by oversubtracting, the speech quality may be compromised as the low energy phonemes are suppressed [4]. An alternative method proposed by Berouti *et al.* [7] suggests that the subtraction factor be made dependent on SNR since the subtraction factor should be made smaller for high SNR situations as compared to the cases of low SNR. They also propose that a limitation of the maximum subtraction be made through a spectral floor $\phi(\omega)$ in order to prevent the deepening of the valley between spectral peaks. Moreover, the spectral floor can be viewed as "filling-in" the valleys to reduce the musical noise due to the distance between the peaks and the valleys.

By incorporating both the SNR based subtraction factor and the spectral floor, the gain function to suppress both the residue noise and echo in (12) can be rewritten as

$$G(\omega,k) = \mathsf{max}\left[\sqrt{\varphi(\omega,k)\frac{P_{z_{\text{noise}}}(\omega,k) + P_{z_{\text{echo}}}(\omega,k)}{P_z(\omega,k)}}, 1 - \sqrt{\varphi(\omega,k)\frac{P_{z_{\text{noise}}}(\omega,k) + P_{z_{\text{echo}}}(\omega,k)}{P_z(\omega,k)}}\right], \tag{13}$$

and

$$\varphi(\omega,k) = \beta\varphi(\omega,k-1) + (1-\beta)\left[\mathcal{S}(\omega)\frac{P_{z_{\text{noise}}}(\omega,k)}{P_{z_{\text{noise}}}(\omega,k) + P_z(\omega,k)}\right], \tag{14}$$

where $\beta$ is the smoothing constant and $P_z(\omega,k)$ is the observed power estimate. Both the residue noise power estimate $P_{z_{\text{noise}}}(\omega,k)$, and the residue echo power estimate, $P_{z_{\text{echo}}}(\omega,k)$ are estimated by using the minimum statistics method [4] and the coherence method [3], respectively. The parameter, $\mathcal{S}(\omega)$ is the $\omega$-th point of an exponential function. The role of the parameter is to exponentially emphasize less on the higher frequency range since most real-world noise mainly concentrate in the low frequency range. Also, the parameter prevents the higher frequency components from being overly weighted, which may result in a whitening effect.

## 4. EXPERIMENTAL RESULTS

The proposed speech enhancement scheme was evaluated in a real room of dimensions $3.5 \times 3.1 \times 2.3$ m$^3$ using a

four-element linear array with a spacing of $0.04$ m, sampled at $8$ kHz. Two loudspeakers emitting babble noise were placed facing the front two corners of the room to create diffuseness and three other loudspeakers (also babble) were randomly placed in the middle of the room facing the array. All simulations were performed with signal to noise ratio (SNR) $= -0.5$ dB and signal to echo ratio (SER) $= 0$ dB. The experimental parameters/settings for the spatio-temporal decorrelator were the same as in [1] with $512$ frequency bins. The parameters $\alpha$ and $\beta$ were set to $0.9$ and $0.95$, respectively.

Figure 2 and Figure 3 plot the noise and echo suppression for both the noisy non double-talk and noisy double-talk situations with varying $L$. Results clearly show that the proposed processor achieves an approximately $5$ dB gain over the spatio-temporal decorrelator in both noise and echo suppression. Also, it is noted that more suppression gain is obtained for the proposed scheme as $L$ decreases. This is because as $L$ reduces, the performance of the spatio-temporal decorrelator drops since there are less spatial degrees of freedom available. Thus, there is more residue noise and echo for the spectral processor to suppress. Clearly, the spectral processor compensates the performance of the decorrelator and it is particularly advantageous if there is a strict constraint on the number of microphones.

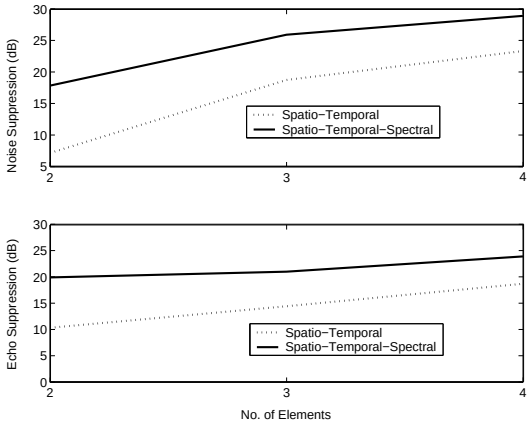


Figure 2: *The noise and echo suppression for the spatio-temporal and the spatio-temporal-spectral with different number of elements for the noisy non double-talk situation.*

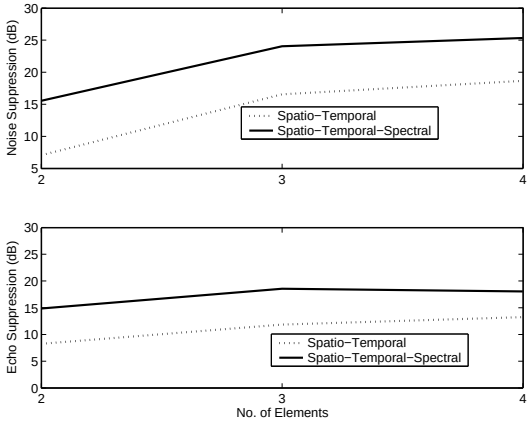


Figure 3: *The noise and echo suppression for the spatio-temporal and the spatio-temporal-spectral with different number of elements for the noisy double-talk situation.*

## 5. CONCLUSIONS

A new spatio-temporal-spectral processor has been presented. Essentially, the processor extends upon the spatio-temporal decorrelator previously proposed by including a non-coherent processor. Results show that the post-processor increases both the noise and echo suppression by more than $5$ dB in both noisy non double-talk and double-talk scenarios.

## 6. REFERENCES


[1] S. Y. Low and S. Nordholm, "A blind approach to joint noise and acoustic echo cancellation," *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, vol. 3, pp. 69–72, March 2005.

[2] R. Martin and J. Altenhöner, "Coupled adaptive filters for acoustic echo control and noise reduction," *IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 5, pp. 3043–3046, May 1995.

[3] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. on Speech and Audio Process.*, vol. 10, no. 5, pp. 245–256, July 2002.

[4] R. Martin, "Spectral subtraction based on minimum statistics," *European Signal Process. Conf.*, pp. 1182–1185, September 1994.

[5] V. Turbin, A. Gilliore, P. Scalart, and C. Beaugeant, "Using pshychoacoustic criteria in acoustic echo cancellation algorithms," *Int. Workshop on Acoustic Echo and Noise Control*, pp. 53–56, September 1997.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 5, pp. 504–512, July 2001.

[7] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 4, pp. 208–211, April 1979.