INTER-CHANNEL DECORRELATION BY SUB-BAND RESAMPLING IN FREQUENCY DOMAIN

Jason Wung, Ted S. Wada, and Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA 30308, USA
{jason.wung, twada, juang}@ece.gatech.edu

ABSTRACT

This paper presents a novel decorrelation procedure by frequency-domain resampling in sub-bands. The new procedure expands on the idea of resampling in the frequency domain that efficiently and effectively alleviates the non-uniqueness problem for a multi-channel acoustic echo cancellation system while introducing minimal distortion to the signal. We show in theory and verify experimentally that the amount of decorrelation in each sub-band, measured in terms of the coherence, can be controlled arbitrarily by varying the resampling ratio per frequency bin. For perceptual evaluation, we adjust the sub-band resampling ratios to match the coherence given by other decorrelation procedures. The speech quality (PESQ) score from the proposed decorrelation procedure remains high at around 4.5, which is about the highest possible PESQ score after signal modification.

***Index Terms*—** Inter-channel decorrelation, time scaling, resampling, multi-channel acoustic echo cancellation

1. INTRODUCTION

The non-uniqueness problem arises during multi-channel acoustic echo cancellation (MCAEC) due to the highly correlated reference signals, *i.e.*, far-end microphone signals, that degrade the convergence rate of the least mean square (LMS) algorithm [1]. A handful of inter-channel decorrelation procedures has been proposed in the past to alleviate such a problem, *e.g.*, [2–4]. As an extension of the decorrelation by resampling technique [5], we proposed in [6] a computationally efficient version based on frequency-domain resampling (FDR) that introduces time-varying delay across channels with negligible audible distortion. When applied to our robust frequency-domain MCAEC system [7], FDR enables faster echo path tracking performance over other decorrelation procedures [6]. This motivates us to further investigate the decorrelation by FDR technique.

We present in this paper a novel approach for inter-channel decorrelation by sub-band resampling (SBR), achieved by varying the resampling ratio across frequencies rather than using the fixed ratio as in FDR. The advantage of SBR is that the amount of decorrelation can be finely controlled for better perceptual quality, *e.g.*, less "resampling," or signal modification, at lower frequencies and vice versa at higher frequencies. Although we have measured the inter-channel coherence before and after several decorrelation procedures in [6], the exact effect of the resampling process on such a measure remains unclear. Since the resampling in discrete time is equivalent to the time scaling in continuous time, we also examine here the change in the coherence after continuous time scaling to analyze the close relationship between resampling and decorrelation. To validate the superior audio quality provided by SBR, we adjust the sub-band resampling ratios to match the coherence to those given by other decorrelation procedures, then compare the processed signals using perceptually objective measures.

2. COHERENCE AS A MEASURE OF CORRELATION

The coherence (or magnitude-squared coherence) [8] is a real-valued function that represents the amount of correlation between two signals in the frequency domain. For the wide-sense-stationary random processes x_t and y_t, the coherence at each frequency ω is given by

$$C_{xy}(\omega) \equiv \frac{|S_{xy}(\omega)|^2}{S_{xx}(\omega)S_{yy}(\omega)}, \quad 0 \le C_{xy}(\omega) \le 1, \tag{1}$$

with $C_{xy} = 1$ being perfectly correlated and $C_{xy} = 0$ being uncorrelated. The cross-spectral density (CSD) $S_{xy}(\omega)$ is given by

$$S_{xy}(\omega) = \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-j\omega\tau}\, \mathrm{d}\tau \equiv \mathcal{F}\{R_{xy}(\tau)\},$$

where $\mathcal{F}\{\cdot\}$ is the continuous time Fourier transform (CTFT) and $R_{xy}(\tau) = \mathrm{E}[x_t y^*_{t-\tau}]$ is the cross-correlation, $\mathrm{E}[\cdot]$ being the mathematical expectation and * denoting the complex conjugation. $S_{xx}(\omega)$ and $S_{yy}(\omega)$ are the power spectral densities (PSDs) of x and y and are calculated by $\mathcal{F}\{R_{xx}(\tau)\}$ and $\mathcal{F}\{R_{yy}(\tau)\}$, respectively.

For $x(t)$ and $y(t)$ as the actual realizations of the stochastic processes in continuous time, the cross-correlation between the two signals is estimated by

$$\tilde{R}_{xy}(\tau) = \int_{-\infty}^{\infty} x(t) y^*(t-\tau)\, \mathrm{d}t,$$

and the CSD is given by $\tilde{S}_{xy}(\omega) = \mathcal{F}\{\tilde{R}_{xy}(\tau)\} = X(\omega)Y^*(\omega)$, where $X(\omega) = \mathcal{F}\{x(t)\}$ and $Y(\omega) = \mathcal{F}\{y(t)\}$. The PSD of $x(t)$ and $y(t)$ is given by $\tilde{S}_{xx}(\omega) = |X(\omega)|^2$ and $\tilde{S}_{yy}(\omega) = |Y(\omega)|^2$, respectively. However, the coherence in this case is equal to one for (1) since only the instant realizations are used for calculation without taking into account the mathematical expectation. Therefore, the CSD is estimated in practice by averaging over short-time evaluations. That is, let $w(t)$ be a window function with the support $t \in [0, T]$, $w_m(t) = w(t - mt_0)$ be the m^{th} window with a delay of mt_0, where $m = 0, 1, \ldots, M-1$, $t_0 \le T$, and M is the number of signal blocks. Then the CSD is estimated by [9]

$$\hat{S}_{xy}(\omega) = \frac{1}{M} \sum_{m=0}^{M-1} X_m(\omega) Y_m^*(\omega),$$

where $X_m(\omega) = \mathcal{F}\{x(t)w_m(t)\}$ and $Y_m(\omega) = \mathcal{F}\{y(t)w_m(t)\}$. The PSD can be similarly estimated. The coherence is estimated as

$$\hat{C}_{xy}(\omega) \equiv \frac{\left|\sum_{m=0}^{M-1} X_m(\omega) Y_m^*(\omega)\right|^2}{\left(\sum_{m=0}^{M-1} |X_m(\omega)|^2\right)\left(\sum_{m=0}^{M-1} |Y_m(\omega)|^2\right)}. \tag{2}$$

3. DECORRELATION BY TIME SCALING

By time expanding a continuous time signal $x(t)$ to $x(t/R)$ with an expansion ratio $R > 1$, the delay is steadily built up over time between the original signal and the time-expanded signal. Intuitively, the cross-correlation between $x(t)$ and $x(t/R)$ should go down due to the delay buildup. We can quantify this effect through the analysis below, which can be similarly applied to time compressing $x(t)$ by choosing $0 < R < 1$.

Let $x(t) = e^{j\omega_0 t}$, $y(t) = x(t/R) = e^{j\omega_0 t/R}$, and $w(t)$ be the rectangular window that is zero outside $t \in [0, T]$. The CTFT of the signals are $X(\omega) = 2\pi\delta(\omega - \omega_0)$ and $Y(\omega) = 2\pi R\delta(\omega - \omega_0/R)$, where $\delta(x)$ is the Dirac delta function. The CTFT of the m^{th} rectangular window is $W_m(\omega) = T\,\text{sinc}(\omega T/2)e^{-j\omega(mt_0+T/2)}$, where $\text{sinc}(x) \equiv \sin(x)/x$. Using the convolution theorem $\mathcal{F}\{x(t)w_m(t)\} = \frac{1}{2\pi}X(\omega) * W_m(\omega)$, the CTFTs of the windowed signals $x_m(t)$ and $y_m(t)$ are given by

$$X_m(\omega) = T\,\text{sinc}\big((\omega-\omega_0)\tfrac{T}{2}\big)e^{-j(\omega-\omega_0)(mt_0+\frac{T}{2})},$$
$$Y_m(\omega) = RT\,\text{sinc}\big((\omega-\tfrac{\omega_0}{R})\tfrac{T}{2}\big)e^{-j(\omega-\frac{\omega_0}{R})(mt_0+\frac{T}{2})}.$$

The frequency contents at ω_0 are given by

$$X_m(\omega)\big|_{\omega=\omega_0} = T,$$
$$Y_m(\omega)\big|_{\omega=\omega_0} = RT\,\text{sinc}\big(\tfrac{\Delta R\omega_0}{R}\tfrac{T}{2}\big)e^{-j\frac{\Delta R\omega_0}{R}(mt_0+\frac{T}{2})}$$
$$= Ae^{-j\frac{\Delta R}{R}\omega_0 t_0 m},$$

where A is a complex constant independent of m and $\Delta R \equiv R - 1$. Using (2), the coherence estimate at ω_0 is

$$\hat{C}_{xy}(\omega)\big|_{\omega=\omega_0} = \frac{\big|\sum_{m=0}^{M-1} TA^* e^{j\frac{\Delta R}{R}\omega_0 t_0 m}\big|^2}{\big(\sum_{m=0}^{M-1} T^2\big)\big(\sum_{m=0}^{M-1}|A|^2\big)}$$
$$= \left[\frac{1}{M}\frac{\sin\big(\frac{\Delta R}{2R}\omega_0 t_0 M\big)}{\sin\big(\frac{\Delta R}{2R}\omega_0 t_0\big)}\right]^2. \quad (3)$$

First of all, we note that (3) is independent of the window size T, which only contributes as a constant factor, and the phase term goes away after taking the absolute value. Second, if $M = 1$, (3) is always equal to one since it is calculated over only a single instance. Third, (3) is always one also if $\Delta R = 0$ since there is no time scaling. Finally, for $M > 1$ and $\Delta R \neq 1$, we can evaluate the reduction in the coherence by the following numerical example.

Suppose the continuous time signal is bandlimited to $f_c = 8$ kHz at the sampling rate $f_s = 16$ kHz. If the coherence measurement frame size is $N = 2048$ samples that is divided into $M = 8$ sub-frames with 50% overlap, then the frame shift in continuous time becomes $t_0 = \frac{N}{M}\frac{1}{f_s} = 16$ ms. We can fix ΔR at certain values and sweep the signal frequency $f_0 = \omega_0/2\pi \in [0, f_c]$ kHz. By doing so with (3) and selecting $\Delta R = 0.0004$, 0.0008, 0.0012, and 0.0016, we obtain the coherence-frequency plot in Fig. 1.

We observe from the plot that for a given ΔR, the coherence is generally inversely dependent on the signal frequency. In particular, we see that before the coherence reaches the first zero, the coherence reduction versus frequency relationship is quite linear. Furthermore, for a fixed frequency before the coherence first reaches zero, *e.g.*, $f_0 = 3$ kHz, the coherence also decreases roughly linearly as a function of ΔR. Thus (3) provides a way to measure the amount of decorrelation at each frequency point for a certain expansion ratio R. Conversely, it allows us to control R for a desired amount of decorrelation in terms of the coherence at certain frequency points, *e.g.*, to minimize the distortion of a signal at low frequencies.

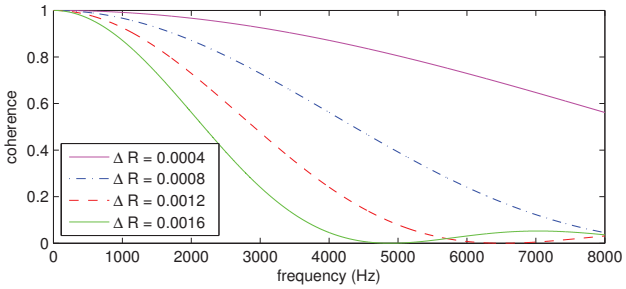


Fig. 1. Coherence-frequency plot obtained from (3).

4. TIME SCALING BY RESAMPLING

For discrete time signals, decorrelation by time expansion/compression is implemented by resampling a signal to a higher/lower sampling rate $\tilde{f}_s$ and playing back the resampled signal at the original rate f_s, where the expansion/compression ratio is related to the resampling ratio as $R = \tilde{f}_s/f_s$. Let $X_N(k)$ be the k^{th} coefficient of the N-point discrete Fourier transform (DFT) of the signal $x[n]$. Given a resampling ratio $0 < R < 2$, the procedure for resampling $x[n]$ by FDR is as follows:

- Zero-extend the signal by a factor of $M = 2^P$, $P \geq 1$.
- Perform MN-point DFT on the extended signal.
- Linearly interpolate between the k^{th} and the $(k+1)^{\text{th}}$ samples

  $$X'_{MN}(k') = R[(1-\alpha)X_{MN}(k) + \alpha X_{MN}(k+1)]$$

  with the constraints $k \leq Rk' \leq k+1$ and $\alpha = Rk' - k$ for each $(k')^{\text{th}}$ new sample to form $2N$ equally spaced samples.
- Perform $2N$-point inverse-DFT on the interpolated samples.
- Discard the samples at the end of the new signal $x'[n]$ to retain the first RN resampled values.

Using the zero-extension factor $M \geq 2$ and taking the $2N$-point inverse-DFT avoids the time domain aliasing after resampling with $R > 1$. We assume M and N to be a power of 2 in general for efficient implementation of DFT via the fast Fourier transform.

4.1. Delay Smoothing

Resampling a frame of N samples introduces the total delay of $N(R-1)$ samples, where time expansion ($R > 1$) and time compression ($0 < R < 1$) introduce positive and negative instantaneous (sub-)sample delay, respectively. Since the discrete time signal is resampled frame by frame without any overlap, there can potentially be a signal discontinuity between the frames if we do not resample each frame correctly.

Although a signal is usually resampled in one direction, *i.e.*, forward in time, it may also be resampled in the backward direction by first time-reversing the signal frame, applying the resampling procedure, and reversing the frame back afterward. Different combinations of the resampling ratio (expansion or compression) and the resampling direction (forward or backward) give rise to four possibilities: forward expansion, forward compression, backward expansion, and backward compression. The change in the delay after resampling a signal frame in four different situations are illustrated in Fig. 2, where the block dots indicate the reference (anchoring) point from which the positive/negative delay starts to grow after resampling.

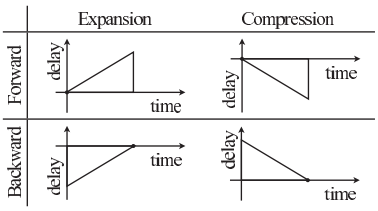
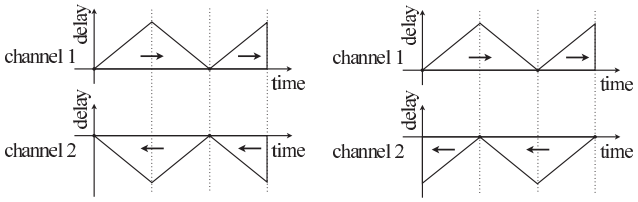


Fig. 2. Signal delay after resampling.



(a) Proposed scheme. (b) Alternative scheme.

Fig. 3. Resampling schemes.

There are basically two constraints for smooth transition between the resampled frames. First, there should be no sudden change in the delay across frames. Second, the reference points of the adjacent frames should be matched. Otherwise, a sudden change in the delay across frames edges introduces a signal discontinuity, or decimation, which in turn causes the undesirable aliasing distortion [4].

4.2. Proper Resampling Schemes

Based on the delay smoothing rules discussed above, there are several possible resampling schemes that achieve the desired decorrelation effect. One of the valid schemes was already covered in [6]. Fig. 3 shows two other schemes that obey the delay continuity constraints, where the dotted lines correspond to the frame boundaries and the arrows indicate the direction of the signal shift after either expansion or compression. In the odd frames of the proposed scheme, forward expansion occurs in channel 1 and forward compression in channel 2, whereas in the even frames backward compression occurs in channel 1 and backward expansion in channel 2. The alternative scheme performs the same process as the proposed scheme in channel 1 while shifting the operation of channel 2 by one frame.

However, although it may appear that the alternative scheme in Fig. 3 achieves the inter-channel decorrelation, it actually fails to do so and thus should be avoided. The reason is that the expansion or the compression occurs in both channels at the same time, with the only difference being resampling in forward or backward direction. That is, expanding or compressing the channels simultaneously with the same resampling ratio R near unity results in a slight shifting of the entire frames in the opposite time direction. Due to the constant amount of induced delay between two frames, the CSD is unchanged and therefore no short-time decorrelation occurs. In other words, the *instantaneous delay difference* between channels is constant in such a case. The entire process becomes much like the input-sliding technique of [4] but with no aliasing distortion at all due to the delay smoothing, hence no decorrelation. For the proposed scheme, the delay difference between channel 1 and channel 2 continuously varies with time. This specifies another design rule, where for a given time, two adjacent channels must not be expanded or compressed with the same R even if the direction of resampling is different. The rest of this paper will only focus on the proposed scheme in Fig. 3(a).

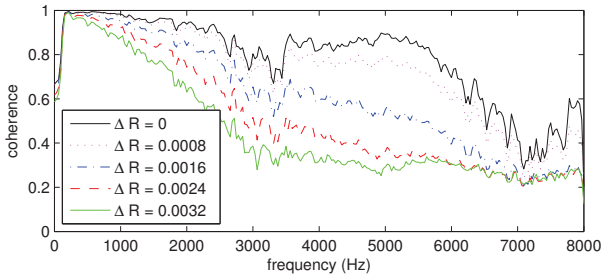


Fig. 4. Inter-channel coherence after resampling (averaged over the entire speech duration after silence removal).

Fig. 4 shows the coherence measured from the actual speech signals after the proposed resampling scheme. Although the coherence reduction is not as large as that in Fig. 1 especially at high frequencies, we observe a similar overall trend. The coherence reduction versus frequency is approximately linear at low frequencies for a fixed ΔR, whereas the coherence reduction is directly proportional to ΔR for a fixed frequency.

5. SUB-BAND RESAMPLING

For the perceptual quality and the actual cancellation performance reasons [5, 6], we may want to modify the signal only in certain sub-bands. For example, the interaural time differences plays an important role for sound localization at low frequencies [10]. A modification of the low sub-band content disturbs the phase information of the signal and ultimately alters the interaural time differences.

To that end, Figs. 1 and 4 point out that for achieving the same overall reduction in the coherence, or equivalently the cross-correlation, the resampling ratio R may be adjusted separately over each sub-band in the frequency domain as if resampling the entire signal frame with a fixed R. This can be done to make sure that the spatial image distortion will be minimized by the resampling process. In addition, a sudden change in R between sub-bands, *e.g.*, $R = 1$ in the low sub-band and $R = R_0 > 1$ in the high sub-band, may introduce the unwanted frequency-domain distortion. It was experimentally verified that the distortion created by such a discontinuity in R has the characteristics of a musical noise. Therefore, we propose to vary the resampling ratio per frequency bin as smoothly across the bins as possible, which simply involves making R a continuous function of frequency, *i.e.*, $R(k)$, and applying the desired $R(k)$ curve to the FDR procedure in Section 4.

6. PERCEPTUAL EVALUATION

To compare the processed speech quality of the proposed SBR scheme against other commonly used decorrelation techniques for MCAEC, the following procedures were tested:

- Additive white Gaussian noise (AWGN) at 25 dB signal-to-noise ratio (SNR, averaged over the entire speech data).
- Nonlinear processor (NLP) [2], given by

$$\tilde{x}_i[n] = x_i[n] + \frac{\alpha}{2}\big(x_i[n] + (-1)^{\mathrm{mod}(i-1),2}|x_i[n]|\big),$$

  where $x_i[n]$ is the reference signal from the i^{th} channel, $\mathrm{mod}(\cdot,\cdot)$ is the modulus function, and $\alpha = 0.5$.
- Phase modulation (PMod) proposed by [3].

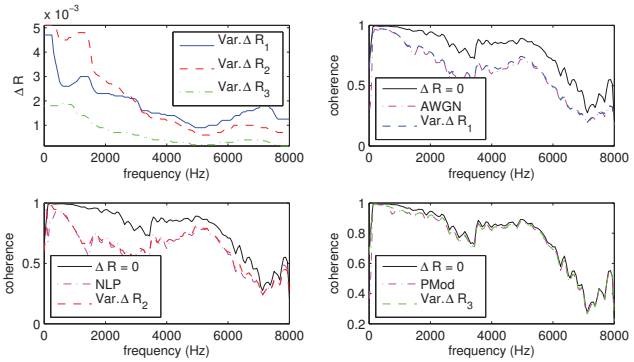


Fig. 5. Variable resampling ratios R_1, R_2, and R_3 and their corresponding coherence plots, which match the coherence from SBR to that of other decorrelation methods.

- Proposed SBR scheme with $N = 512$ and variable resampling ratios R_1, R_2, and R_3 as shown in Fig. 5.

A stereo reference signal of 30 seconds was used for the evaluation. Silences were removed prior to calculating the coherence. As SBR allows us to fine-tune the coherence at each frequency bin, R_1 is used to achieve the same coherence given by AWGN, R_2 to achieve that by NLP, and R_3 to achieve that by PMod to form the same basis for measuring the processed speech quality and comparing against other decorrelation procedures. Fig. 5 also shows how well the coherence can be controlled by SBR. Thus by properly choosing ΔR, the average degree of decorrelation, measured in terms of the coherence, by SBR can be matched to that of AWGN, NLP, and PMod.

For objective evaluation of the quality of the processed signals, segmental signal-to-noise ratio (SSNR), log-spectral distortion (LSD), and perceptual evaluation of speech quality (PESQ) score were used. SSNR measures the deviation of the processed signal from the original signal in the time domain while LSD measures that in the frequency domain. Both narrowband and wideband modes were used for the PESQ score (NB- and WB-PESQ), which is an objective measurement that predicts the results of mean opinion score (MOS) in subjective listening tests. NB-PESQ-LR and WB-PESQ-LR correspond to the evaluations obtained after averaging the measures taken individually from the left and the right channels.

Table 1 summarizes the quality of the processed speeches reflected by the objective measurements. AWGN has the worst performance in terms of the SSNR. SBR always has better SSNR than others since the delay is varied smoothly in time. The LSD from AWGN may be the smallest since the locations of the spectral peaks are unaffected while only the spectral valleys are filled with more white noise. Still, the distortion introduced by AWGN is quite audible, especially at the SNR of 25 dB. NLP leads to the largest LSD due to the non-linear processing (half-wave rectification), and the resulting frequency-domain distortion can be easily perceived when $\alpha = 0.5$. SBR produces some LSD since the frequency coefficients are modified by the resampling process, but such a distortion in the frequency domain is almost negligible for SBR when ΔR is very small. This is expected since no audible distortion should be produced after the proper resampling of a signal. Most importantly, the PESQ score clearly demonstrates the superiority of SBR. We note that PMod has the worst PESQ score as the evaluation was performed on a stereo signal, *i.e.*, possibly due to the distortion of the sound image by the phase modulation since the PESQ score is still high for each channel. SBR with R_3, on the other hand, does not have this issue as the PESQ score remains high at around 4.5. Overall, the proposed

Table 1. Processed speech quality comparison.

method	AWGN	SBR R_1	NLP	SBR R_2	PMod	SBR R_3
SSNR	8.59	**14.22**	9.06	**14.00**	5.38	**23.03**
LSD	**0.29**	0.45	2.38	**0.51**	0.42	**0.18**
NB-PESQ-LR	4.04	**4.52**	4.09	**4.50**	4.52	**4.54**
NB-PESQ	3.85	**4.37**	4.36	**4.48**	3.83	**4.54**
WB-PESQ-LR	3.64	**4.61**	3.80	**4.58**	4.61	**4.63**
WB-PESQ	3.46	**4.18**	4.07	**4.36**	2.15	**4.62**

SBR scheme introduces mostly imperceptible frequency-domain and spatial distortions to the reference signal and has the highest speech quality measures among all the other decorrelation methods while achieving the same degree of decorrelation.

7. CONCLUSION

We presented in this paper a novel approach for inter-channel decorrelation by sub-band resampling (SBR) in the frequency domain. Specifically, we are able to smoothly vary the resampling ratio per frequency bin for achieving the desired coherence in each sub-band. By enforcing the continuity in delay across frames during frame-wise resampling, we are also able to avoid any undesirable signal discontinuity. The end result is the smoothness in both the time and the frequency domains. Perceptual evaluation shows that the proposed SBR scheme delivers consistently higher signal quality after the processing than other existing decorrelation methods.

8. REFERENCES

[1] M.M. Sondhi, D.R. Morgan, and J.L. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *Signal Processing Letters, IEEE*, vol. 2, no. 8, pp. 148–151, 1995.

[2] T. Gänsler and J. Benesty, "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," *International Journal of Adaptive Control and Signal Processing*, vol. 14, pp. 565–586, 2000.

[3] J. Herre, H. Buchner, and W. Kellermann, "Acoustic Echo Cancellation for Surround Sound using Perceptually Motivated Convergence Enhancement," in *Proc. IEEE ICASSP*, 2007.

[4] A. Sugiyama, Y. Joncour, and A. Hirano, "A stereo echo canceler with correct echo-path identification based on an input-sliding technique," *Signal Processing, IEEE Transactions on*, vol. 49, no. 11, pp. 2577–2587, 2001.

[5] T.S. Wada and B.-H. Juang, "Multi-channel acoustic echo cancellation based on residual echo enhancement with effective channel decorrelation via resampling," in *Acoustic Echo and Noise Control, 2010. IWAENC '10. International Workshop on*, 2010.

[6] T.S. Wada, J. Wung, and B.-H. Juang, "Decorrelation by resampling in frequency domain for multi-channel acoustic echo cancellation based on residual echo enhancement," in *Proc. IEEE WASPAA*, 2011.

[7] T.S. Wada and B.-H. Juang, "Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement," *Proc. IEEE ICASSP*, pp. 205–208, 2009.

[8] G.C. Carter, "Coherence and time delay estimation," in *Proceedings of the IEEE*, 1987, pp. 236–255.

[9] P.D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70–73, June 1967.

[10] F.L. Wightman and D.J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.