ON THE PERFORMANCE OF THE ROBUST ACOUSTIC ECHO CANCELLATION SYSTEM WITH DECORRELATION BY SUB-BAND RESAMPLING

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

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

ABSTRACT

This paper examines the effect of inter-channel decorrelation by sub-band resampling (SBR) on the performance of the robust acoustic echo cancellation (AEC) system based on the residual echo enhancement technique. Due to the flexibility of SBR, the decorrelation performance as measured by the coherence can be matched with other conventional decorrelation procedures. Given the same degree of decorrelation, we have shown previously that SBR achieves superior audio quality compared to other procedures. We show in this paper that SBR also provides higher stereophonic AEC performance in a very noisy condition, where the performance is evaluated by decomposing the true echo return loss enhancement and the misalignment per sub-band to better demonstrate the superiority of our decorrelation procedure over other methods.

Index Terms— Inter-channel decorrelation, resampling, multi-channel acoustic echo cancellation, residual echo enhancement

1. INTRODUCTION

We have proposed recently a novel approach for inter-channel decorrelation procedure by sub-band resampling (SBR) [1] to alleviate the non-uniqueness problem that arises during multi-channel acoustic echo cancellation (MCAEC) due to the highly correlated reference signals. Figure 1 shows a conventional stereophonic AEC (SAEC) setup with a decorrelation procedure inserted before loudspeaker playback to mitigate the problem. SBR is an extension of the decorrelation by resampling (DBR) technique [2, 3] that introduces time-varying delay across channels with minimal distortion. Instead of fixing the resampling ratio in DBR, SBR allows selective variation of the resampling ratio across frequencies to finely control the amount of decorrelation as measured by the coherence [4].

Unlike the traditional decorrelation procedures [5–7] where the inter-channel coherence is usually fixed throughout the frequencies for a given parameter, SBR is highly flexible and can be fine-tuned for better perceptual quality, *e.g.*, less "resampling" at lower frequencies and vice versa at higher frequencies. We have shown that given the same degree of decorrelation, SBR outperforms the conventional methods in terms of the audio quality [1]. In light of such advantages of SBR over other decorrelation procedures, the effect of SBR on the performance of MCAEC has not yet been fully addressed, although DBR has been shown to be superior to other procedures in terms of the echo path tracking performance [3]. This motivates us to further investigate the effect of SBR on the MCAEC performance.

We examine in this paper the effect of SBR on the performance of our robust acoustic echo cancellation (R-AEC) system based on the residual echo enhancement (REE) technique [8]. The overall goal is to achieve the decorrelation process for MCAEC through the *system approach*, *i.e.*, proper integration of individual components, or algorithms, for mutual interaction to benefit the system as a whole. That is, all the signal enhancement problems should be best dealt with as a system issue, unlike the conventional single-algorithm solutions with limited real-world applicability. Specifically our aim is to integrate the decorrelation procedure not simply as a separate preprocessor but as a part of the AEC system, capable of controlling both the echo cancellation and the tracking performances while introducing the least amount of audio distortion possible. To that end, we evaluate the "true" (*i.e.*, noise-free) echo return enhancement (tERLE) and the misalignment through sub-band decomposition and demonstrate the superiority of our approach over other methods.

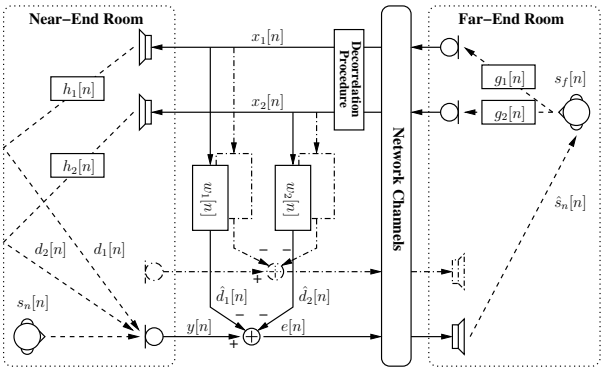

Fig. 1. A conventional stereophonic AEC (SAEC) setup.

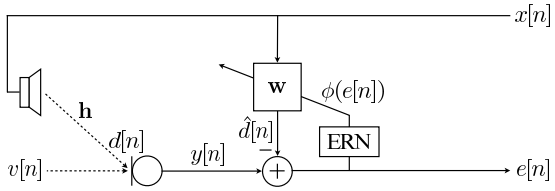

Fig. 2. A robust AEC (R-AEC) system with an adaptive filter $\mathbf{w}$ and the error recovery nonlinearity (ERN).

2. ROBUST ACOUSTIC ECHO CANCELLATION

A single-channel R-AEC system with REE is illustrated in Figure 2. Let y be the near-end microphone signal, which consists of the near-

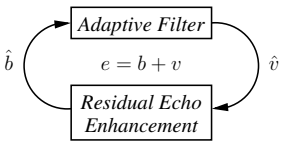

Fig. 3. System integration of adaptive filter and residual echo enhancement that make up the R-AEC component.

end noise or speech v mixed with the acoustic echo $d = \mathbf{h}^{\mathrm{T}}\mathbf{x}$, where $\mathbf{h}$ is the room impulse response (RIR) vector (a truncated version of the actual impulse response), $\mathbf{x}$ is the far-end reference signal vector, and $\{\cdot\}^{\mathrm{T}}$ is the transposition operator. The adaptive filter coefficients vector $\mathbf{w}$ models the RIR, and the filtered output $\hat{d} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ approximates the echo d. The observed estimation error e of the AEC is given by

$$\begin{aligned} e[n] &= d[n] - \hat{d}[n] + v[n] \\ &= b[n] + v[n], \end{aligned}$$

where b is the *true* error (residual echo) that results from the misalignment between the RIR and the adaptive filter coefficients, *i.e.*, $(\mathbf{h}^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}})\mathbf{x}$. By "true" we mean a noise-free quantity, *i.e.*, $v = 0$. However, strong v during double talk, for example, may corrupt the estimation error and cause the adaptive filter to diverge. The error recovery nonlinearity (ERN) reduces such a disturbance remaining in the estimation error and enables the adaptive filter to better estimate the linear part of the system response, where block-iterative adaptation (BIA) (*i.e.*, batch adaptation) permits the recovery of lost convergence speed due to the aggressive step-size control [8].

The REE technique is represented systematically in Figure 3 that recursively refines the estimations $\hat{b}$ and $\hat{v}$ contained in the corrupted residual echo $e = b+v$. This built-in dual re-estimation process, carried out via BIA, makes the R-AEC component less sensitive to misspecification of the system parameters and mis-estimation of the signal statistics [8]. The system approach, for which the REE paradigm is just one realization, enables the encompassing of many traditional signal enhancement techniques in analytically consistent yet practically effective manner for solving the enhancement problem in a very noisy and disruptive acoustic mixing environment.

3. SYSTEM ASPECT OF DECORRELATION

As illustrated in Figure 1, the acoustic echo estimate at one of the microphones for MCAEC with P channels is give by

$$\hat{d}[n] = \mathbf{w}^{\mathrm{T}}[n]\mathbf{x}[n] = \sum_{i=0}^{P-1} \mathbf{w}_i^{\mathrm{T}}[n]\mathbf{x}_i[n],$$

were $\mathbf{w} = [\mathbf{w}_1^{\mathrm{T}}, \ldots, \mathbf{w}_P^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{x} = [\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_P^{\mathrm{T}}]^{\mathrm{T}}$ are the concatenated filter coefficient and reference signal vectors, respectively. Thus MCAEC is traditionally approached as a single-channel problem, which consequently leads to the non-uniqueness problem for the least-square-based adaptive algorithms along with other noise-robustness issues associated with single-channel AEC.

The least mean square (LMS) algorithm iteratively and stochastically solves the normal equation

$$\mathbf{R}_{\mathbf{xx}}\mathbf{w} = \mathbf{r}_{\mathbf{x}y},$$

where $\mathbf{R}_{\mathbf{xx}} = E\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ is the auto-correlation matrix of $\mathbf{x}$ and $\mathbf{r}_{\mathbf{x}y} = E\{\mathbf{x}y\}$ is the cross-correlation vector between $\mathbf{x}$ and y. For such a dynamic solution, a mismatch in the sampling rate between $\mathbf{x}$ and y on the order of merely 0.01% is enough to break down the correlation structure of $\mathbf{r}_{\mathbf{x}y}$ for significant decrease in the LMS algorithm's cancellation performance [9]. Conversely, we should be able to induce the same effect for the decorrelation purpose, *i.e.*, to improve the conditioning of $\mathbf{R}_{\mathbf{xx}}$, by resampling x frame-wise. This close and dynamic systematic relationship between the sampling rate and the LMS algorithm is one major reason for DBR's effectiveness as revealed in [2, 3].

One other crucial system aspect is that BIA enables a natural recovery of the convergence speed and hence reduces the need for aggressive decorrelation applied directly to x, which subsequently minimizes the audio distortion and also the potential interference with the adaptive cancellation process. We have observed that the DBR and the R-AEC components complement one another such that low coherence, or equivalently low misalignment, over the entire frequency range is not necessary for high cancellation performance during MCAEC. Conventional wisdom instead favors as large decrease in the coherence as possible over all frequencies, which most likely causes the degradation of the actual cancellation performance [2, 3].

4. DECORRELATION BY SUB-BAND RESAMPLING

Decorrelation by time expansion/compression is implemented by resampling a signal to a higher/lower sampling rate $\bar{f}_s$ and playing back the resampled signal at the original rate f_s, where the expansion/compression ratio is related to the resampling ratio as $R = \bar{f}_s/f_s$. 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. For best audio quality, however, the resampling ratio is adjusted smoothly across frequencies, which simply involves making R a continuous function of frequency, *i.e.*, $R(k)$.

Let $X_N(k)$ be the k^{th} coefficient of the N-point discrete Fourier transform (DFT) of the signal x. Given a resampling ratio $0 < R(k) < 2$, the procedure for resampling x by frequency-domain resampling (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

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

  with the constraints $k \leq R(k)k' \leq k+1$ and $\alpha = R(k)k' - k$ for each $(k')^{\text{th}}$ new sample to form $2N$ equally spaced samples.
- Perform $2N$-point inverse-DFT on the interpolated samples.
- Discard the appropriate amount samples at the end of the new signal x' based on the function $R(k)$.

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

Figure 4 illustrates the resampling scheme that is used in this paper (refer to [1] for more detail). The continuity in the delay across resampled frame ensures better audio quality, as it eliminates the distortion due to signal decimation. Many other framing options are possible to introduce varying decorrelation effect across time and channels suitable for directly assisting the LMS adaptation process.

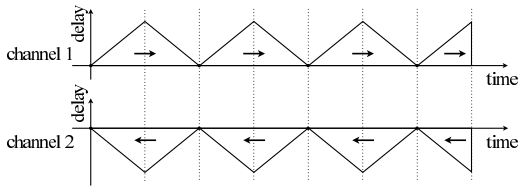

Fig. 4. The resampling scheme used in this paper.

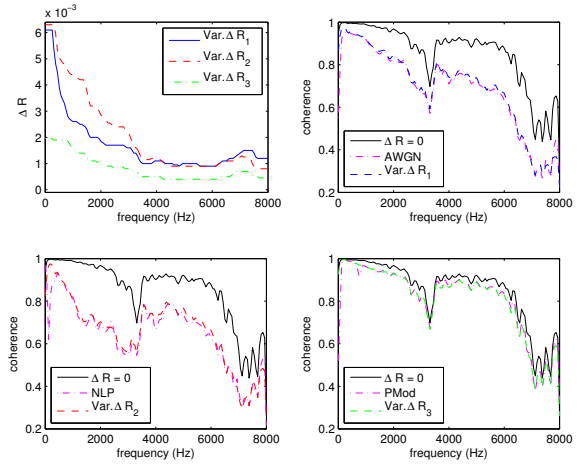

Fig. 5. Variable resampling ratios R_1, R_2, and R_3 and their corresponding coherence plots, which match the coherence from SBR to that of other decorrelation methods.

5. SIMULATION RESULTS

To compare the R-AEC performance with the proposed decorrelation by SBR against other commonly used decorrelation procedures, the following methods were tested:

- Additive white Gaussian noise (AWGN) at 15 dB segmental signal-to-noise ratio (SSNR).
- Nonlinear processor (NLP) [5], 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 [7].
- FDR with fixed resampling ratio $R = 1.0028$.
- SBR with $N = 512$ and variable resampling ratios R_1 to R_5.

5.1. Quality Evaluation

A stereo reference signal of 30 seconds was used for the evaluation of speech quality after decorrelation. Silences were removed prior to calculating the coherence. Figure 5 shows how well the coherence can be controlled by SBR, where R_1 is adjusted at each frequency bin 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. Thus by properly choosing $\Delta R = R - 1$, the average degree of decorrelation, measured in terms of the coherence, by SBR can be matched to that of AWGN, NLP, and PMod. Also to demonstrate the ability of SBR to control the AEC performance per sub-band, the coherence is matched with regular FDR only in the mid to high bands while leaving the low band unmodified. The two other SBR coherence-matching schemes with the variable resampling ratios R_4 and R_5 used for this purpose are shown in Figure 6.

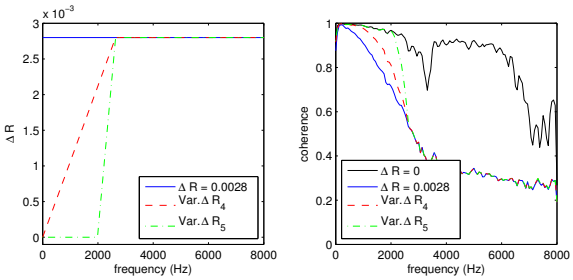

Fig. 6. FDR with fixed $\Delta R = 0.0028$ and the corresponding variable resampling ratios R_4 and R_5 that match the coherence of FDR in the mid to high frequency bands.

Table 1. Processed speech quality comparison.

method	AWGN	R_1	NLP	R_2	PMod	R_3	FDR	R_4	R_5
SSNR	**15.00**	8.68	3.24	**7.22**	1.48	**13.62**	6.62	8.39	**9.51**
LSD	**0.07**	0.51	2.45	**0.66**	0.37	**0.24**	0.79	0.66	**0.59**
$\text{PESQ}^{\text{NB-LR}}$	3.67	**4.50**	4.03	**4.49**	4.53	**4.53**	4.52	4.55	**4.55**
PESQ^{NB}	3.57	**4.51**	4.39	**4.51**	3.94	**4.55**	**4.25**	4.24	4.24
$\text{PESQ}^{\text{WB-LR}}$	3.56	**4.59**	3.76	**4.57**	4.62	**4.63**	4.61	4.63	**4.63**
PESQ^{WB}	3.52	**4.48**	3.96	**4.47**	2.56	**4.63**	3.73	3.74	**3.74**

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

Table 1 summarizes the quality measures. SBR generally outperforms the other conventional methods in terms of the sound quality as reflected by the PESQ score [1]. We note that even though the SSNR and the LSD of AWGN are better than SBR with R_1, the distortion introduced by AWGN is quite audible as indicated by the PESQ score. The distortion by SBR, on the other hand, is almost negligible when ΔR is very small. We also note that the resampling ratios in the high band for FDR, SBR with R_4, and SBR with R_5 are set to large values to demonstrate the effect of highly decorrelated signal after DBR on the tERLE and the misalignment performances. As a result, the PESQ^{WB} score suffers due to large distortion in the high frequency region. Still, even though the PESQ scores are quite similar in those cases, better SSNR and LSD can be achieved by avoiding the resampling of the low band as reflected by SBR with R_4 and R_5.

5.2. SAEC Evaluation

The same measured RIRs from [2, 3] were used to simulate SAEC. The number of talkers, simulated with TIMIT speech corpus, at each

Table 2. Average tERLE (dB, higher is better).

band	*none*	AWGN	SBR R_1	NLP	SBR R_2	PMod	SBR R_3
low	*24.5*	**24.4**	23.3	22.6	**23.1**	21.5	**24.1**
mid	*20.1*	19.8	19.8	18.7	**19.7**	18.7	**19.4**
high	*17.1*	15.5	**17.5**	14.3	**17.3**	**17.3**	15.8
all	*23.8*	**23.7**	22.9	22.2	**22.7**	21.1	**23.3**

Table 3. Average misalignment (dB, lower is better).

band	*none*	AWGN	SBR R_1	NLP	SBR R_2	PMod	SBR R_3
low	*-8.1*	-8.0	**-9.5**	-9.9	**-10.1**	**-9.6**	-8.5
mid	*-10.9*	-11.1	**-16.8**	-13.6	**-16.8**	**-18.7**	-12.1
high	*-7.9*	-8.3	**-15.7**	-10.5	**-15.3**	**-15.5**	-9.8
all	*-8.4*	-8.7	**-13.7**	-10.7	**-13.8**	**-13.6**	-9.9

Table 4. Average tERLE (dB, higher is better).

band	*none*	FDR	SBR R_4	SBR R_5
low	*24.5*	23.6	24.2	**24.3**
mid	*20.1*	21.7	21.7	21.7
high	*17.1*	21.3	21.3	21.3
all	*23.8*	23.7	24.2	**24.3**

Table 5. Average misalignment (dB, lower is better).

band	*none*	FDR	SBR R_4	SBR R_5
low	*-8.1*	**-9.4**	-8.7	-8.1
mid	*-10.9*	-21.7	-21.7	-21.7
high	*-7.9*	-21.0	-21.0	-21.0
all	*-8.4*	**-15.8**	-15.2	-14.8

end was set equal to that of microphones and loudspeakers. Talkers randomly took turns to speak exactly one utterance per sequence, where at most two spoke simultaneously with an overlap of 2 seconds at each end and 1 seconds between both ends (*i.e.*, double-talk duration), and the pattern was repeated for 30 seconds. The near-end RIRs were switched at 15 seconds to enact a sudden disruption to the RIR. The RIRs were truncated to 128 ms before convolution. The near-end RIRs were scaled to produce the echo return loss (ERL) of 10 dB, where the signal energy after decorrelation was normalized to match that of the original to ensure consistent ERL control. White Gaussian noise with 100 dB and 40 dB signal-to-noise ratios (SNRs) were added to x and d, respectively. Air-conditioner noise and speech with the echo-to-noise ratios (ENRs) of 20 dB and 0 dB, respectively, were also added to d.

For the SAEC performance comparison, the tERLE and the misalignment were decomposed into three sub-band components (low, mid, and high) through the Fourier series expansion (which gives far better reconstruction accuracy than using the DFT filter banks formed from a prototype filter). For tERLE, y and e were decomposed with 50% overlap of the analysis frames. For misalignment, $\mathbf{h}$ and $\mathbf{w}$ were mirrored in time and concatenated to extend their size by a factor of two prior to the decomposition.

Tables 2 and 3 show the SAEC results (averaged only over echo duration for tERLE and over entire time for misalignment) from the REE-based frequency-block LMS algorithm with $\mu = 0.12$, $\beta = 0.998$, $\gamma = 10$, $\eta = 5$, $iter = 4$, $B = L$, and the overlap factor $of = 4$ ($of = 1$ was used in [2, 3]) along with exponential weighting and scaling of the step-size μ by half during double talk [2]. First, although AWGN is able to provide the tERLE closer to when no decorrelation is used than SBR, it leads to much worse misalignment. Second, NLP tends to hurt the low-band tERLE more than SBR when compared to no decorrelation. The performance gain by SBR against NLP is even larger in the mid and the high bands for the tERLE and especially for the misalignment. Finally, PMod is capable of providing lower misalignment over all bands than SBR when its coherence is matched by SBR, but it does not necessarily translate to higher tERLE, which is less in the low to mid bands for PMod than SBR. Poor misalignment by SBR in this case is expected since the coherence is not reduced much after the matching.

The results in Tables 4 and 5 indicate that a substantial gain in the tracking capability appears in the mid to high bands for FDR and SBR when compared to no decorrelation and other decorrelation procedures. The tERLE is also increased especially in the high band. Such an improvement is attributed largely to the DBR's ability to continuously instill both short and long-time decorrelation for a direct benefit of the LMS algorithm and not simply due to the coherence reduction. Furthermore, SBR is able to provide higher tERLE than FDR in the low band by selectively not modifying the signal components in that region, in which case the tERLE is recovered naturally through BIA of the REE technique. This results in higher overall tERLE for SBR than without decorrelation.

6. CONCLUSION

We applied sub-band resampling (SBR) for inter-channel decorrelation with the robust AEC system based on the residual echo enhancement (REE) technique to evaluate the stereophonic AEC performance through the sub-band analysis of the true echo return enhancement (tERLE) and the misalignment. Simulation results show that a significant recovery of the performance lost due to the non-uniqueness problem in the mid to high bands is possible via a combination of SBR and REE's block-iterative adaptation (BIA) and that SBR allows selective decorrelation to maintain high tERLE in the low band naturally through BIA without the need for aggressive decorrelation. SBR thus provides the flexibility to leave the low band untouched and only resample the high band according to the desired AEC performance and the sound quality requirement.

7. REFERENCES

[1] J. Wung, T.S. Wada, and B.-H. Juang, "Inter-channel decorrelation by sub-band resampling in frequency domain," *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, Mar. 2012.

[2] T.S. Wada and B.-H. Juang, "Multi-channel acoustic echo cancellation based on residual echo enhancement with effective channel decorrelation via resampling," *Acoustic Echo and Noise Control (IWAENC), 2010 International Workshop on,*, Sept. 2010.

[3] 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," *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 289–292, Oct. 2011.

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

[5] 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.

[6] 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.

[7] J. Herre, H. Buchner, and W. Kellermann, "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2007 IEEE International Conference on*, Apr. 2007, pp. 17–20.

[8] T.S. Wada and B.-H. Juang, "Enhancement of residual echo for robust acoustic echo cancellation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 175–189, Jan. 2012.

[9] E. Robledo-Arnuncio, T.S. Wada, and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2007 IEEE Workshop on*, pp. 34–37, Oct. 2007.