MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION BASED ON RESIDUAL ECHO ENHANCEMENT WITH EFFECTIVE CHANNEL DECORRELATION VIA RESAMPLING

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

Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA, 30332, USA

{twada,juang}@ece.gatech.edu

ABSTRACT

The residual echo enhancement (REE) procedure proposed in [1] is able to improve the acoustic echo cancellation (AEC) performance in a very noisy acoustic mixing environment by utilizing the natural learning ability of the least-mean square (LMS) algorithm without precise estimation of the signal statistics. We demonstrate in this paper that the technique can also be applied effectively in a multi-channel AEC (MCAEC) setting to indirectly assist in the recovery of lost AEC performance due to the non-uniqueness problem. In addition, we incorporate other techniques to further boost the REE-based MCAEC performance. One of the techniques is a new channel decorrelation procedure based on resampling that directly alleviates the non-uniqueness problem while introducing minimal distortion to signal quality and statistics.

Index Terms— residual echo enhancement, multi-channel acoustic echo cancellation, semi-blind source separation, non-uniqueness problem, sampling rate mismatch

I. INTRODUCTION

There are two main issues that plague the least-mean square (LMS) algorithm when applied to multi-channel acoustic echo cancellation (MCAEC) (see Fig. 1 for stereo AEC (SAEC)). First, the LMS algorithm by itself has difficulty converging to the optimal solution in the presence of local noise (e.g., double-talk) since each noisy sample directly perturbs the single-sample estimate of the mean-square error (MSE) gradient $\nabla_{\mathbf{w}} E[e^2] = -2E[e\mathbf{x}] \approx -2e\mathbf{x}$, thereby leading to "noisy" update of the filter coefficients vector $\mathbf{w}$. A traditional MCAEC strategy performs single-channel MSE optimization, hence the MSE-based MCAEC inherits the same noise-sensitivity of a single-channel counterpart. Second, the non-uniqueness problem [2] occurs in a MCAEC framework when the filter length L is longer than or equal to the far-end room impulse response length M [3], where the near-end echo path solution depends also on the far-end room acoustic characteristics. Although such a condition is very rare since in reality the acoustic impulse response is infinite in length, the high correlation between the reference signals (i.e., far-end microphone signals) and the effect of near-end noise would cause the convergence rate to decrease so greatly that the solution practically behaves as if non-unique. The far-end room dependency remains likewise whether or not $L \geq M$.

The noise-robustness issue can be effectively solved by residual echo enhancement (REE) [1] that applies a noise-suppressing memoryless nonlinearity to "enhance" the filter estimation error before updating the filter coefficients (see Fig. 2). Both the steady-state and the convergence behaviors of the LMS algorithm are improved significantly through REE and multiple recursive filtering and adaptation on a batch of noisy data (in contrast to the usual sample-wise or block-online adaptation). However, while the REE technique may be readily extended to MCAEC, it does not directly address the non-uniqueness issue to reduce the dependency of the Wiener solution on the far-end room response. Some form of decorrelation must still be applied to the reference channels before playback and adaptation (see Fig. 1) for improved tracking of echo path changes but often with a side effect of audible distortion.

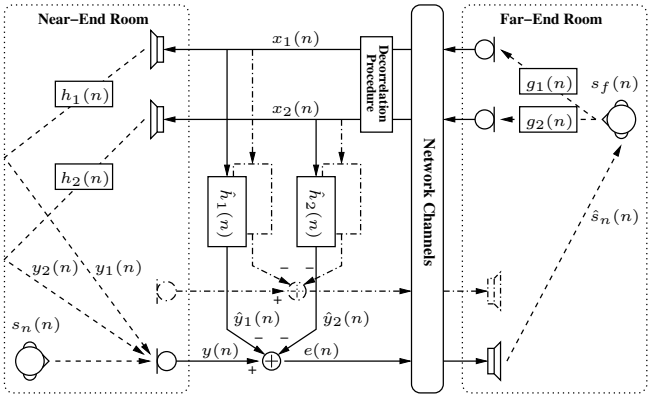

Fig. 1. Conventional stereo AEC setup. The MSE optimization is performed separately in each microphone channel.

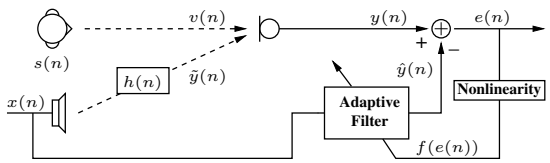

Fig. 2. REE with a noise-suppressing memoryless nonlinearity in the error-feedback loop to reduce the effect of a local noise v.

Key criteria for an ideal decorrelation procedure are as follows.

- Retains original audio quality and image of far-end sources.
- Retains original excitation characteristics of echo paths.
- Retains original signal statistics used for adaptive filtering.
- Extendable to large number of channels.
- Requires low computational complexity.

Many conventional techniques, e.g., a nonlinear "half-wave rectifying" processor [4] and comb filtering [5], do not entirely satisfy the first two requirements. They may not likely meet the third condition necessary for optimal steady-state performance by an MSE-based adaptive filter, and they also tend to be incompatible for the case of more than two audio channels.

We demonstrate in this paper that the REE technique can be used to improve the MCAEC performance in very noisy acoustic conditions as it allows the recovery of lost convergence rate due to both the noise-robustness control enforced by REE and the non-uniqueness problem. Several ways to further boost the performance are also presented, one of which is a novel decorrelation procedure based on resampling that exploits the effect of sampling rate mismatch examined in [6]. The new decorrelation procedure shares essential features from successfully implemented schemes in [7] and [8] while satisfying all of the five requirements listed above.

II. RESIDUAL ECHO ENHANCEMENT

REE may be viewed as a generalization of adaptive step-size and regularization procedures for non-Gaussian signals characterized by not only second but also higher-order statistics [1]. The technique can be derived directly from the natural gradient algorithm and independent component analysis (ICA). More importantly, a loss in the convergence rate incurred as a tradeoff for a gain in the adaptation stability can be compensated by batch-wise adaptation, usually reserved for ICA learning, without precise estimation of the signal statistics. The REE procedure is a special case of semi-blind source separation (SBSS) [9] without the source separation. Another major difference from SBSS is that the optimization is performed per microphone channel as opposed to simultaneous optimization across all channels during SBSS. Simply put, the combined technique takes advantage of the inherent ability of the LMS algorithm in converging to the optimal solution as it instills just enough noise-robustness control to consistently maintain stability.

III. NON-UNIQUENESS PROBLEM

Let $y_i(n) = \sum_j \sum_k h_{ij}(k) x_j(n-k) = \sum_j \mathbf{h}_{ij}^T \mathbf{x}_j(n)$ be the noise-free recording from i^{th} microphone, where "T" denotes vector transposition, $\mathbf{x}_j(n)$ is the reference vector from j^{th} loudspeaker, $\mathbf{h}_{ij}$ is the time-invariant room response vector, $1 \le i \le P$, $1 \le j \le Q$, and $0 \le k \le N-1$. Assuming $L = N$, a set of filter coefficients corresponding to the echo paths between all Q loudspeakers and i^{th} microphone is obtained from the normal equation $\mathbf{R}\mathbf{w}_i = \mathbf{r}_i$, where $\mathbf{R} = \{E[x_j(n-k)x_{j'}(n-k')]\}$ is an $LQ \times LQ$ matrix, $\mathbf{w}_i = \{w_{ij}(k)\}_i$ and $\mathbf{r}_i = \{E[y_i(n)x_{j'}(n-k')]\}_i$ are $LQ \times 1$ vectors, and $E[\cdot]$ is the expectation operator. The equation indicates that even if the uniqueness condition of $L < M$ is met (i.e., $\mathbf{x}_j$ is linearly independent of $\mathbf{x}_{j'}$ for $j \ne j'$ [3]), the problem remains ill-conditioned if $E[x_j x_{j'}] \ne 0$. The convergence behavior of a stochastic gradient descent algorithm is then assisted by a decorrelation procedure $\phi(\cdot)$ such that $E[\phi(x_j)\phi(x_{j'})] \approx 0$ for $j \ne j'$. Still, any extra processing, linear or nonlinear, will inevitably change the statistics of non-stationary random processes $x_j(n)$ and $x_{j'}(n)$ and modify the steady-state (or near steady-state) solution, where the effect may be significant for the LMS algorithm that uses a very rough estimate of the gradient. A decorrelation procedure should be designed to minimize such an effect.

IV. DECORRELATION BY RESAMPLING

A very small mismatch in the sampling rate between audio channels of few hundred parts per million ($\sim 0.01\%$) is enough to break down the correlation structure necessary for sufficient AEC performance [6]. Conversely, we should be able to induce a similar effect on the highly correlated reference signals by resampling them to instead improve the MCAEC performance while minimizing the distortion on signal quality and statistics.

Let f_1 and f_2 be the current and the new sampling rates, respectively. The resampling ratio is defined as

$$R = \frac{f_2}{f_1} = 1 + r, \tag{1}$$

where the mismatch ratio is defined as $r = f_\Delta / f_1$, $f_\Delta = f_2 - f_1$. Assuming WOLOG a real-valued $R > 1$ (or $r > 0$), sampling rate expansion gives the identity relationship

$$x(nR^{-1}) \longleftrightarrow X(Z^R), \tag{2}$$

where $x(n)$ and $X(Z)$ are the discrete time sequence of a continuous time signal $x(t)$ and the corresponding Z-transform, respectively, after which the upsampled signal is obtained by lowpass filtering (interpolation) [10]. By equating $x(n-d) = x(nR^{-1})$,

$$d = n\left(\frac{r}{1+r}\right), \tag{3}$$

which is the fractional delay of expanded samples with respect to the original samples. Thus after upsampling, (2) implies spectral warping (i.e., frequency-dependent modulation) is applied to the original signal, and (3) means the delay grows progressively in time (i.e., samples gradually accumulate over the current time scale).

A time-varying phase shift in subbands was applied as a decorrelation procedure for MCAEC in [7] with larger modulation at higher bands to perceptually hide the signal distortion after synthesis, whereas one-sample delay was inserted periodically across channels into frames with half delay period per frame and quarter-period shifting during SAEC in [8]. We propose combining the resampling approach with the alternating projection technique of [11]. Such a combination takes on key features from [7] and [8] as it periodically imparts smoothly increasing modulation (in frequency) and delay (in time) across channels. The main drawback is the computational cost of resampling at the rate $R \simeq 1$ ($r \simeq 0$), which requires very large integer-valued resampling ratios for the ideal upsampling and downsampling scheme. Such a problem can be solved by the resampling-by-interpolation strategy proposed in [6], which drastically reduces the computation time by omitting the downsampling process and reusing a short interpolation filter (sinc function) per block. Therefore, the decorrelation is achieved simply by lowpass filtering in an appropriate manner.

V. COMBINED MCAEC PROCEDURE

Just as in [1], frequency-block LMS (FBLMS) [12] is combined with the REE procedure using a "compressive" nonlinearity for double-talk robustness and the regularized normalization factor [13]

$$\frac{S_{x_j}(k,l)}{S_{x_j}^2(k,l) + \gamma S_{v_i}^2(k,l)} \tag{4}$$

for stability when the echo path is weakly excited, where the reference and the noise power spectrums S_{x_j} and S_{v_i} for k^{th} frequency bin at l^{th} block index are determined per j^{th} and i^{th} channels, respectively. The same simplified statistics estimation strategy in [1] is utilized, where S_{v_i} in (4) is estimated directly by the residual echo power spectrum S_{e_i} and the over-suppression factor $\eta \ge 1$ is employed this time with REE such that the signal-to-noise ratio (SNR) = η for improved stability (η = 1 in [1]).

The following modifications are included for extra improvement in the overall MCAEC performance. First, double-talk detection (DTD) [14] is used to decrease the step-size by half during double-talk to maintain as much stability as possible. Second, exponential weighting (EW) [15] is applied to the time-domain filter coefficients during the FBLMS's gradient constraint procedure for increased convergence rate and also adaptation stability. Finally, a truly batch-wise adaptation is carried out for a batch of B samples, where the filtering and the adaptation steps are performed per block size L (same as the filter size) and repeated across blocks of $L < B$ in the same batch for *iter* iterations ($B = L$ in [1]).

In order to verify the decorrelation-by-resampling idea, the following decorrelation procedures are tested with simulated SAEC.

- Nonlinear processor (NLP) [4].
- Additive white Gaussian noise (AWGN).
- One-sample delay (OSD) [8].
- Resampling by upsampling and downsampling (RUD).
- Resampling by interpolation (RBI) [6].

The last three methods are applied to non-overlapping frames in each channel. In order to eliminate the audible "pops" between the processed frames due to discontinuity, the following simplified smoothing schemes are used, where $x_j(m,n)$ and $\hat{x}_j(m,n)$ are the original and the resampled values, respectively, of n^{th} sample in m^{th} (current) frame of size N_f from j^{th} loudspeaker channel.

- OSD: The average of $x_j(m-1, N_f)$ and $x_j(m,1)$ is inserted between the two samples to create one-sample delay. In order to avoid the accumulation of delay, $x_j(m, N_f)$ is overlapped and averaged with $x_j(m+1, 1)$.

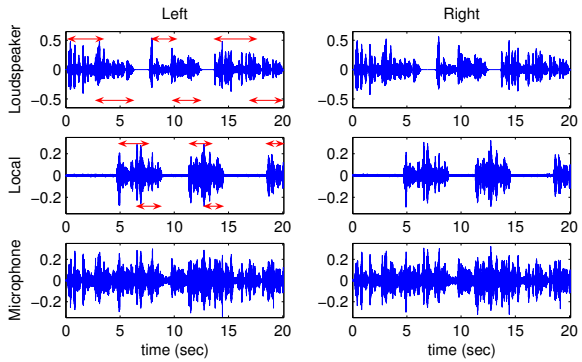

Fig. 3. Near-end loudspeaker, local, and microphone signals. Double-arrows indicate individual speech activity.

- RUD: $R > 1$ is chosen such that one extra sample is produced by resampling. Afterward, $\hat{x}_j(m, 1)$ is averaged with $x_j(m-1, N_f)$, and the extra sample $\hat{x}_j(m, N_f+1)$ is overlapped and averaged with $x_j(m+1, 1)$.
- RBI: After resampling-by-interpolation that produces the same number of samples as before, $\hat{x}_j(m, 1)$ is averaged with $x_j(m-1, N_f)$, and $\hat{x}_j(m, N_f)$ is averaged with $x_j(m+1, 1)$.

Only one extra sample delay (look-ahead) is incurred by the above strategies for real-time playback. More advanced framing and smoothing are possible, e.g., [8], albeit with longer delay.

VI. SAEC SIMULATION RESULTS

A pair of microphones 2 cm apart were placed 50 cm away from the middle of a pair of loudspeakers 50 cm apart (P = Q = 2). The configuration was used to record three sets of impulse responses, two for the near- and far-end talkers and one for the near-end loudspeakers, with average reverberation time of T_{60} = 250 ms. Two talkers (male and female speeches sampled at 16 kHz) were placed at both ends, where at most two talkers at either end were simultaneously speaking with overlap of about two seconds (see Fig. 3). The impulse responses were truncated to 128 ms (L = M = N = 2048) before convolution, and the near-end impulse responses were scaled to produce an echo return loss (ERL) of 10 dB. 40 dB SNR AWGN was applied to the far-end microphone signals, and an air-conditioner noise and local speeches (double-talk) with echo-to-noise ratios of 20 dB and 0 dB, respectively, were mixed with the acoustic echo to comprise the near-end microphone signals. The signal energy after decorrelation was normalized to match that of the original for as equal ERL for all cases as possible. α = 0.15, β = 0.99, γ = 1 [1], and η = 5 were used for FBLMS and REE.

Fig. 4 indicates that a batch-wise adaptation, permitted by the REE procedure in a noisy environment, accelerates the convergence rate significantly especially at the beginning of adaptation (in contrast to the common approach of simply using a fast-converging adaptive algorithm to combat the effect of the non-uniqueness problem). Using DTD to decrease the step-size during double-talk assists in increasing the true ERLE (tERLE), i.e., the echo return loss enhancement (ERLE) calculated without the local noise. The effectiveness of EW during both single- and double-talk is also observed. Continuous, noise-robust adaptation afforded by REE is crucial for MCAEC since the far-end room response change may occur during double-talk, e.g., far-end speech activity switches at $t \approx 7.5$ sec in Fig. 3.

However, Fig. 5 reveals that only a minor improvement in the overall misalignment is possible without a decorrelation procedure. The effect the far-end speech activity transition is clearly visible at $t \approx 2.5$ sec in Fig. 6 without decorrelation even when all other techniques are employed. Some improvement is displayed after decorrelation in Figs. 7, 8, and 9 for NLP (the nonlinearity

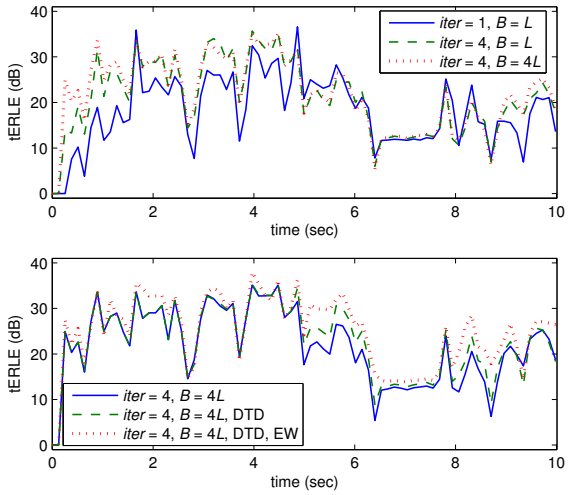

Fig. 4. True ERLE (averaged over left and right channels) for combinations of *iter*, B, DTD, and EW without decorrelation.

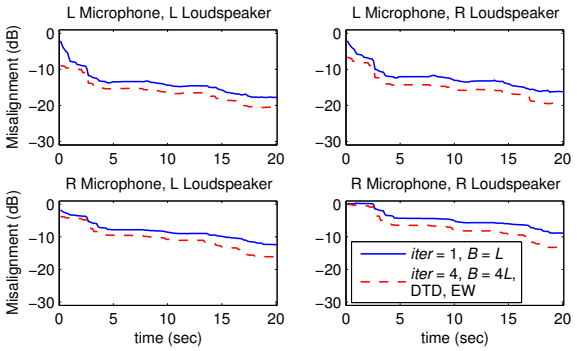

Fig. 5. Improvement in misalignment without decorrelation.

parameter was set at 0.5 [4]), AWGN (30 dB SNR), and OSD (N_f = L), respectively, but with limitations, e.g., degradation of the near steady-state performance by NLP is quite apparent.

On the other hand, Figs. 10 and 11 confirm the effectiveness of the resampling technique. RBI provides virtually the same results as RUD (`MATLAB`'s `resample` function was used for RUD, the reuse-block size and `sinc` function length of 64 was used for RBI [6], and N_f = L and R = 1.0004 were used for both). Informal listening tests indicated no loss in perceptual quality after decorrelation by RUD or RBI, whereas remaining distortion in the residual echo was obvious for NLP and AWGN. Remnants of discontinuity may still be noticeable to a very attentive listener with a headset, but they can be easily smoothed out further by using more samples.

VII. CONCLUSION

We successfully applied the residual echo enhancement (REE) technique to stereo AEC in very noisy acoustic conditions. Other traditional techniques, such as double-talk detection and exponential weighting, were integrated into the AEC system to further improve the noise robustness and the overall cancellation performance. We also proposed a new decorrelation procedure with minimal signal distortion via resampling that effectively alleviates the non-uniqueness problem. The combined approach is computationally feasible and can be extended readily to more than two channels and higher sampling rates. A full evaluation, e.g., measurement of coherence after decorrelation, is warranted for future work.

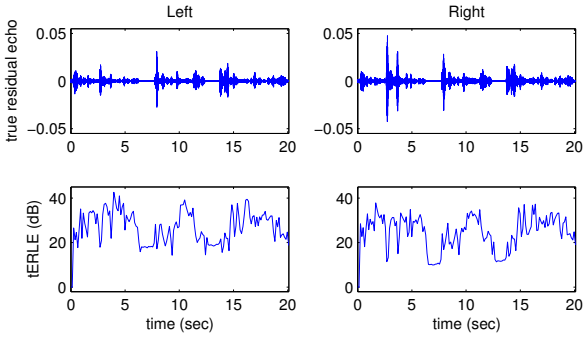

Fig. 6. True residual echo and tERLE without decorrelation.

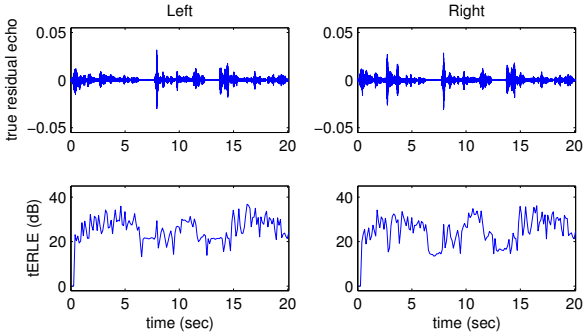

Fig. 7. True residual echo and tERLE with NLP.

VIII. REFERENCES

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

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

[3] K. Ikeda and R. Sakamoto, "Convergence analyses of stereo acoustic echo cancelers with preprocessing," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1324–1334, May 2003.

[4] T. Gänsler and J. Benesty, "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," in *J. Adapt. Control Signal Process.*, vol. 14, 2000, pp. 565–586.

[5] J. Benesty, D. R. Morgan, J. L. Hall, and M. M. Sondhi, "Stereophonic acoustic echo cancellation using nonlinear transformations and comb filtering," in *Proc. IEEE ICASSP*, vol. 6, May 1998, pp. 3673–3676.

[6] E. Robledo-Arununcio, T. S. Wada, and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," in *Proc. IEEE WASPAA*, Oct. 2007, pp. 34–37.

[7] J. Herre, H. Buchner, and W. Kellermann, "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *Proc. IEEE ICASSP*, vol. I, Apr. 2007, pp. 17–20.

[8] A. Sugiyama, Y. Mizuno, A. Hirano, and K. Nakayama, "A stereo echo canceller with simultaneous input-sliding and sliding-period control," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 325–328.

[9] F. Nesta, T. S. Wada, S. Miyabe, and B.-H. Juang, "On the non-uniqueness problem and the semi-blind source separation," in *Proc. IEEE WASPAA*, Oct. 2009, pp. 101–104.

[10] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, 1999.

[11] S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," in *Proc. IEEE ICASSP*, May 1995, pp. 3059–3062.

[12] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Magazine*, vol. 9, no. 1, pp. 14–37, Jan. 1992.

[13] A. Hirano and A. Sugiyama, "A noise-robust stochastic gradient algorithm with an adaptive step-size for mobile hands-free telephones," in *Proc. IEEE ICASSP*, vol. 2, May 1995, pp. 1392–1395.

[14] T. Gänsler, M. Hansson, C. J. Ivarson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Trans. Commun.*, vol. 44, no. 11, pp. 1421–1427, Nov. 1996.

[15] S. Makino, Y. Kaneda, and N. Koizumi, "Exponentially weighted step-size NLMS adaptive filter based on the statistics of a room impulse response," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 1, pp. 101–108, Jan. 1993.

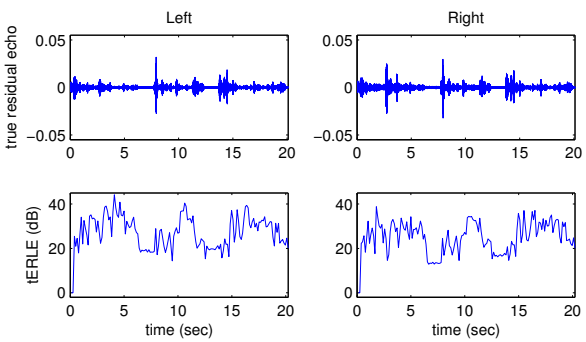

Fig. 8. True residual echo and tERLE with AWGN.

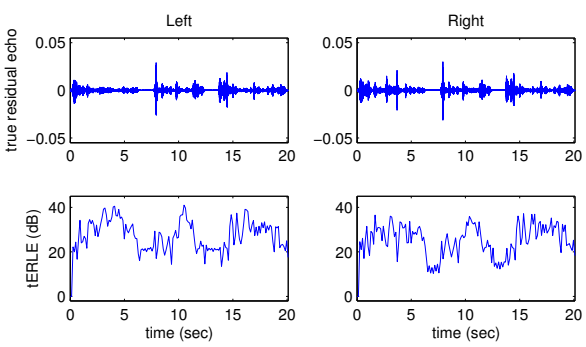

Fig. 9. True residual echo and tERLE with OSD.

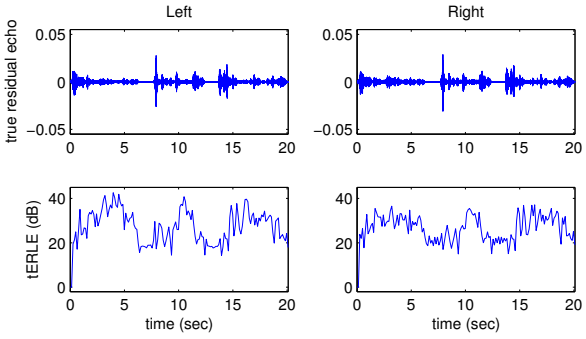

Fig. 10. True residual echo and tERLE with RUD or RBI.

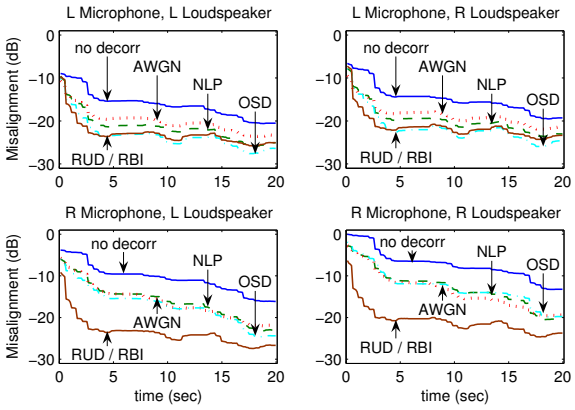

Fig. 11. Improvement in misalignment after decorrelation.