EUSIPCO 2013 1569744425

ADAPTIVE MULTICHANNEL LINEAR PREDICTION BASED DEREVERBERATION IN TIME-VARYING ROOM ENVIRONMENTS

Jae-Mo Yang and Hong-Goo Kang

Dept. of Electrical and Electronic Eng., Yonsei Univ., Korea
⟨jaemo2879, hgkang⟩@dsp.yonsei.ac.kr

ABSTRACT

This paper proposes a real-time acoustic channel equalization method with an adaptive multi-channel linear prediction technique. The proposed method takes a theoretically perfect channel equalization algorithm that also solves the problems existed in the actual implementation stage. The linear-predictive multi-input equalization (LIME) is an appropriate attempt for performing blind dereverberation with assuring the theoretical basis, however, it requires a huge computational cost for calculating large dimension of covariance matrix and its inversion. The proposed equalizer is formed as the multichannel linear prediction (MLP) oriented structure with a new formula that is optimized to time-varying acoustical room environments. Experimental results under both artificially generated and real room environments with a conversational speech confirm the superiority of the proposed method.

Index Terms— Dereverberation, multichannel linear prediction, real time equalization, adaptive filter.

1. INTRODUCTION

In typical speech communication systems such as hands-free telephones and voice-controlled systems, the characteristic of received speech signal is changed by room reverberation and background noise. A number of recent literatures on dereverberation techniques can be divided into two classes: reverberation reduction or perfect cancellation [1]. Methods in the former class are usually designed by single channel spectral enhancement schemes that utilize the statistical characteristic of RIR, e.g. reverberation time (RT60). Specific algorithms vary depending on the way of estimating (late) reverberation power [1, 2]. It is advantageous that they can work online by employing a short time Fourier transform. The other class derives an inverse filter to multichannel RIRs [3, 4, 5, 6, 7]. Several literatures have tried to show a possibility of adopting single-channel equalization techniques, however, there are no clear explanations on how to overcome the minimum phase problem [6]. From a theoretical perspective point of view, reverberation can be completely removed using a multiple microphone based technique, i.e. the multiple-input/output inverse theorem (MINT), if it meets specific conditions including:the co-primeness between channels and a sufficient filter length [7]. In practice, however, the MINT is computationally expensive and should be combined with a blind system identification (BSI) method which is an another challenging task. It is very sensitive to even small errors of channel impulse responses which is unavailable in the estimation process. The blind dereverberation approach assumes that the target signal be independent and identically distributed (i.i.d), but this hypothesis does not hold for speech-like signals [5]. Recent studies use multichannel linear prediction (MLP) techniques to remove reverberant components for the primary input signal [3, 4, 5, 8]. An over-whitening problem should be also considered in speech-like signals. There are several attempts to manage the over-whitening problem using a delayed linear prediction (DLP) [8], pre-whitening/post-compensation method [3], and the LIME algorithm [5].

The LIME algorithm provides a theoretically perfect equalization filter that is similar to the MINT solution, and also supports a blind method for colored target signals. The algorithm consists of two stages. In the first stage, the speech residual is estimated by MLP filters which are derived by the minimum mean squared error (MMSE) criterion. Therefore, MLP filters can be considered as a specific form of the MINT solution that can be applied to the situation where the target signal characteristic varies. In the second stage, the LIME algorithm estimates average speech characteristic to compensate for the whitening effect of the prediction filter. However, this approach requires a huge computational cost for calculating large covariance matrix and its inversion. It is also not suitable for a practical system since the RIRs may change during the observation interval. In addition, the equalization fails if the covariance matrix inversion is not accurate [9].

In this paper, we verify that the process of the conventional LIME structure can be reformulated by taking an adaptive way, which is suitable for time varying situations. Principle drawbacks caused by generating a large covariance matrix is avoidable and various filter estimation techniques can be adopted in the proposed method. Reliable adaptive filters (ADFs) are first tested in artificially generated environments to verify the feasibility of the proposed equalization method. Finally, the dereverberation result to conversational speech in a real room environment is presented.

2. LINEAR-PREDICTIVE MULTI-INPUT EQUALIZATION

2.1. Basic assumptions

The room impulse responses (RIRs) between the source and i-th microphone can be modeled by L_h-th order time-invariant polynomials. Using the matrix formulation, the captured signal vector of length L_w can be written as

$$\mathbf{x}_i(n) = \mathbf{H}_i^T \mathbf{s}(n), \quad i = 1, ..., P \ , \tag{1}$$

where $\mathbf{H}_i$ is an $(L_h + L_w - 1) \times L_w$ convolution matrix expressed in [5] and P is the number of channels. The minimum length of the signal vector is low-bounded to $L_w \geq (L_h - 1) \,/\, (P - 1)$ to obtain the unique LIME solution in accordance with the assumption that RIRs do not share common zeros [5].

A single target signal, $s(n)$, is assumed to be generated by an L_{lp}-th order auto regressive (AR) process as follows;

$$\mathbf{s}(n) = \mathbf{C}^T \mathbf{s}(n-1) + \mathbf{e}(n), \tag{2}$$

where $\mathbf{C}$ is the transpose of the $L_{lp} \times L_{lp}$ Frobenius companion matrix whose first column consists of AR coefficients, $\mathbf{s}(n) = [s(n), ..., s(n - L_{lp} + 1)]^T$ and $\mathbf{e}(n) = [e(n), 0, ..., 0]^T$ [5]. A matrix-vector representation given in (2) is advantageous while deriving optimal filters for the LIME algorithm. Details of the LIME algorithm is faithfully described in the next subsection using two signal models given in (1) and (2).

2.2. LIME algorithm

The LIME algorithm consists of two principle filters: multichannel linear prediction (MLP) filter and speech synthesis filter $1/a(z)$. The prediction residual, $e_w(n)$, for the first channel input as a primary signal is given as

$$e_w(n) = x_1(n) - \mathbf{x}^T(n-1)\mathbf{w}, \tag{3}$$

where $\mathbf{w} = \left[\ \mathbf{w}_1^T, \quad \cdots \quad \mathbf{w}_P^T \ \right]^T$ is an MLP filter and $\mathbf{x}(n-1)$ is a $PL_w \times 1$ reference signal vector with a single tap delay. Using (1) and (2), the optimal MLP filter that minimizes the mean squared error of the residual signal, (3), is expressed as [5]

$$\begin{aligned} \mathbf{w} &= \left(\mathbf{H}^T \mathbf{R}_{\mathbf{s}(n-1)\mathbf{s}(n-1)} \mathbf{H}\right)^+ \mathbf{H}^T \mathbf{R}_{\mathbf{s}(n-1)\mathbf{s}(n)} \mathbf{h}_1 \\ &= \mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T\right)^{-1} \mathbf{C}\mathbf{h}_1, \end{aligned} \tag{4}$$

where the covariance matrix $\mathbf{R}_{\mathbf{s}(n)\mathbf{s}(n)} = E\left\{\mathbf{s}(n)\mathbf{s}^T(n)\right\}$, $\mathbf{H} = [\mathbf{H}_1, ..., \mathbf{H}_P]$ and $\mathbf{A}^+$ represents the Moore-Penrose generalized inverse matrix $\mathbf{A}$. The covariance matrix can be canceled out in (4) if we assume a positive definite matrix. By replacing the column vector $\mathbf{h}_1$ with the matrix $\mathbf{H}$ in (4), we can define the matrix $\mathbf{Q}$ as

$$\mathbf{Q} = \mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T\right)^{-1} \mathbf{C}\mathbf{H}, \tag{5}$$

then the first column of $\mathbf{Q}$ becomes the optimal MLP filter $\mathbf{w}$.

The MLP process generates whitened residual signal $e_w(n)$ by removing not only reverberant components but also its own speech characteristic. This undesired over-whitening effect should be compensated by a speech synthesis filter, $1/a(z)$. Note that the coefficients of polynomial $a(z)$ are equivalent to the characteristic polynomial of the companion matrix $\mathbf{C}$ [10]. Let us consider the nonzero eigenvalues of matrix $\mathbf{Q}$, then the following equality is satisfied

$$\lambda\left(\mathbf{Q}\right) = \lambda\left(\mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T\right)^{-1} \mathbf{C}\mathbf{H}\right) = \lambda\left(\mathbf{C}\right), \tag{6}$$

where $\lambda\left(\mathbf{Q}\right)$ represents coefficients of characteristic polynomial.

In conclusion, two principle filters of the LIME algorithm, MLP and $1/a(z)$, are completely reconstructed by the matrix $\mathbf{Q}$. In a practical system, however, a prior knowledge of either RIR or the target speech signal are *blind*. Thus, an approximation of the matrix $\mathbf{Q}$ which can be calculated with the signals received at the microphones can be formulated using (1) and (4);

$$\hat{\mathbf{Q}} = \mathbf{R}^+_{\mathbf{x}(n-1)\mathbf{x}(n-1)} \mathbf{R}_{\mathbf{x}(n-1)\mathbf{x}(n)}. \tag{7}$$

Here, the covariance matrix is estimated by time-averaging;

$$\mathbf{R}_{\mathbf{x}(n)\mathbf{x}(n)} = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{x}(n)\mathbf{x}(n)^T. \tag{8}$$

3. THE PROPOSED W-ORIENTED ADAPTIVE LIME

To solve the Eq. (7) and (8) a huge computation of constructing large covariance matrix and solving its pseudo-inversion is needed. It makes the LIME algorithm be impractical despite its theoretically perfect equalization performance. Note that the solution of the filter that is completely based on the $\mathbf{Q}$ matrix can not provide a feasible solution, especially when the RIR term is varied. In this section, we propose a novel MLP filter-oriented scheme that can simplify the calculation process of constructing $\mathbf{Q}$ matrix. Since the proposed method takes a multichannel adaptive filtering approach, it can be also applied to the circumstance of channel variation. Note that, we assume a dual-microphone system in the following subsections for the simplicity of equations.

3.1. Q matrix analysis

Let us consider last multiplication of two matrices in (5), $\mathbf{CH}$, with decomposing $\mathbf{C}$ into sum of two simple matrices as fol-

lows;

$$\mathbf{CH} = \left(\begin{bmatrix} a_1 & 0 & \cdots & 0 \\ a_2 & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & 0 \\ a_{L_{lp}} & 0 & \cdots & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \ddots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \right) \mathbf{H} = \left(\mathbf{A} + \tilde{\mathbf{I}} \right) \mathbf{H}. \tag{9}$$

Then, the first and the second matrix multiplications can be simplified as follows;

$$\mathbf{AH} = \begin{bmatrix} \mathbf{a}h_1(0) & \mathbf{0}_{mtx} & \mathbf{a}h_2(0) & \mathbf{0}_{mtx} \end{bmatrix}, \tag{10}$$

$$\tilde{\mathbf{I}}\mathbf{H} = \begin{bmatrix} \mathbf{h}_1(1) & \tilde{\mathbf{H}}_1 & \mathbf{h}_2(1) & \tilde{\mathbf{H}}_2 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{11}$$

where $\mathbf{a} = \left[a_1, ..., a_{L_{lp}}\right]^T$, $h_i(0)$ is the first component of $\mathbf{h}_i$ and $\mathbf{0}_{mtx}$ is an $L_{lp} \times (L_w - 1)$ zero matrix. In (11), the square matrix $\tilde{\mathbf{I}}$ operates as throwing out the first row of the latter matrix and filling all the last row with zeros. Therefore, $\mathbf{h}_i(1)$ becomes $L_{lp} - 1$ vector which is equivalent to the $((i-1) \times L_w + 1)$ -th column of the matrix $\mathbf{H}$ except for the first row component. The $(L_{lp} - 1) \times (L_w - 1)$ matrix $\tilde{\mathbf{H}}_i$ is equivalent to the convolution matrix of i-th channel, $\mathbf{H}_i$, excluding the last column and row. Sequentially, (9) can be reformulated as

$$\mathbf{CH} = \begin{bmatrix} \mathbf{a}h_1(0) \\ +\mathbf{h}_1(1) \end{bmatrix} \quad \mathbf{0}_{mtx} \quad \begin{bmatrix} \mathbf{a}h_2(0) \\ +\mathbf{h}_2(1) \end{bmatrix} \quad \mathbf{0}_{mtx} \Big] + \begin{bmatrix} \mathbf{0}_{vec} & \tilde{\mathbf{H}}_{(L_{lp}-1)\times(L_{lp}-1)} \\ 0 & \mathbf{0}_{vec}^T \end{bmatrix}, \tag{12}$$

where $\mathbf{0}_{vec}$ is an $(L_{lp} - 1) \times 1$ zero vector and $\tilde{\mathbf{H}}$ becomes $\tilde{\mathbf{H}} = \left[\tilde{\mathbf{H}}_1; \mathbf{0}_{vec}; \tilde{\mathbf{H}}_2\right]$. Note that, $\mathbf{H}$ and $\tilde{\mathbf{H}}$ are pretty much similar except only for the fact that L_w-th column is filled with zero components and the last row and the last column are omitted. Finally, the ideal $\mathbf{Q}$ in the LIME algorithm, (5), can be expressed in a new way as

$$\mathbf{Q}_w = \begin{bmatrix} \mathbf{w}^{(1)} & \mathbf{0}_{mtx} & \mathbf{w}^{(2)} & \mathbf{0}_{mtx} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{vec} & \hat{\mathbf{I}}_{(L_{lp}-1)\times(L_{lp}-1)} \\ & \mathbf{0}_{vec}^T \end{bmatrix}, \tag{13}$$

where $\mathbf{w}^{(i)} = \mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{C}\mathbf{h}_i$ is the ideal MLP filter which takes the current input of i-th channel as a primary signal. Second term of (13) is produced by multiplying $\mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T\right)^{-1}$ with the second term of (12) and obviously $\hat{\mathbf{I}}$ becomes an approximation of the identity matrix. Figure 1 depicts the ideal $\mathbf{Q}$, (5), and its approximation, (7) as a squared matrix form. We confirm that the matrix is composed of two MLP filter vectors and diagonal entries in both cases that the suggested $\mathbf{Q}_\mathbf{w}$ is sufficient to substitute the ideal matrix.

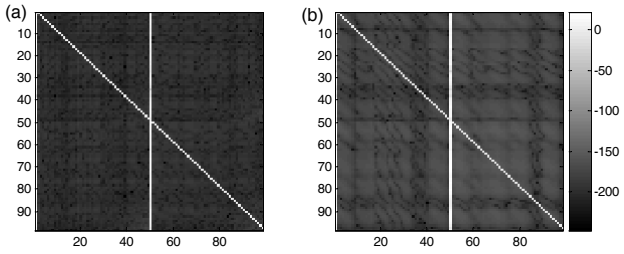

Fig. 1. Squared components of the matrix $\mathbf{Q}$ in log-scale [dB] when $L_h = 50$ and $P = 2$, (a) ideal $\mathbf{Q}$ (b) approximated $\hat{\mathbf{Q}}$.

3.2. Adaptive LIME method

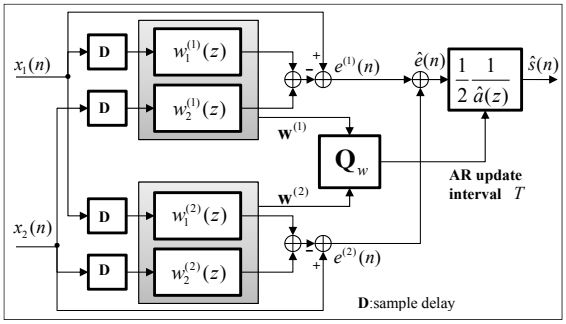

Fig. 2. Block diagram of the proposed $\mathbf{w}$-oriented real time adaptive equalization method ($P = 2$ case).

From the above analysis, we showed that the estimation of MLP filters for each channel should be taken prior to the estimation of $\mathbf{Q}$ matrix. The proposed method requires very low complexity and can be applied to time varying conditions using an online processing. Figure 2 depicts the schematic diagram of the proposed $\mathbf{w}$-oriented real time adaptive equalization method. At first, the prediction filters, $\mathbf{w}^{(i)}$ is estimated by utilizing an adaptive filter with an MMSE criterion as follows;

$$\mathbf{w}^{(i)} = \arg \min_{\mathbf{w}_1^{(i)}, \mathbf{w}_2^{(i)}} E\left\{ \left| e^{(i)}(n) \right|^2 \right\}, i = 1, 2. \tag{14}$$

We can further enhance the residual signal by averaging MLP outputs because the prediction error has a characteristic of white noise. Finally, the speech compensation filter, $1/\hat{a}(z)$, is derived by the characteristic polynomial of $\mathbf{Q}_w$ at every T

intervals. Last but not least, assumptions that there are no time difference of arrival (TDOA) between channels should be satisfied to estimate MLP filters in the proposed method. A number of TDOA compensation techniques can be presented, however, we assume that the target speaker is located in the front direction in this paper.

4. EXPERIMENTAL RESULTS

A number of reliable ADFs are introduced to implement MLP such as normalized least mean square (NLMS), variable step size LMS (VS-LMS), steepest descent (SD), conjugate gradient (CG) and recursive least square (RLS) methods [11, 12]. The segmental prediction-to-distortion ratio (PDR), (15), is employed to evaluate the prediction performance.

$$PDR = 10\log_{10}\left(\frac{\sum|e(n)|^2}{\sum|e(n)-\hat{e}(n)|^2}\right)[dB]. \qquad (15)$$

The accuracy of speech synthesis filter is evaluated by the log-spectral distance (LSD) between the ideal and estimated coefficients. We consider both a synthesized RIRs (Polack's model) and real RIR measured in a room.

4.1. Performance evaluation for ADF-LIME

At first, the performance of MLP and $\hat{a}(z)$ is verified in a controlled environment. The target signals are generated by a time-invariant AR process with an input of white noise. The model order of the AR process is given by $L_{lp} = L_h + L_w - 1$ and it is extracted from speech signal. The RIRs are generated by the Polack's model and truncated them to 100 taps corresponding to a short duration of 12.5ms. A dual-channel system is implemented ($P = 2$) and the length of MLP filter is set to minimum, $L_h - 1$ [see (1)]. The forgetting factor, β, in different ADF methods (VS-LMS, SD, CG, RLS) are set to 0.9999 to equalize the observation interval for past input signals. The best step-sizes which show the fastest convergence with ensuring the stability of ADFs are determined by lots of iterative simulations in both fixed step-size (NLMS, SD) and variable step-size (VS-LMS, CG) cases.

Figure 3 shows the segmental PDR of MLP filters using the proposed ADFs and the LIME schemes with Monte Carlo experiments. In this simulation, RIR had been changed at 4s point by $\mathbf{h}_i = \alpha\mathbf{h}_i + (1-\alpha)\mathbf{h}_{i,new}$ where the smoothing factor is set to 0.9. The covariance matrix in the LIME method is calculated in different ways such that LIME1 utilizes whole signal and LIME2, the ideal case, uses different covariance matrix after RIR is changed. Among the ADFs, the RLS method shows the best prediction performance, especially it is superior to the conventional LIME method (LIME1) in both tracking capability and the accuracy of stabilized filter perspectives. The CG method has better performance than LIME2, however, the rest of ADFs are failed to minimize

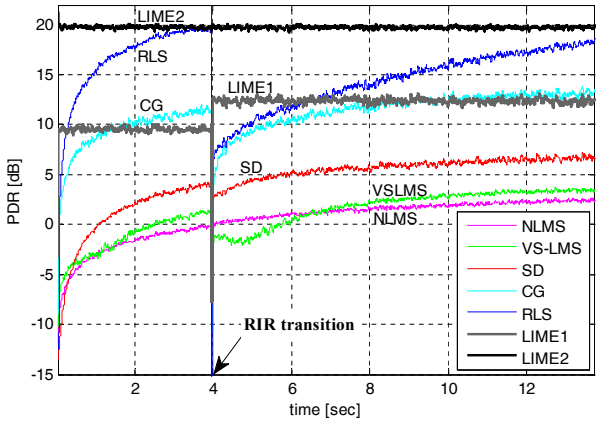

Fig. 3. Segmental PDR of MLP filters using the proposed ADFs and LIME method considering RIR transition at 4.0s.

the prediction error. The reason can be found from the fact that the eigen-value spread ratio of the input signal synthesized by AR process is very high [11]. Therefore, only the methods which are not affected by the characteristic of input signal (RLS, CG) can successfully find the global minimum in an adaptive way. The accuracy of estimated AR polynomial $\hat{a}(z)$ is summarized in Fig.4. Not surprisingly, it shows similar trend to the prediction performance explained in Fig.3. Consequently, we confirm that the estimated AR coefficients which are obtained from the proposed $\mathbf{Q}_\mathbf{w}$ can exactly compensate the input signal.

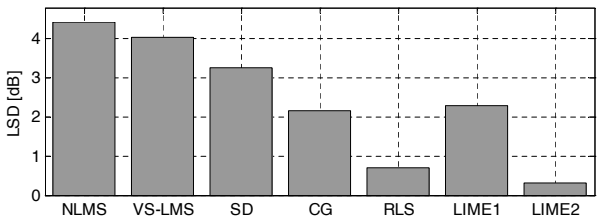

Fig. 4. LSD between ideal $a(z)$ and estimated $\hat{a}(z)$.

4.2. Real room experiment for a conversational speech

To confirm the performance of the proposed method in a real environment, a conversational speech is convolved with measured RIR in Aachen impulse response database (AIR-DB) [13]. The AIR-DB is targeting a dual-channel system whose microphone spacing is fixed to $0.17m$ and the RIRs are measured in various room environments for arbitrary speaker locations. Two-channel RLS-MLP with the same length of RIR as the previous experiment are used to produce residual signals. In this case, the AR process can be considered as time-varying, thus $\hat{a}(z)$ is updated at every $T = 0.5s$. The whole process is carried out in a sample-by-sample manner not to

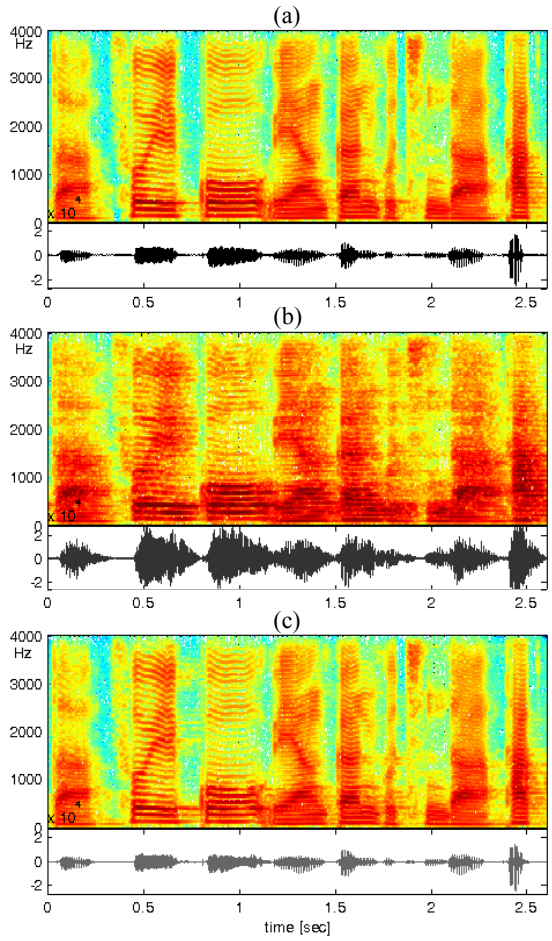

Fig. 5. Example result of the proposed RLS-MLP method for a male speech in office room (speaker to microphone distance= $3m$, RT60= $0.48sec$, $L_h = 1000taps$ and $P = 2$) (a) clean signal $s(n)$ and its spectrogram (b) observed signal $x_1(n)$ (c) output signal $\hat{s}(n)$.

bring system delay. Figure 5 shows the result of the proposed method in the office room. A conversational male speech, length of 16s, is used and the signals in last 2.5s interval are depicted when the ADFs are stabilized. Figure 5 (c) shows that the reverberant components are removed, which is verified in both time and frequency domain results. There remain some leakage terms which are shown to be audible in some frequency bands though. This result is occasionally happened for a conversational speech signal because the non-stationarity of the signal can disturb the filter adaptation. Subjective tests confirm that it sounds like a low level late reverberation. Note that, it can be easily removed by the late reverberation suppression method for direct listening or by the cepstral mean normalization (CMN) for a automatic speech recognition (ASR) process [1][5].

5. CONCLUSION

We have presented a new adaptive scheme for the LIME algorithm by reformulating the covariance matrix, which is suitable for time-varying acoustic environment. Experimental results in real room environments showed that the proposed method successfully removed the reverberation even for conversational speech. Future work involves both adopting the fast version of ADFs and improving the performance of the adaptive multichannel prediction filter by considering the (near-) common zeros problem between channels and background noise.

6. REFERENCES

[1] E. A. P. Habets, "Singe- and multi-microphones speech dereverberation using spectral enhancement," *Ph.D thesis*, 2007.

[2] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," *Proc. ICASSP*, pp. 817–820, 2006.

[3] T. Okamoto, Y. Iwaya, and Y. Suzuki, "Wide-band dereverberation method based on multichannel linear prediction using pre-whitening filter," *Applied acoustics*, vol. 73, pp. 50–55, Jan 2012.

[4] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signal with speaker-position change detection," *Proc. ICASSP*, pp. 3733–3736, 2009.

[5] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. ASLP*, vol. 15, pp. 430–440, Feb 2007.

[6] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. ICASSP*, pp. 3701–3704, 2001.

[7] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, vol. 36, pp. 145–152, Feb 1988.

[8] T. Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance normalized delayed linear prediction," *IEEE Trans. ASLP*, vol. 18, pp. 1717–1731, Sep 2010.

[9] I. Ram, E. Habets, Y. Avargel, and I. Cohen, "Multi-microphone speech dereverberation using lime and least squares filtering," in *Proceedings of European Signal Processing Conference (EUSIPCO'08)*, 2008.

[10] S. Rombouts and K. Heyde, "An accurate and efficient algorithm for the compuption of the characteristic polynomial of a general square matrix," *Journal of Computational Physics*, vol. 140, pp. 453–458, Mar 1998.

[11] S. Haykin, "Adaptive filter theory," 2002, Prentice Hall.

[12] K. P. C. Edwin, "An introduction to optimization," 2001, Jhon Wiley and Sons Inc.

[13] M. Jeub, M. Schafer, H. Kruger, C. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?," in *Proc. 20th Int. Congress on Acoust*, 2010.