ON THE USE OF LIME DEREVERBERATION ALGORITHM IN AN ACOUSTIC ENVIRONMENT WITH A NOISE SOURCE

Marc Delcroix [† ‡] , Takafumi Hikichi [†] and Masato Miyoshi [† ‡]

[†]NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho (Keihanna Science City), Soraku-gun, Kyoto 619-0237 Japan
[‡] Graduate School of Information Science and Technology, Hokkaido University,
Kita 14, Nishi 9, Kita-ku, Sapporo, 060-0814 Japan

ABSTRACT

This paper addresses the speech dereverberation problem in the presence of a noise source. We show that the previously presented LInear-predictive Multi-input Equalization (LIME) algorithm can achieve both dereverberation and noise reduction. Experiments show that, for a reverberation time of 0.2 seconds, precise dereverberation is possible in the presence of a colored noise source of SNR of 5 dB, with a dereverberation of the room impulse response by more than 20 dB.

1. INTRODUCTION

Room reverberations degrade the characteristics and the audible quality of speech recorded by distant microphones. It is a severe problem for applications such as automatic speech recognition, hearing aids and hands-free telephony [1]. The dereverberation problem consists in recovering a target speech from observed reverberant signals. Much research have been undertaken on the dereverberation problem using both single [2] and multi-microphone techniques [3]. However, it seems that there have been few reports of dereverberation working successfully in noisy conditions [4], [5].

Dereverberation and noise reduction using MINT [6] has been proposed [4]. If the sources are spatially independent and each source is temporally independent and identically distributed (i.i.d.), multi-channel inverse filters can be blindly obtained, and these filters would perform both dereverberation and noise reduction. However, the temporally i.i.d. hypothesis does not hold for speech-like signals. When applied to speech, such methods degrade the speech characteristics by causing an excessive whitening of the recovered signal.

We have already proposed the LInear-predictive Multi-input Equalization (LIME) algorithm to solve the whitening problem [7] [8] [9]. In [9], we showed that LIME could achieve the precise dereverberation of a single speech signal. In this paper, we expand the LIME algorithm for performing both dereverberation and noise reduction in a multi-source scenario. We also present results for experiment of speech dereverberation in the presence of a colored noise source.

2. PRINCIPLES

We consider the acoustic system shown in Fig. 1, with P microphones and two sources; a target speech signal, $s_1(n)$, and a noise source, $s_2(n)$. The target signals observed at the microphones are degraded by the noise source and reverberation. The microphone signals, $u_j(n)$, can be expressed as:

$$u_j(n) = \sum_{k=0}^{M-1} h_{1,j}(k)s_1(n-k) + \sum_{k=0}^{M-1} h_{2,j}(k)s_2(n-k). \quad (1)$$

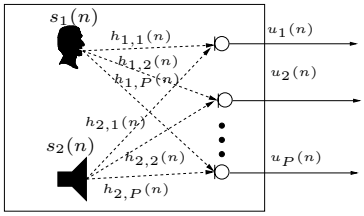


Fig. 1. *Acoustic system. $s_1(n)$ is the target signal, $s_2(n)$ is a noise source, $h_{i,j}(n)$ is the room impulse response between the i^{th} source and the j^{th} microphone and $u_j(n)$ is the signal observed at the j^{th} microphone.*

Here, we expand the previously reported LIME algorithm to recover a target signal $s_1(n)$ precisely from the P observed signals $u_j(n)$, where $j = 1, \ldots, P$. First we calculate filters which linearly predict the microphone signal, $u_1(n)$, from the past samples of P microphone signals $u_j(n)$ ($j = 1, \ldots, P$). These filters would cancel out the room reverberation and suppress the noise signal $s_2(n)$, if we assume that the target source is closer to microphone 1 than the noise source, i.e. $h_{1,1}(0) \neq 0$ and $h_{2,1}(0) = 0$ and that the noise source is closer to another microphone [7]. However, the filters degrade the target-source characteristics causing excessive whitening. In the second step we estimate the target-source characteristics to recover the target signal precisely.

2.1. Hypotheses

We construct the following hypotheses:

- First, we assume that the source signals $s_i(n)$, $i = 1, 2$, are modeled by autoregressive (AR) processes applied to white noise $e_i(n)$. The Z-transform of the AR process of the i^{th} source is $1/a_i(z)$ where $a_i(z)$ is an AR polynomial of order N_i given by:

$$a_i(z) = 1 - \{a_{i,1}z^{-1} + \ldots + a_{i,N_i}z^{-N_i}\}. \quad (2)$$

Using a matrix formulation we can write [10]:

$$\mathbf{s}_i(n) = \mathbf{C}_i^T \mathbf{s}_i(n-1) + \mathbf{e}_i(n), \quad (3)$$

where $\mathbf{s}_i(n) = [s_i(n), \ldots, s_i(n-(N_i+1))]^T$, $\mathbf{C}_i$ is the $N_i \times N_i$ companion matrix defined as:

$$\mathbf{C}_i = \begin{pmatrix} a_{i,1} & 1 & 0 & \ldots & 0 \\ a_{i,2} & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 \\ a_{i,N_i} & 0 & \ldots & \ldots & 0 \end{pmatrix}, \quad (4)$$

and $\mathbf{e}_i(n) = [e_i(n), 0 \ldots, 0]^T$, $i = 1, 2$.
We can write an equation for the sources as:

$$\mathbf{s}(n) = \mathbf{C}^T \mathbf{s}(n-1) + \mathbf{e}(n), \tag{5}$$

where $\mathbf{s}(n) = \left[\mathbf{s}_1^T(n), \mathbf{s}_2^T(n)\right]^T$, $\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & 0 \\ 0 & \mathbf{C}_2 \end{pmatrix}$, and $\mathbf{e}(n) = \left[\mathbf{e}_1^T(n), \mathbf{e}_2^T(n)\right]^T$.
Moreover, we assume that noise source $s_2(n)$ is stationary.

- We model the room transfer functions by using time invariant polynomials that do not share common zeros.
  Using a matrix form, we can re-write Equation (1) as:

$$\mathbf{u}(n) = \mathbf{H}^T \mathbf{s}(n). \tag{6}$$

where $\mathbf{u}(n) = [\mathbf{u}_1^T(n), \ldots, \mathbf{u}_P^T(n)]^T$,
$\mathbf{u}_j(n) = [u_j(n), \ldots, u_j(n-(L-1))]^T$, $\mathbf{H}$ is a $2(M+L-1) \times PL$ convolution matrix expressed as
$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{1,1} & \ldots & \mathbf{H}_{1,P} \\ \mathbf{H}_{2,1} & \ldots & \mathbf{H}_{2,P} \end{pmatrix}$,
$\mathbf{H}_{i,j} = \begin{pmatrix} \mathbf{h}_{i,j} & 0 & \ldots & 0 \\ 0 & \mathbf{h}_{i,j} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \mathbf{h}_{i,j} \end{pmatrix}$,
$\mathbf{h}_{i,j} = [h_{i,j}(0), \ldots, h_{i,j}(M-1)]^T$, $i = 1, 2$ and $j = 1, \ldots, P$.

2.2. LIME for multi-sources

Conventional LIME is expanded hereafter to perform both dereverberation and noise reduction in a multi-source scenario. Let us consider that the first microphone signal, $u_1(n)$, can be linearly predicted from the past samples of P microphone signals $u_j(n)$ ($j = 1, \ldots, P$). The prediction error can be defined as [11]:

$$\hat{e}(n) = u_1(n) - \mathbf{u}^T(n-1)\mathbf{w} \tag{7}$$

where $\mathbf{w}$ is a prediction filter set of length PL. Minimizing the mean square value of the prediction error gives us:

$$\mathbf{w} = \mathbf{R}^+ \mathbf{r} \tag{8}$$

where $\mathbf{A}^+$ is the Moore-Penrose generalized inverse of matrix $\mathbf{A}$ [12], $\mathbf{R} = E\{\mathbf{u}(n-1)\mathbf{u}^T(n-1)\}$ is the covariance matrix, $\mathbf{r} = E\{\mathbf{u}(n-1)u_1(n)\}$ is the correlation vector and $E\{\}$ is an expectation operator. By replacing the scalar $u_1(n)$ in $\mathbf{r}$ with the observation vector $\mathbf{u}^T(n)$, we define the prediction matrix, $\mathbf{Q}$, as:

$$\mathbf{Q} \triangleq \mathbf{R}^+ \tilde{\mathbf{R}}, \tag{9}$$

where $\tilde{\mathbf{R}} = E\{\mathbf{u}(n-1)\mathbf{u}^T(n)\}$ is a one step shifted covariance matrix. By definition, the first column of $\mathbf{Q}$ is equivalent to the prediction filter set $\mathbf{w}$. Using Equation (6), and the fact that the covariance matrix of the source signals may be considered positive definite, we can express the prediction filter set and prediction matrix as [8]:

$$\mathbf{Q} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}\mathbf{H}, \tag{10}$$

and

$$\mathbf{w} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}\mathbf{h}_1, \tag{11}$$

where $\mathbf{h}_1 = [\mathbf{h}_{1,1}^T, \mathbf{h}_{2,1}^T]^T$ is the first column of $\mathbf{H}$. Using Equations (5), (6) and (11) the prediction error becomes:

$$\begin{aligned} \hat{e}(n) &= \mathbf{s}^T(n)\mathbf{h}_1 - \mathbf{s}^T(n-1)\mathbf{H}\mathbf{w} \\ &= \mathbf{s}^T(n)\mathbf{h}_1 - \mathbf{s}^T(n-1)\mathbf{H}\mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{C}\mathbf{h}_1 \\ &= (\mathbf{s}^T(n) - \mathbf{s}^T(n-1)\mathbf{C})\mathbf{h}_1 \\ &= \mathbf{e}^T(n)\mathbf{h}_1 \\ &= h_{1,1}(0)e_1(n) + h_{2,1}(0)e_2(n). \end{aligned} \tag{12}$$

If we assume that the target source is closer to microphone 1 than the noise source, i.e. $h_{1,1}(0) \neq 0$ and $h_{2,1}(0) = 0$, the prediction error becomes:

$$\hat{e}(n) = h_{1,1}(0)e_1(n). \tag{13}$$

Equation (13) shows that the effect of room reverberation is canceled out as well as the interference coming from the noise source. However, the obtained signal is whitened.

For a single source, the target signal could be recovered by filtering the prediction error with an estimate of the AR polynomial of the target signal [8]. Such an estimate was obtained from the characteristic polynomial of the prediction matrix. Indeed, we proved that the characteristic polynomial of the prediction matrix $\mathbf{Q}$ was equivalent to the characteristic polynomial of the companion matrix, $\mathbf{C}$:

$$f_c(\mathbf{Q}, \lambda) = f_c(\mathbf{C}, \lambda), \tag{14}$$

where $f_c(\mathbf{A}, \lambda) = det(\mathbf{A} - \lambda\mathbf{I})$ is the characteristic polynomial of matrix $\mathbf{A}$. With a single source, the characteristic polynomial of $\mathbf{C}$ is equivalent to the AR polynomial of the single source.

When there are two sources, however, the characteristic polynomial of $\mathbf{C}$ is the product of the AR polynomials of the two sources. The estimated AR polynomial $\hat{a}(z)$ therefore becomes [12]:

$$\begin{aligned} \hat{a}(z) &= f_c(\mathbf{Q}, \lambda) \\ &= f_c(\mathbf{C}, \lambda) \\ &= f_c(\mathbf{C_1}, \lambda) f_c(\mathbf{C_2}, \lambda) \\ &= a_1(z)a_2(z). \end{aligned} \tag{15}$$

If such a biased AR polynomial were used, the recovered signal would be degraded. To avoid such degradation, we propose estimating the noise AR polynomial, $a_2(z)$, and using it to filter the LIME output. Note that if the noise source is white, $a_2(z) = 1$ and $\hat{a}(z)$ is equivalent to the AR polynomial of the target source.

2.3. Estimation of noise AR polynomial

With a colored noise source, the noise AR polynomial should be estimated. A technique for extracting the target source AR polynomial from Equation (15) has been proposed [7]. The method first estimates the room transfer functions $h_{1,i}(z)$ from the cross-correlation between the prediction error $\hat{e}(n)$ and the output of a microphone $u_i(n)$. The AR polynomial of the target source, $a_1(z)$, is then obtained as the greatest common divisor of the estimate of $h_{1,i}(z)$ and the AR polynomial $\hat{a}(z)$ shown in Equation (15). However, the computation of the greatest common divisor is hard to perform in practice.

To avoid such complex computation, in this paper, we estimate the noise AR polynomial by using the LIME algorithm when the speaker is silent. In that case, only the noise source will be active and the problem is simplified to a single source dereverberation problem. LIME can thus be used to estimate the noise AR process $a_2(z)$ without bias. However, such techniques would require the use of voice activity detection [13].

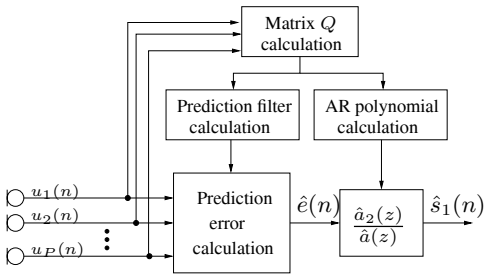


Fig. 2. *Schematic diagram of proposed method.*

2.4. Algorithm of proposed method

Figure 2 is a schematic diagram of the proposed method. The algorithm can be summarized as follows:

1. First, an estimate of the noise AR polynomial $\hat{a}_2(z)$ is obtained using the LIME algorithm when only the noise source is active.
2. Prediction matrix $\mathbf{Q}$ is obtained using Equation (9).
3. The prediction error is calculated using Equation (7) and the prediction filter set, $\mathbf{w}$, given by the first column of matrix $\mathbf{Q}$.
4. The target signal is recovered by filtering the prediction error with $\frac{\hat{a}_2(z)}{\hat{a}(z)}$, where $\hat{a}(z)$ is given by the characteristic polynomial of matrix $\mathbf{Q}$.

3. EXPERIMENT

We used the proposed method for the dereverberation of 4 seconds of reverberant speech in the presence of a colored noise source. The experimental conditions are summarized in Table 1. The room impulse responses were generated by the Image method [14] and the reverberation time was 0.2 sec. The distance between the speaker and microphone 1 and the distance between the noise source and microphone 1 were $1.67m$ and $1.89m$, respectively. The computational complexity for this simulation is about the same as the complexity involved in performing single-source dereverberation in a room with a reverberation time of 0.4 sec. The signal to noise ratio (SNR) at the microphone was 5 dB. In the experiment, we used a 1 second reverberant noise observation to evaluate the AR polynomial of the noise with the LIME algorithm.

Table 1. *Experimental conditions.*

Order of $a_2(z)$	30
Room impulse response duration	0.2 sec
Data length	4 sec
Sampling frequency	8 kHz
Room size	$8m \times 5m \times 3m$
Number of microphones	8

Figure 3 plots the energy decay curve of an original and equalized room impulse response. The equalized impulse response was obtained by filtering the original room impulse responses between the target source and the microphones, with the prediction filter set, $\mathbf{w}$, and the estimated source AR process, $\frac{\hat{a}_2(z)}{\hat{a}(z)}$. We observed that using the proposed algorithm, the original room impulse response was attenuated by more than 20 dB. Figure 4 (a)-(b)-(c)-(d) plot the waveforms and spectrograms of the target signal, the reverberant signal, the observed signal and the signal processed with LIME, respectively. The Itakura-Saito Distance (ISD) [15] was used to measure the speech quality. The recovered signal is composed of the dereverberated speech and interference coming from the noise source. Figure 4 (b)-(d) reveal that the room reverberation effect is successfully eliminated after processing with the proposed method. To evaluate the dereverberation performance, we used the signal to distortion ratio (SDR) defined as:

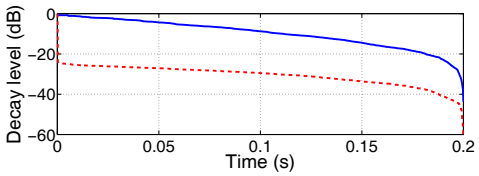


Fig. 3. *Energy decay curve of original (solid line) and equalized (dotted line) room impulse response.*

$$SDR = 10\log_{10}\left(\frac{\sum |s_1(n)|^2}{\sum |s_1(n) - \hat{d}_1(n)|^2}\right), \qquad (16)$$

where $s_1(n)$ is the target signal and $\hat{d}_1(n)$ is the dereverberated signal obtained by applying the prediction filters and the estimated AR process to the reverberant speech. The target signal is well recovered with an SDR value of 21 dB. As shown in Table 2, similar results could be obtained for female and male speakers for input SNR value of 5 and 10 dB. In theory, as long as the post-processing filter, $\frac{\hat{a}_2(z)}{\hat{a}(z)}$ shown in Fig. 2, is estimated precisely, the dereverberation performance should not be affected by the SNR at the microphone.

Table 2. *Value of the SDR for a female and a male speaker.*

Input SNR		5 dB	10 dB
SDR	Female	21 dB	21 dB
	Male	19 dB	21 dB

In Fig. 4(d), we observe the presence of additive stationary colored noise coming from the interference from the noise source. To evaluate the noise reduction performance, we calculate the SNR defined as:

$$SNR = 10\log_{10}\left(\frac{\sum |s_1|^2}{\sum |s_1(n) - \hat{s}_1(n)|^2}\right), \qquad (17)$$

where $\hat{s}_1(n)$ is the recovered signal obtained with LIME. The SNR value is 11 dB. If prediction filter set $\mathbf{w}$ were precisely calculated, the prediction error $\hat{e}(n)$ should be exactly equivalent to the target speech generating white noise $e_1(n)$, as shown in Equation (13). However, due to numerical errors in the calculation of the prediction filters, the prediction error contains some remaining noise. In practice, the noise is reduced but not completely eliminated.

To reduce the interference further, we used spectral subtraction based noise reduction [16]. The target signal is then better recovered, as seen in Fig. 4(e), with an SNR value of 13 dB.

4. CONCLUSION

In this paper, we presented an extension of the LIME dereverberation algorithm for use in the presence of a noise source. We showed that if a microphone is closer to the target source than to the noise source and another microphone is closer to the noise source than to the target source, the LIME algorithm could be directly applied. For a colored noise source, however, the recovered signal would suffer from distortion caused by the noise AR process. We remove such distortion by estimating the noise characteristics by applying the LIME algorithm to noise only observations. The proposed method could

achieve both precise dereverberation and suppression of the interference coming from the noise source. The performance of the interference suppression could be improved further by using a conventional noise reduction technique as post-processing.

5. REFERENCES

[1] Gillespie, B.W. and Atlas, L., "Acoustic diversity for improved speech recognition in reverberant environments," Proc. ICASSP02, vol. I, pp. 557-560, 2002.

[2] Nakatani, T. and Miyoshi, M., "Blind dereverberation of single channel speech signal based on harmonic structure," Proc. ICASSP'03 , vol 1, pp. 92-95, April, 2003.

[3] Gannot, S. and Moonen, M., "Subspace methods for multimicrophone speech dereverberation," Proc. IWAENC, pp. 47-50, 2001.

[4] Furuya, K., "Noise reduction and dereverberation using correlation matrix based on the multiple-input/output inverse filtering theorem (MINT)," HSC 2001, pp. 59-62, 2001.

[5] Bobillet, W., Grivel, E., Guidorzi, R. and Najim, M., "Cancelling convolutive and additive colored noises for speech enhancement," Proc. ICASSP04, vol. II, pp. 777-780, 2004.

[6] Miyoshi, M. and Kaneda, Y., "Inverse filtering of room acoustics," IEEE Trans. ASSP, vol. 36, no. 2, pp. 145-152, 1988.

[7] Miyoshi, M., "Estimating AR parameter-sets for linear-recurrent signals in convolutive mixtures," Proc. ICA'03, pp. 585-589, 2003.

[8] Delcroix, M., Hikichi, T. and Miyoshi, M., "Blind dereverberation algorithm for speech signals based on multi-channel linear prediction," Acoustical Science and Technology, vol. 26, no. 5, pp. 432-439, 2005.

[9] Delcroix, M., Hikichi, T. and Miyoshi, M., "Improved blind dereverberation performance by using spatial information," Proc. Interspeech'05, pp. 2309-2312, 2005.

[10] Kailath, T., Sayed, A.H. and Hassidi, B., "Linear Estimation," NJ: Prentice Hall, 2000.

[11] Haykin, S., "Adaptive filter theory," 3rd ed., NJ: Prentice-Hall, 1996.

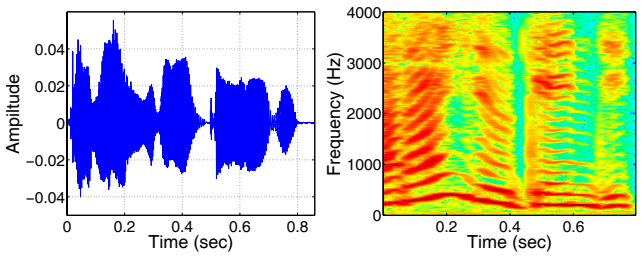
[12] Harville, D. A., "Matrix algebra from a statistician's perspective," Springer-Verlag, 1997.

[13] Li, K., Swamy, M.N.S. and Ahmad, M.O., "An improved voice activity detection using higher order statistics," IEEE Trans. SAP, vol. 13, no. 5, pp. 965 - 974, 2005.

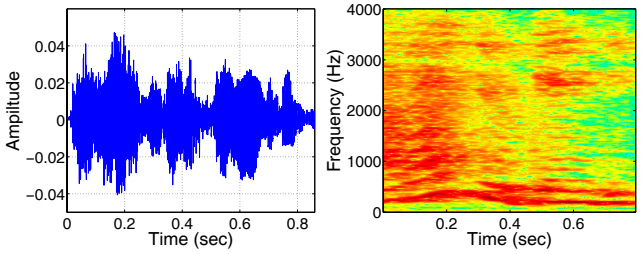
[14] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., vol. 65, no. 4, pp. 943-950, 1979.

[15] Quackenbush, S. R., Barnwell, T. P. and Clements, M. A., "Objective Measures of Speech Quality," NJ : Prentice Hall, 1988.

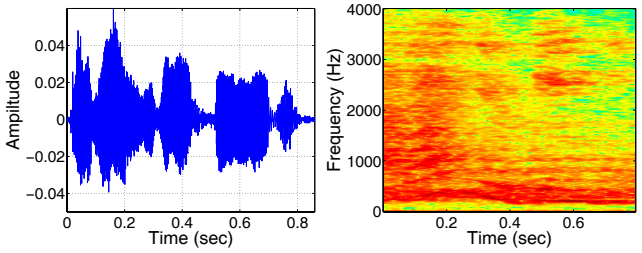
[16] Benesty, J., Makino, S. and Chen, J., "Speech Enhancement," Springer-Verlag, 2005.



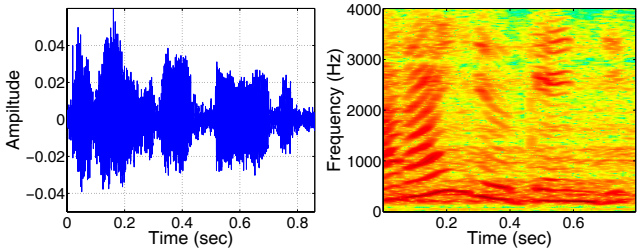
(a) Target signal



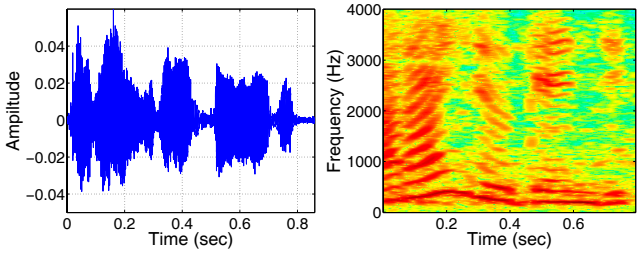
(b) Reverberant signal



(c) Observed signal (ISD 2.04)



(d) Signal processed with LIME (ISD 1.60)



(e) Signal processed with LIME and noise reduction (ISD 1.36)

Fig. 4. *Waveforms and spectrograms of target signal, reverberant signal, observed signal, signal processed with LIME and signal processed with LIME and noise reduction.*