MULTI-MICROPHONE SPEECH DEREVERBERATION USING LIME AND LEAST SQUARES FILTERING

Idan Ram, Emanuël A.P. Habets, Yekutiel Avargel, and Israel Cohen

Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

ABSTRACT

In this paper a multi-microphone speech dereverberation algorithm is presented. The developed algorithm is based on the LIME algorithm proposed by Delcroix et al. in [IEEE Trans. Audio, Speech and Language Process., vol. 15,no. 2, pp. 430-440, Feb. 2007]. The LIME algorithm is shown to be signal dependent and to either produce a very good or a very poor estimate of the source signal. Two non-intrusive methods are proposed to assess the performance of the LIME algorithm for an unknown source signal. These methods can be used to detect errors in the source signals' estimation. Least squares filters are calculated using signal segments that are successfully recovered by LIME, and are used to dereverberate other signal segments for which LIME produced poor estimates. Experimental results demonstrate that the signal segments for which LIME fails can successfully be detected and dereverberated using the least squares filters.

1. INTRODUCTION

In general, acoustic signals captured by distant microphones in a room suffer from distortions caused by reverberation. The captured signals are the sum of direct signals (traveling directly from the source to the microphones) and delayed signals (arriving at the microphones after being reflected by the walls of the room). Reverberation degrades the fidelity and intelligibility of speech and causes severe problems for applications such as automatic speech recognition, hearing aids and hands free telephony [1]. The dereverberation problem consists of recovering a source signal from observed reverberant signals. Although much effort has been devoted to the dereverberation problem using both single- and multi-microphone techniques, speech dereverberation remains a challenging problem (see for example [2] and the reference therein). Multi-microphone techniques appear particularly interesting because theoretically perfect inverse filtering can be achieved provided that the room acoustics are known [3].

When no *a priori* knowledge of the room acoustics is available, the problem is referred to as blind dereverberation [4, 5]. An important blind dereverberation technique is called blind deconvolution. In general, it is assumed that the source signal is independent and identically distributed (i.i.d.). However, this assumption does not hold for speech-like signals. The speech generating process is deconvolved when applying such deconvolution techniques to speech. Consequently, the speech signal is excessively whitened.

One of the algorithms which address the whitening problem is called LInear-predictive Multi-input Equalization (LIME) [4, 5]. This algorithm uses multi-channel linear prediction to calculate a set of prediction filers and a compensation filter. Firstly, the filters are calculated from the captured microphone signals. Secondly, the prediction residual signal is estimated using the prediction filters. Finally, the source signal is recovered by applying the compensation filter to the prediction residual signal. It should be noted that the LIME algorithm requires a relatively large speech segment to estimate the required filters. The minimum segment length depends on the length of the room impulse responses, and the number of microphones. In [5] a segment of 2 seconds was used for a reverberation time of 0.48 seconds. The authors suggest [4] that the calculated filters might be used to dereverberate future segments. Unfortunately, the slightest change in the room might render these filters useless.

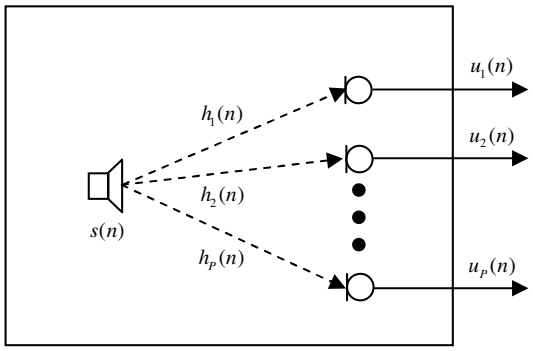


Figure 1: Acoustic system.

In this article we show that even when all of LIME's preliminary conditions hold, LIME does not always produce a perfect reconstruction of the source signal. Moreover, for certain signal segments LIME's output might be much more distorted than its reverberant input. We propose two methods to identify those signal segments and a method to dereverberate them using a set of filters that was calculated utilizing segments for which LIME produced good results.

This paper is organized as follows. In Section 2 we formulate the dereverberation problem. In Section 3 we introduce the LIME algorithm and analyze its performance. In Section 4 we develop two methods to detect those segments for which LIME does not perform well. In addition, we propose a method which allows us to dereverberate these segments in an alternative way. Experimental results are presented and discussed in Section 5. Finally, Section 6 contains our conclusions.

2. PROBLEM FORMULATION

We consider an acoustic system with one source and P microphones as shown in Fig. 1. The room impulse response between the source and the i-th microphone is called $h_i(n)$ where $h_1(n)$ is chosen to be the response of a unit impulse to the microphone closest to the source. The signal received by

the i-th microphone $u_i(n)$ can be modeled by the input signal convolved with $h_i(n)$, $i \in \{1,\ldots,P\}$, i.e.,

$$\begin{aligned} u_i(n) &= h_i(n) * s(n) \\ &= \sum_{k=0}^{M-1} h_i(k)s(n-k), \end{aligned} \tag{1}$$

where M is the number of taps of the room impulse response.

The blind dereverberation problem consists in recovering the source signal $s(n)$ from the P observed microphone signals $u_i(n)$ $i \in \{1,\ldots,P\}$.

3. LINEAR-PREDICTIVE MULTI-INPUT EQUALIZATION

Firstly, the required assumptions for the LIME algorithm are provided. Secondly, the LIME algorithm proposed in [4, 5] is summarized. Finally, we discuss a number of limitations of the LIME algorithm.

3.1 Hypotheses

The LIME algorithm assumes the following hypotheses [4, 5]:

1. The room transfer functions (RTF) are modeled using time invariant polynomials and assumed to share no common zeros. The RTFs are defined as

$$H_i(\mathrm{z}) = \sum_{k=0}^{M-1} h_i(k)\mathrm{z}^{-k} \quad i \in \{1,\ldots,P\}.$$

Using matrix formulation (1) can be rewritten as:

$$\mathbf{u}_i(n) = \mathbf{H}_i^T \mathbf{s}(n)$$

where $\mathbf{u}_i(n) = [u_i(n),\ldots,u_i(n-L+1)]^T$, $\mathbf{H}_i$ is a $(M+L-1)\times L$ convolution matrix expressed as

$$\mathbf{H}_i = \begin{pmatrix} h_i(0) & 0 & \ldots & 0 \\ h_i(1) & h_i(0) & \ddots & \vdots \\ \vdots & & \ddots & \\ h_i(M-1) & & & \\ 0 & h_i(M-1) & & h_i(0) \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & 0 & h_i(M-1) \end{pmatrix},$$

and $\mathbf{s}(n) = [s(n),\ldots,s(n-N+1)]^T$.
The length of the signal vector $\mathbf{u}_i(n)$ is denoted by L, and its minimum length is derived from the condition

$$L \geq \frac{M-1}{P-1}.$$

2. The input signal $s(n)$ is assumed to be generated from a finite AR process applied to a white noise $e(n)$. The Z-transform of the AR process is $1/a(\mathrm{z})$, where $a(\mathrm{z})$ is the AR polynomial

$$a(\mathrm{z}) = 1 - \{a_1\mathrm{z}^{-1} + \ldots + a_N\mathrm{z}^{-N}\},$$

where N denotes the length of the AR polynomial. Here the long-term AR process of the speech is modeled by $1/a(\mathrm{z})$, rather than the short-term AR process. The length of N is given by

$$N = M + L - 1.$$

Using matrix formulation, one can write

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

where $\mathbf{C}$ is the $N \times N$ companion matrix defined as:

$$\mathbf{C} = \begin{pmatrix} a_1 & 1 & 0 & \ldots & 0 \\ a_2 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 \\ a_N & 0 & \ldots & \ldots & 0 \end{pmatrix}$$

and $\mathbf{e}(n) = [e(n),0,\ldots,0]^T$.

3.2 Algorithm

The LIME algorithm consists of the following steps:

1. Both the prediction filter $\mathbf{w}$ and the AR polynomial $a(\mathrm{z})$ are estimated from a matrix $\mathbf{Q}$ which is defined as [4, 5]

$$\mathbf{Q} = \left(E\{\mathbf{u}(n-1)\mathbf{u}^T(n-1)\}\right)^{\dagger} E\{\mathbf{u}(n-1)\mathbf{u}^T(n)\}.$$

where $\mathbf{u}(n) = [\mathbf{u}_1^T(n),\ldots,\mathbf{u}_P^T(n)]^T$, $\mathbf{A}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{A}$, and $E\{\cdot\}$ denotes the time averaging operator. Here, the covariance matrix is estimated using

$$E\{\mathbf{x}(n)\mathbf{y}^T(n)\} = \frac{1}{N_s}\sum_{n=0}^{N_s-1} \left(\mathbf{x}(n)-\mathbf{m}_{\mathbf{x}}\right)\left(\mathbf{y}(n)-\mathbf{m}_{\mathbf{y}}\right)^T,$$

where N_s denotes the length of the reverberant signal segment, $\mathbf{x}(n) = [x(n),\ldots,x(n-L+1)]^T$, $\mathbf{y}(n) = [y(n),\ldots,y(n-L+1)]^T$, and $\mathbf{m}_{\mathbf{x}}$, $\mathbf{m}_{\mathbf{y}}$ are their mean vectors respectively. The mean vectors are calculated using

$$\mathbf{m}_{\mathbf{x}} = \frac{1}{N_s}\sum_{n=0}^{N_s-1}\mathbf{x}(n) \text{ and } \mathbf{m}_{\mathbf{y}} = \frac{1}{N_s}\sum_{n=0}^{N_s-1}\mathbf{y}(n).$$

The first column of $\mathbf{Q}$ gives the prediction filter $\mathbf{w}$, and an estimate of the AR polynomial $a(\mathrm{z})$ is obtained from the characteristic polynomial of $\mathbf{Q}$.

2. The prediction residual is defined as [4, 5]

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

The residual signal is free from the effect of room reverberation but is also excessively whitened. Filtering the prediction residual with $1/\hat{a}(\mathrm{z})$ produces the recovered input signal multiplied by a factor of $h_1(0)$.

3.3 Limitations

The computation of large covariance matrices causes LIME to be a computationally exhaustive algorithm. This problem might have been eased by the fact that in theory, if all LIME's preliminary assumptions hold, it should perform perfect dereverberation. In practice, even when all these hypothesizes are valid, LIME may not perform well, and produce

signals whose audible quality is worse than that of the reverberant input signals. Experimental results demonstrate that the algorithm's performance is signal dependent, i.e., applying the algorithm to different segments of the same speech signal produces different performance levels.

An example illustrating the signal dependent performance of LIME is depicted in Fig. 2. This figure shows the result of an experiment in which LIME has been applied to different segments of a reverberant speech signal captured by two microphones in a room. The room dimensions were 5.04 m×6.35 m×4 m (length×width×height) the source location was (2.3 m, 1.7 m, 1.95 m) and the two microphone locations were (2.5 m, 3.15 m, 2.08 m) and (2.1 m, 3.25 m, 2.1 m). The room impulse responses were 800 taps long, and were calculated using the image method [6]. The male speech signal was 40960 samples long and its sampling rate was 16 kHz. Each segment for dereverberation was 14000 samples long, and the overlap between the segments was 14000-256 samples. LIME's performance was assessed using the log spectral distortion (LSD) [7] and segmental SNR (segSNR) [7] measures. Fig. 2(a) and (b) show that in most segments the segSNR is high and the LSD between LIME's estimate and the source signal is much lower than the LSD between the reverberant signal and the source signal. This indicates that LIME removes most of the reverberation effects in those segments. On the other hand in other segments the segSNR is low and the LSD between LIME's estimate and the source signal is much higher than that between the reverberant signal and the source signal. This indicates that LIME produces an estimate which is very different from the source signal in these segments.

In practice, the LSD and segSNR measures cannot be used to assess LIME's performance since the input signal is unknown. Thus a different method is required to assess whether LIME produces good estimates of the source signal.

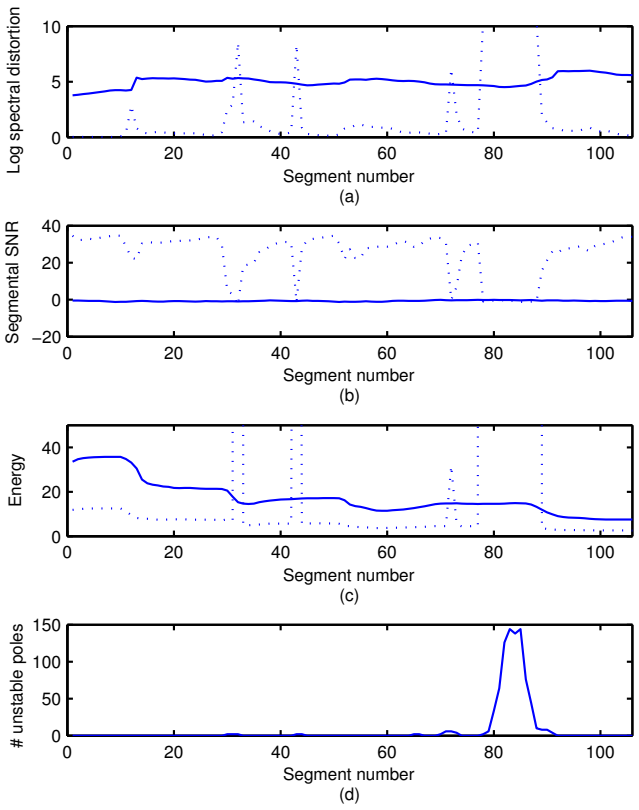


Figure 2: (a) LSD between the reverberant and source signal (solid) and between LIME's estimates and the source signal (dotted), (b) segSNR of the reverberant signal segments (solid) and segSNR of LIME's estimates (dotted) (c) energy of the reverberant (solid) and the estimated (dotted) segments (d) number of unstable poles of each segment's estimated AR process.

4. PROPOSED SOLUTION

In this section we propose two non-intrusive methods to assess the performance of the LIME algorithm. In addition, we propose to construct a set of filters that can be applied to the received signals or to those signal segments for which LIME fails to dereverberate the received signal.

4.1 Performance assessment

One way to assess the performance of the algorithm's output is to look at its energy. In general, when comparing two signal segments, the segment with less energy contains less reverberation. Thus, if the energy of LIME's estimate is lower than the reverberant signal's energy, there is high probability that LIME's output is less reverberant than its input signal. Fig. 2(c) shows that the energy of the estimated signal is indeed higher than that of the reverberant signal in segments where LIME did not perform well.

Another non-intrusive method to asses the performance is obtained by analyzing the cause of the problem. Recall that LIME estimates the long-term AR process. This estimate is used to construct the compensation filter $1/a(\mathrm{z})$, which is applied to the prediction residual signal $\hat{e}(n)$. When analyzing the number of unstable poles of $1/a(\mathrm{z})$, i.e., the number of zeros of $a(\mathrm{z})$ for which the amplitude is larger than 1, it was found that LIME does not perform well in case the number of unstable poles is larger than zero. In Fig. 2(d) the number of unstable poles of $1/a(\mathrm{z})$ is shown for each segment. While all the AR processes calculated in segments where LIME didn't perform well contain unstable poles, the number of unstable poles is zero or small when LIME did perform well. Therefore, by discarding LIME's results in segments where the AR polynomials have zeros outside the unit circle we avoid using bad estimates of LIME at the cost of loosing a small number of estimates from segments where LIME performance was satisfactory despite the unstable poles.

In the sequel we use both methods to assess the performance of LIME. We will now propose a method to dereverberate signals for which LIME failed to produce good estimate of the source signal.

4.2 Dereverberation using least squares filters

When LIME produces a poor estimate of the source signal or a certain part of it, a different approach is required in order to perform dereverberation.

Here we assume that we have at least one segment of the same reverberant speech signal which LIME successfully dereverberated. Using this segment, a set of filters is calculated which minimizes the least squares (LS) error between LIME's estimate and the filtered microphone signals. These LS filters are then used in segments where LIME failed to dereverberate the received microphone signals.

Alternatively, the LS filters can be computed adaptively.

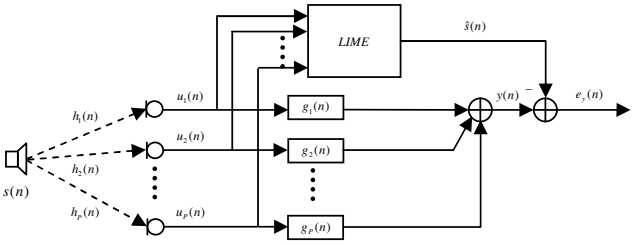


Figure 3: Proposed system to estimate the least squares filters.

In this case the filters can be updated when LIME is able to correctly estimate the source signal. The use of an adaptive estimation technique is beyond the scope of this paper.

The system used to estimate the LS filters is shown in Fig. 3. Let us define the LS filters of length $\widetilde{L}$ as $\mathbf{g}_i(n)$ $i \in \{1,\ldots,P\}$. To some extent the length of the LS filters $\widetilde{L}$ can be used to control the amount of reverberation that is reduced. Then we define the error between LIME's estimate $\hat{s}(n)$ and the filtered microphone signals as

$$e_y(n) = \sum_{i=1}^{P} \mathbf{u}_i^T(n)\mathbf{g}_i - \hat{s}(n).$$

In matrix form we get

$$\mathbf{e}_y(n) = \mathbf{U}(n)\,\mathbf{g} - \hat{\mathbf{s}}(n),$$

where $\mathbf{e}_y(n) = [e_y(n),...,e_y(n+\widetilde{N}_s)]^T$ with $\widetilde{N}_s = N_s - M$, $\mathbf{g}_i = [g_i(0),\ldots,g_i(\widetilde{L}-1)]^T$, $\mathbf{g} = [\mathbf{g}_1^T,\ldots,\mathbf{g}_P^T]^T$, $\hat{\mathbf{s}}(n) = [\hat{s}(n),...,\hat{s}(n+\widetilde{N}_s)]^T$, $\mathbf{U}_i(n)$ is a $(\widetilde{N}_s+1)\times\widetilde{L}$ matrix expressed as

$$\mathrm{U}_i(n) = \begin{pmatrix} u_i(n) & 0 & \ldots & 0 \\ u_i(n+1) & u_i(n) & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ u_i(n+\widetilde{L}-1) & u_i(n+\widetilde{L}-2) & \ddots & u_i(n) \\ \vdots & \vdots & \ddots & \vdots \\ u_i(n+\widetilde{N}_s) & u_i(n+\widetilde{N}_s-1) & \ldots & u_i(n+\widetilde{N}_s+\widetilde{L}-1) \end{pmatrix},$$

and $\mathbf{U}(n) = [U_1(n),\ldots,U_P(n)]$.

Minimizing the square error $|e_y(n)|^2$ gives us

$$\begin{aligned}\hat{\mathbf{g}} &= \arg\min_{\mathbf{g}} \left(\mathbf{U}(n)\,\mathbf{g} - \hat{\mathbf{s}}(n)\right)^2 \\ &= (\mathbf{U}^T(n)\mathbf{U}(n))^{-1}\mathbf{U}^T(n)\hat{\mathbf{s}}(\mathbf{n}).\end{aligned}$$

Applying these filters to the microphone signals produces an estimate of the source signal

$$y(n) = \sum_{i=1}^{P} \tilde{\mathbf{u}}_i^T(n)\mathbf{g}_i,$$

where $\tilde{\mathbf{u}}_i(n) = [u_i(n),\ldots,u(n-\widetilde{L}+1)]^T$.

Using one or both detection methods introduced in the previous subsection, segments where LIME did not perform well can be identified, and dereverberated using the LS filters.

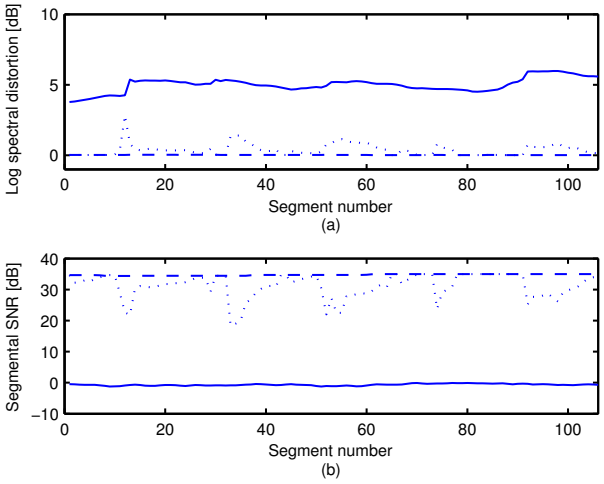


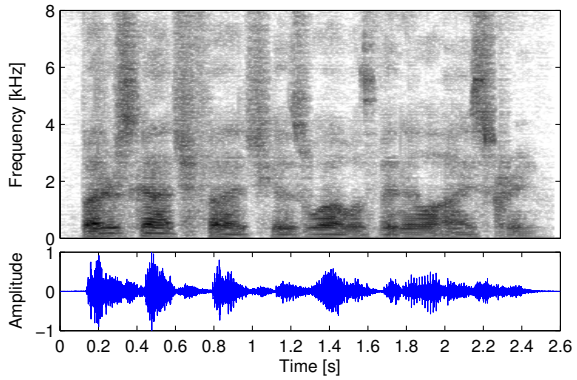
Figure 4: (a) LSD between the reverberant and source signal (solid), between the LS filters output and the source signal (dashed) and between the combined method's output and the source signal (dotted), (b) segSNR of the reverberant signal segments, segSNR of estimated segments obtained using the LS filters (dashed) and segSNR of the segments obtained using LIME and LS filters (dotted).

It should be noted that two estimates are acquired for each signal segment, viz., the first estimate is produced by LIME and the second estimate is produced by the LS filters. Here, two options are considered to combine the estimates. The first option is to use LIME's output in case the estimate obtained by LIME is assumed to be accurate, and to use the LS filters' output in case the estimate obtained by LIME is assumed to be inaccurate. In case the LS filters produce an accurate estimate of the source signal the concatenated signal segments sound well. However, in case the LS filters' output contains some reverberation, which could happen with long reverberation times, there might be a (subjectively disturbing) discontinuity between the signal segments. The second option is to continuously use the output of the LS filters to acquire a more continuous, but possibly more reverberant estimate of the source signal. Results of an informal listening test indicated that a continuous reduction level results in a higher subjective preference.

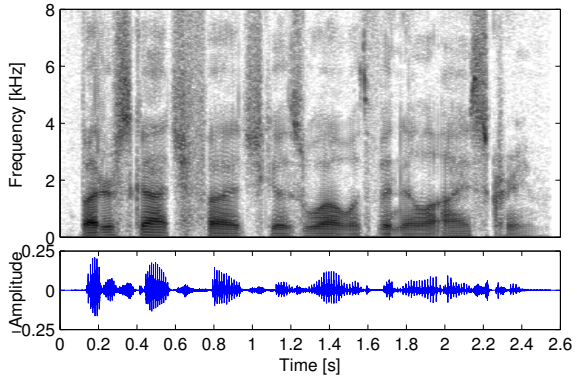
5. EXPERIMENTAL RESULTS

In order to assess the performance of the proposed algorithm we performed several experiments. The setup is described in Section 3.3. The length of the LS filters $\widetilde{L}$ was chosen to be $\widetilde{L} = \frac{M-1}{P-1} + 10$.

In the first experiment we chose the 81-st segment (shown in Fig. 2) as the signal to dereverberate and termed it segment II. The LSD of LIME in segment II was worse than the performance in any other segment. Firstly, we applied LIME to the part of the signal which precedes segment II, which we termed segment I. Both performance assessment methods indicated that LIME performed well. Therefore, we were able to use LIME's output to calculate the LS filters. Secondly, we employed the LS filters to dereverberate segment II. The LSD between the source signal and the re-



(a) Reverberant signal $u_1(n)$.



(b) Processed signal $y(n)$.

Figure 5: Spectrogram and waveform of the reverberant signal $u_1(n)$ and the output signal $y(n)$ of the LS filters.

verberant signal was 4.52 dB and the segSNR was -0.15 dB. The LSD between the estimated signal and the source signal achieved by LIME was 11.13 dB and the segSNR was -0.58 dB. The LSD between the output of the LS filters and the source signal was 0.004 dB and the segSNR was 35 dB, showing great improvement compared to LIME.

In the second experiment we applied the two proposed methods to all the speech signal's segments. While the first method uses the output of the LS filters continuously, the second method only uses the output of the LS filters in case LIME is assumed to produce a poor estimate of the source signal. Applying LIME to the first segment produced an estimate which was approved by both performance assessment methods. Therefore, LIME's estimate was used to calculate the LS filters used in every segment that was badly recovered by LIME according to either of the performance assessment methods. It is shown in Fig. 4 that the LSDs between the output of the proposed methods and the source signal is always lower than the LSD between the reverberant signal and the source signal. In addition, the segSNRs of processed signals is always higher than the segSNR of the reverberant signal. It can be seen that once the LS filters are calculated using an accurate estimate of the source signal the LS filters provide a more stable solution compared to LIME. In Fig. 5 the spectrogram and waveform of the output signal of the LS filters and the reverberant signal are shown. It can be seen that the smearing of the signal along the frequency and time axes that result from the reverberation is significantly reduced.

In the third experiment we applied the LS filters calculated in the previous experiment for a male speech signal to a female speech signal reverberating in the same room conditions. The LSD between the original signal and the reverberant signal was 4.11 dB and the segSNR was -0.71 dB. The LSD between the output of the LS filters and the original signal was 0.01 dB and the segSNR was 34.09 dB which indicates a significant reverberation reduction. This result indicates that when the spectral content of LIME's estimate is diverse enough the LS filters calculated from one speech signal can be used to dereverberate another signal that is observed under the same room conditions.

6. CONCLUSIONS

In this paper we presented a multi-microphone speech dereverberation algorithm based on the LIME algorithm and LS filtering. It was shown that the performance of LIME is signal dependent. We proposed two non-intrusive methods to evaluate the performance of LIME. Firstly, the energy of the output signal can be compared to the energy of the input signal. Secondly, the roots of the AR process estimated by LIME can be used to assess LIME's performance. Signal segments successfully dereveberated by LIME were used to calculate LS filters. The LS filters were used to dereverberate other signal segments for which LIME failed to recover the source signal. Experimental results show that the signal segments for which LIME produces poor estimates of the source signal can be detected successfully. In addition, the LS filters calculated from one signal segment were shown to reduce reverberation in different signals under the same room conditions.

7. ACKNOWLEDGEMENTS

The authors express their gratitude to dr. M. Delcroix for his support in the implementation of LIME.

REFERENCES

[1] P.A. Naylor and N.D. Gaubitch, "Speech dereverberation,"*Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sept. 2005.

[2] E.A.P. Habets, "Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement," *Ph.D. Thesis*, Technische Universiteit Eindhoven, Jun. 2007.

[3] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics,"*IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no 2, pp. 145-152, Feb. 1988.

[4] M. Delcroix, T. Hikichi, and M. Miyoshi, "Blind dereverberation algorithm for speech signals based on multichannel linear prediction,"*Acoust. Sci. Technol.*, vol. 26, no. 5, pp. 432-439, 2005.

[5] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction,"*IEEE Trans. Audio, Speech and Language Process.*, vol. 15, no. 2, pp. 430-440, Feb. 2007.

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

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