LPC-BASED SPEECH DEREVERBERATION USING KALMAN-EM ALGORITHM

Boaz Schwartz¹, Sharon Gannot¹, and Emanuel A.P. Habets²

¹Bar-Ilan University, Faculty of Engineering, Ramat-Gan, 52900, Israel
²International Audio Laboratories Erlangen*, Am Wolfsmantel 33, 91058 Erlangen, Germany

ABSTRACT
An algorithm for multichannel speech dereverberation is proposed that simultaneously estimates the clean signal, the linear prediction (LP) parameters of speech, and the acoustic parameters of the room. The received signals are processed in short segments to reduce the algorithm latency, and several expectation-maximization (EM) iterations are carried out on each segment to improve the signal estimation. In the expectation step, the fixed-lag Kalman smoother (FLKS) is applied to extract the clean signal from the data utilizing the estimated parameters. In the maximization step, the parameters are updated according to the output of the FLKS. Experimental results show that multiple EM iterations and the application of the LP model improve the quality of the output signal.

1. INTRODUCTION
Speech signals are commonly degraded by reverberation, and to overcome this problem, numerous approaches were developed in the last few decades. The approaches can be broadly classified into those that mainly exploit the properties of the acoustic channel and those that mainly exploit the properties of the source signal. Only few approaches exploit both properties.

Algorithms that focus on dereverberation of the residue of a LP analysis were proposed in [1] and [2]. The residual signal is obtained from the reverberant signal by the estimated LP parameters, and then processed in order to reduce the reverberant component. The dereverberated signal is then reproduced from the processed residual and the LP parameters.

In [3] [4], an EM algorithm for dereverberation and noise reduction in the time-frequency domain is presented. In the E-step the Wiener filter, calculated using the current values of the room and LP parameters, is used to estimate the clean speech signal. In the M-step, the current signal estimate is used to update parameters. Weinstein et al. [5] presented an EM algorithm for noise reduction in time-domain. The Kalman smoother is used in the E-step to estimate the speech and noise signals from the mixed measurements, and the room impulse response (RIR) is updated in the M-step. Since this is an iterative EM algorithm, we refer to this method as the iterative Kalman-EM (KEM) method. In [6], the iterative-KEM scheme was used for single-microphone speech enhancement in the time domain. A recursive version of the KEM scheme was also proposed in [6], where the Kalman smoother was substituted by a fixed-lag Kalman smoother (FLKS), adjusting the scheme to online speech enhancement in the time domain. We refer to this method as the recursive-KEM method.

Recently, we developed an algorithm for dereverberation in the short-time Fourier transform (STFT) domain [7] that is based on the iterative-KEM scheme. In the E-step, the Kalman smoother is applied to extract the clean signal from the data utilizing the estimated parameters. In the M-step, the parameters are updated according to the output of the Kalman smoother. We refer to this algorithm as the Kalman-EM for dereverberation (KEMD). For simplicity, the statistical model for speech assumed no correlation between frequency bands. Nevertheless, experimental results show significant dereverberation capabilities of the proposed algorithm with only low speech distortion. Recently, a recursive version of the KEMD algorithm was developed in [8].

In this paper we develop a new KEM algorithm for dereverberation in the STFT domain. In contrast to [7, 8] we use a LP model for speech to exploit the smoothness of speech power spectral density (PSD). In contrast to [3], [4], and [7], in which iterations are performed on the entire data, the proposed algorithm iterates on short segments, allowing a reasonable online performance.

2. PROBLEM FORMULATION
A common representation for speech signals is the short-time LP model [9],
\[ x_n = \sum_{q=1}^{Q} x_{n-q} a_{\ell,q} + w_n , \]  
where \( w_n \sim \mathcal{N} \{ 0, \phi_w(\ell) \} \)

\[ x(\ell, k) \sim \mathcal{N}_{\mathbb{C}} \{ 0, \phi_s(\ell, k) \} . \]

The speech variance is a function of the LP parameters, i.e.,
\[ \phi_s(\ell, k) = \phi_w(\ell) \cdot \left| 1 - a_\ell^T e_k \right|^{-2} , \]

where \( a_\ell = [ a_{\ell,1}, \ldots, a_{\ell,Q} ]^T \), \( e_k = \left[ e^{-1} 2\pi jk \frac{\Delta f}{K}, e^{-2} 2\pi jk \frac{\Delta f}{K}, \ldots, e^{-Q} 2\pi jk \frac{\Delta f}{K} \right]^T \) is the transpose operator, and \( K \) is the number of frequency bins.

This research was supported by a Grant from the GIF, the German-Israeli Foundation for Scientific Research and Development.

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.
The noisy and reverberant observations are modelled by using the convolutive transfer function (CTF) approximation \[11\]. Using this model, the \( j \)-th microphone signal can be expressed as
\[
\begin{align*}
z_j(\ell, k) & \approx \sum_{l=0}^{L-1} h_{j,l}(k) \cdot x(\ell - l, k) + v_j(\ell, k),
\end{align*}
\] (4)
where \( L \) is the CTF length that depends on the reverberation time. We further assume that \( v_j(\ell, k) \) are independent and stationary complex-Gaussian random variables:
\[
v_j(\ell, k) \sim \mathcal{N}\{0, \phi_v(\ell, k)\}. \tag{5}
\]

While the spectral coefficients \( x(\ell, k) \) and \( x(\ell', k') \) are statistically dependent, they are independent given \( \phi_v(\ell, k) \) and \( \phi_v(\ell', k') \) \[12\]. According to (3), the speech variances, \( \phi_v(\ell, k) \), for all \( \ell \) and \( 0 \leq k \leq K - 1 \) are fully defined by \( \alpha_c \) and \( \phi_v(\ell) \).

To calculate the minimum mean square error (MMSE) estimates of the clean speech signal (i.e., the dereverberated signal), we define the state-space equations\(^1\)
\[
x_t = \Phi x_{t-1} + u_t, \quad z_t = H x_t + v_t, \tag{6}
\]
where the vectors are defined by
\[
x_t \equiv [x(\ell - L + 1), \ldots, x(\ell)]^T, \quad u_t \equiv [0, \ldots, x(\ell)]^T, \quad z_t \equiv [z_1(\ell), \ldots, z_J(\ell)]^T, \quad v_t \equiv [v_1(\ell), \ldots, v_J(\ell)]^T,
\]
and \( J \) is the number of microphones. The process and measurement matrices are, respectively, equal to
\[
\Phi \equiv \begin{pmatrix} 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \cdots & 0 & 1 \\ 0 & \cdots & \cdots & 0 \end{pmatrix}, \quad H \equiv [h_1, \ldots, h_J]^T,
\]
where \( h_j = [h_{j,L-1}, \ldots, h_{j,0}]^T \).

The second-order statistics of \( u_t \) and \( v_t \) are defined as
\[
F(\ell) \equiv E\{u_t u_t^H\} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_v(\ell) \end{pmatrix},
\]
and
\[
G \equiv E\{v_t v_t^H\} = \begin{pmatrix} \phi_v & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_v \end{pmatrix},
\]
where \((\cdot)^H\) is the conjugate transpose operator. As the noise is assumed to be stationary and spatially uncorrelated, the matrix \( G \) is time-invariant and diagonal.

### 3.1. Review of the KEMD Algorithm
In \[7\] we presented the KEMD algorithm that estimates both the clean speech signal and the acoustic parameters using the complete data set denoted by \( Z \) that is defined as
\[
Z \equiv \{z_j(\ell, k), \ell \in T, k \in K, j \in J\}, \tag{7}
\]
where \( T = \{0, \ldots, T - 1\} \) are time indices in the STFT domain, \( K = \{0, \ldots, K - 1\} \) are the frequency indices, and \( J = \{0, \ldots, J - 1\} \) are the microphone indices. The goal of the KEMD algorithm is to estimate the clean speech signal \( x(\ell, k) \) for \( \ell \in T \) and \( k \in K \), given the noisy and reverberant observations \( Z \).

A maximum likelihood (ML) approach was adopted to estimate the clean speech signal and the acoustic parameters that are denoted by \( \Theta \). Since the clean speech signal is unobservable, it is defined as a latent data set and the EM algorithm is applied. In the E-step, the Kalman smoother is used to extract the clean speech signal using the available estimates of the acoustic parameters. In the M-step, the estimate of \( \Theta \) is updated by using the previously estimated signal. The algorithm iterates until convergence. It was shown that the KEMD algorithm significantly reduces reverberation while preserving the speech quality.

In \[7, 8\] we assumed that the clean speech samples are independent, i.e., \( x(\ell, k) \) is independent of \( x(\ell', k') \) for every \( (\ell, k) \neq (\ell', k') \). The ML estimates for \( \phi_v(\ell, k) \) in \[7\] are given by
\[
\hat{\phi}_v^{(p+1)}(\ell, k) = E\{|x(\ell, k)|^2 | Z, \hat{\Theta}^{(p)}\}, \tag{8}
\]
where \((\cdot)^{(p)}\) denotes the estimate calculated at the \( p \)-th iteration, and \( E\{|Z, \hat{\Theta}^{(p)}\} \) is the expected value given the observation \( Z \) and the current parameter estimate \( \hat{\Theta}^{(p)} \). According to the statistical model of speech signal, the estimates of \( \phi_v(\ell, k) \) and \( \phi_v(\ell, k') \) are also independent.

### 3.2. Segmental Iterative EM Algorithm
In this contribution we propose to calculate the variances of the clean speech assuming a LP model. The number of parameters is consequently decreased and the estimated PSD becomes smoother, leading to a more naturally sounding output signal.

Due to the nature of the reverberant signal, the correlation between \( x(\ell, k) \) and \( z_j(\ell + \Delta, k) \) rapidly decays for \( \Delta > L \). We therefore propose to process short segments in a sequential manner. This way, the latency of the algorithm is reduced without significantly impairing the performance. The data in the \( d \)-th segment is defined as
\[
\mathcal{Z}_d \equiv \{z_j(\ell, k), k \in K, j \in J, \ell \in T_d\}, \tag{9}
\]
where \( T_d \) is the set of STFT time-frames of the \( d \)-th segment.

Instead of the iterative-KEM and the Kalman smoother used in \[7\], we adopt the recursive-KEM scheme \[6\] and apply the FLKS in each iteration. Accordingly, each iteration starts with the recursive estimation of both clean speech signal and the acoustic parameters. Then, the next iteration is carried out on \( \mathcal{Z}_d \), utilizing the estimated parameters from the previous iteration. After \( N_p \) iterations are completed, the subsequent segment is processed.

The E-step for time-frame \( \ell \) comprises a single step of FLKS to obtain the new state vector \( \hat{x}_t(\ell) \). One step of the FLKS is summarized in Algorithm 1, where \((\cdot)^{(p)}_\ell\) denotes a parameter estimator at time-frame \( \ell \) and iteration \( p \).

\(^1\)When applicable, the frequency index is omitted for brevity.
Algorithm 1: One step of fixed-lag Kalman smoother.

Predict:
\[ \hat{\mathbf{z}}_{\ell|\ell-1} = \Phi \hat{\mathbf{x}}_{\ell-1|\ell-1} + \mathbf{F}(p-1) \]
\[ \mathbf{P}_{\ell|\ell-1} = \Phi \mathbf{P}_{\ell-1|\ell-1} \Phi^T + \mathbf{F}(p-1) \]

Update:
\[ \mathbf{K}_{\ell|\ell} = \mathbf{P}_{\ell|\ell-1} \left( \mathbf{H}_{\ell|\ell-1}^T \mathbf{P}_{\ell|\ell-1} \mathbf{H}_{\ell|\ell-1} + \mathbf{G}(p) \right)^{-1} \]
\[ \hat{\mathbf{x}}_{\ell|\ell} = \mathbf{P}_{\ell|\ell-1} \mathbf{H}_{\ell|\ell-1}^T \hat{\mathbf{z}}_{\ell|\ell-1} \]
\[ \mathbf{P}_{\ell|\ell} = \left( \mathbf{I} - \mathbf{K}_{\ell|\ell} \mathbf{H}_{\ell|\ell-1} \right) \mathbf{P}_{\ell|\ell-1} \]

Algorithm 2: Summary of the segKEMD-LP algorithm.

1. Initialize with previous segment values:
   \[ \hat{\Omega}(0)(dD) \leftarrow \Omega(0)(dD-1), \]
   for
   \[ \Omega \in \{ \Omega_j, \Omega_{zz}, \Omega_{xz}, \Omega_{rz}, \Omega_{rz} \} \].

2. For each \( \ell \in \mathcal{T}_d \):
   - Estimate \( |x(\ell, k)|^2 \) using (17).
   - Calculate the LP parameters of clean speech, using (15), and then (13).
   - Initialize the variance of speech \( \hat{\mathcal{G}}(\ell) \) using (3).

3. For \( p = 1 \) to \( N_p \) do
   - Reset \( \hat{\mathbf{x}}_p(dD-1|dD-1), \mathbf{P}_p(dD-1|dD-1) \) to their values in the last iteration of the previous segment.
   - Initialize parameters and statistics for current iteration:
     \[ \hat{\Omega}(p)(dD) \leftarrow \Omega(0)(dD-1) \]
   - For \( \ell \in \mathcal{T}_d \) do
     - E-step
       - Execute one step of the FLKS using Algorithm 1, to obtain \( \hat{\mathbf{x}}_{\ell|\ell}, \hat{\mathbf{P}}_{\ell|\ell} \).
     - M-step
       - Update the statistics using (12).
       - Propagate the acoustic parameters,
         \( \hat{\mathbf{R}}_p(\ell|\ell) \) and \( \hat{\mathbf{p}}(\ell|\ell) \), using (10)-(11).
       - if \( \ell \geq dD + L - 1 \) then
         - Estimate \( |x(\ell, L + 1)|^2 \) using (16).
         - Calculate the LP parameters of clean speech, using (15), and then (13).
         - Estimate speech variance, \( \hat{\mathcal{G}}(\ell - L + 1) \), using (3) in the next iteration.
         - end
   - end

3.3. Estimation of the LP Parameters of Speech

As described in Algorithm 2, the LP parameters, \( \alpha_t(\ell) \) and \( \phi_\omega(\ell) \), are initially estimated at the beginning of the segment and re-estimated in the M-step at time-frame \( \ell + L - 1 \). In both cases, the estimators are obtained by solving the Yule-Walker equations

\[ \alpha_t = C_t^{-1} c_t \]
\[ \phi_\omega = c_t^T c_t , \]

where \( C_t \) and \( c_t \) are given by

\[ C_t = \begin{bmatrix} c_t(0) & \cdots & c_t(Q-1) \\ \vdots & \ddots & \vdots \\ c_t(Q-1) & \cdots & c_t(0) \end{bmatrix} , \]
\[ c_t = \begin{bmatrix} c_t(1) \\ \vdots \\ c_t(Q) \end{bmatrix} , \]

and \( c_t(q) \) is the \( q \)-th correlation coefficient of the clean speech signal at time-frame \( \ell \):

\[ c_t(q) = E \{ x_n x_{n-q} \} , n \in \mathcal{N}_t, 0 < q < Q , \]
where $N_e$ are the time-domain samples that are analyzed in the $\ell$-th STFT frame, and $x_n$ is the speech signal at time sample $n$ as in (1). The correlation coefficients are obtained using

$$
\hat{\rho}(p) = \frac{1}{K} \sum_{k=0}^{K-1} |x(\ell, k)|^2 e^{j \frac{2\pi k p}{N}}.
$$

Similarly to the first-order statistics estimator, we use fixed-lag smoothing for the second-order statistics, i.e.,

$$
[x(\ell - L + 1, k)]^2 = \hat{R}^{(p)}_{xx}(\ell, k)_{1,1},
$$

where $[i, j]$ denotes the $(i, j)$-th element. When we approach the end of the segment we gradually reduce the lag size to estimate $|x(\ell, k)|^2$ as the fixed-lag data becomes unavailable.

The estimator in (16) is computed using the FLKS output. This estimator, however, is only available for iterations $p > 1$, for which the FLKS output was already calculated. In [13], it is shown that when multiple microphones are used, spatial averaging of the correlation coefficients improves the accuracy of the LP estimators. Applying this idea to the STFT domain, we estimate the PSD at time-frame $\ell$ using spatial averaging and spectral subtraction, i.e.,

$$
|x(\ell, k)|^2 = \frac{1}{J} \sum_{j=1}^{J} \left[ |z_j(\ell, k)|^2 - \phi_{\nu_j} (dD - 1, k) \right].
$$

### 4. PERFORMANCE EVALUATION

The dereverberation performance of the segKEMD-LP algorithm was evaluated using reverberant speech signals recorded in a var-echoic chamber at Bar-Ilan University. The room dimensions are $6 \times 5.9 \times 2.3$ m (length $\times$ width $\times$ height). Two different reverberation times ($T_{60}$) were tested: 480 and 630. Different speech signals related to eight different speakers, from the TIMIT database were played from Fostex 6301BX loudspeakers. Each loudspeaker was located in one of six positions, at distances between 1.3 and 4 m from the microphones, and with different angles. The reverberant signals were captured by four AKG CK32 omni-directional microphones, placed on a linear array with total length of 25 cm. The measured reverberant signal to noise ratio (RSNR) is 40 dB. A 60 s clean speech signal was constructed by using eight different utterances from the same speaker taken from the TIMIT database. The total number of experiments therefore equals 96, comprising 96 minutes of reverberant speech.

The STFT analysis window was set to a 32 ms Hamming window with 50% overlap, the system length $L$ was set to 20 frames, the segment length was set to 1.6 sec, and the LP order $Q$ was set to 12.

The proposed algorithm (segKEMD-LP) was compared with a variant (segKEMD-P), which utilizes the segmental scheme of segKEMD-LP, without applying the LP model. Specifically, $\phi_{\nu}(\ell, k)$ was estimated using (17) and (16) rather than (15), and (13). The two variants were tested using one or three iterations. Sonograms for the reverberant, segKEMD-P and segKEMD-LP output signals (both with 3 iterations) are depicted in Fig. 1. Some of the residual reverberation in segKEMD-P is reduced due to the LP estimates. In addition, informal hearing tests revealed that in segKEMD-LP, the musical noise is lower than in segKEMD-P, and the output sounds more natural.

![Fig. 1. Sonograms of the reverberant signal (top), segKEMD-P (middle), and segKEMD-LP (bottom), for $T_{60} = 630$ ms.](image)

### Table 1. Average performance obtained for segKEMD-P and segKEMD-LP.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Algorithm</th>
<th>$N_p$</th>
<th>Input</th>
<th>Output</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSRNR</td>
<td>segKEMD-P</td>
<td>1</td>
<td>7.03</td>
<td>9.72</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td>segKEMD-LP</td>
<td>1</td>
<td>9.84</td>
<td>9.84</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>segKEMD-P</td>
<td>3</td>
<td>10.25</td>
<td>10.25</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>segKEMD-LP</td>
<td>3</td>
<td>10.47</td>
<td>10.47</td>
<td>0.00</td>
</tr>
<tr>
<td>MFCC</td>
<td>segKEMD-P</td>
<td>1</td>
<td>10.61</td>
<td>9.68</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>segKEMD-LP</td>
<td>1</td>
<td>9.55</td>
<td>9.55</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>segKEMD-P</td>
<td>3</td>
<td>9.66</td>
<td>9.66</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>segKEMD-LP</td>
<td>3</td>
<td>9.54</td>
<td>9.54</td>
<td>0.00</td>
</tr>
</tbody>
</table>

We used two objective measures to evaluate performance: the frequency-weighted signal to noise ratio (WSNR) [14] and the distance between mel-frequency cepstrum coefficients (MFCC) [15]. Here we refer to the WSNR as the frequency-weighted signal to reverberation and noise ratio (WSRNR). Each of the variants was tested using the entire data set described above, and the WSRNR and MFCC were calculated for every signal. The average results are summarized in Table 1. In all cases, the segKEMD-LP outperforms the segKEMD-P. Furthermore, we notice that the additional iterations (i.e., $N_p = 3$ instead of $N_p = 1$), increases the WSRNR and decreases the MFCC distance.

### 5. CONCLUSIONS

An EM algorithm for dereverberation in the STFT domain was presented, where the reverberant signal is processed in short segments. Several iterations are carried out for each segment in order to improve the estimated parameters. The recursive-KEM scheme was utilized to estimate acoustic parameters of the room and the LP model was used for the estimation of speech variances. Splitting the signal into segments decreases the algorithm latency, while the LP model was shown to improve the quality of the output signal. Repeated iterations were shown to improve the accuracy and the quality of the processed signal.
6. REFERENCES


