Adaptive Time Delay Estimation Using Filter Length Constraints for Source Localization in Reverberant Acoustic Environments

Daniele Salvati and Sergio Canazza

Abstract—Adaptive time delay estimation based on blind system identification (BSI) focuses on the impulse responses between a source and a microphone to estimate the time difference of arrival (TDOA) in reverberant environments. In this letter, we consider the adaptive eigenvalue decomposition (AED) BSI method based on the normalized multichannel frequency-domain least mean square (NMCFLMS) algorithm. We show that the use of filter length constraints (FLC) based on the maximum TDOA between microphones improves the performance of the NMCFLMS filter for the localization of different sound types in highly reverberant environments. The experimental results demonstrate the improvement of the proposed method for reverberation times (RT₅₀) of up to 2 s. Applications for this method include teleconferencing systems, musical interfaces, videogames, and monitoring systems.

Index Terms—Blind system identification, filter length constraint, microphone array, reverberant environment, time delay estimation.

I. INTRODUCTION

ACOUSTIC source localization (ASL) is rapidly gaining importance in a growing number of applications. In highly reverberant environments, accurate ASL may be attractive for telecommunication systems, musical control interfaces, monitoring systems, and videogames. The aim of an ASL system is to estimate a sound source position in space by analyzing the sound field using a microphone array, which is a set of microphones arranged to acquire spatial information for sounds. Time delay estimation (TDE) methods include estimating the time difference of arrival (TDOA) between a pair of microphones. The most used techniques are generalized cross-correlation (GCC) [1] and adaptive eigenvalue decomposition (AED) [2] based on the blind system identification (BSI), which focuses on impulse responses between a source and the microphones. The AED fully considers reverberation, and one of its advantages is that it enhances performance under highly reverberant conditions. AED extension for multiple microphones has been proposed [3]. By introducing the channel impulse response \( h_m \), from the source to the microphone \( m \),

the reverberant model for discrete-time signals received by \( M \) sensors can be expressed as

\[
    x_m(k) = h_m * s(k) + v_m(k) \quad m = 1, 2, \ldots, M
\]

(1)

where \( s[k] \) is the source signal, \( v_m[k] \) the uncorrelated noise signal, and \( * \) denotes convolution. Assuming that the system is linear and time invariant, and neglecting the influence of noise in the (1), the cross-relation for a microphone pair can be written as

\[
    x_i[k] * h_j = x_j[k] * h_i \quad i, j = 1, 2, \ldots, M.
\]

(2)

Adaptive TDE based on the AED method aims at minimizing the following cost function using the least mean square (LMS) filter

\[
    J(k + 1) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} e_{ij}^2(k+1) / \| \hat{h}_i[k] \|^2
\]

(3)

where \( e_{ij}(k+1) = (x_i^T(k+1)\hat{h}_i(k) - x_j^T(k+1)\hat{h}_i(k)) \) is the signal error (in which \( x_m(k) = [x_m(k), x_m(k-1), \ldots, x_m(k-L+1)]^T \) is the observed vector at time \( k \), and \( \hat{h}_m[k] = [\hat{h}_m,\hat{h}_m,\ldots,\hat{h}_m^-L+1]^T \) is the estimated impulse response for length \( L \), and \( \hat{h}_i[k] = [\hat{h}_i,\hat{h}_i,\ldots,\hat{h}_i^-L] \) \( \cdot \) denotes the unit norm).

However, the time-domain LMS convergence rate is slow, and to accelerate convergence and to provide an efficient implementation, a normalized multichannel frequency-domain LMS (NMCFLMS) algorithm was proposed in [4]. Finally, the TDOA between two microphones can be estimated by comparing direct paths of the responses and by assuming that they are related to the highest peak

\[
    \hat{r}_{ij} = \arg \max_i |\hat{h}_{ij}(k)| - \arg \max_i |\hat{h}_{ij}(k)|
\]

(4)

To enhance NMCFLMS algorithm robustness, we propose herein to use filter length constraints (FLC) based on the maximum TDOA between a pair of microphones. In a highly reverberant condition, the NMCFLMS has the disadvantage of estimating the highest value of the responses, which does not always correspond to the direct path (due to the temporal whitening), and the FLC has the advantage to reduce the area of convergence increasing the likelihood that the highest value of the responses is related to the direct path of the acoustic source. In fact, the goal of adaptive TDE is to estimate the direct response paths. The reference response is initialized with
in the middle of the filter, and this position is maintained during the filter update. Considering that the maximum TDOA is given in samples by \( \tau_{\text{max}} = \frac{d s_f}{v} \) (where \( d \) is the distance between the microphones, \( f_s \) is the sampling frequency and \( v \) is the speed of sound), the search for a direct response path can then be constrained to a length between \(-\tau_{\text{max}}\) and \(\tau_{\text{max}}\) samples. Thus, the sensor responses must be positioned in the filter block consistent with a direct path for the source to the reference sensor. FLC error reduction is an advantage, especially when the reverberation time is high, or when a wide range of the spectrum of the acoustic source is corrupted by ambient reverberation noise (in particular for pseudo-periodic sources).

II. NMCFLMS ALGORITHM

The NMCFLMS algorithm is aimed at minimizing the cost function of (3) efficiently using fast Fourier transform (FFT) and the overlap-save technique [5]. The filter operates on a block-by-block basis and it achieves fast convergence.

Considering block signals with length \( N \) and response filters with length \( L \) (with \( L \leq N/2 \)), updated filter coefficients at block time \( b \) \((b = 0, 1, \ldots \text{and with } m, i, j = 1, 2, \ldots, M)\) can be written as

\[
X_m(b + 1) = \text{IFFT}_N \{ x_m(b + 1) \} \\
H_m(b) = \text{IFFT}_N \{ h_m(b)^T 0_{1 \times (N - L)}^T \}. \tag{5}
\]

Referring to (2), the signal error of the generic microphone pair becomes

\[
E_{ij}(b + 1) = X_i(b + 1) \odot H_j(b) - X_j(b + 1) \odot H_i(b) \tag{6}
\]

where \( \odot \) denotes element-by-element multiplication. In the time-domain, the signal error is

\[
e_{ij}(b + 1) = \text{IFFT}_N \{ E_{ij}(b + 1) \} \tag{7}
\]

where IFFT is the inverse FFT. Using the last \( N - L \) elements of \( e_{ij}(b + 1) \), we generate the frequency-domain block error sequence in the following way

\[
E'_{ij}(b + 1) = \text{IFFT}_N \{ 0_{1 \times L} e_{ij}(b + 1)_{1 \times (N - L)}^T \}. \tag{8}
\]

The power spectrum can then be written with a recursive scheme for more stable implementation

\[
P_m(b + 1) = \lambda P_m(b) + (1 - \lambda) \cdot \sum_{i=1, i \neq m}^M X_i(b + 1)^* \odot X_i(b + 1) \tag{9}
\]

where \( \lambda \) is a forgetting factor and \((\cdot)^*\) denotes the complex conjugate operator. The filter update formula is given by

\[
\Delta H_m(b + 1) = \sum_{i=1}^M X_i(b + 1)^* \odot E'_{im}(b + 1) \odot (P_m(b + 1) + \delta I_{N \times 1}) \tag{10}
\]

where \( \odot \) denotes element-by-element division and \( \delta \) is a small and positive number. Hence, the filter response update in the time-domain is

\[
\Delta h_m(b + 1) = \text{IFFT}_N \{ \Delta H_m(b + 1) \}. \tag{11}
\]

We finally derive the NMCFLMS algorithm by considering the first \( L \) elements of the (11)

\[
h_m(b + 1) = h_m(b) - \Delta h_m(b + 1). \tag{12}
\]

The unit norm constraint is applied to the model filter coefficients in the time-domain to avoid a trivial solution with all-zero elements; the estimated responses at block \( b + 1 \) are

\[
\hat{h}(b + 1) = \frac{h(b + 1)}{\| h(b + 1) \|}. \tag{13}
\]

An improved NMCFLMS method [6] imposes sparse priors on the responses to reduce the temporal whitening, which may interfere with the detection of desired direct paths. Direct impulses and early reflections of the acoustic impulse responses are quite sparse. Therefore, an NMCFLMS method with sparse prior imposition (NMCFLMS-SPI) is proposed to mitigate whitening effects from temporally correlated natural sounds.

III. FILTER LENGTH CONSTRAINTS

To better understand the proposed improvement based on FLC, it is important to highlight the NMCFLMS-SPI’s performance. The NMCFLMS filter coefficients in the time-domain are initialized at 0 except for a middle tap on the reference sensor filter as follows

\[
h_0(0) = [0_{1 \times L/2-1} \ 1 \ 0_{1 \times L/2}]^T. \tag{14}
\]

The first observation for NMCFLMS-SPI is that the initialization of \( h_0(0) \) is maintained during the filter update, and the direct path is forced into the position of index \( L/2 \) in the impulse response. This facilitates the imposition of a length constraint on the filter because the direct path of the second response is within the range \( [\tau_{12,\text{max}} < \tau_{12} < \tau_{12,\text{max}}] \), which holds true for the remaining sensors. Therefore, the imposed filter length for each response is

\[
L_m = 2 \tau_{1m,\text{max}} m = 2, 3, \ldots, M \tag{15}
\]

where \( \tau_{1m,\text{max}} = d_{1m} f_s / v \) is the maximum TDOA in samples between the microphones with a \( d_{1m} \) distance.

In addition, for the experiments in Section IV, we also observed that for the NMCFLMS-FLC algorithm the initialization \( h_0(0) \) is maintained during the filter update; therefore, the proposed approach is generally valid even without SPI, and the initial filter response of the reference microphone can be maintained during the filter update. We must consider that the length of the filters is the same for two sensors; thus, in the (3) the response position is generated in the initial \( L \) elements. However, for multiple sensors the filters have different lengths, which must be adapted such that each filter’s center is aligned with the reference filter for initialization. If the microphones of the array are positioned at increasing distances \((\tau_{1M,\text{max}} > \cdots > \tau_{12,\text{max}})\), we can modify the (5) such that...
the frequency impulse responses for the NMCFLMS-FLC algorithms are

$$H_M(b) = FFT_N \{ [h_M(b)^T \cdot 0_{1 \times (N - L_M)}]^T \}$$

$$H_m(b) = FFT_N \{ [0_{1 \times 1}, h_m(b)^T \cdot 0_{1 \times L_m}]^T \}$$

$$m = 2, 3, \ldots, M - 1$$

(16)

where \(L_1 = (L_M - L_m)/2\) and \(L_2 = (N - L_M + (L_M - L_m))/2\).

The second observation for NMCFLMS-SPI is that it provides improved performances especially during sound onsets (which are notoriously less affected by reverberation), and its effectiveness is limited in following the changes during the filter update, which limits the operation for moving sources.

The Section IV compares experiments using NMCFLMS, NMCFLMS-SPI, NMCFLMS-SPI-FLC, NMCFLMS-FLC, and GCC with the phase transform (PHAT) weighting function to evaluate the performance of the proposed implementation.

**IV. EXPERIMENTAL RESULTS**

The performance of the proposed NMCFLMS-FLC has been evaluated using three types of sound sources: a speaker, a singing voice, and a clarinet. Different reverberant conditions have been used up to a 2 s reverberation time (RT\(_{50}\)). Evaluation in highly reverberant environments may be appealing for different applications, including a speaker for teleconferencing systems or monitoring systems, a singing voice for videogames or musical applications, and musical instruments for audio control interfaces [7]. The image-source method (ISM) was used to simulate reverberant audio data in room acoustics [8]. The ISM assumes that source and microphones are omnidirectional; it provides an approximation of the acoustic energy decay in room impulse responses generated using the image-source technique, and the impulse response and sound sources are convoluted to produce reverberant signals. A room of \((3.5 \times 4.5 \times 3)\) meter was used. The distance between microphones was 20 cm, and the sources were located in five different positions. Fig. 1 illustrates the room setup. Under these conditions, the source at \(P_1\) reached the array with a -21 samples TDOA, at \(P_2\) with -9 samples, at \(P_3\) with 0 samples, at \(P_4\) with 7 samples, and at \(P_5\) with 23 samples. The tests were conducted with a 35 dB signal-to-noise ratio (SNR), which was obtained by adding mutually independent white Gaussian noise to each channel. Therefore, each performance was repeated 100 times for each source, each position and each reverberation to support a robust statistical analysis, because the white Gaussian sequence changes every time. The sampling frequency was 44.1 kHz, and the block size with Hann analysis windows was 2048 samples. The three sound sources recorded in anechoic environments were an alternating female and male speaker signal of a duration of 43 s, a male singing voice\(^1\) signal of 97 s, and a clarinet\(^2\) signal of 101 s. Calculating the mean power for two signal blocks (which detects whether sound is present) facilitates filter initialization and TDE activation-deactivation.

The threshold value used for the power was \(10^{-4}\). The powers at the first block \(P_1(1)\) and \(P_2(1)\) were initialized with a mean power of two signals. The \(\lambda\) was set to 0.5, and the \(\varepsilon\) was set to \(5 \cdot 10^{-2}\) for the speaker sound and \(5 \cdot 10^{-5}\) for pseudo-periodic sounds (singing voice and clarinet). The filter length for

---

\(^1\)http://acustica.ing.unife.it/eng-ver/
\(^2\)http://theremin.music.uiowa.edu/
TABLE I

<table>
<thead>
<tr>
<th>SPEAKER</th>
<th>NMCFLMS</th>
<th>NMCFLMS-SPI</th>
<th>NMCFLMS-SPI-FLC</th>
<th>NMCFLMS-FLC</th>
<th>GCC-PHAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGING VOICE</td>
<td>PSR</td>
<td>RMS</td>
<td>PSR</td>
<td>RMS</td>
<td>PSR</td>
</tr>
<tr>
<td>RT00</td>
<td>0.25</td>
<td>96.82</td>
<td>81.81</td>
<td>87.48</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>92.06</td>
<td>81.81</td>
<td>87.48</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>83.92</td>
<td>1.54</td>
<td>78.66</td>
<td>1.63</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>79.76</td>
<td>7.68</td>
<td>73.69</td>
<td>3.49</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>51.66</td>
<td>60.74</td>
<td>78.41</td>
<td>2.28</td>
</tr>
</tbody>
</table>

NMCFLMS and NMCFLMS-SPI was set to 256 samples in accordance with [6], [9]. The filter length for NMCFLMS-FLC based on the maximum TDOA was set to 52 samples (with a 343 m/s sound speed). The NMCFLMS, NMCFLMS-SPI, NMCFLMS-SPI-FLC, NMCFLMS-FLC, and GCC-PHAT performances were evaluated for the percentage success rate (PSR) and root mean square (RMS) error of all positions. The PSR is defined as

$$PSR = 100 \left( \frac{\text{number of correct TDOA estimations}}{B} \right)$$

(17)

and the RMS error is

$$RMS = \sqrt{\frac{\sum_{i=1}^{5} \sum_{b=1}^{B} (\tau_{i} - \hat{\tau}_{b})^2}{B}}$$

(18)

where $B$ is the total number of analysis blocks, $B_{i}$ is the total number of analysis blocks for the $i$-th position, $\tau_{i}$ is the real TDOA for the $i$-th position and $\hat{\tau}_{b}$ is the estimated TDOA at block $b$. Table I shows the experimental results for the three sound sources used. For each experiment, NMCFLMS-FLC had considerably enhanced performance and produced a PSR increase and RMS decrease compared with NMCFLMS and NMCFLMS-SPI. In particular, the advantage of using FLC is clear for increasing reverberation times. In this condition of SNR, the SPI does not provide significant improvements during sound onsets, in which the noise can be a problem for the correct convergence of the filter. Using FLC with SPI produced no substantial improvements in the PSR. For the speaker, the FLC performed at up to 2 s $\text{RT}_{50}$ with a very low RMS error. However, we observe some inversions of the PSR behavior when increasing the reverberation time due to the nonstationary characteristics of speech signal: the value of FLC at 1.5 s $\text{RT}_{50}$ and the SPI's performance, which heavily depends on the sound onset. For a singing voice and a clarinet, which are pseudo-periodic sounds, the improvement due to the FLC is clearer if compared with NMCFLMS and NMCFLMS-SPI; however, the RMS error for a clarinet is not comparable to vocal performance. Fig. 2 shows the comparison of TDOA estimates at a specific analysis block $b$ for a singing voice.

V. CONCLUSIONS

Using FLC based on the maximum TDOA between microphones in the NMCFLMS algorithm facilitates accurate TDE in highly reverberant environments. As it has been pointed out, through the observation that reference response initialization does not change during the filter update, searching the direct path for the other responses in a filter length twice the maximum TDOA can reduce direct path estimation error. Experimental results show enhanced NMCFLMS-FLC performance if compared with NMCFLMS and NMCFLMS-SPI. In particular, FLC robustness is demonstrated by increasing reverberation times. The proposed improvement is suitable for the following applications, in which the integration of ASL systems may be attractive: teleconferencing systems (speaker), musical interfaces (singing voice and clarinet), videogames (speaker and singing voice), and monitoring systems (speaker).

REFERENCES