A subband based acoustic source localization system for reverberant environments

Tobias Wolff, Markus Buck, Gerhard Schmidt
Harman/Becker Automotive Systems, Acoustic Signal Processing Research
Soeflinger Str. 100, D-89077 Ulm, Germany
E-Mail: {twolff,mbuck,geschmidt}@harmanbecker.com

Abstract

In this paper an efficient acoustic source localization system is described which works within a subband processing framework and is robust against reverberation. The proposed system is related to the adaptive eigenvalue decomposition algorithm proposed by Benesty et al., and uses adaptive subband filters to model the room impulse responses. The subband filters are analyzed in order to estimate the relative time delay between the direct sound paths. The subband processing permits for special extensions in the filter adaptation rule which allow for an efficient extraction of the time delay information and tracking of moving sources. The proposed system has been evaluated in real acoustic environments with different reverberation times, and is compared to the generalized cross correlation method.

1 Introduction

Acoustic source localization is important for many audio based applications like human/machine interfaces or conference systems. Once the position of the speaker is known, a multi-channel speech acquisition system can be steered towards the talker while noises coming from other directions can be suppressed effectively. The speaker localization task is basically achieved by estimating the time delay between a pair of microphone signals. The most prominent technique for time delay estimation is the generalized cross correlation (GCC) method [1]. However, localization systems, which are based on this method work well as long as there is only moderate reverberation [2, 3]. Benesty et al. proposed the adaptive eigenvalue decomposition algorithm for robust source localization in acoustical environments with large reverberation times [4]. This approach incorporates two adaptive filters which permit for an explicit modeling of the room reverberation. In the system described in [4], the adaptive filters are implemented in the frequency domain. A subband domain version however, is not known yet. As a matter of fact, a great number of speech and audio applications is actually implemented in the subband domain. It is therefore desirable to develop source localization algorithms that work reliably in typical subband processing frameworks. In the following we present a subband domain system for adaptive source localization, that can locate and track a moving speaker in reverberant environments.

The method proposed here is inspired by the adaptive eigenvalue decomposition algorithm [4], and exploits the same underlying principle. Therefore this localization principle is outlined at first. Subsequently the proposed system is described. In Section 4 some results from a comparative study between the generalized cross correlation and the proposed system are presented. Finally some concluding remarks are given in the last section.

Figure 1: Basic structure of the localization system using adaptive filters.

2 Localization principle

Let \( x_m(t), m \in \{1,2\} \), be the real value of the \( m \)-th microphone signal \( x_m(t) \) at time \( t = t \cdot T_s, t \in \mathbb{Z}, T_s \in \mathbb{R} \) being the sampling interval. If the room is modeled as a linear time-invariant (LTI) system, the microphone signals can be written as

\[
x_m(l) = \sum_{n=-\infty}^{\infty} h_m(n)s(l-n) + b_m(l),
\]

where \( s(l) \) and \( b_m(l) \) are the source signal and additive noise components respectively. The discrete-time representations of the room impulse responses, with time lag \( n \), are denoted as \( h_m(n) \). If the room impulse responses were known, the source could easily be localized by comparison of the temporal difference between their maxima:

\[
\hat{\tau} = \arg\max_n \left\{ h_1(n) - h_2(n) \right\},
\]

because these maxima are associated with sound wave propagating on the direct path between the source and the center of the microphone array. The difference \( \hat{\tau} \) hardly depends on the reverberation, since reverberation effects are represented by the rest of the impulse responses. Eq. (2) is an estimate of the normalized time shift \( \tau \), which relates to the direction of arrival (DOA) of the direct sound, hence the azimuth angle \( \theta \), according to

\[
\tau = f_s \cdot \frac{d_{\text{mic}}}{c} \cos(\theta) = \tau_{\text{max}} \cdot \cos(\theta).
\]

Here, \( d_{\text{mic}} \) is the microphone distance, \( c \) is the speed of sound, and \( f_s \) denotes the sampling frequency.

Thus, the goal is to blindly estimate the impulse responses. This can be accomplished using two adaptive FIR filters \( \hat{h}_m(n) \). The basic processing structure is depicted in Figure 1. In the noise free case \( (b_m(l) = 0) \) it becomes obvious that the output \( e(l) \) will be zero if

\[
\hat{h}_1(n) = h_2(n) \quad \text{and} \quad \hat{h}_2(n) = h_1(n),
\]
because the convolution is commutative for LTI-systems. Hence, if the power of the difference $e(l)$ between both filter outputs is minimized in the presence of noise, subject to the constraint $h_m(n) \neq 0$, the impulse responses $h_m(n)$ can be considered as estimates for the actual room impulse responses. Precisely speaking, this is only true up to a convolutive term, and if it is assumed that the room impulse responses $h_m(n)$ do not share any common zeros. Minimization of the power of the error signal $e(l)$, and thus estimation of the room impulse responses, can be achieved using a constrained least-mean-squares (LMS) algorithm [4].

The problem may also be formulated as finding the eigenvector that corresponds to the smallest eigenvalue of the covariance matrix of the two microphone signals. For this reason the method has been called "adaptive eigenvalue decomposition algorithm" [4].

3 Proposed subband system

The main reasons why many signal processing tasks are performed in the subband domain, rather than in the time domain, are reduction of the computational complexity and faster adaptation of adaptive filters. These features, however, come at the expense of an increased memory load and an additional signal delay which is inevitable. Whenever this can be tolerated a subband processing framework is an attractive way to implement practical systems.

An efficient uniform subband decomposition can be achieved using a DFT-modulated filterbank [5]. For each microphone signal $x_m(l)$, the corresponding complex valued subband signals are obtained as:

$$X_m(e^{j\Omega_m}l, k) = \sum_{p=0}^{N_p-1} \sum_{n=0}^{N-L-1} w_n + p N x_m(k D - p N - n) e^{-j \Omega_m n}.$$  

(5)

Here $\Omega_m$ denotes the normalized center frequency of the $\mu$-th subband, $\mu \in \{0, \ldots, N-1\}$, $N$ is the DFT-length, and $k$ denotes the block index obtained through $D$-fold downsampling, $D \in \mathbb{N}^+$. The weights of the length $N_p \cdot N$ prototype analysis filter are denoted by $w_n$. A very popular way to do the subband decomposition is to use $N_p = 1$. This special case is usually referred to as Short-Time-Fourier Transform (STFT) and offers a good trade-off between temporal resolution and complexity reduction. The STFT will be considered throughout this paper. The proposed method however is not bound to this particular filterbank.

In brief, we propose to use two subband FIR filters, which are updated according to a modified constrained LMS algorithm. Due to this particular adaptation rule, only a subset of the filter coefficients needs to be analyzed to estimate the direction of arrival. This analysis is also done in a non-standard way. Please see Figure 2 for an overview of the proposed system. Details on each part are given next.

3.1 Subband filtering

As all operations described in the following are performed in each subband, the subband index will be omitted as long as it is not needed. According to the processing structure proposed by Benesty et al. we apply a length-$L$ FIR filter

$$\tilde{H}_m(k) = [\tilde{H}_{m,0}(k), \ldots, \tilde{H}_{m,L-1}(k)]^T.$$  

(6)

$\tilde{H}_1(e^{j\Omega_m}l)$

Coefficient Selection

$\tilde{H}_{i_0}(e^{j\Omega_{i_0}}l)$

Localization

$X_1(e^{j\Omega_m}l)$

Analysis Filterbank

$X_2(e^{j\Omega_m}l)$

Adaptation, Normalization, Phase-Shifting

$x_1(l)$

Figure 2: Processing structure of the proposed system.

where

$$\tilde{H}_1(e^{j\Omega_m}l) = \begin{bmatrix} \tilde{H}_{i_0}(e^{j\Omega_{i_0}}l) \\ \vdots \\ \tilde{H}_{i_0}(e^{j\Omega_{i_0}}l) \end{bmatrix}.$$  

(7)

in each subband and each channel respectively. The vector of input signals $X_m(k)$ is denoted as

$$X_m(k) = [X_m(k), \ldots, X_m(k-L+1)]^T.$$  

(8)

Both, the input signals and the filter coefficients are arranged in stacked vectors

$$\tilde{H}(k) = \begin{bmatrix} H_1^T(k) \\ -H_2^T(k) \end{bmatrix}^T$$  

(9)

and

$$X(k) = \begin{bmatrix} X_1^T(k) \\ X_2^T(k) \end{bmatrix}^T.$$  

(10)

The complex valued error signal is then given as

$$E(k) = \tilde{H}^H(k) X(k).$$  

In the following the subband filter adaptation rule is described.

3.2 Filter adaptation

The filters are updated in a three-step process. The first stage changes the filters according to the normalized least-mean-square (NLMS) adaptation rule [6] given as

$$H(k) = H(k-1) - \beta \frac{X(k-1) E^*(k-1)}{X^H(k-1) X(k-1)},$$  

(11)

where $\beta$ denotes the stepsize. The second step implements the contraint. To this end the result from the first update step is normalized to avoid the trivial solution:

$$H(k) = H(k) \left( H^H(k) \tilde{H}(k) \right)^{-\frac{1}{2}}.$$  

(12)

Up to this point the method closely corresponds to the original time domain version described in [4]. The third update step however adds a useful novelty. We introduce another modification that shifts all phase values by the phase of the $i_0$-th coefficient of the second filter:

$$\tilde{H}(k) = \tilde{H}(k) \begin{bmatrix} \tilde{H}^*_2(k) \\ \tilde{H}^*_2(k) \end{bmatrix} C.$$  

(13)
At block $k = 0$ the filter coefficients are initialized as

$$
\hat{H}_{m,i}(0) = \begin{cases} 
1 & \text{for } m = 2, \text{ and } i = i_0, \\
0 & \text{else,}
\end{cases}
$$

which means that Eq. (12) keeps $\hat{H}_{2,i_0}(k)$ dominant for all times $k$ and Eq. (13) enforces that its phase is kept at zero.\footnote{The step size is assumed to be sufficiently small.} As a consequence the relative phase information is moved to the first filter, whereas $\hat{H}_{2,i_0}(k)$ acts as a reference with regard to the phase. Additionally, Eq. (13) introduces a much smaller than the length of the synthesis window. In

In the following the subband notation is re-introduced and the contraction vector

$$
C = \begin{bmatrix} \epsilon \cdot 1_{1 \times L}, 1_{1 \times L} \end{bmatrix}^T,
$$

where $1_{1 \times L}$ denotes a length-$L$ identity vector and $\epsilon$ is a positive factor smaller than 1. The contraction factor $\epsilon$ fades out the coefficients of the first filter in case of missing excitation. Since this mechanism makes the filter forget its convergence state, tracking of moving sources is enabled.

### 3.3 Extracting the DOA Information

In the following the subband notation is re-introduced and the frame index $k$ is omitted. Both sets of subband filters $H_m(e^{j\Omega_m})$ correspond to a time domain impulse response that is obtained by applying the overlap-add method to the subband filters:

$$
\tilde{h}_m(n) = \frac{1}{N} \sum_{i=0}^{L-1} \tilde{w}_{n-iD} \sum_{\mu=0}^{N-1} \hat{H}_{m,\mu}(e^{j\Omega_m}) e^{j\Omega_m(n-iD)},
$$

where $\tilde{w}_n$ denotes the synthesis window that is zero for $n < -N/2$ and $n > N/2 - 1$. Thus, $\tilde{h}_m(n)$ shows $N/2$ non-causal values, which reflects a property of subband filters. It should be noted however, that

$$
\arg\max_n \{ \tilde{h}_2(n) \} = i_0 D
$$

holds at any time, which is enforced by the adaptation rules given in Eq. (12) and Eq. (13). As a consequence, only $\tilde{h}_1(n)$ has to be analyzed. The direct sound peak in $\tilde{h}_1(n)$ must be found in the $\pm \tau_{\text{max}}$ vicinity of $n = i_0 D$. Therefore all values outside the interval

$$
i_0 D - \tau_{\text{max}} \leq n \leq i_0 D + \tau_{\text{max}}
$$

can be discarded. Within this interval, $\tilde{w}_n$ can be approximated to be constant, because for a practical setup $\tau_{\text{max}}$ is much smaller than the length of the synthesis window. In particular, $\tilde{w}_n$ can be set to $\tilde{w}_n \approx 1$, by scaling the analysis window accordingly. By neglecting the components that overlap with the $i_0$-th part, Eq. (16) is reduced to

$$
\tilde{h}_1(n) \approx \frac{1}{N} \sum_{\mu=0}^{N-1} \hat{H}_{1,\mu}(e^{j\Omega_{\mu}}) e^{j\Omega_{\mu}(n-i_0 D)},
$$

which is the inverse DFT of $\hat{H}_{1,\mu}(e^{j\Omega_{\mu}}) e^{-j\Omega_{\mu}i_0 D}$. For our purpose it is sufficient to transform only $\hat{H}_{1,\mu}(e^{j\Omega_{\mu}})$:

$$
\tilde{h}_1(n + i_0 D) \approx \frac{1}{N} \sum_{\mu=0}^{N-1} \hat{H}_{1,\mu}(e^{j\Omega_{\mu}}) e^{j\Omega_{\mu}n},
$$

which shifts the maximum into $-\tau_{\text{max}} \leq n \leq \tau_{\text{max}}$. If Eq. (20) is employed before the maximum search, the obtained resolution depends on the sampling frequency and the microphone distance and is much too coarse for most applications. A reasonable interpolation is obtained if the filter coefficients $\hat{H}_{1,\mu}(e^{j\Omega_{\mu}})$ are considered as the coefficients of the Fourier series of the corresponding real-valued continuous time signal $g(t)$, which is the $N/2$-periodic interpolation of $\hat{H}_1(n + i_0 D)$:

$$
g(t) = \hat{H}_{1,i_0}(e^{j\Omega_{\mu}}) + 2 \sum_{\mu=1}^{N/2} \text{Re} \left\{ \hat{H}_{1,i_0}(e^{j\Omega_{\mu}}) e^{j\Omega_{\mu}ft} \right\}.
$$

Substituting Eq. (3) into Eq. (21), and discarding the offset $\hat{H}_{1,i_0}(e^{j\Omega_{\mu}})$, a DOA measure which directly depends on the azimuth angle $\theta$ is obtained:

$$
\phi(\theta) = \text{Re} \left\{ \sum_{\mu=1}^{N/2} \hat{H}_{1,i_0}(e^{j\Omega_{\mu}}) e^{j\Omega_{\mu} \tau_{\text{max}} \cos(\theta)} \right\}.
$$

The final DOA-estimate is obtained by choosing the angle that maximizes Eq. (22):

$$
\hat{\theta} = \arg\max_{\theta} \{ \phi(\theta) \}.
$$

### 3.4 Discussion

The localization measure in Eq. (23) evaluates only those subband filter coefficients that contain the information of interest. Due to the phase-shifting (Eq. (13)) this information is contained only in a subset of the first subband filter. Therefore only $N/2$ coefficients have to be analyzed.

Because of the contraction mechanism, the magnitude of the filter coefficients of the first filter is decreased if there is no excitation. Therefore their contribution to the localization measure is reduced from frame to frame. Without introducing this mechanism, successful tracking would not be possible, because the peak of the angular spectrum would stick at the old position and re-adaptation would take too long. In [3] the adaptive filters have to be reset periodically to enable tracking of moving sources. In principle the contraction mechanism proposed here is comparable to this proceeding but here the filters do not loose their complete information at once.

The angular spectrum in Eq. (22) can be evaluated for any angle $\theta$. The spatial resolution can therefore be chosen as desired for the respective application and must not even be fixed to some predefined grid. For each tested angle, only a scalar product of length $N/2$ has to be processed, which keeps the computational complexity at a reasonable level.

The original adaptive eigenvalue decomposition algorithm described in [4] uses a frequency domain implementation of the adaptive filters. Since the normalization is applied in the time domain, seven FFTs are required in each frame. In [3] it is argued that two of these FFTs can be saved by dropping the norm-constraint. The subband method proposed here requires only two FFTs for the filterbank. Since source localization algorithms are usually used in combination with other signal processing tasks like beamforming for instance, those FFTs will be processed anyway in practical systems. Another advantage of the described method is the scalability of the length of the adaptive filters without the need of changing the surrounding signal processing framework.
power. A contraction factor of

We fix the second coefficient, rather than the first, because

GCC-PHAT and the proposed subband algorithm are com­

ronments with reverberation times

reasonable trade-off between tracking speed and good con­

ters. In order avoid miss-detections in speech pauses, the

Figure 3: Comparison of the proposed subband algorithm

and the GCC-PHAT. The dashed line indicates the actual

speaker location at \( \theta = 60 \) degree. The histograms show

that the proposed method leads to much fewer false esti­mations than the GCC-PHAT method.

4 Results

The system described above has been implemented using a

DFT-length of \( N = 256 \), a subsampling factor of \( D = 64 \),

and a sampling frequency of \( f_s = 11025 \) Hz. The analysis

window is a Hann window and the microphone distance

has been set to \( d_{\text{mic}} = 29.4 \) cm.

The length of the adaptive filters is \( L = 5 \), and the sec­cond filter coefficient \( (i_0 = 1) \) of the second filter is fixed.

We fix the second coefficient, rather than the first, because

thereby we allow for non-causal components to be repre­sented by the first filter tap which leads to reduced error

dower. A contraction factor of \( \epsilon = 0.9 \) appeared to be a

reasonable trade-off between tracking speed and good con­vergence. The angular spectrum \( g_\theta(\theta, k) \) is evaluated at 181

equidistantly spaced points.

The performance of the proposed system is compared

the that of the GCC-PHAT algorithm, which was imple­mented in the same framework. The GCC-PHAT function was

computed from the PHAT-weighted cross power spec­tral density (CPSD) using Eq. (22) as well. The CPSD is esti­mated using recursive smoothing with a smoothing constant

of \( \alpha = 0.7 \).

Both systems are evaluated for three acoustical envi­ronments with reverberation times \( T_{60} \approx 60 \) ms, \( T_{60} \approx 500 \) ms

and \( T_{60} \approx 1800 \) ms respectively. We used impulse

responses that were actually measured in these rooms, whereas the distance to the microphone array was 3 meters. In order avoid miss-detections in speech pauses, the same SNR-dependent recursive smoothing was applied as post-processing to the results both algorithms under test. In Fig3 an extraction of the analysis can be seen. The GCC-PHAT and the proposed subband algorithm are com­pared in terms of histograms of the respective results. It can be seen that the proposed algorithm produces only very few false estimations though there is strong reverberation. The performance of the GCC-PHAT, on the contrary, is de­graded as the reverberation increases.

5 Concluding remarks

In this paper a new method for acoustic source localization

has been proposed that is robust against reverberation. The

method is related to the adaptive eigenvalue decomposition

algorithm, but works in a subband processing framework. Experiments indicate that the proposed method is more ro­bust against reverberation than the traditional GCC-PHAT

method.

References


