Robust Speaker Localization through Adaptive Weighted Pair TDOA (AWEPAT) Estimation

Nilesh Madhu, Rainer Martin

Institute of Communication Acoustics
Ruhr-Universität Bochum
44480, Bochum
Germany
{firstname}.{lastname}@rub.de

Abstract
Time delay of arrival (TDOA) estimation between signals input to two or more microphones plays an important role in speaker localization. Most methods employ a linear array of two or more microphones and use the generalized cross correlation method or eigenspace analysis (AEA) methods. TDOA estimation with linear arrays, however, is highly sensitive to estimation errors when the signals arrive from an endfire direction. In this paper we propose a novel adaptive algorithm which makes use of a three-microphone planar array. This algorithm exhibits a much smaller estimation error over the complete azimuth range of 0–360 degrees as compared to other algorithms. The computational complexity of this approach is comparable to other state-of-the-art algorithms.

1. Introduction
Speaker localization in adverse environments finds application in many areas like source separation, telephone and video conferencing systems etc. Most methods of speaker localization involve – in some way or the other – estimation of the time delay of arrival (TDOA) between the signals input to a pair (or a linear array) of microphones. Most methods can be considered as a special case of the GCC [1] (generalized cross correlation) method with different weighting functions.

Other methods estimate TDOA by approximating the transfer functions from the source to the microphones [3], [4], [5]. In this case, the TDOA is given by the time difference of the peaks in the impulse responses corresponding to the direct path to the microphones. This scheme is shown in Figure 1. With respect to this figure, s(n) is the source signal, x1(n) and x2(n) are the input signals to the two microphones, h1(n) represents the transfer function from the source to the ith microphone and ẑn is the estimate of the corresponding transfer function. As can be seen, this requires the estimation of N filters for N microphones.

The two microphone system can locate a speaker reasonably well if the azimuth angle θ lies between 40° and 140°. Below and above these values, the localization is subject to a large variance due to the high sensitivity to the estimation error of the TDOA in these regions. To see this more clearly consider the following: for an inter-microphone distance d of 4 cm and an azimuth of 0°, the TDOA is \( \tau = \frac{d \cos(\theta)}{c} \approx 1.176 \times 10^{-4} \sec \) seconds or 3.7 samples at a sampling rate of 32kHz. Due to the noise in the environment, the estimation of the TDOA will not be perfect. Assume that the estimated TDOA is in error by –10%. This would mean an estimated angle of \( \theta = 27.8^\circ \). If now, the same percentage of error were to occur at an azimuth of 60°, we would have an estimated azimuth of 63°, an error of only 3°.

Further, linear arrays implicitly alias the θ and –θ directions. This makes linear arrays impractical in systems where the complete azimuth range of 0–360° is required. An example of such an application would be a tele/video conferencing system where people sit around a table.

To combat localization difficulties along the endfire directions and to use the complete azimuth range, a system using a planar array, with three microphones, and adaptive weighting over two pairs in the array is proposed. Results of experiments conducted in an anechoic chamber and a normal (reverberant) room corroborate the efficacy and applicability of the system.

2. Two microphone TDOA estimation
First consider the simple 2 microphone system as shown in Figure 2. The two input signals x1(n) and x2(n) are related as in (1) (where * denotes the convolution operator).

\[
x_1(n) = h_1(n) * s(n) \tag{1}
\]
\[
x_2(n) = h_2(n) * s(n) \tag{2}
\]

where \( h_1(n) \) represents the transfer function between the source and the ith microphone. In Figure 2, T is a delay to take positive and negative relative shifts into account. Since the transfer functions may be approximated by FIR filters, the same equation can be expressed as:

\[
x_1(n) = h_1^T s(n) \tag{1}
\]
\[
x_2(n) = h_2^T s(n) \tag{2}
\]
where
\[
\begin{align*}
\mathbf{h}_1^T &= (h_1(0), h_1(1), \ldots, h_1(L)) \\
\mathbf{h}_2^T &= (h_2(0), h_2(1), \ldots, h_2(L)) \\
\mathbf{s}(n) &= (s(n), s(n-1), \ldots, s(n-L))^T.
\end{align*}
\]

Consider now, the equalization of \( x_2(n) \) with respect to \( x_1(n) \). The simplest way is to estimate an equalization filter \( \mathbf{w}_{21} \) such that it minimizes the mean square error between the two signals:
\[
\mathbf{w}_{21}^{(opt)} = \arg\min_{\mathbf{w}_{21}} J
\]
\[
J = E\left\{ \left( x_1(n-T) - \mathbf{w}_{21}^T x_2(n) \right)^2 \right\}
\]
where \( \mathbf{w}_{21}^T = (w_{21}(0), w_{21}(1), \ldots, w_{21}(K)) \), with \( K \leq L \) being the filter order and \( x_2(n) = (x_2(n), x_2(n-1), \ldots, x_2(n-K)) \).

In the light of estimating the TDOA, transforming (3) into the (discrete) frequency domain yields, for an optimal \( W(\omega_k) \), the Wiener solution:
\[
W_{opt}(\omega_k) = E\left\{ \frac{X_1(\omega_k) X_2^*(\omega_k)}{|X_2(\omega_k)|^2} \right\}
\]
An inverse discrete Fourier transform of (4) then yields:
\[
\mathbf{w}_{opt}(\ell) = \text{IDFT}\left\{ \frac{E\left\{ X_1(\omega_k) X_2^*(\omega_k) \right\}}{E\left\{|X_2(\omega_k)|^2\right\}} \right\}, \quad \ell = 0, \ldots, K
\]
which can be seen as the GCC equation with a Wiener weighting filter (or the Roth processor) [1, 2]. Therefore, it suffices to estimate the position of the peak of this filter with respect to the delay \( T \) to obtain the TDOA information.

3. Adaptive Weighted pair TDOA (AWePaT) estimation

As already explained, the two microphone system of the previous section is inefficient in its coverage of the azimuth range. Therefore an adaptive weighted pair TDOA (AWePaT) method is proposed where use is made of a planar array with 3 microphones, each located at the apex of an equilateral triangle of base 4 cm. The block diagram is as shown in Figure 3. The source is assumed to be in the far field, justifying the assumption of plane wavefronts. This assumption only serves to simplify the calculations for the angle of arrival. It does not have any bearing on the algorithm itself. In the system, \( T \) represents,

![Figure 3: TDOA estimation by transfer function approximation](image)

as before, a delay in order to account for negative shifts of the peak. Filters \( \mathbf{w}_{13} \) and \( \mathbf{w}_{23} \) are the equalization filters and the \( \alpha_i \) represent the measure of ‘trust’ placed upon the individual TDOA measurements. Channels 1 and 2 are equalized with respect to the reference channel 3.

The equalization is performed using adaptive block NLMS-type [6] algorithms and, for a time domain implementation, is summarized as:
\[
e(nL-l) = [\alpha_1(n) \alpha_2(n)] \left[ \begin{array}{c} \mathbf{w}_{13}^T(n) x_1(nL-l) \\ \mathbf{w}_{23}^T(n) x_2(nL-l) \end{array} \right]
\]
\[
-x_3(nL-l-T)
\]
\[
\mathbf{w}_{23}(n+1) = \mathbf{w}_{23}(n) - \mu (\sum_{i=1}^{L} e(nL-l) x_i(nL-l) )
\]
where \( n \) is the frame number, \( L \) is the block length and \( \alpha_1(n) \) and \( \alpha_2(n) \) are the ‘trust’ weights for the different pairs, indicating the reliability of the respective pair, \( \mu \) is the step size and \( x_i(nL) = (x_1(n), x_2(n-1), \ldots, x_l(n-K)) \). The TDOA is estimated for each pair (1, 3), (2, 3) where, as before, it suffices to estimate the position of the peaks of the filters with respect to the delay \( T \) to obtain the TDOA information. The adaptive weights \( \alpha_i \) reduce the sensitivity to estimation errors in the adaptation loop of the equalization filters.

Using the estimated TDOA, \( \tau_i(n) \), the individual azimuths can be calculated as:
\[
\hat{\theta}_i(n) = \cos^{-1} \left( \frac{c \tau_i(n)}{d} \right).
\]

Further, using the knowledge of the array geometry, resolution of ambiguities is easily achieved. At the end of each frame the estimated azimuth for that frame is calculated as a weighted sum of the individual estimates as
\[
\hat{\theta}_{avg}(n) = \alpha \tau(n) \hat{\theta}(n)
\]
where \( \alpha(n) = [\alpha_1(n) \alpha_2(n)]^T \) is the weighting vector and \( \hat{\theta}(n) = [\hat{\theta}_1(n) \hat{\theta}_2(n)]^T \) is the vector of azimuth estimates.

Additionally, a first order smoothing is incorporated on the estimated azimuth as
\[
\hat{\theta}_{avg}(n) \leftarrow \gamma \hat{\theta}_{avg}(n-1) + (1-\gamma) \hat{\theta}_{avg}(n) \quad (0 \leq \gamma < 1)
\]

Following this, the trust values (weights) are updated for the next frame.

4. Evaluation of the trust weights

The aim of using the trust weights is to minimize the average error in the estimated azimuth. Thus, minimizing the mean square error is a criterion that can be used to calculate the optimal weights \( \alpha_i \) (where the time index has been dropped for convenience):
\[
J = E\left\{ (\hat{\theta}_{avg} - \theta)^2 \right\}
\]
The optimization on \( \alpha \) is constrained as \( \alpha^T 1 = 1 \), where \( 1 \) represents a column vector of ones. Thus we have, from (7) or (8) and (9):
\[
J = E\left\{ \alpha^T \left[ \begin{array}{c} \hat{\theta}_1 - \theta \\ \hat{\theta}_2 - \theta \end{array} \right]^2 \right\}
\]
To solve this optimization problem, we shall make the following assumptions:
- the azimuth estimates are asymptotically unbiased (i.e., \( E\{\hat{\theta}\} = \theta \))
- the errors in the azimuth estimates are uncorrelated (i.e., \( E\{ (\hat{\theta}_i - \theta)(\hat{\theta}_k - \theta) \} = 0 \quad (i \neq k) \)
Under these assumptions, the cost function for optimization simplifies to

$$\mathcal{J} = \arg\min_{\alpha} \alpha^T \text{diag} \left( \hat{\theta}_1^2, \hat{\theta}_2^2 \right) \alpha + \lambda (\alpha^T 1 - 1)$$

(11)

where $\hat{\theta}_k^2 = \text{var} \left( \hat{\theta}_k - \theta \right)$ represents the respective azimuth estimation error variance and $\lambda$ is the Lagrangian multiplier. Differentiating (11) with respect to $\alpha$ and equating to zero yields:

$$\lambda = \frac{2}{1^2 \hat{\theta}_1^{-1} \hat{\theta}_2^{-1}}$$

(12)

with $\mathbf{R}_\theta = \text{diag} \left( \hat{\theta}_1^2, \hat{\theta}_2^2 \right)$. Combining the above with the constraint gives us:

$$\alpha = \frac{\mathbf{R}_\theta^{-1}}{1^2 \hat{\theta}_1^{-1} \hat{\theta}_2^{-1}} = \frac{1}{\hat{\theta}_1^{\frac{1}{2}} + \hat{\theta}_2^{\frac{1}{2}}} \left[ \begin{array}{c} \hat{\theta}_2^{\frac{1}{2}} \\ \hat{\theta}_1^{\frac{1}{2}} \end{array} \right]$$

(13)

As can be seen from (13), the weighting depends upon the variance of the error in the estimate. This variance is, in general, unavailable. Therefore, we propose the use of the Cramér-Rao bound on the error variance in the above equation.

The CRLB for the azimuth angle estimate can be calculated as \cite{17, 8}

$$\hat{\theta}_k^2 = \left( 2 \sum_r \omega_r^2 \sigma_{\omega_r}^2 \right)^{-1} \left( \frac{d \sin(\theta_k) / c}{\sigma_{\omega_r}^2} \right)^2$$

(14)

In the above, the ratio $\sigma_{\omega_r}^2 / \sigma_{\omega_k}^2$ represents the signal to noise ratio for the discrete frequency $\omega_k$. If we assume that all sensors have the same SNR, an assumption that is not so unlikely, the optimum trust weights simplify to:

$$\alpha = \frac{1}{\sin^2(\theta_1) + \sin^2(\theta_2)} \left[ \begin{array}{c} \sin^2(\theta_1) \\ \sin^2(\theta_2) \end{array} \right]$$

(15)

To calculate the trust weights for the $(n + 1)^{th}$ frame, the azimuth estimate $\hat{\theta}_{avg}(n)$ of the previous frame is used. We define the predicted azimuth

$$\tilde{\theta}(n + 1) = \left[ \begin{array}{c} \tilde{\theta}_1(n + 1) \\ \tilde{\theta}_2(n + 1) \end{array} \right] = \mathcal{F} \left( \hat{\theta}_{avg}(n) \right)$$

(16)

where $\mathcal{F} \()$ is a simple (vector) function of the geometry of the array. For example, in the case of the equilateral placement indicated in Figure 4, $\mathcal{F} \()$ is computed as:

$$\mathcal{F} \left( \hat{\theta}_{avg}(n) \right) = \left[ \begin{array}{c} \hat{\theta}_{avg}(n) \\ \hat{\theta}_{avg}(n) - \frac{\pi}{3} \end{array} \right]$$

The $\tilde{\theta}_1$ values from (16) are used in (15) and the trust weights are updated as

$$\alpha(n + 1) = \frac{1}{\sin^2(\tilde{\theta}_1(n + 1)) + \sin^2(\tilde{\theta}_2(n + 1))} \left[ \begin{array}{c} \sin^2(\tilde{\theta}_1(n + 1)) \\ \sin^2(\tilde{\theta}_2(n + 1)) \end{array} \right].$$

(17)

Note that this method is easily extensible to more than two microphone pairs or to more than two linear arrays.

5. Experimental results

To evaluate the AWePaT method, experiments with varying degrees of complexity were conducted. The results were compared to a two microphone case, using microphones 1 and 3, for which the AEDA algorithm \cite{3} was used and to a three microphone case where the steered-response power (SRP-PHAT) algorithm \cite{10} was used. The azimuth was measured with respect to the array arrangement as in Figure 4. Measurements were made for azimuth angle values in steps of $10^\circ$ from 0 to 180°. Because of the unresolved angle ambiguity of the two-microphone setup, we show results only for the azimuth range of 0 to 180°. The three-microphone setup performs equally well on the full range.

![Figure 4: Array configuration](image)

5.1. Evaluation in an anechoic chamber

The setup consisted of speech signals drawn from the TIMIT database, upsampled to 32kHz and then broadcast via a Genelec speaker. The filters $w_{1,3}$ and $w_{2,3}$ were of 96 taps in length. The block length was also set to 96 (for the proposed method). The average root mean square error (RMSE) vs. azimuth is plotted in Figure 5. It can be seen that whereas the two microphone approach has a widely varying profile, the AWePaT algorithm tends to have a relatively flat RMSE profile over the azimuth range. Since an anechoic chamber is a rather ideal situation, a comparison with the SRP-PHAT was not performed for this case.

5.2. Evaluation in a reverberant room

To evaluate the robustness of the approach, it was also subjected to tests in a reverberant room. The speakers were positioned around a table at an approximate distance of 1m from the center of mass of the array. The SNR was around 10 dB and the reverberation time was about 600 ms. With reference to Figure 6 it can be seen again that while the AWePaT algorithm presents a (mostly) flat RMSE profile, the high sensitivity to estimation errors in the endfire directions of the two-microphone approach is clearly visible. Additionally, in order to compare the algorithms used for TDOA estimation, the figure also shows the performance of our algorithm (see Section 2) using 2 microphones for localization. It may be seen that the performance of
this algorithm closely matches the performance of the AEDA approach. We further tested our algorithm against the comparable SRP-PHAT 3 microphone system, using the same two microphone pairs as for the AWePaT algorithm. As can be seen, the AWEPAT approach again has the better performance.

6. Interplay between the weights

Intuitively, the logic behind selecting the weights should be such that an array which perceives the source close to its endfire direction should have a low weight, forcing a larger dependency upon the other array. Similarly, for azimuthal directions that generate the same (absolute) TDOA at both the arrays, the weights should be uniformly distributed. It is worthwhile to examine if the trust weights selected do indeed conform to our intuition. Accordingly, Figures 7(a) and 7(b) show the time evolution of the trust values for two specific azimuth angles (for an experiment in a reverberant room). It may be seen that after a short period of adaptation the trust weights converge to a steady state.

Indeed, when one array perceives a source in its endfire direction, its weight drops and the other array is trusted to a higher extent. For the second case, both arrays estimate the same (absolute) TDOA leading to a uniform weight distribution between the arrays. Thus, the weighting function does conform to our intuitive expectations and provides a graceful trust degradation for endfire directions and scalable trust weights for other azimuth values.

7. Summary and conclusions

This paper demonstrated a method to localize speakers via estimation of the time delay of arrival. In contrast to other approaches for TDOA estimation, this method focused upon the usage of a planar array with weighted pair estimates that flatten the error profile and make the localization robust for the complete azimuth range. A comparison with a two microphone algorithm with similar computational requirements indicates that the AWePaT method provides a more robust and consistent result at (approximately) the same computational cost. Further, a comparison was also made with the SRP-PHAT approach using the three microphone setup and, as demonstrated, results favor the AWePaT algorithm.

As a final note, the method is not confined to a 3-microphone setup but can be extended to any number of microphones. The derivation for the trust weights follows the same pattern and its extension is trivial.

8. References


