Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 287167, 14 pages
doi:10.1155/2008/287167

Research Article

Localization of Directional Sound Sources Supported by A Priori Information of the Acoustic Environment

Zoltán Fodróczi[1] and András Radványi[2]

[1] *Faculty of Information Technology, Pázmány Péter Catholic University, Práter u. 50/A, 1058 Budapest, Hungary*
[2] *Analogic and Neural Computing Laboratory, Computer and Automation Research Institute, Hungarian Academy of Sciences, Lagymanyosi u. 11, 1111 Budapest, Hungary*

Correspondence should be addressed to Zoltán Fodróczi, fodroczi@digitus.itk.ppke.hu

Received 6 November 2006; Revised 6 March 2007; Accepted 11 July 2007

Recommended by Douglas B. Williams

Speaker localization with microphone arrays has received significant attention in the past decade as a means for automated speaker tracking of individuals in a closed space for videoconferencing systems, directed speech capture systems, and surveillance systems. Traditional techniques are based on estimating the relative time difference of arrivals (TDOA) between different channels, by utilizing crosscorrelation function. As we show in the context of speaker localization, these estimates yield poor results, due to the joint effect of reverberation and the directivity of sound sources. In this paper, we present a novel method that utilizes a priori acoustic information of the monitored region, which makes it possible to localize directional sound sources by taking the effect of reverberation into account. The proposed method shows significant improvement of performance compared with traditional methods in "noise-free" condition. Further work is required to extend its capabilities to noisy environments.

1. INTRODUCTION

The inverse problem of localizing a source by using signal measurements at an array of sensors is a classical problem in signal processing, with applications in sonar, radar, and acoustic engineering. In this paper, we focus on a subset of these efforts, where the speaker is to be localized in a conference environment. Brandstein's book [1] provides a comprehensive introduction to the state-of-the-art methods in this field. Generally, three classes of source localization algorithms are taken into account: (i) high-resolution spectral estimation [2, 3], (ii) steered beamformer energy response [4, 5], and (iii) estimation of time difference of arrivals (TDOA) [6–10]. Some algorithms combine features from more than one class such as the accumulated correlation method [11] which has shown [12] how to combine the accuracy of beamforming and the computational efficiency of TDOA-based techniques [6–10].

In 1976, Knapp and Carter [13] proposed the generalized cross-correlation (GCC) method that was the most popular technique for TDOA estimation. Since then, many new ideas have been proposed to deal more effectively with noise and reverberation by taking advantage of the nature of a speech signal [14, 15] or by utilizing redundant information from multiple sensor pairs [11, 16–18]. Another interesting approach is to utilize the impulse response functions from the source to the microphones. There exist two branches which follow this strategy. The first one is the high-resolution spectral estimation technique [2, 3] where the transfer functions are estimated blindly by an adaptive algorithm intended to find the eigenvalues of the cross-correlation matrix. The more accurate this estimate is, the better the relative delay between the two microphone signals can be estimated. Unfortunately, in practical applications, this estimate is still not usable because of its high sensitivity to noise. The second method is termed the "matched filter array-" (MFA-) based algorithm [19, 20] in which the impulse response functions are precomputed by exploiting the known geometric relationship between the sound source and an array of sensors, based on the image model method [21, 22]. By convolving the captured signal with the precomputed impulse responses, the signal-to-noise ratio (SNR) of a delay-and-sum beamformer could be significantly increased [19, 20], however, its computational demand is also significant. Due to the high

computational requirement, the real-time application of this method requires a special hardware system [23], thus it has not become widely used.

In this paper, we propose a novel method that integrates the fundamental idea of MFA-based methods into a computationally efficient framework. Our algorithm utilizes precomputed impulse response functions to integrate the effect of reverberation as an additional cue. The hypothetical source location is determined on the basis of matching between the precomputed and the observed map. A similar concept was utilized in [24], where synthesized response patterns of beamformer were compared to observed patterns. In our study, we consider the effect of source directivity on source localization performance; thus our system can more accurately localize nonisotropic sound sources (e.g., human sources) as well, without being limited by their orientation.

2. THE ACOUSTIC MODEL

The source localization problem has led to several proposed signal models which are discussed in [2]. In our work, we utilize a similar signal model that was previously used by Renomeron and his colleagues in [20]. We assume a sound source of point like spatial extent at location s, where $s \in C$and C is a set of discrete points in three-dimensional space, related to possible sound source locations. In addition, we assume that the sound source directivity is given by function $\xi_s(\phi, \theta)$, where ϕ is the azimuth and θ is the elevation angle. There are N microphones located at $m_i (m_i \in C, i = 1 \cdots N)$ with directivities given by function $\xi_m(\phi, \theta)$. The acoustic environment is taken into account as a set of surfaces with given spatial extent and with their independent acoustic absorbing coefficient (β). The effect of reverberation is modeled by frequency-independent specular reflections where the reflected path of sound propagation can be constructed by the image model method [21, 22]. In more complex environments, this can also be done, by more efficiently computable techniques such as ray tracing [25] or beam tracing [26, 27]. The set of sound propagation paths between the source and microphone i is denoted by P_i. In Figure 1, a simplified two-dimensional example can be seen with two reflecting surfaces where a direct path (solid line), two first-order reflection paths (dashed line), and one second-order reflection path (dotted line) are depicted for each microphone. The azimuth angle of the sound source is interpreted as shown in the figure.

According to the above model, the signal recorded by the ith microphone can be written as

$$x_i(t) = \sum_{p \in P_i} a(\tau_p, R_p) \cdot u(t - \tau_p) + \eta_i(t), \tag{1}$$

where u is the signal emitted by the source (s), t is time, τ_p is the time required for the sound to travel through path p, and η_i is additive mutually uncorrelated Gaussian white noise. The list of reflecting surfaces that act along a specified propagation path p is denoted by R_p. Function α represents the

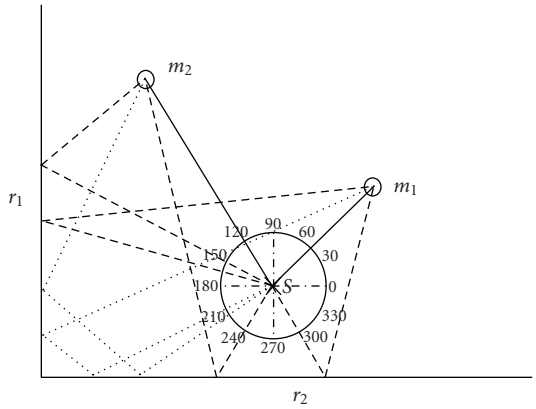

Figure 1: An example of a simple acoustic environment.

effect of attenuation, which in the case of direct propagation is given as

$$a(\tau_p, \{\}) = \frac{1}{\tau_p \cdot v_{\text{sound}}} \cdot \xi_s(\phi_{s,p}, \theta_{s,p}) \cdot \xi_m(\phi_{m,p}, \theta_{m,p}), \tag{2}$$

while in case of reverberant path,

$$a(\tau_p, R_p) = \frac{1}{\tau_p \cdot v_{\text{sound}}} \cdot \xi_s(\phi_{s,p}, \theta_{s,p}) \cdot \xi_m(\phi_{m,p}, \theta_{m,p}) \cdot \prod_{r \in R_p} (1 - \beta(r)) \tag{3}$$

where v_{sound} is the velocity of sound, r an element of R_p, $\beta(r)$ the absorbing coefficient of the reflecting surface r, $\phi_{s,p}$ and $\theta_{s,p}$ the azimuthal and elevation angles of the propagation path p when leaving the source, while $\phi_{m,s}$ and $\theta_{m,s}$ are the azimuthal and elevation angles of the same path measured at microphone i.

3. THE EFFECT OF THE ACOUSTIC ENVIRONMENT ON THE CROSS-CORRELATION FUNCTION

The traditional method of TDOA estimation is based on the well-known cross-correlation function which is computed between two recorded signals as

$$R_{x_i,x_j}(k) = E[x_i(t) \cdot x_j(t - k)], \tag{4}$$

where E denotes expectation. The argument k that maximizes (4) provides an estimate of the TDOA. Because of the finite observation time, however, $R_{x_i,x_j}(k)$ can only be estimated. A widely used estimation method is the computation of

$$c_{x_i,x_j}(k) = \int_{-W}^{W} x_i(t) \cdot x_j(t + k) dt, \tag{5}$$

where $2 \cdot W$ is the time length of window on which the correlation is computed. The range of potential TDOA is restricted to an interval, $k = [-D_{..} + D]$, which is determined by the physical separation between the microphones from

$$D = \frac{\|m_i - m_j\|}{v_{\text{sound}}}, \tag{6}$$

where $\|m_i - m_j\|$ is the length of the vector that interconnects the microphones.

In an anechoic chamber, the highest peak of the cross-correlation function unambiguously assigns the TDOA; however, in everyday acoustic environments, reverberation makes the estimation unreliable, since the delayed replicas of the original signal add unwanted peaks to the correlation function. In our model, the height and place of unwanted peaks can be predicted. In order to make this estimation possible, we substitute (1) into (5) and after some algebraic manipulations which are detailed in the appendix, we obtain the following form:

$$c_{x_i,x_j}(k) = \sum_{(p,q)\in P_i\times P_j} a(\tau_p, R_p)\cdot a(\tau_q, R_q)\cdot c_{u,u}(\tau_p - \tau_q - k), \tag{7}$$

where P_i and P_j are sets of propagation paths from the source to microphones i and j, respectively. The $c_{u,u}(\tau_p - \tau_q - k)$ is the autocorrelation function of signal u with lag k, shifted by $(\tau_p - \tau_q)$ along the time axis and $\times$ denotes the Cartesian product, where (p, q) assigns a 2-*tuple*, where $p \in P_i$ and $q \in P_j$. The cross-correlation function without the joint effect of two specified paths $f \in P_i$ and $g \in P_j$ is denoted by

$$c_{x_i,x_j\setminus(f,g)}(k) = \sum_{(p,q)\in P_i\times P_j\setminus(f,g)} a(\tau_p, R_p)\cdot a(\tau_q, R_q)\cdot c_{u,u}(\tau_p - \tau_q - k). \tag{8}$$

Unfortunately, the computation of (7) is not possible, since the original signal (u) is not available, thus its autocorrelation function ($c_{u,u}$) is not computable. On the other hand, by examining the properties of the autocorrelation function, we can have assumptions regarding certain features of the cross-correlation function.

The autocorrelation function has its highest peak with the steepest slope at zero lag (i.e., *zero-peak*). There are also other smaller peaks with less steep slopes, caused by the periodicity of the signal. The less periodic the signal is, the smaller the further peaks will be. By assuming an aperiodic signal such as Dirac delta, peaks, that is, local maxima of the cross-correlation function can be exactly predicted, since the autocorrelation function ($c_{u,u}$) has only one peak. This observation is valid in case of other aperiodic signals too. In those cases the term "peak" refers to high correlation value, higher than the multiple of the mean of the two signals. When the incoming signal is not completely aperiodic, as happens in case of speech signals, local maximum caused by reverberation appears in the cross-correlation function if there exist paths f and g such that

$$\begin{aligned} a(\tau_f, R_f)\cdot a(\tau_g, R_g)\cdot c_{u,u}(0)'_+ &> c_{x_i,x_j\setminus(f,g)}(\tau_f - \tau_g)'_+, \\ a(\tau_f, R_f)\cdot a(\tau_g, R_g)\cdot c_{u,u}(0)'_- &> c_{x_i,x_j\setminus(f,g)}(\tau_f - \tau_g)'_-, \end{aligned} \tag{9}$$

where $c_{u,u}(0)'_-$ and $c_{u,u}(0)'_+$ indicate the leftward and rightward derivatives of the autocorrelation function at zero lag. The $c_{x_i,x_j\setminus(f,g)}(\tau_f - \tau_g)'_-$ and $c_{x_i,x_j\setminus(f,g)}(\tau_f - \tau_g)'_+$ are the leftward and rightward derivatives of the cross-correlation function without considering the joint effect of paths f and g.

The exact determination of cases when the above conditions hold is not possible without knowing the spectral content of the incoming signal. Nevertheless, the probability of occurrence of local maxima increases if

$$a(\tau_f, R_f)\cdot a(\tau_g, R_g)\cdot c_{u,u}(0) \gg c_{u,u}(h), \tag{10}$$

where $h \neq 0$, that is, the attenuation of a given reverberation path is small, and the nonzero peaks of autocorrelation function are small compared to the height of the zero peak. By using the well-known phase transformation (PHAT) weighting [13], the incoming signal can be whitened and the second condition can be fulfilled.

As a consequence of the above properties, we can define the predicted local maxima function of the cross-correlation function as

$$p_{x_i,x_j}(k) = \sum_{p\in P_i}\sum_{q\in P_j} a(\tau_p, R_p)\cdot a(\tau_q, R_q)\cdot \delta(\tau_p - \tau_q - k), \tag{11}$$

where $\delta(\tau_p - \tau_q - k)$ is the shifted Dirac delta function at lag k. This function does not predict every local maximum of the cross-correlation function. Additional local maxima might exist, owing to the periodicity of the incoming signal, while at the same time, weak reflections do not necessarily produce local maxima. For this, $p_{x_i,x_j}(k)$ can also be referred to as the probability of existence of local maxima at $c_{x_i,x_j}(k)$, although the term "probability" is used loosely (i.e., not in its strict sense). In Figure 2, the cross-correlation function (upper diagram) and the predicted local maxima function (bottom diagram) are illustrated for an omnidirectional source located in the environment shown in Figure 1, and when u is equal to "k" as uttered by a male speaker in an anechoic chamber. It can be seen in Figure 2 that at the places, where $p_{x_1,x_2}(k)$ predicts local maxima with relatively high probability, local maxima appear in the cross-correlation function. Figure 2 illustrates the effect of PHAT weighting as well. Correlation computation on the whitened signals (dotted line in Figure 2) highlights the reverberation effects by suppressing correlation peaks caused by signal periodicity. In Figure 2, squares on the cross-correlation function indicate places of supposed local maxima where reverberation takes effect.

Local maxima of cross-correlation function (either PHAT weighted or not) in Figure 2 are identified by a two-digit code. The first digit identifies the code of the path which has reached m_1, while the second digit identifies the path which has reached m_2. The path code 1 indicates the direct path (solid line in Figure 1); codes 2 and 3 are the first-order reflections from reflectors r_1 and r_2, respectively (dashed lines in Figure 1); while code 4 is the second-order reflection path (dotted line in Figure 1).

The probability function of local maxima in the cross-correlation function ($p_{x_i,x_j}(k)$) depends on the properties of the acoustic configuration, that is, the location of the sound source and the location of reflector surfaces. Thus, by assuming that the reflecting surfaces are fixed, in order to indicate the source location, an additional suffix s has to be affixed to $p_{x_i,x_j}(k)$. Thus, $p_{s,x_i,x_j}(k)$ refers to $p_{x_i,x_j}(k)$ when the source is at location s.

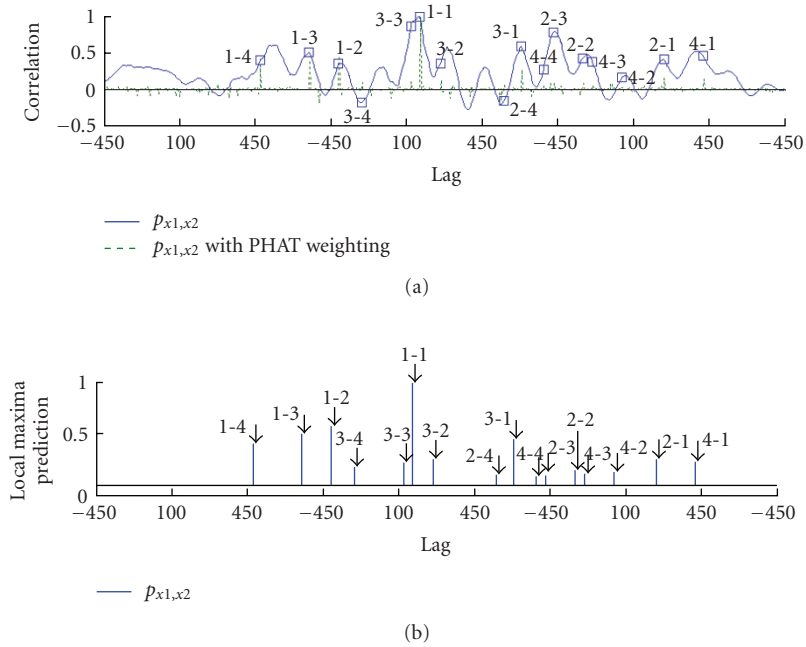

Figure 2: The cross-correlation function (upper) and its prediction of local maxima (lower).

3.1. *Effect of source directivity*

Until now, earlier studies about source localization have not considered the directional characteristics of the source; however, by examining the effect of source directivity, several phenomena can be explained. The relatively weak performance of TDOA-based speaker localization systems used currently is interpreted as the consequence of reverberation that causes spurious peaks in the cross-correlation function, since two reflected paths with the same propagation delay to the microphone may add leading to a higher peak, resulting in false TDOA estimation. By taking source and microphone directivity into account, the coincidence of time difference of reverberation paths is not a necessary condition for the occurrence of false TDOA estimation. Due to the joint effect of the source and microphone directivity, a less attenuated reverberation path may result in a peak higher than that of the direct path. Although in speaker localization systems the application of omnidirectional microphones is widely spread, the directional characteristic of mouth [28] may lead to a difference of several dB in the level of attenuation between different paths. The current attenuation level depends on the spectral content of the speech uttered from the mouth. Even so, as stated in the second section, we apply a frequency-independent model, thus the directivity of mouth is modeled by a function which is independent of the frequency. The attenuation to a given direction is considered to be the average of attenuation computed in the spectral region of interest. Using this simplification, we can state when

$$\alpha(\tau_d, \{\}) < \alpha(\tau_r, R_r) \tag{12}$$

holds, the highest peak will not assign the true source location. In expression (12), indices r and d denote any reflected and direct path, respectively.

In Figure 3, the effect of source directivity of a human speaker in the environment in Figure 1 is illustrated. The cross-correlation function and the probabilities of local maxima in $c_{x_1,x_2}(k)$ for 270° head direction are depicted in Figure 3. As it can be seen, the highest peak of the cross-correlation function (3-3) gives a false TDOA, resulting in bad location estimates in traditional TDOA-based algorithms [6–11].

To find the correct TDOA, the directivity of nonisotropic sound sources should be considered and the definition of predicted local maxima function has to be extended to a direction-specific form. The latter is given by $p_{s,\phi,\theta,x_i,x_j}(k)$, where s is the location of sound source, x_i and x_j refer to the signals recorded by microphone i, and j, ϕ, and θ are the azimuthal and elevation orientations of the source, respectively.

A predicted local maxima function is to be created for each microphone pair based on the given acoustic configuration, that is, the location of sound source and microphones, the direction of sound source, and the acoustic properties of the environment. In fixed acoustic environment, the number of predicted local maxima functions is $\binom{N}{2} \cdot |C_A|$, where N denotes the number of microphones and $|C_A|$ is the cardinality of the set of possible acoustic configurations. C_A contains triplets with general structure (s, ϕ, θ), where s is the location of the sound source ($s \in C$), ϕ and θ are the azimuth and elevation degrees of different source orientations. Obviously, in case of an isotropic sound source, orientation does not need to be distinguished, that is, $|C_A| = |C|$.

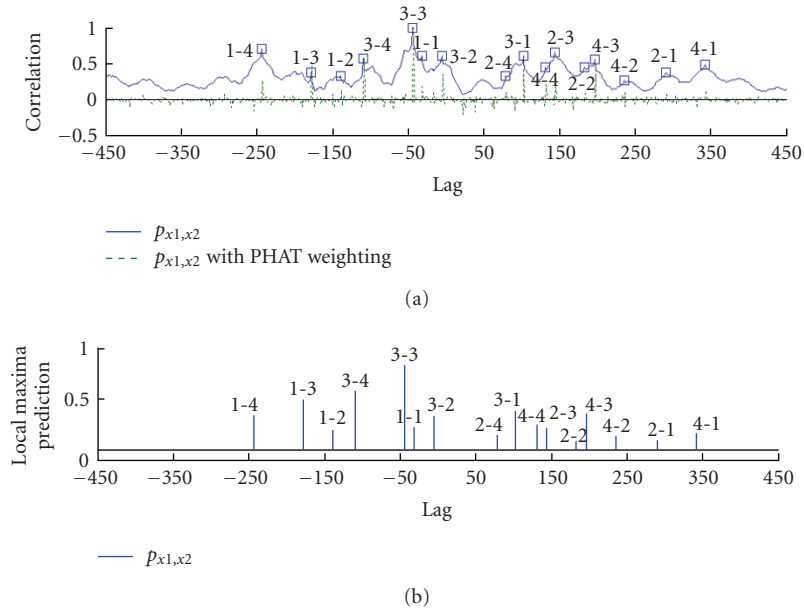

Figure 3: The effect of mouth directivity. The true TDOA is at (1-1).

4. AGGREGATE EFFECT OF THE ACOUSTIC ENVIRONMENT

The proper accumulation of the local maxima predictions of microphone pair combinations is essential for constructing a robust and computationally efficient algorithm. An effective method was published in [11], which follows the principle of least commitment. It is effective as it delays the decision as long as possible, resulting in more robust behavior. The idea is to map the PHAT-weighted cross-correlation functions to a common coordinate system according to

$$£(l) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} c_{x_i,x_j}(\tau_{i,l} - \tau_{j,l}), \tag{13}$$

where $£(l)$ is the likelihood that the source is at location $l (l \in C)$; $\tau_{i,l}$ and $\tau_{j,l}$ are the travel times of the sound wave from location l to microphones i and j, respectively. In this paper, we apply this idea to accumulate the local maxima predictions of the cross-correlation functions, thus we define

$$p^{\mathrm{RM}}_{s,\phi,\theta}(\mathrm{l}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_{s,\phi,\theta,x_i,x_j}(\tau_{i,l} - \tau_{j,l}), \tag{14}$$

where $p^{\mathrm{RM}}_{(s,\phi,\theta)}(l)$ is the accumulated prediction of local maxima at location l for the acoustic setup $(s, \phi, \theta) \in A_C$, in which s is the location of the sound source, ϕ and θ its azimuth and elevation angles. Note that the probability of local maxima in $c_{x_i,x_j}(k)$ depends on the attenuation of delayed replicas caused by reverberation, thus $p^{\mathrm{RM}}_{s,\phi,\theta}(l)$ could also be referred to as the accumulated effect of reverberation at location l, By computation of $p^{\mathrm{RM}}_{s,\phi,\theta}(l)$ for every possible source location point, the so-called accumulated predicted reverberation-effect map (later referred to as predicted reverberation map) can be created, which is denoted by $p^{\mathrm{RM}}_{s,\phi,\theta}$.

Figure 4 shows two predicted reverberation maps: one for the arrangement in Figure 1 (left) and the other for the same arrangement but with an additional microphone (right). The source in this example is assumed to be omnidirectional.

The outstanding features of these maps are their local maxima points. Thus a subset of local maxima points of predicted reverberation map is referred to as

$$\widehat{\widehat{p^{\mathrm{RM}}_{s,\phi,\theta}}} = \left\{ m \in \widehat{p^{\mathrm{RM}}_{s,\phi,\theta}} \mid p^{\mathrm{RM}}_{s,\phi,\theta}(m) > T_r \cdot \max_{c \in C} \left\{ p^{\mathrm{RM}}_{s,\phi,\theta}(c) \right\} \right\}, \tag{15}$$

where T_r is a parameter denoting the lowest level of the predicted reverberation effect that needs to be considered, $\widehat{p^{\mathrm{RM}}_{s,\phi,\theta}}$ is the set of local maxima points. Note that, in the following space, we will use "hat" sign $(\hat{\cdot})$ to denote the local maxima of an arbitrary map, while "double-hat" sign $(\hat{\hat{\cdot}})$ will be used to refer to the local maxima points which are above a certain limit.

5. SOLVING THE INVERSE PROBLEM

In source localization practice, the inputs are records of microphone signals from which a set of cross-correlation functions can be computed. The cross-correlations can be mapped to the monitored region as shown in (13). By computing the likelihood for every possible source location point, the accumulated correlation map (£) [11] can be created, where $£(l)$ refers to the likelihood of source at location l. In [11], the location with the highest probability is selected as the hypothetical source location point. In our approach, we utilize this probability map but we defer the decision and integrate the effect of reverberation as an additional cue to make our estimation robust, as far as speaker direction is concerned.

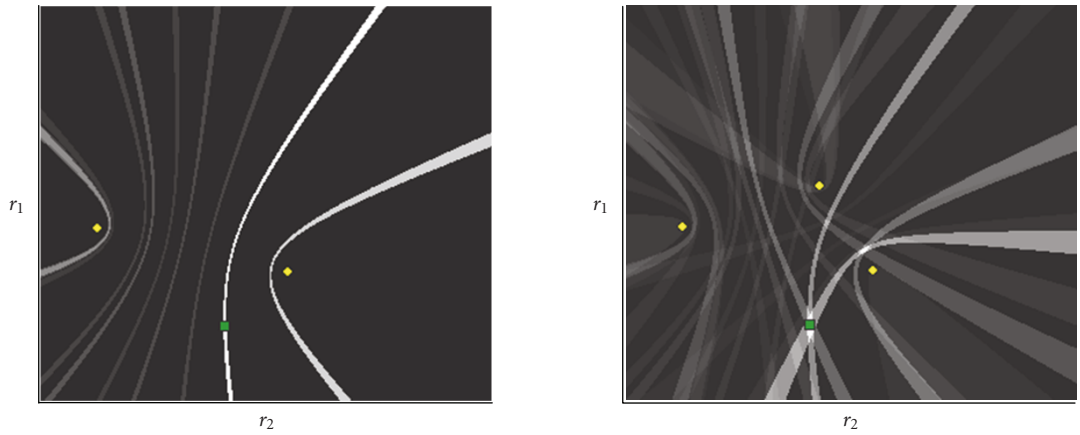

Figure 4: The predicted reverberation map. Rhombi show the places of microphones, and squares indicate the source location.

As we have shown, earlier reverberation causes local maxima in the cross-correlation function. This information is highlighted by applying PHAT weighting during cross-correlation computation. Thus, by finding the local maxima of the accumulated correlation map, the effect of reverberation can be summed up to define

$$\hat{\hat{\pounds}} = \{m \in \hat{\pounds} \mid \pounds(m) > T_r \cdot \pounds_{\max}\}, \tag{16}$$

where $\hat{\pounds}$ indicates the local maxima points of the accumulated correlation map, T_r is the parameter of the lowest limit of significant reverberation effect, and $\pounds_{\max} = \max_{l \in C}\{\pounds(l)\}$.

5.1. Finding the prestored configuration which fits observations best

In the previous sections, we have considered a method for creating predictions and have discussed how to extract the effect of reverberation from our measurement. In the following section, a similarity measure between predictions and observation is analyzed.

First, based on the accumulated correlation map ($\pounds$), the so-called feasible configuration set (f_C) is created. The members of the feasible configuration set ($f_C = \{(z, \phi, \theta) \in C_A\} \subset C_A$) are configurations, such that the accumulated correlation value at the predicted maximum location ($m \in C$, $p^{\mathrm{RM}}_{z,\phi,\theta}(m) = \max_{l \in C}\{p^{\mathrm{RM}}_{z,\phi,\theta}(l)\}$) is close to the maximum of the accumulated correlation map ($\pounds_{\max} \cdot T_c < \pounds(m)$), where T_c controls the acceptable difference compared to the maximum of accumulated correlation map ($\pounds_{\max}$). In the following steps, selection of the most probable configuration among these feasible configurations (f_C) will be discussed.

Note that both the selected local maxima of the predicted reverberation maps ($\widehat{\widehat{p^{\mathrm{RM}}_{s,\phi,\theta}}}$), which are stored for every possible configuration ($(s, \phi, \theta) \in C_A$), and the selected local maxima of the accumulated correlation map ($\hat{\hat{\pounds}}$), which is computed from the cross-correlation function, contain points from the monitored region (C). In both cases, a value is assigned to every location of these maps ($(p^{\mathrm{RM}}_{z,\phi,\theta}(l) \mid l \in \widehat{\widehat{p^{\mathrm{RM}}_{z,\phi,\theta}}})$, $(\pounds(l) \mid l \in \hat{\hat{\pounds}})$) describing their reliability. The number of predicted local maxima points ($|\widehat{\widehat{p^{\mathrm{RM}}_{s,\phi,\theta}}}|$) varies between different configurations. The number of observed local maxima points ($|\hat{\hat{\pounds}}|$) could also vary due to noise, thus the similarity of these two point sets should be measured through global properties such as the center of gravity (P_{cg}). As a consequence, the matching of an observation to the elements of f_c is computed as

$$D(z, \phi, \theta) = \left\| P_{\mathrm{cg}}\left(\widehat{\widehat{p^{\mathrm{RM}}_{z,\phi,\theta}}}\right) - P_{\mathrm{cg}}\left(\hat{\hat{\pounds}}\right) \right\| + \left\| P_{\mathrm{icg}}\left(\widehat{\widehat{p^{\mathrm{RM}}_{z,\phi,\theta}}}\right) - P_{\mathrm{icg}}\left(\hat{\hat{\pounds}}\right) \right\|, \tag{17}$$

where the first term shows the distance from the center of gravities of the prediction (z, ϕ, θ) to that of the observation. The computation of center of gravity on any $M \in \{\widehat{\widehat{p^{\mathrm{RM}}_{z,\phi,\theta}}} \mid (z, \phi, \theta) \in f_C\} \cup \{\hat{\hat{\pounds}}\}$ map can be carried out by evaluating

$$P_{\mathrm{cg}}(M) = \frac{\sum_{m \in M}(M(m) \cdot T_{\mathrm{TDOA}}(m))}{\sum_{m \in M} M(m)}, \tag{18}$$

where $M(m)$ is the value of map M at location $m \in M$ and $T_{\mathrm{TDOA}}(m)$ assigns an $\binom{N}{2}$-dimensional vector that corresponds to m in the TDOA space ($\mathbb{S}_{\mathrm{TDOA}}$), ($T_{\mathrm{TDOA}}(m) \in \mathbb{S}_{\mathrm{TDOA}} \subset \mathbb{R}^{\binom{N}{2}}$). $T_{\mathrm{TDOA}}(\cdot)$ assigns an operator that projects an arbitrary location from C to $\mathbb{S}_{\mathrm{TDOA}}$ as given by

$$T_{\mathrm{TDOA}}(m) = \left(\chi_1, \chi_2, \ldots, \chi_{\binom{N}{2}}\right)^T, \tag{19}$$

where ^{T} assigns the transpose operation, $\chi_k (k = 1 \ldots \binom{N}{2})$ is the kth coordinate in $\mathbb{S}_{\mathrm{TDOA}}$, which is equal to

$$\chi_k = \tau_{i,m} - \tau_{j,m}, \tag{20}$$

where $\tau_{i,m}$ and $\tau_{j,m}$ are the travel times of the sound wave from location m to microphones i and j, respectively. The index pairs of the microphones (i, j) are selected as the kth element of the list of all combinations of the microphone indices.

The result of $P_{\text{cg}}(M)$ is a point in $\mathbb{S}_{\text{TDOA}}$ which assigns the center of gravity of map M. The second term in (17) is thedistance between the so-called inverse center of gravity (P_{icg}) points where the inverse center of gravity of map (M) is computed from

$$P_{\text{icg}}(M) = \frac{\sum_{m\in M}\left[\left(M_{\max} - M(m)\right)\cdot T_{\text{TDOA}}(m)\right]}{\sum_{m\in M}\left(M_{\max} - M(m)\right)}, \quad (21)$$

where $M_{\max}$ is the maximum value of map M.

In (17), $\|\cdot\|$ denotes the length of a vector in the TDOA space which interconnects the points arising from either P_{icg} or P_{cg}, and can be computed as

$$\|v_{\text{TDOA}}\| = \sum_{k=1}^{\binom{N}{2}} \sqrt{v_k^2}, \quad (22)$$

where $v_{\text{TDOA}} \in \mathbb{S}_{\text{TDOA}}$ and v_k is the kth coordinate of v_{TDOA}.

The hypothetical source location point determined by the proposed method is the best matching configuration and is selected as

$$\min_{(z,\phi,\theta)\in f_C}\{D(z,\phi,\theta)\}. \quad (23)$$

To sum up what is mentioned in the previous sections, we extended the accumulated correlation algorithm for acoustic localization. We have built offline maps that store the reverberation effect of different acoustic configurations. The observation gathered from the microphone records were compared to these prestored maps to find the best match, which yields the most likely source location.

6. EFFECT OF DISCRETIZATION

The above equations assume continuous time and an infinitely dense grid of possible source location points, which are obviously not applicable in practice. By assuming that all delays ($\tau_{i,c}$) can be adequately represented by an integer number of sampling periods and by considering the Nyquist-theorem, the continuous-time variables can be replaced by their discretized equivalents. The question of spatial resolution of the accumulated correlation maps leads to the problem of time-delay imprecision or misalignment of beamformers [29]. The energy map of a beamformer is the visual representation of variations in beamformer output energy versus the coordinates of the point which the beamformer is steered to. The source manifests itself as a peak in the energy map. The map depends on the array geometry and on the spectral content of the signal. The width of the peak in the energy map is, generally, smaller for higher-frequency sources. In [29], it is shown that there exists an inverse relationship between the peak width in the energy map and the sound wavelength (λ); and it is conservatively estimated that an error in the source position of less than $\lambda/5$ will still result in a coherent gain in the beamformed signal. This result is referred to as *imprecision heuristic*. Since the accumulated correlation map is essentially the same as the energy map of beamformers [12], the *imprecision heuristic* can be applied in our case as well. Based on this rule and by considering the maximum allowable spatial resolution, the maximum frequency of the sound signal usable for localization can be determined. The same concept can be applied to mapping the predicted local maxima functions in (14). In this case, $p_{x_i,x_j}(k)$ should be redefined as

$$p_{x_i,x_j}(k) = \sum_{p\in P_i}\sum_{q\in P_j} a(\tau_p, R_p)\cdot a(\tau_q, R_q)\cdot\Pi(\tau_p - \tau_q - k), \quad (24)$$

where $\Pi(\tau_p - \tau_q - k)$ is the value of the lowpass filtered and shifted Dirac delta function at lag k. Lowpass filtering of Dirac delta is carried out in compliance with *imprecision heuristic*.

Using this modified version of predicted local maxima function, the $p^{\text{RM}}_{s,\phi,\theta}$ maps can be created for the required resolution in (14).

7. PERFORMANCE EVALUATION

7.1. The test environment

In an attempt to evaluate the performance of the proposed algorithm in a real-reverberant acoustic environment, an acoustic model was built for an auditorium in Pázmány Péter Catholic University (Budapest, Hungary) using the CATT [30] Acoustic simulation software. In the three-dimensional acoustic model of the auditorium (Figure 5) a two-dimensional so-called source location plane was defined parallel to the floor at 1.7 m, the average height of common speakers. In practical applications where the height of speakers varies, it could be necessary to define several source location planes parallel to each other. However, in this paper, we do not consider this a problem and assume the height of the speaker to be constant at 1.7 m. The most significant energy portion of speech is around 500 Hz for male and around 700 Hz for female speakers, thus we choose 700 Hz as the highest frequency used for localization. The spatial resolution was determined from *imprecision heuristic* [29] with resolution of 0.1 m. The set containing the possible source location points (C) was created as nodes of a grid of 0.1 m density defined on the source location plane.

The creation of the predicted local maxima functions requires a priori the impulse response functions from every possible source location points to the microphones. Determination of these impulse response functions by measurements, due to their high number, could be problematic. There are several acoustic modeling softwares [30, 31] available that can be used for predicting the impulse response functions even in a very complex environment. In this work, we have utilized the CATT Acoustic software. The elaboration of the model can be determined along the guidelines described in Section 8.1 by considering the highest frequency

(a) (b)

Figure 5: In the left figure, the 3D model of the simulated acoustic environment of the auditorium is depicted. The right figure is the photo of the modeled auditorium.

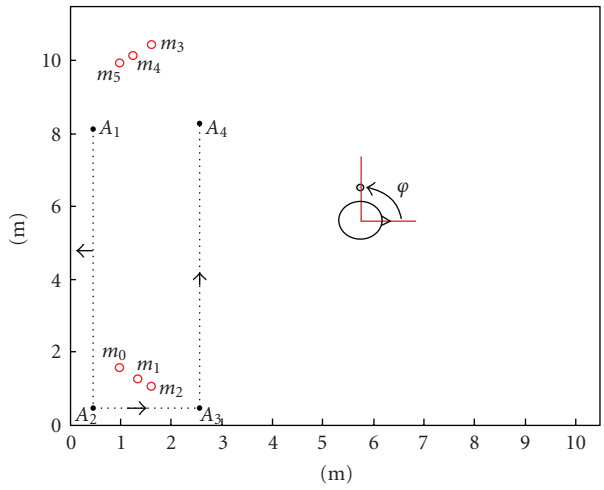

Figure 6: Positions of microphones and the azimuth degree of the speaker direction in the monitored auditorium.

used for localization. Based on these assumptions, we took each object of spatial extent more than 1 m in any direction into consideration. In each possible source location point, we distinguished four different speaker directions, with 90° rotations of the azimuthal degree. The human mouth directivity data used for creating the impulse response functions was created according to the results published in [28] by averaging the directivity data below 1 kHz. According to [28], we may say that this approximation gives good results for several speakers of different sex. Since the variation of the attenuation level of the mouth is relatively independent of the elevation angle of the head in the region of interest, we did not distinguish different elevation angles, and it was fixed at 0° to the source location plane. The location of the omnidirectional microphones and the interpretation of the head direction are shown in Figure 6.

The above procedure resulted in 53891 different acoustic configurations and 323346 impulse response functions. The impulse responses were generated with a maximum of four orders of specular reflections and the predicted local maxima functions were created by considering the fifty strongest reflection paths based on (24) by assuming 25 kHz sampling frequency. The $\widehat{p^{\mathrm{RM}}}$ and $\hat{\mathfrak{L}}$ sets were developed by applying a series of gradient searches. For each run, the initial point of the gradient search was chosen from a subset of C, whose 1077 points were equally distributed in the source location plane. The calculation of all the impulse response functions and the 53891 predicted reverberation-effect maps ($\widehat{\widehat{p^{\mathrm{RM}}}}$) required less than one day for a Pentium IV class computer. In each experiment, the maximum acceptable accumulated correlation difference was set to 5%, and thus the value of T_c was 0.95 at the selection of feasible configuration set (f_C). Performances of the algorithms were compared on a hypothetical speaker path shown by a dashed line in Figure 6. In the first part of the path (A_1-A_2), the speaker turns to the wall and moves to point A_2. This part aims at modeling a lecturer when writing on the blackboard, while speaking to the audience. In the second (A_2-A_3) and the third part (A_3-A_4), speech is directed to the direction of movement. On some parts of this path, condition (12) holds which highlights the extended capabilities of the proposed method; while other parts aim at comparing performance in classical cases when (12) does not hold.

7.2. Optimal level of considerable reverberation effect

In order to check the performance of the proposed method, we divided the 27-second-long anechoic recording of an English male speaker into 40 segments. The sample rate of the signal was 25 kHz, the length of each segment was 32768 samples, and the adjacent segments were overlapped with 16384 samples. The microphone signals were synthesized by convolving these recordings with the generated impulse responses of points on the path shown in Figure 6. The impulse responses used in convolution were generated with eight orders of specular reflections. Performances of the accumulated correlation and the proposed method were measured by using the 700 Hz lowpass filtered versions of the selected segments. In order to examine the global properties of different T_r parameters, we computed the root mean square (RMS) localization error along 178 points of the path, and have shown the results in Figure 7.

Results show that the proposed method decreased the RMS localization error compared with the accumulated correlation method. The optimal value of the considered

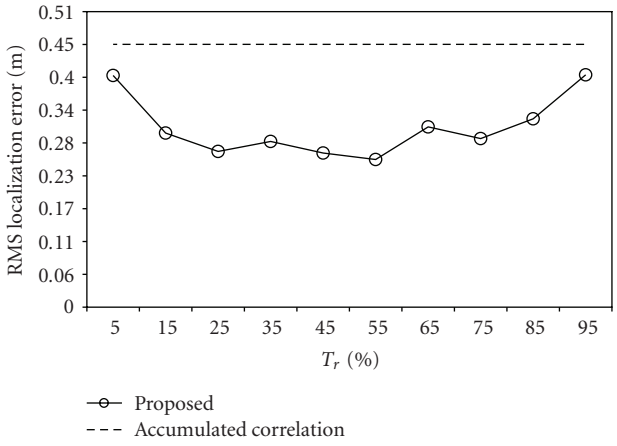

Figure 7: Performance of sound source localization algorithms related to path in Figure 6.

Table 1: Performance of the accumulated and the proposed method on different parts of the path.

	Equation (12) holds	Equation (12) Does not hold
Number of locations	134	44
RMS error of the accumulated correlation [m]	0.58	0
RMS error of the proposed method (T_r = 55%) [m]	0.25	0.1
RMS error of the proposed method (T_r = 25%) [m]	0.3	0.06

reverberation effect is below 55%, because, above this limit, it identifies the source location with more uncertainty. Below this limit, the remaining localization error is caused by the limited capabilities of the applied match measurement induced by the information loss of center of gravities (see Section 5.1). Taking even the smallest peaks into account (below T_r = 15%), the performance decreases because the peaks caused by the deviation of the correlation values of the signals are considered to be the effects of reverberation.

Examining the results in Figure 8, a remarkable performance difference can be observed between the two methods, which originates from the parts of the path given when the speaker faces the wall and the condition in (12) holds. On the remaining portion of the path, both methods perform basically the same as detailed in Table 1. The slightly worse performance of the proposed method when (12) does not hold can be attributed to the imperfections of match measurement detailed in Section 5.1.

7.3. Performance in noisy condition

The robustness of source localization algorithms in noisy conditions is an important feature. Several previous studies [2, 9, 32] on source localization, including this paper, assume that noise is uncorrelated across the array although this assumption does not hold in real environments. Correlating noise fields lead to the improved model of the effect of real-world pointlike noise sources such as computer fans, projectors, and ceiling fans. However, few works [33, 34] succeeded in extending the capabilities of existing methods to spatially correlated noise with known statistics, due to its challenging complexity. The current work does not consider the correlated noise problem but examines the robustness of the proposed method applied to uncorrelated noise fields. We have added mutually uncorrelated Gaussian white noise to the microphone inputs which were used in the previous section. The resulting signals with 30 to −10 dB signal-to-noise-ratio (SNR) were used to compare the performance of the accumulated correlation method with the performance of the proposed one with T_r = 0.55 and T_r = 0.25.

The results in Figure 9 show that for low-SNR values, the proposed method gives slightly worse results. The reason is that added noise causes additional local maxima in the cross-correlation function. Since the effect of reverberation is considered through local property (i.e., local maximum), additional local maxima caused by added noise make the estimation less reliable. A possible solution to this problem could be the integration of the effect of reverberation in certain areas (see the lighter areas in Figure 4). However, the proper integration of the effect of reverberation at acceptable speed is not a trivial task, and it is not discussed in this work.

7.4. Performance in different acoustic environment

The performance evaluation of localization algorithms in different reverberation conditions is a common practice [1–14]. In this paper, we use reverberation as an additional cue to make the localization more robust; thus in our case, this task is interpreted as to evaluate localization performance in varying acoustic conditions. The acoustic environment may alter due to the effect of several factors [35] such as humidity, temperature, location of reverberant/absorption surfaces. By considering the typical application area of our algorithm, the first two effects can be ignored since these parameters in everyday conference environment are considered to be constant together with location and wrapping, that is, absorption coefficient of walls and furniture. However, the number of people in the hall may vary from one person to full capacity of the room, thus we have to evaluate the performance of our algorithm as the function of the density of listeners in the auditorium. To analyze the effect of the audience size on the localization performance, we used the acoustic model discussed earlier. We have synthesized records based on the same path (see Figure 6), but the absorption coefficient of the audience area was changed to the measured values published in [36]. Using this method, we simulated a density of 2 person/m^2 in the audience area with changing reverberation time (T_{30}) of the auditorium from 3.5 seconds to 1.5 seconds. The localization was performed on microphone signals which were synthesized by impulse responses of the altered room. The results of this experiment are shown in Figure 10 where the RMS localization error ratio of the proposed method with T_r = 55% to accumulated correlation is depicted. The figure shows that the proposed method tolerates moderate changes

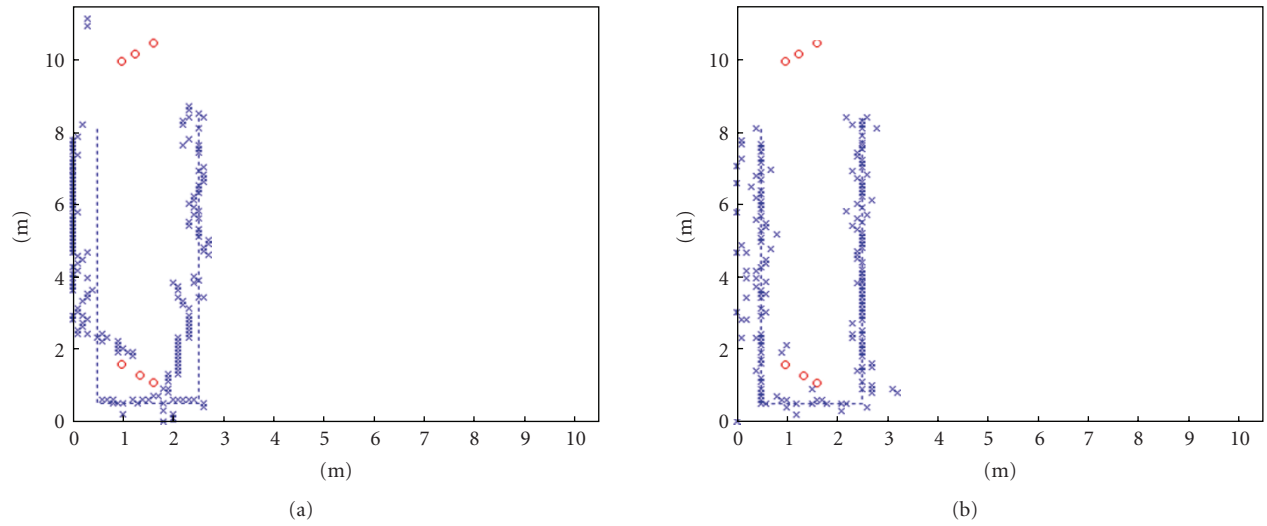

Figure 8: Localization results. The left figure shows results by the accumulated correlation method, while the right figure shows the results through the proposed method with $T_r = 55\%$.

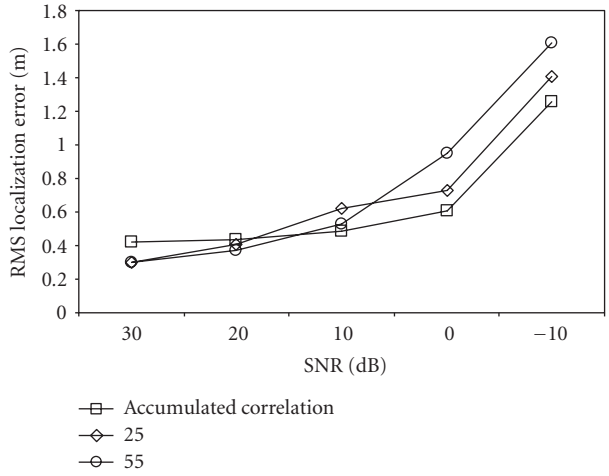

Figure 9: Effect of added Gaussian white noise on localization performance.

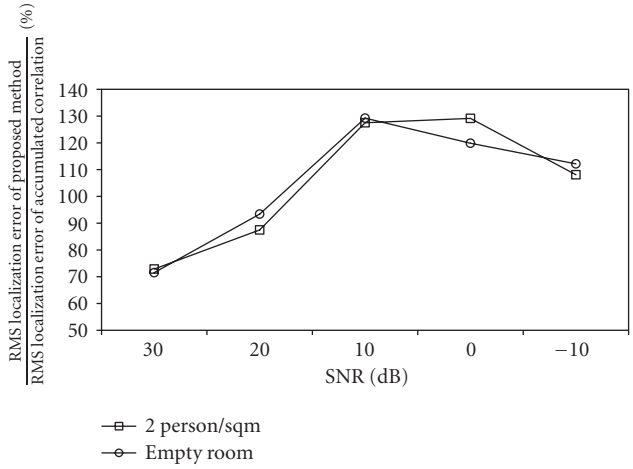

Figure 10: Localization performance in different acoustic conditions.

in the acoustic environment, due to the fact that its performance basically does not alter.

7.5. Speed of convergence

A conventional way of obtaining more reliable location estimates is to aggregate the results of several measurements. The speed of convergence of estimates to the true source location could be an important issue in case of low-quality measurements. In case of the algorithms in question, the accumulation of results of different measurements is done through the aggregation over time of accumulated correlation maps, thus we redefine the notation of $£(l)$ as

$$£(l) = \sum_{i=L-S}^{L} £_i(l) \quad \forall l \in C\,, \tag{25}$$

where $£_i(l)$ is the accumulated correlation map of the ith measurement computed according to (13) at location l, and L is the sequence number of the last measurement. S controls the number of previous measurements to be considered. The value of S should be set according to the several parameters of application such as the maximum velocity of the moving speaker, the sampling rate, or the length of window on which correlation is computed ($2 \cdot W$). In our experiments, we set $S = L$ to examine the convergence speed of the proposed method. The results of localization algorithms were checked at each location of the path shown in Figure 6. The microphone signals applied in this experiment were synthesized by applying the same anechoic recordings we used earlier. In order to examine the evaluation of estimates along the time axis, 27-second-long signals were created for each location (i.e., the speaker spent 27 seconds in each location on the path). The results of both methods were determined after every 32768 samples of the microphone signals for each location on the path. The RMS localization errors computed for each location were averaged along the path in each time instance with the results shown in Figure 11.

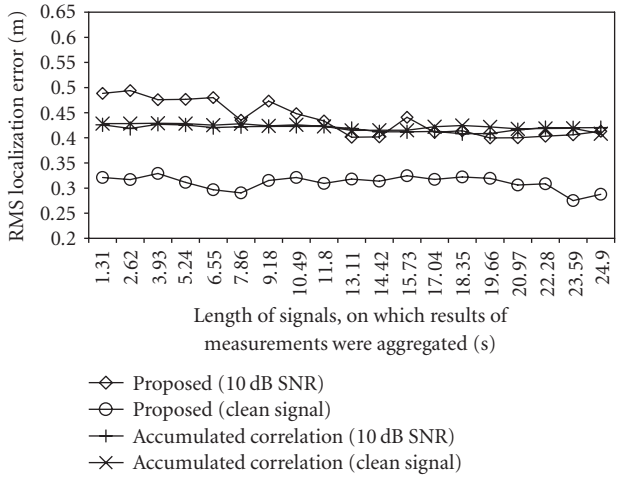

Figure 11: Evaluation of location estimates by aggregating the results of several measurements.

The evaluation of location estimates was performed for both clean and noisy signals with a 10 dB SNR.

The results show that the summation of results of several measurements does not change the performance of the accumulated correlation method, since the error caused by joint effect of reverberation and source directivity holds for each measurement. Moreover, it can be seen that a 10 dB SNR in the recorded signals does not alter the performance of accumulated correlation method. The examination of the results of the proposed method on clean signals suggests that localization error is independent on the signal length as its performance does not change during the evaluation. Moreover, it proves that the error of the proposed method is caused by the imperfections of matching measurement between observation and predictions and the error incurred from spatial discretization. Performance evaluation on noisy signals shows that by averaging several measurements, the error introduced by the added noise can be decreased, and the performance of the proposed method can be slightly improved. Nevertheless, it does not exceed the performance of the accumulated correlation method and the speed of convergence is too slow to track speakers in practical applications.

8. DISCUSSION

8.1. Validity of the applied acoustic model

In our work, we have considered the frequency-independent specular sound reflection model that has certain limitations. Using the model in question, good approximations can be achieved when the following conditions hold.

(i) The wavelength of the sound signal is significantly shorter than the extent of the reflector.
(ii) The surface of the reflector can be considered to be planar compared to the sound wavelength.
(iii) There is no obstructing object along the propagation path the extent of which would be comparable to the wavelength.

Table 2: Approximation of frequencies where the applied acoustic model gives good approximation of real-world effects.

Application environment	Typical dimensions (height · width · depth)	Lowest frequency range
Office room	3 m · 5 m · 5 m	2 kHz
Class room	3 m · 10 m · 6 m	1.5 kHz
Small Auditorium	5 m · 15 m · 10 m	600 Hz
Conference Hall	8 m · 30 m · 30 m	200 Hz

In cases when the first and the third conditions fail, the edge diffraction effects should be considered, while failure of the second condition implies the importance of modeling sound scattering. By considering a typical conference environment and application profile, we can assume that the third condition holds. The investigation of the remaining factors, however, is an active research area in computational acoustics. Studies related to the problem [37–39] suggest that the early part of reverberation can be well characterized by the specular reflection model. Since early reflections contain the main portion of energy of the reverberated sound, the applied model is considered suitable for predicting the most disturbing peaks in the cross-correlation function that is caused by delayed replicas of sound.

Based on the typical dimensions of the monitored region, the validity of specular reflection model can be evaluated by considering the frequency of the sound signals which are present. In our study, we considered four application environments, listed in Table 2. In this table, we indicate the lowest-frequency range above which the specular reflection model can be considered to be a good approximation of the real world. For smaller enclosures, the results show that only the higher frequency portion of speech can fit this method. Consequently, the proposed method is more suitable for speaker localization in auditoriums and conference halls.

8.2. Computational requirement

The speed of source localization algorithms is a crucial factor, because the typical application profile requires real-time processing. In Table 3, we summarized the offline and real-time computational requirement of the proposed procedure, the accumulated correlation and the MFA-based methods.

The distinct advantage of the proposed method compared to MFA-based ones is that there is no need to deconvolve the input signal in real-time at each location of the search space, since the effect of reverberation is offline evaluated. On the other hand, this method carries moderate computational overhead compared to the accumulated correlation, owing to local maxima extraction and match measurement. The effect of this latter factor can be controlled through parameter T_c creating a feasible configuration set. In our experiments, the number of configurations in this set was

Table 3: Estimated computational requirement of different algorithms.

Algorithm	Offline computation	Real-time computation
Accumulated correlation [11]	—	- Computation of cross-correlation function - Mapping to a common coordinate system.
Delay-and-sum beamformer with MFA [20]	- Computation of impulse response function for each possible source location point	- Deconvolution of the input signal with the appropriate impulse response function for every possible source location point - Computation of beamformer response with the deconvolved signals
Proposed method	- Computation ofimpulse response function for each possible configuration - Creation of predicted reverberation map for each configuration - Search for local maxima in each predicted map	- Computation of cross-correlation functions - Mapping to a common coordinate system - Search for local maxima - Match measurement with a subset of stored predictions

less than 100 in every case, leaving most of the computational overhead to the gradient search process. Even though, creating the accumulated correlation map can be performed more efficiently than the energy response map of beamformer, the computational requirement of the proposes method is still questionable in cases when three-dimensional search space is considered.

9. CONCLUSION

In this work, a novel TDOA-based sound source localization algorithm was presented which integrates a priori information of the acoustic environment for the localization of directional sound sources in reverberant environments. The algorithm utilizes the redundant information provided by multiple sensors to enhance the TDOA performance. By the support of the specular reflection model of sound waves, more reliable localizations can be achieved in the cases when the joint effect of source directivity and reverberation causes traditional methods to fail. The proposed method results in significantly better estimates in case of noise-free signals, while it performs worse in SNR's signals lower than 20 dB. We showed that integration of information from the acoustic environment could be carried out offline with reasonable real-time computational overhead. The validity of the acoustic model applied and the performance of the proposed algorithm in various simulated acoustic conditions were discussed suggesting its usability in conference environment. Although this work demonstrated the importance of directional properties of sound sources and showed an alternative localization framework where a matching of observations to predicted quantities was considered, speaker localization remains to be a difficult problem in a practical noisy environment. Further research is stimulated by the results obtained in this direction to increase performance of source localization algorithms using a priori information of the acoustic environment.

APPENDIX

By substituting (1) to (4), we get

$$\begin{aligned} R_{x_i,x_j}(k) &= E\left[\left(\sum_{p\in P_i} a(\tau_p,R_p)\cdot u(t-\tau_p)+\eta_i(t)\right)\right. \\ &\qquad \left.\cdot\left(\sum_{q\in P_j} a(\tau_q,R_q)\cdot u(t-\tau_q-k)+\eta_j(t-k)\right)\right], \\ R_{x_i,x_j}(k) &= E\left[\left(\sum_{p\in P_i} a(\tau_p,R_p)\cdot u(t-\tau_p)\right)\right. \\ &\qquad \left.\cdot\left(\sum_{q\in P_j} a(\tau_q,R_q)\cdot u(t-\tau_q-k)\right)\right] \\ &\qquad + E\left[\left(\sum_{p\in P_i} a(\tau_p,R_p)\cdot u(t-\tau_p)\right)\cdot\eta_j(t-k)\right] \\ &\qquad + E\left[\left(\sum_{q\in P_j} a(\tau_q,R_q)\cdot u(t-\tau_q-k)\right)\cdot\eta_i(t)\right] \\ &\qquad + E[\eta_i(t)\cdot\eta_j(t-k)]; \end{aligned} \tag{A.1}$$

as we assumed, the components are mutually uncorrelated, thus the second, third, and the fourth expressions are approximately equal to zero if both signal and noise have zero-mean and the equation can be rewritten as

$$\begin{aligned} R_{x_i,x_j}(k) &= \sum_{p\in P_i}\sum_{q\in P_j} a(\tau_p,R_p)\cdot a(\tau_q,R_q) \\ &\qquad \cdot E[u(t-\tau_p)\cdot u(t-\tau_q-k)], \end{aligned} \tag{A.2}$$

which is equal to the weighted sum of the time-shifted autocorrelation functions whose shift is equal to $(\tau_p-\tau_q)$:

$$c_{x_i,x_j}(k) = \sum_{p\in P_i}\sum_{q\in P_j} a(d_p,R_p)\cdot a(d_q,R_q)\cdot c_{u,u}(\tau_p-\tau_q-k). \tag{A.3}$$

ACKNOWLEDGMENTS

Special thanks are due to Daniel V. Rabinkin for his audio records that helped us during the development of the algorithm, to Bengt-Inge Dalenback, head of CATT Acoustic, for supporting the software for research, and to Dr. A. C. C. Warnock for supplying directional data of the human mouth. We also thank the anonymous reviewers for their valuable comments. This project has been supported by the Hungarian Scientific Research Fund OTKA-TS40858.

REFERENCES

[1] J. H. DiBiase, H. F. Silverman, and M. S. Branstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Branstein and D. B. Ward, Eds., chapter 8, pp. 157–180, Springer, New York, NY, USA, 2001.

[2] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[3] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.

[4] G. C. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 922–926, 1977.

[5] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[6] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 6, pp. 998–1003, 1982.

[7] M. Brandstein, J. E. Adcock, and H. Silverman, "Practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech and Language*, vol. 9, no. 2, pp. 153–169, 1995.

[8] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, 1997.

[9] M. Brandstein and H. Silverman, "Robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 1, pp. 375–378, Munich, Germany, April 1997.

[10] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 1, pp. 231–234, 1997.

[11] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian acoustic localization," in *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP '02)*, vol. 2, pp. 1793–1796, Orlando, Fla, USA, May 2002.

[12] S. T. Birchfield, "A unifying framework for acoustic localization," in *Proceedings of the 12th European Signal Processing Conference (EUSIPCO '04)*, Vienna, Austria, September 2004.

[13] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[14] M. S. Brandstein, "Pitch-based approach to time-delay estimation of reverberant speech," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ASSP '97)*, New Paltz, NY, USA, October 1997.

[15] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[16] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ASSP '01)*, pp. 71–74, New Paltz, NY, USA, October 2001.

[17] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 549–557, 2003.

[18] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 509–519, 2004.

[19] E. E. Jan, *Parallel processing of large scale microphone arrays for sound capture*, Ph.D. thesis, Rutgers the State University of New Jersey, New Brunswick, NJ, USA, 1995.

[20] R. J. Renomeron, D. V. Rabinkin, J. C. French, and J. L. Flanagan, "Small-scale matched filter array processing for spatially selective sound capture," in *Proceedings of the 134th Meeting of the Acoustical Society of America*, San Diego, Calif, USA, December 1997.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[22] J. Borish, "Extension of the image model to arbitrary polyhedra," *Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.

[23] H. F. Silverman, W. R. Patterson III, J. L. Flanagan, and D. V. Rabinkin, "Digital processing system for source location and sound capture by large microphone arrays," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 1, pp. 251–254, Munich, Germany, April 1997.

[24] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell, "A probabilistic framework for multi-modal multi-person tracking," in *Proceedings of IEEE Workshop on Multi-Object Tracking (WOMOT '03)*, Madison, Wis, USA, June 2003.

[25] A. Krokstad, S. Strom, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[26] P. Heckbert and P. Hanrahan, "Beam tracing polygonal objects," in *Proceedings of the 11th Annual Conference on Computer Graphics (SIGGRAPH '84)*, vol. 18, no. 3, pp. 119–127, Minneapolis, Minn, USA, July 1984.

[27] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West, "Beam tracing approach to acoustic modeling for interactive virtual environments," in *Proceedings of the Annual Conference on Computer Graphics (SIGGRAPH '98)*, pp. 21–32, Orlando, Fla, USA, July 1998.

[28] W. T. Chu and A. C. C. Warnock, "Detailed directivity of sound fields around human talkers," IRC Research Report IRC-RR-104, National Research Council of Canada, Ottawa, Ontario, Canada, 2002.

[29] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 499–508, 2004.
[30] CATT-Acoustic http://www.catt.se.
[31] Odeon Room Acoustic Software http://www.odeon.dk.
[32] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Processing Letters*, vol. 12, no. 8, pp. 561–564, 2005.
[33] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 2, pp. II133–II136, Montreal, Canada, May 2004.
[34] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, 2003.
[35] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, John Wiley & Sons, New York, NY, USA, 1962.
[36] L. Karlen, *Akustik i rum och byggander*, Svensk Byggtjänst, Stockholm, Sweden, 1983.
[37] N. Tsingos, I. Carlbom, G. Elko, R. Kubli, and T. Funkhouser, "Validating acoustical simulations in the Bell Labs Box," *IEEE Computer Graphics and Applications*, vol. 22, no. 4, pp. 28–37, 2002.
[38] M. Kleiner, R. Orlowski, and J. Kirszenstein, "Comparison between results from a physical scale model and a computer image source model for architectural acoustics," *Applied Acoustics*, vol. 38, no. 2–4, pp. 245–265, 1993.
[39] L. L. Beranek, *Concert and Opera Halls: How They Sound*, American Institute of Physics, Woodbury, NY, USA, 1996.