A Phase-Modified Approach for TDE-Based Acoustic Localization

Georgios Athanasopoulos⋆, Werner Verhelst†

⋆Vrije Universiteit Brussel, Dept. of Electronics and Informatics, Pleinlaan 2, Brussels, Belgium
†iMinds, Dept. of Future Media and Imaging, Gaston Crommenlaan 8, Ghent, Belgium

gathanas@etro.vub.ac.be, wverhels@etro.vub.ac.be

Abstract

A novel approach is presented for acoustic localization based on Time Delay Estimation (TDE). The proposed method relies on the observation that the spectrum of many real-world ambient noise types sufficiently satisfies the assumption of sinusoidal stationarity. By employing noise phase spectrum projection, we approximate the phase information of the equivalent clean sources received by the microphone array. The phase-modified signals can be effectively used with different methods of conventional TDE. The proposed approach is evaluated with real noise data and for different reverberation times. Experimental results confirm the suitability of the sinusoid stationarity concept and show that the proposed method performs well under both reverberant and noisy acoustic conditions.

Index Terms: time delay estimation, generalized cross correlation, adaptive eigenvalue decomposition, noise phase spectrum

1. Introduction

Sound Source Localization is crucial for a large number of applications such as hearing aid devices, hands-free communication systems, robot audition systems, etc. The most commonly applied methods rely on Time Delay Estimation (TDE) of the audio signal arriving at the different sensors of a microphone array. Different techniques for performing the TDE have been proposed [1]. The Generalized Cross-Correlation (GCC) [2] is a well-established method, frequently used due to its simplicity and computational efficiency. In the GCC, the TDE is facilitated by the introduction of properly selected filters, known as weighting functions. Over the past few years new TDE techniques have been proposed such as the Steered Response Power (SRP) [3] and the Oriented Global Coherence Field (OGCF) [4]. The SRP and OGCF essentially rely on the GCC and are known to be suitable for multiple speakers localization [5, 6]. Hence, improvement of the GCC performance also benefits a wider class of GCC-based techniques. A more recent type of TDE utilizes the eigenvector space of the signals’ covariance matrix. The Adaptive Eigenvalue Decomposition (AED) is a commonly used method of this type [7]. The AED uses an estimation of the impulse responses between the acoustic source and the microphones for performing the TDE and can cope with both noisy and reverberant environments [8]. Despite the improvements proposed over the recent years, in practice, the performance of both GCC and AED appears to decrease in the presence of prominent ambient noise.

On the other hand, recent research in the area of speech enhancement for speech recognition has demonstrated the potential of Phase Estimation via Delay Projection which is based on the concept of sinusoid stationarity [9]. Motivated by the important role of the phase spectrum in the TDE and having in mind the detrimental effect of ambient noise, we propose a novel phase-modified approach which relies on the observation that the spectrum of many real-life noise types can be approximated by stationary or slowly time-varying sinusoids. This approach allows to project the estimated noise phase information of non-speech periods to subsequent speech periods and thus approximate the phase information of the equivalent clean source signal received by the microphone array.

Addressing the effect of noise and reverberation has already been in the epicenter of TDE for several years. To this respect, the authors of [10] developed a maximum likelihood estimator for both ambient noise and reverberation. More recently, noise suppression techniques were fused to the TDE problem [11, 12]. A wavelet based pre-filtering was introduced in [13], while [14] proposes a modified weighting function for joint noisy and reverberant situations. As opposed to the previous studies in which only the magnitude of the spectrum is considered, this paper proposes an approach that takes both the magnitude and the phase spectra into account. Moreover, we show that the phase-modified approach can be effectively applied to improve the performance of different existing TDE methods.

The paper is organized as follows. In section 2, we formulate the theory and review the GCC and AED focusing on how reverberation and especially ambient noise introduce errors in the TDE. In section 3 we discuss the characteristics of ambient noise and look at how the noise phase can be estimated via delay projection. Section 4 motivates and describes the proposed algorithm. Finally, in section 5 we present the experimental evaluation and discuss the results.

2. Time Delay Estimation

2.1. The Role of Reverberation and Noise

Let us consider a microphone array of N sensors and denote any microphone pair of the array as \((m_i, m_j)\). We assume \(s(n)\) to be the unknown source signal in the far-field. We also accept the room to be linear and time invariant. Considering only the direct path, the signal received by microphone \(m_i\) can be written as

\[
x_i(n) = a_i s(n - \tau_i) + w_i(n)
\]

where \(a_i\) is the attenuation factor, \(\tau_i\) the propagation time, \(w_i(n)\) the additive noise and \(n\) the time index in samples. The time delay \(\tau_{i,j}\) between the two sensors of a microphone pair will be

\[
\tau_{i,j} = \tau_i - \tau_j.
\]

Although Eq. (1) can be expanded to accommodate delayed components due to multipath propagation (reverberation), in real-world settings a more accurate model can be considered. Suppose that \(h_i(n)\) is the room’s impulse response (of finite length \(L\)) for the given source and microphone positions. The received signal by microphone \(m_i\) can be now expressed as the
linear convolution between the source signal and the room’s impulse response
\[ x_i(n) = h_i(n) * s(n) + w_i(n). \]  
Although Eq. (3) does not explicitly note the propagation time, this model allows to take the room’s reverberation into account.

An interesting observation in both models is that the estimation of the time delay \( \tilde{\tau}_{i,j} \) from the available signals will be disturbed by the presence of ambient noise and reverberation. Rewriting Eq. (3) in the frequency domain yields
\[ X_i(k) = |H_i(k)||S(k)|e^{j\angle H_i(k) + \angle S(k))} + |W_i(k)|e^{j\angle W_i(k)} \]  
with \( X_i(k) \), \( H_i(k) \), \( S(k) \) and \( W_i(k) \) the Discrete Fourier Transforms of \( x_i(n) \), \( h_i(n) \), \( s(n) \) and \( w_i(n) \) respectively and \( k \) denoting the frequency bin index. From Eq. (4), it is evident that the presence of ambient noise and reverberation influences not only the magnitude but also the phase of the received signals.

### 2.2. Generalized Cross-Correlation

A well established method for computing the TDE for a microphone pair is to locate the highest peak of the GCC function
\[ r_{i,j}(n) = \sum_{k=0}^{K-1} \Psi(k)X_i(k)X_j^*(k)e^{-j\frac{2\pi nk}{N}} \]  
where \( X_i(k) \), \( X_j(k) \) are the Short-Time Fourier Transform (STFT) of the signals \( x_i(n) \), \( x_j(n) \) from microphones \( m_i \) and \( m_j \) respectively, and \( * \) denotes the complex conjugate. The TDE \( \tilde{\tau}_{i,j} \) can then be found as [2]
\[ \tilde{\tau}_{i,j} = \arg \max_n r_{i,j}(n). \]  

In Eq. (5), \( \Psi(k) \) is a frequency dependent weighting function used to de-emphasize specific frequency components that are expected to contribute unreliable information to the GCC function. The selection of a suitable weighting function depends on several criteria, such as the presence of ambient noise, reverberation and even the properties of the signals (e.g., periodicity). Choosing the most appropriate weighting function is of great importance in practical implementations. Various weighting functions have been proposed in literature. The most commonly used examples are the Maximum Likelihood (ML) and Phase Transform (PHAT) [2].

The ML weighting function is defined as
\[ \Psi_{\text{ML},i,j}(k) = \frac{|X_i(k)||X_j(k)|}{|N_i(k)|^2|X_i(k)|^2 + |N_j(k)|^2|X_j(k)|^2} \]  
where \( N_i(k) \), \( N_j(k) \) denote the noise spectra at the respective microphones \( m_i \), \( m_j \). The ML function of Eq. (7) can be seen as a frequency dependent filter that increases the contribution of regions of the spectrum that exhibit a high signal-to-noise ratio (SNR). The ML function is derived analytically. Although it is optimal for high levels of uncorrelated noise, its performance can degrade for signals exhibiting a high SNR (i.e., \( \sim 20 \text{dB} \)) [15]. The ML weighting assumes the propagation model of Eq. (1) and a stationary noise model which is uncorrelated with the source signal and with the noise signals observed at other microphones. However, in real-world environments the performance of the GCC-ML is suboptimal since at that case reverberation is present, the ambient noise picked by the microphones is (partly) correlated and the noise spectra need to be estimated as they are not known a priori [15].

The Phase Transform is the empirical weighting function
\[ \Psi_{\text{PHAT},i,j}(k) = \frac{1}{|X_i(k)||X_j(k)|}. \]  
In the absence of noise, the PHAT weighting is known to perform well under reverberant conditions as it depends only on the room’s response [16]. By substituting Eq. (6) in Eq. (5), it is apparent that in GCC-PHAT all spectral components have identical contribution. As a consequence, in the presence of noise, the spectral components that are dominated by noise receive equal importance and thus introduce errors in the GCC.

Summarizing, the GCC is a simple and computationally efficient method whose implementation (Eq. (5)) requires only three FFTs per frame. As it relies on the signal model of Eq. (1), its performance is satisfactory in the absence (or moderate presence) of noise and reverberation. Under these conditions and as can be observed from Eq. (5), the TDE depends only on the phase differences between \( X_i(k) \) and \( X_j(k) \), while magnitude differences do not affect the peak location. Higher reverberation and noise levels influence the phase differences between \( X_i(k) \) and \( X_j(k) \) and introduce errors in the TDE. Finally, the weighting functions of Eqs. (7) and (8), although to some extent alleviate the effects of noise and reverberation, make use only of the signals’ magnitude while they do not take into account the phase information.

### 2.3. Adaptive Eigenvalue Decomposition

The AED algorithm proposed in [7] overcomes some of the drawbacks of the GCC. This approach assumes the signal model of Eq. (3) and relies on the rough estimation of the impulse responses between the acoustic source and the microphone array signals. Assuming that the main peaks of the impulse responses correspond to the direct path between the source and each microphone, the TDE can be calculated as the time difference between these two main peaks. In the absence of noise, the following relation holds
\[ h_j(n) * x_j(n) = h_j(n) * h_i(n) * s(n) = h_i(n) * x_j(n) \]  
where \( h_i(n) \), \( h_j(n) \) are the two impulse responses and \( s(n) \) the unknown source signal. Let us assume the covariance matrix \( \mathbf{R} \) of a microphone pair signals
\[ \mathbf{R} = \begin{bmatrix} \mathbf{R}_{x_i x_i} & \mathbf{R}_{x_i x_j} \\ \mathbf{R}_{x_j x_i} & \mathbf{R}_{x_j x_j} \end{bmatrix} \]  
where
\[ \mathbf{R}_{x_i x_j} = E[x_i(n)x_j^T(n)] \]  
with \( E[\cdot] \) denoting the expected value operator. From Eqs. (9) and (10) it can be shown that
\[ \mathbf{R} \mathbf{u} = 0 \]  
where \( \mathbf{u} = [h_i - h_j] \). From Eq. (12) follows that \( \mathbf{u} \) is the eigenvector of the covariance matrix \( \mathbf{R} \) corresponding to the eigenvalue 0.

In the presence of noise, Eq. (12) does not hold strictly and therefore the two impulse responses can be estimated as the eigenvector corresponding to the smallest eigenvalue. The estimated eigenvector \( \hat{\mathbf{u}} = [\hat{h}_i - \hat{h}_j] \) can be obtained through the adaptive algorithm proposed (along with the pertinent assumptions) in [7], or the frequency domain approach of [17]. The latter requires four FFTs per frame and hence reaches a computational efficiency similar to the GCC.
The AED algorithm is known to perform well in reverberant environments as it takes fully into account the reverberation effect during the TDE [15]. In the presence of prominent ambient noise however, the microphone signals \(x_i(n)\) are contaminated (as per Eq. (4)) and the performance of the AED degrades.

### 3. Noise Phase Projection

In practical applications, the signals of a microphone array contain ambient noise from the environment in which the array is installed. Ambient noise is the result of fast pressure variations in the fluid medium, i.e., the air. The origin of these variations spans from mechanical vibrations to aerodynamic forces causing pressure perturbations. The observed ambient noise is a composite of all variations from many sources. Since many noise sources consist of repetitive pressure pulses, it is intuitive to assume that in periods during which the ambient noise is steady (i.e., non-rapidly varying), it exhibits somewhat certain frequency content and periodic characteristics.

Let us consider a non-speech period. According to Eq. (3), the signal \(x_i(n)\) from microphone \(m_i\) contains only the ambient noise \(w_i(n)\). From the STFT analysis and assuming that the STFT center angular frequencies represent the actual noise frequencies so that the effect of spectral smearing is minimal, it follows that the observed ambient noise for the given frame can be expressed as a sum of sinusoids

\[
w_i(n) = \sum_{k=0}^{K-1} a_{i,k} \cos(\omega_{i,k} n + \varphi_{i,k})
\]

where \(k\) is the frequency bin index, \(K\) the length of the STFT, \(f_s\) the sampling frequency and \(a_{i,k}, \omega_{i,k}, \varphi_{i,k}\) are the sinusoidal amplitude, angular frequency and phase respectively. We further accept, as discussed above, that the sinusoids remain stationary or are slowly time-varying between sufficient number of frames. The expected phase \(\varphi_k\) for each sinusoid (corresponding to frequency bin \(k\)) at the start of a frame \(\lambda\) can be expressed as

\[
\varphi_k = \varphi_k^\lambda + \tau\omega_k\delta
\]

where \(\tau\) is the time between the start of subframes frames and \(\delta\) is the number of frames prior to the current frame \(\lambda\) in which the reference phase is assumed known. In practical implementations, the reference phase can be calculated during non-speech periods while presuming sinusoid (quasi)stationarity during speech periods.

As shown in [9], for sufficiently small frame shifts between subframes, the sinusoid stationarity assumption is satisfied for both typical noise sources and clean speech. In the scope of this paper, it is not the intention to accurately estimate the phase spectra of the entire speech signal (e.g., for performing speech enhancement addressed to human listeners). Our objective is to improve the TDE by employing an approximation of the equivalent clean speech phase for as many frequency bins and processing frames as feasible.

### 4. The Phase-Modified Approach

Motivated by the importance of the phase information in the TDE, we developed a phase-modified approach. The proposed algorithm essentially uses the noise phase information of non-speech periods and projects it to subsequent speech periods by adopting the sinusoid stationarity assumption. The estimated speech phase is then used as the signal’s phase in the conventional TDE in order obtain a more accurate estimation. Figure 1 shows an overview of the algorithm.

In practice, estimating the room’s impulse response \(h_i(n)\) is not trivial, especially since no a priori knowledge is available about the source signal \(s(n)\). Moreover, as discussed in section 2, both GCC-PHAT and AED perform well under reverberant and noise-free conditions. Bearing this in mind, we rewrite Eq. (4) taking only the effect of noise into account

\[
x_i(n) = |X'_i(k)|e^{j\angle X'_i(k)} + |W_i(n)|e^{j\angle W_i(n)}
\]

where \(X'_i(k)\) represents the equivalent clean (i.e., noise-free) but reverberant signal that is captured by microphone \(m_i\).

Having calculated the noise phase spectrum by performing STFT analysis during the latest non-speech frame (i.e., reference frame), we use Eq. (14) to project an estimate of the noise phase spectrum \(\angle W_i(k)\) for subsequent frames that contain speech. Depending on the spectral content of the ambient noise, this operation can be performed selectively for a subset of \(\xi\) frequency bins of interest, leaving intact the phase of the remaining bins. The parameter \(\xi\) can be adjusted at each reference frame. Solving Eq. (15) with respect to \(\angle X'_i(k)\) gives the equivalent clean source signal’s phase spectrum for the current speech frame

\[
\angle X'_i(k) = \text{atan2}(\sin \angle X_i(k), \cos \angle X_i(k) - \angle W_i(k))
\]

where the function \(\text{atan2}(y, x)\) expresses the principal value of the arctangent of \(y/x\) in the range \((-\pi, \pi]\). In practice, a smoothed estimate of the noise spectral magnitude is used in Eq. (16), which is calculated during non-speech periods using the single-pole recursion

\[
|W_i^m(k)| = \gamma |W_i^{m-1}(k)| + (1 - \gamma)|W_i^m(k)|
\]

with \(0 < \gamma < 1\) and \(m\) denoting the non-speech frame index.

Finally, we use the estimated via Eq. (16) phase spectra \(\angle X'_i(k)\) and \(\angle X'_j(k)\) to synthesize the signals \(|X_i(k)|e^{j\angle X'_i(k)}\) and \(|X_j(k)|e^{j\angle X'_j(k)}\).
and \(|X_j(k)|e^{jX_j(k)}\) based on which the TDE can be calculated as discussed in section 2.

5. Experiments & Discussion

The phase-modified algorithm was evaluated under noisy and reverberant acoustic conditions. The image method [18] was used to simulate different reverberant conditions. The simulated environment was a rectangular room with dimensions 5m × 4m × 3.8m. The reflection coefficients for all surfaces (walls, ceiling, floor) were uniform and independent of the frequency. An omnidirectional point source was located at (1m, 1.5m, 1.5m) and a microphone pair consisting of omnidirectional point sensors was placed at (3m, 2.4m, 1.5m) and (3.08m, 2.4m, 1.5m). The reflection coefficients were chosen so that different reverberation times (RT60) were achieved, spanning from non-reverberant (0ms) to mild (120ms), moderate (350ms), and high (580ms). Clean speech sentences (20s long, female speaker) were convolved with the computed impulse responses to obtain the reverberant speech. Three typical real-world broadband noise types were used: (i) automobile traffic noise, (ii) office noise primarily caused by an air-conditioner, and (iii) internal robot noise caused by the robot’s instruments (e.g., the processor’s ventilation fan, gyroscopes, etc.). Noise types (i) and (ii) were recorded using a pair of omnidirectional microphones, located 8cm apart. The robot noise was recorded with the commercially available NAO robot, using the left and right embedded microphones. All noise types were properly scaled and added to the reverberant speech, resulting in different SNRs. All signals were sampled at 48KHz.

Processing was performed using 64ms frames with no overlap for computing the TDE and 10% frame shift for the phase projection and the calculation of the equivalent clean source phase spectrum (Eqs. (14) and (16)). The subset of \(\xi\) frequency bins was empirically selected for each noise type. Hann windowing was applied to prevent spectral leakage. No other pre/post-processing was applied. The TDE performance was assessed in terms of hit-rate, defined as the percentage of the accurate TDE (i.e., exact time delay in samples) over the total estimates. The frames containing speech were identified using a simple energy threshold. In order to decouple the performance of the TDE from that of the speech detector, especially in low SNR, the speech detection was performed based on the clean signals.

Figure 2 presents the results of our evaluation for different SNR and RT60. Although consistent results were obtained for all noise types, due to space constrains, only one noise type per reverberant condition is shown. In the non-reverberant case (Figure 2(a)), the performance of all TDE methods is satisfactory. For lower SNR, the GCC-PHAT performance slightly decreases compared to GCC-ML, which is in line with previous observations [14]. When reverberation is considered (Figure 2(b),(c)), the performance of the GCC-ML degrades drastically, as anticipated (see section 2). When the other shift, although the AED performs significantly better than the GCC-PHAT, they both decrease proportionally at lower SNR. When the phase-modified approach is applied, their performance considerably improves for all noise types and the majority of SNRs.

Figure 3 illustrates the TDE performance of all considered noise types under moderate noise (SNR = 15dB) and high reverberation (RT60 = 580ms). An interesting observation is that at high RT60, different TDE performance is observed for each noise type. Nevertheless, our results demonstrate that the phase-modified approach outperforms the conventional GCC-PHAT and AED for all noise types. On the contrary, at higher

\[
\text{Figure 2: TDE performance for different SNR, noise types and reverberation times: (a) traffic noise, non-reverberant conditions; (b) office noise, RT60 = 120ms; (c) robot noise, RT60 = 350ms.}
\]

\[
\text{Figure 3: TDE performance for different noise types in reverberant (RT60 = 580ms) and noisy (SNR = 15dB) conditions.}
\]

\[
\text{RT60 (Figures 2(c) and 3), the proposed approach has no significant effect on the GCC-ML. This is justified by the fact that the effect of reverberation is known to be the major weakness of the ML weighting function, while neither Eq. (7) nor Eq. (15) take this effect into account. Concluding, the observed results suggest that in the context of TDE, the sinusoid stationarity assumption is sufficiently satisfied for typical real-world noise types and confirm that the proposed phase-modified approach outperforms the conventional TDE methods under both reverberant and noisy conditions.}
\]

6. Acknowledgements

Parts of this research were performed in the context of the FP7 project ALIZ-e (ICT-248116).
7. References


