Locating an Acoustic Source Using a Mutual Information Beamformer

Osama N. Alrabadi, Fotios Talantzis, and Anthony G. Constantinides

Abstract Beamforming remains one of the most common methods for estimating the Direction Of Arrival (DOA) of an acoustic source. Beamformers operate using at least two sensors that look among a set of geometrical directions for the one that maximizes received signal power. In this paper we consider a two-sensor beamformer that estimates the DOA of a single source by scanning the broadside for the direction that maximizes the mutual information between the two microphones. This alternative approach exhibits robust behavior even under heavily reverberant conditions where traditional power-based systems fail to distinguish between the true DOA and that of a dominant reflection. Performance is demonstrated for both algorithms with sets of simulations and experiments as a function of different environmental variables. The results indicate that the newly proposed beamforming scheme can accurately estimate the DOA of an acoustic source.

1 Introduction

Locating an acoustic source in a reverberant and noisy enclosure using an array of microphones remains an open problem in a class of different applications. Typical examples include smart environments [1] and security systems [2]. Such systems are typically required to identify the location of the active speech source in physical space, from a short time frame on which speech is considered as stationary (typically 10 to 30 ms). Most solutions to the problem require employment of arrays in the enclosure and the use of an Acoustic Source Localization (ASL) system. ASL is based on the asynchrony between the various microphones and the corresponding cross-correlation between their signals. The various methods are based on two approaches: time delay estimation (TDE) [3], and direct methods with the latter shown to be more robust [4].

Please use the following format when citing this chapter:

Alrabadi, O.N., Talantzis, F. and Constantinides, A.G., 2009, in IFIP International Federation for Information Processing, Volume 296; *Artificial Intelligence Applications and Innovations III*; Eds. Iliadis, L., Vlahavas, I., Bramer, M.; (Boston: Springer), pp. 283–291.

The basic component of direct methods is a beamformer that scans a set of candidate directions for the one that exhibits the maximum power [4]. This process is known as estimation of the Direction Of Arrival (DOA). Tuning the beamformer to scan different directions refers to simply delaying the outputs of its microphones by a different amount and then multiplying each of them by a set of appropriate coefficients. In presence of noise and reverberation though the DOA estimate provided could be spurious due to ensuing reflections and noise. Methods to overcome these effects have been presented [5, 6] but still suffer significantly in heavily reverberant environments.

In the present work we present a new criterion for choosing the direction from which the acoustic source emits. We extend the work that was presented in [1] for TDE and use a two-microphone array to look for the DOA that maximizes the marginal Mutual Information (MI) at the output of the beamformer. Information theory concepts in beamforming have been used before [7] but have no mechanisms to deal with reverberation. The approach presented in this paper involves a framework that takes into account the effects of the spreading the information into samples neighboring to the one that maximizes the MI comparing function. Through experiments and extensive simulations we demonstrate that this novel MI based beamformer resolves to a great degree the reverberation problem and generates robust DOA estimations. To verify our mathematical framework we test and compare it with the traditional power-based for a set of different environmental variables.

The rest of the paper is organized as follows. In Section II we formulate the DOA estimation problem under the beamformer constraint and present the typical power-based method which is used at a later stage for comparison purposes. The MI based alternative is presented in Section III. Section IV examines the performance of the two systems under different criteria such as reverberation level, array geometry and other requirements imposed by real-time systems. Section V discusses briefly the conclusions of this study.

2 System Model

A DOA estimation system is typically employed in a reverberant environment and it considers at least two microphones. The sound source that the system attempts to locate and track is assumed to be in the far field of the microphones. Therefore, we can approximate the spherical wavefront emanating from the source as a plane wavefront of sound waves arriving at the microphone pairs in a parallel manner. Let $\mathbf{r}_m, m = 1,2$ denote the positions of the two microphones with their distance being d meters. The discrete signal recorded at the m^{th} microphone at time k is then:

$$x_m(k) = h_m(k) * s(k) + n_m(k), \tag{1}$$

where $s(k)$ is the source signal, $h_m(k)$ is the room impulse response between the source and m^{th} microphone, $n_m(k)$ is additive white Gaussian noise, and $*$ denotes

convolution. The length of $h_m(k)$, and thus the number of reflections, is a function of the reverberation time T_{60} (defined as the time in seconds for the reverberation level to decay to 60 dB below the initial level) of the room and expresses the main problem when attempting to track an acoustic source. Data for DOA estimation is collected over frames of L samples which for the t^{th} frame we denote as $\mathbf{x}_{tm} = [x_{tm}(0)\ldots x_{tm}(L-1)]$ with $x_{tm}(k) = x_m(L(t-1)+k)$.

Estimating the DOA using a traditional beamformer involves scanning a set of geometrical directions and choosing the one that maximizes the beamformer output power. Typically this is performed in the frequency domain. As in the time-domain, processing is performed in frames with the use of an L-point Short Time Fourier Transform (STFT) over a set of discrete frequencies ω. Thus, the output of the beamformer at frame t and frequency ω is:

$$Y_t(\theta,\omega) = \frac{1}{2}\sum_{m=1}^{2} H_{tm}(\theta,\omega)X_{tm}(\omega) \tag{2}$$

where $X_{tm}(\omega)$ is the ω^{th} element of frame $\mathbf{X}_{tm}$ i.e. the STFT of $\mathbf{x}_{tm}$. $H_{tm}(\theta,\omega)$ is the weight applied to the m^{th} microphone when the beamformer is steered toward direction θ. The beamformer weights are calculated as:

$$H_{tm}(\theta,\omega) = e^{-\frac{j\omega d_m}{c}\sin\theta} \tag{3}$$

where d_m is the Euclidean distance of the m^{th} microphone from the origin. Without loss of generality we can consider $\mathbf{r}_1$ as the origin i.e. $d_1 = 0$ and $d_2 = d$. Thus, in the case of the power-based beamforming the estimated direction $\theta_s^{[P]}$ from which the source emits at frame t can be estimated as:

$$\theta_s^{[P]} = \arg\max_{\theta} |\widehat{Y}_t(\theta)|^2 \tag{4}$$

where $|\widehat{Y}_t(\theta)|^2 = \sum_\omega W(\omega)|Y_t(\theta,\omega)|^2$ is the average beamformer output power over the L discrete frequencies ω. $W(\omega)$ denotes any frequency weighting that is used. In a reverberant environment though, the true source location is not always the global maximum of the power function and thus the above approach often generates wrong estimates.

3 Mutual Information Beamforming

The MI of two variables is an information theoretical measure that represents the difference between the measured joint entropy of the two variables(in our case these are the microphone signals) and their joint entropy if they were independent. Since the analysis will be independent of the data frame we can drop t to express frames simply as $\mathbf{X}_m$ for any t. So for any set of frames, the MI at the output of the beamformer when steered toward an angle θ is [8]:

$$I_N = -\frac{1}{2}\ln\frac{\mathbf{det}[\mathbf{C}(\theta)]}{\mathbf{det}[\mathbf{C}_{11}]\mathbf{det}[\mathbf{C}_{22}]} \tag{5}$$

the joint covariance matrix $\mathbf{C}(\theta)$ is a concatenation of frames $\mathbf{X}_1$ and $\mathbf{X}_2$ shifted by different amounts in samples:

$$\begin{aligned}
&\mathbf{C}(\theta) \approx \\
&\Re\left\{\begin{bmatrix} \mathbf{X}_1 \\ \mathcal{D}(\mathbf{X}_1,1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_1,N) \\ \mathcal{D}(\mathbf{X}_2,\frac{d\sin\theta}{cf_s}) \\ \mathcal{D}(\mathbf{X}_2,\frac{d\sin\theta}{cf_s}+1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_2,\frac{d\sin\theta}{cf_s}+N) \end{bmatrix}\begin{bmatrix} \mathbf{X}_1 \\ \mathcal{D}(\mathbf{X}_1,1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_1,N) \\ \mathcal{D}(\mathbf{X}_2,\frac{d\sin\theta}{cf_s}) \\ \mathcal{D}(\mathbf{X}_2,\frac{d\sin\theta}{cf_s}+1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_2,\frac{d\sin\theta}{cf_s}+N) \end{bmatrix}^H\right\} \\
&= \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12}(\theta) \\ \mathbf{C}_{21}(\theta) & \mathbf{C}_{22} \end{bmatrix}
\end{aligned} \tag{6}$$

where the $\Re\{.\}$ operation returns only the real part of its argument. Function $\mathcal{D}(\mathbf{A},n)$ shifts the frequency components contained in frame $\mathbf{A}$ by n samples. This is typically implemented by using an exponential with an appropriate complex argument.

If N is chosen to be greater than zero the elements of $\mathbf{C}(\theta)$ are themselves matrices. In fact for any value of θ, the size of $\mathbf{C}(\theta)$ is always $2(N+1)\times 2(N+1)$. We call N the *order* of the beamforming system. N is really the parameter that controls the robustness of the beamformer against reverberation. In the above equations and in order to estimate the information between the microphone signals, we actually use the marginal MI that considers jointly N neighboring samples (thus the inclusion of delayed versions of the microphone signals). This way function (5) takes into account the spreading of information due to reverberation and returns more accurate estimates.

The estimated DOA $\theta_s^{[MI]}$ is then obtained as the angle that maximizes (5), i.e.

$$\theta_s^{[MI]} = \arg\max_{\theta}\{I_N\} \tag{7}$$

4 Performance Analysis

In order to demonstrate the improved robustness of the MI based beamformer we conducted DOA estimation experiments and simulations for a single source and a two-microphone system. We used a speech signal of duration 10 sec sampled at

$f_s = 44.1$ kHz which was broken into overlapped frames using a hamming window and an overlap factor of $1/2$. The source was placed at the geometrical angles of $\theta_s = -60^o, -30^o, 0^o, 30^o, 60^o$ (so as to validate the performance under different arrivals), and at a distance $R_o = 2$ m, from the mid-point between the two microphones. The test scenario involves scanning the broadside of the array i.e. from -90^o to $+90^o$ in steps of 3^o and looking for the values that maximize functions (4) and (7). For each frame of data processed, the beamforming systems return a different DOA estimate. The squared error for frame t is then computed as:

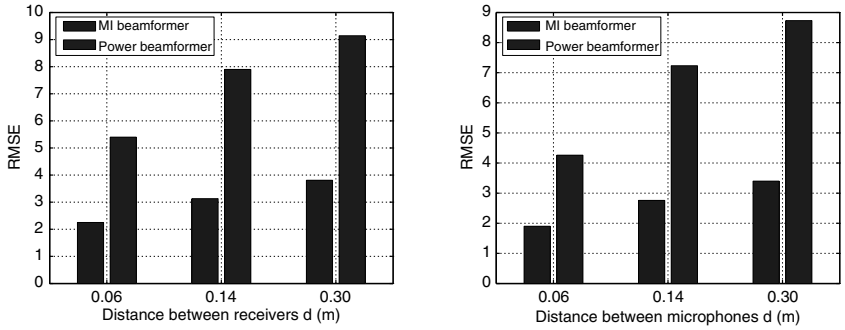
$$\sigma_t = (\theta_s - \widehat{\theta}_t)^2 \tag{8}$$

where θ_s is the actual DOA and $\widehat{\theta}_t$ is either $\theta_s^{[P]}$ or $\theta_s^{[MI]}$, depending on the beamforming system used. The Root Mean Squared Error (RMSE) metric is the performance measure used to evaluate the systems. For a single experiment or simulation this is defined to be the square root of the average value of σ_t over all frames. This is calculated separately for the two beamforming systems. Thus, the lower the average RMSE value, the better the performance of the estimating system.

4.1 Real Experiments

First we look into a set of real experiments performed in a typical reverberant room of size $[5, 3.67, 2.58]$ m equipped with a speaker playing the test signal and a microphone array in which we can change the microphone distances. We repeated the playback of the test signal for 30 random displacements of the overall relative geometry between the source and microphone array inside the room. For each of these displacements we examined the performance of the system for three different inter-microphone distances. The reverberation time of the room was measured to be approximately 0.3 s. In the figures to follow we present the average RMSE over all 30 experiments. It's also worth noting that experiments are conducted in presence of ambient noise from both air-conditioning and personal computers, estimated to be 15 dB.

Fig. 1(a) shows the average RMSE of the beamforming systems for different distances d between the sensors. Effectively, changing the inter-microphone distance changes the resolution of the array. It is evident that the MI based beamformer remains more robust in estimating the correct DOA for all distances. The improvement of performance for both beamforming systems as the spacing decreases can prove misleading since it is caused by the decreased resolution. Safe conclusions were drawn by observing the comparative performance of the two systems for each spacing.



(a) Effect of d during experiments

(b)Effect of d during simulations

Fig. 1 Average RMSE for the two beamforming systems during experiments and simulations. Values are shown for three different inter-microphone distances. $L = 0.5 \times T_{60} f_s$ and $N = 4$.

4.2 Simulations

Simulations where performed for three different environments differentiated by their reverberation times T_{60}. For the used sampling rate f_s these result in impulse responses $h(k)$ of different lengths. The impulse responses are generated using the image model [9] modified to allow for non-integer sample delays. The simulated room dimensions are identical to the ones of the room used in the experiments. These where then convolved with the speech signal to create the microphone signals. Moreover, 15dB of additive noise was also introduced to the signals. The process was repeated for 30 random displacements and rotations of the relative geometry between the source and the receivers inside the room.

4.2.1 Effect of system order

Choosing the order N of the MI beamforming system affects performance significantly. Fig. 2 shows the RMSE for varying N for all three environments. L is chosen to be $0.5 \times T_{60} \times f_s$. Since by increasing N we include more information about reverberation, the MI calculations became more accurate and the estimation of the correct DOA becomes more robust. Thus, the effect of N is more evident for higher reverberation times.

4.2.2 Effect of reverberation

The most limiting factor in designing a robust beamformer is the effect of reverberation. As someone might expect, as the room becomes more reverberant the performance of the estimating systems degrades because reflections enforce the

power or the MI at a wrong DOA. Fig. 3 summarizes the effect for the case when $L = 0.5 \times T_{60} f_s$, $N = 4$. The MI beamformer exhibits a more robust behavior in all environments when compared to the power-based beamformer of the corresponding order.

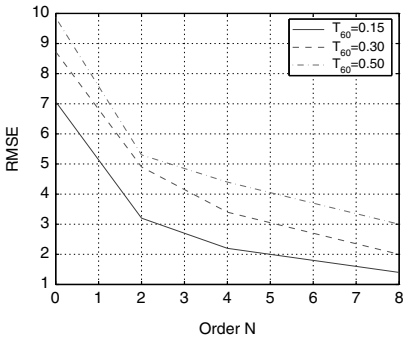


Fig. 2 RMSE of MI system with increasing order N for different values of T_{60}. $L = 0.5 \times T_{60} f_s$. Microphone spacing is 0.30 m.

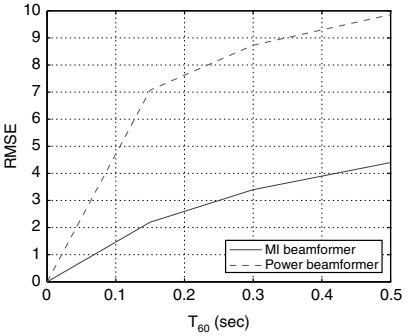


Fig. 3 RMSE of MI and power systems for varying $T60$. $L = 0.5 \times T_{60} f_s$. Shown for microphone spacing of 0.30 m.

4.2.3 Effect of inter-microphone distance

We also investigate the effect of changing the distance between the microphones for $T_{60} = 0.30$ sec, in order to compare the simulation results with those of the experiments in Fig. 1.(a). Fig. 1.(b) shows the resulting RMSE as the distance of the microphones increases. The MI system remains better for any spacing. The values between Fig. 1(b) and Fig.1(a) are not identical but their differences remain small.

These can be explained by noting that the experimental room is far from the idealized version of the simulations. In reality, the experimental environment contains furniture and walls of different texture and materials that a explain to a great degree the differences. Additionally, the image model used in the simulations is subject to a set of assumptions [9].

4.2.4 Effect of frame size

Beamforming systems are normally used in real-time applications so their response time is crucial. In terms of our DOA estimation system this translates into the number of samples L that are needed to produce a robust estimate. Thus, we examine the effect of the value of L by considering a series of different block sizes. To keep these a function of the reverberation level in the room we examine L=[0.25, 0.5, 0.75, 1]$\times T_{60} f_s$ in samples. Fig.4 expresses the effect of L on the performance of the MI beamformer as compared to the classical power-based for $T_{60} = 0.15$ sec and $T_{60} = 0.30$ sec. This shows that, for the chosen parameters, the MI based method is more robust than its counterpart, where $N = 4$ in all cases. In real-time systems where small block sizes are required, the presented system would obviously be preferable since it requires far fewer data to perform satisfactory.

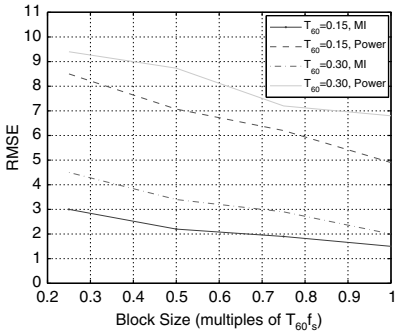


Fig. 4 RMSE of MI and power systems for varying value of L. Shown for $T_{60} = 0.15$ sec and $T_{60} = 0.30$ sec. Microphone spacing is 0.30 m.

5 Conclusions

In this paper a novel beamforming system has been introduced that detects the presence of an acoustic source based on information theory concepts. We demonstrated that such an approach can take into account information about reverberation and thus return DOA estimations those are more robust. This was demonstrated by a set

of experiments and simulations under similar conditions. The MI-based beamformer showed improved robustness for all examined scenarios and for any combination of environmental and system variables like reverberation time, inter-microphone spacing and frame size.

Acknowledgment

This work has been partly sponsored by the European Union, under the FP7 project HERMES.

References

1. A. Pnevmatikakis, F. Talantzis, J. Soldatos, L. Polymenakos, "Robust Multimodal Audio-Visual Processing for Advanced Context Awareness in Smart Spaces", *Springer Journal on Personal and Ubiquitous Computing*, DOI: 10.1007/s00779-007-0169-9, April 2007.
2. Y. Wang, E. Chang, K. Cheng, "A Video Analysis Framework for Soft Biometry Security Surveillance", *Proc. ACM Workshop on Video Surveillance and Sensor Networks*, pp. 71-78, Singapore, 2005.
3. C.H. Knapp, G.C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoust. Speech and Sig. Proc.*, vol. 24, no. 4, pp. 320-327, 1976.
4. E.A. Lehmann and A.M. Johansson, "Particle Filter with Integrated Voice Activity Detection for Acoustic Source Tracking," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 50870, 2007.
5. B. Yoon, I. Tashev, A. Acero. "Robust Adaptive Beamforming Algorithm Using Instantaneous Direction of Arrival with Enhanced Noise Suppression Capability", *Proc. ICASSP 2007*, Honolulu, USA, April 2007.
6. M.S. Brandstein, H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms", *Proc. ICASSP 1997*, pp. 375-378, 1997.
7. L.C. Parra, C.V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech Proc.*, vol. 10, no. 6, pp. 352-362, 2002.
8. T.M. Cover, J.A. Thomas, "Elements Of Information Theory", *Wiley*, 1991.
9. J.B. Allen, D.A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, 1979.