Maximum *a posteriori* Multimodal 3D Object Localization With a Depth Sensor and Stereo Microphones

Bowon Lee
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA, USA
bowon.lee@hp.com

Kar-Han Tan
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA, USA
kar-han.tan@hp.com

ABSTRACT

We propose an algorithm for multimodal object localization with a depth sensor and stereo microphones. For this we formulate a joint probability distribution of object locations conditioned upon depth and acoustic observations. Then we use the maximum *a posteriori* estimation for object localization. For multimodal fusion, we map likelihood of acoustic observation given time difference of arrival information to that given object location in a three dimensional space. Our method offers a principled way to fuse information from microphones and depth sensors, and experimentally we find that it reliably locates the object without requiring careful calibration of the sensors.

1. INTRODUCTION

Multimedia telecommunication systems capable of capturing and rendering audio-visual scenes of local and remote participants have drawn significant attention in recent years. For capturing audio-visual scenes, most existing systems rely on controlled environments, which are expensive to build because it requires acoustic treatment and/or controlled lighting. In uncontrolled environments, the quality of captured audio-visual scenes deteriorates dramatically and hinders the system's ability to support seamless collaboration among remote participants. For effective interactions it is important to capture audio-visual scenes in higher quality and extract useful information such as human faces or gestures. Multimodal sensor arrays consisting of cameras and microphones have been the most popular choices for this purpose [5, 8, 13].

The main advantage and motivation to use multimodality is to achieve better performance than each individual modality. For example, harsh illumination may cause face detection algorithms to fail while it does not affect performance of acoustic source localization. Multimodal object localization with audio-visual fusion is one of the most well-studied topics in the literature [5, 8, 9, 15, 14].

For acoustic modality, audio signals captured at a pair of microphones can be used for finding time difference of arrival through the generalized cross correlation method [11], which can be extended to the steered response power method (SRP) for an array with more than two microphones [6]. One of the most challenging problems with multimodal sensor arrays is to fuse information from heterogeneous sensors [15].

Fusion of the source location found through the acoustic and visual information has been achieved through state-space model [17]. State-space provides means to tie observations from different modalities and suitable for tracking sources because we can dynamically update states from multimodal observations. In particular, Kalman filtering [9, 15] and Particle filtering [17, 8] have been the most popular methods for the state-space approach in the literature.

In this paper, we investigate how observations from a microphone array can be fused with those from a depth camera, an emerging class of sensors. Depth cameras are active devices that measure the time-of-flight (TOF) of infrared light pulses and are able to measure 3D depth in real time. Until recently, TOF depth sensors are large, exotic, and expensive devices but in the past few years they have become cheaper, smaller, and therefore viable in many applications. Researchers have started to study this new class of sensors and it has been shown that fusing depth sensors synergistically with multiple conventional cameras improves the performance of stereo 3D reconstruction algorithms [19].

For multimodal fusion between a depth sensor and stereo microphones, our method finds a joint probability distribution function (PDF) of having an object in a three dimensional space based on acoustic and depth observations. With the proposed probabilistic model we use the maximum *a posteriori* (MAP) estimation method for multimodal object localization.

2. ACOUSTIC OBJECT LOCALIZATION

This section describes a method for estimating time difference of arrival (TDOA) with a pair of microphones. Our approach is to apply generalized cross correlation (GCC) between signals prefiltered via frequency weighting in the Fourier domain. We show that under some conditions and with the appropriate frequency weighting, the GCC method becomes an approximation of the Maximum-Likelihood (ML) TDOA estimation.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6235
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6235

2.1 Time Difference of Arrival Estimation

If we consider two microphones capturing sound from an object, signal at each microphone during an observation interval $t \in [(n-1)T, nT]$, for $n = 1, 2, \cdots$ can be modeled as

$$\begin{aligned} x_1^n(t) &= a_1^n(t) * s(t) + v_1(t) \\ x_2^n(t) &= a_2^n(t) * s(t) + v_2(t), \end{aligned} \tag{1}$$

where $a_1^n(t)$ and $a_2^n(t)$ denote the impulse response from the object to each microphone, $s(t)$ is the signal from the object, $v_1(t)$ and $v_2(t)$ denote noise signals at each microphone, $*$ denotes the convolution operator, and superscript ^{n} denotes frame number. For mathematical formulation, we assume $x_1^n(t)$ and $x_2^n(t)$ are quasi-stationary, i.e., within a frame with interval T, the impulse responses are deterministic and both signal and noise are stationary random processes each with zero mean.

In multipath environments where there exist reflections against walls and other objects, we can consider the impulse response as a direct path followed by a series of reflections, which are treated as components in noise. Then Eq. (1) becomes

$$\begin{aligned} x_1^n(t) &= \alpha_1^n s(t - \tau_1^n) + v_1(t) \\ x_2^n(t) &= \alpha_2^n s(t - \tau_2^n) + v_2(t), \end{aligned} \tag{2}$$

where τ_1^n and τ_2^n represent the propagation delays of the direct paths and α_1^n and α_2^n are signal attenuation due to propagation, all for the nth frame. With this model, TDOA estimation is a problem to find $\tau^n = \tau_2^n - \tau_1^n$, a relative propagation delay between two microphones.

2.2 Generalized Cross Correlation Method

Generalized cross correlation (GCC) method computes the cross correlation of prefiltered signals $y_1(t) = h_1(t) * x_1(t)$ and $y_2(t) = h_2(t) * x_2(t)$ and find the time delay which maximizes the cross correlation $R_{y_1y_2}(\tau) = E[y_1(t)y_2(t+\tau)]$

$$\hat{\tau} = \arg\max_{\tau} R_{y_1y_2}(\tau). \tag{3}$$

For simplicity, superscript ^{n} has been omitted for the remainder of this paper.

The cross power spectral density at the kth frequency bin, in terms of the N-point discrete Fourier transform (DFT) of the cross correlation can be expressed as

$$\phi_{y_1y_2}[k] = \psi[k]\phi_{x_1x_2}[k], \tag{4}$$

where $\psi[k] = H_1[k]H_2^*[k]$ is referred to as a generalized prefilter [11]. The purpose of using $\psi[k]$ is to make the cross correlation have a distinctive peak at the true time delay when noise and reflections are present. We typically estimate the cross power spectral density using the periodogram, i.e., $\phi_{x_1x_2}[k] = \frac{1}{N}X_1[k]X_2^*[k]$. Then Eq. (3) can be expressed in terms of the inverse DFT (IDFT) of Eq. (4)

$$\hat{\Delta} = \arg\max_{\Delta} \frac{1}{N} \sum_{k=0}^{N-1} \psi[k]X_1[k]X_2^*[k]e^{j\frac{2\pi k}{N}\Delta}, \tag{5}$$

where $\Delta = f_s\tau$ is a TDOA in the discrete time domain with f_s denoting the sampling frequency. Various frequency weightings are well summarized in [11]. Among those, the phase transform (PHAT) frequency weighting

$$\psi_{PHAT}[k] = \frac{1}{|X_1[k]X_2^*[k]|} \tag{6}$$

has been the most popular choice among the prefilters due to its robustness in reverberant environments [4].

2.3 Maximum-Likelihood TDOA Estimation

Knapp and Carter [11] showed that the ML-TDOA can be considered as a GCC method by assuming that the source and noise are uncorrelated random processes with Gaussian distribution in the time domain. Using the central limit theorem [16], we can relax the Gaussian distribution assumption to any probability distribution in the time domain [7].

With those assumptions, we can formulate the joint probability distribution function (PDF) of the DFT coefficients in the kth frequency bin as [11]

$$p(X_1[k], X_2[k]\big|\Delta) = \frac{1}{\pi^2|\mathbf{Q}_k|} e^{-\mathbf{X}[k]^H\mathbf{Q}_k^{-1}\mathbf{X}[k]}, \tag{7}$$

where $\mathbf{X}[k] = [X_1[k]X_2[k]]^T$ and $\mathbf{Q}_k$ is a covariance matrix of signals $X_1[k]$ and $X_2[k]$ defined as

$$\begin{aligned} \mathbf{Q}_k &= E[\mathbf{X}[k]\mathbf{X}[k]^H] \\ &= \begin{bmatrix} E[X_1[k]X_1[k]^*] & E[X_1[k]X_2[k]^*] \\ E[X_2[k]X_1[k]^*] & E[X_2[k]X_2[k]^*] \end{bmatrix} \\ &= N \begin{bmatrix} \phi_{x_1x_1}[k] & \phi_{x_1x_2}[k] \\ \phi_{x_1x_2}^*[k] & \phi_{x_2x_2}[k] \end{bmatrix} \\ &= N \begin{bmatrix} \phi_{ss}[k] + \phi_{v_1v_1}[k] & \phi_{ss}[k]e^{-j\frac{2\pi k}{N}\Delta} \\ \phi_{ss}[k]e^{j\frac{2\pi k}{N}\Delta} & \phi_{ss}[k] + \phi_{v_2v_2}[k] \end{bmatrix}, \end{aligned} \tag{8}$$

where $\phi_{ss}[k]$, $\phi_{v_1v_1}[k]$, and $\phi_{v_2v_2}[k]$ denote power spectral densities of source and noise at each microphone and superscripts ^{T} and ^{H} denote transpose and complex conjugate transpose respectively. Note that the last equality is based on the assumption that $V_1[k]$ and $V_2[k]$ are not correlated with each other nor with the source signal $S[k]$, and that attenuation due to propagation is negligible.

Provided that we know the covariance matrix $\mathbf{Q}_k$, which is a function of the time delay Δ and cross spectral densities of signal and noise according to Eq. (8), the ML estimation $\hat{\Delta}_{ML}$ of the time delay is

$$\hat{\Delta}_{ML} = \arg\max_{\Delta} \prod_{k=0}^{N-1} p(X_1[k], X_2[k]\big|\Delta) \tag{9}$$

and it has been shown that [11]

$$\hat{\Delta}_{ML} = \arg\max_{\Delta} \frac{1}{N} \sum_{k=0}^{N-1} G_{ML}[k]e^{j\frac{2\pi k}{N}\Delta}, \tag{10}$$

where

$$G_{ML}[k] = \frac{|\phi_{x_1x_2}[k]|}{\phi_{x_1x_1}[k]\phi_{x_2x_2}[k] - |\phi_{x_1x_2}[k]|^2} X_1[k]X_2^*[k]. \tag{11}$$

According to Eq. (5), we can consider

$$\psi_{\mathrm{ML}}[k] = \frac{|\phi_{x_1x_2}[k]|}{\phi_{x_1x_1}[k]\phi_{x_2x_2}[k] - |\phi_{x_1x_2}[k]|^2} \tag{12}$$

as a ML prefilter in the GCC framework [11].

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6235
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6235

Even though the ML-TDOA estimate in Eq. (12) is optimal and achieves the Cramér-Rao lower bound [11], its optimality is dependent upon the availability of cross spectra, which can only be estimated for quasi-stationary processes. In practice their inaccurate estimates degrade the accuracy of the TDOA. In the following section, alternative frequency weightings for ML-TDOA will be presented.

2.4 ML-TDOA Based on Phase Variance Estimation

Knapp and Carter [11] also showed that the ML frequency weighting in Eq. (12) can be expressed as a function of variance of the cross-spectrum phase, $\mathrm{var}[\theta_k]$ [10, p. 379]

$$\psi_{\mathrm{ML}}[k] \approx \frac{1}{|X_1[k]X_2^*[k]|\mathrm{var}[\theta_k]}, \tag{13}$$

where $\theta_k = \angle X_1[k]X_2^*[k]$. Note that a frequency component with $\mathrm{var}[\theta_k] = 0$ allows for a perfect TDOA, in correspondence with the infinite weight as given by Eq. (13).

Based on this, Brandstein *et al.* [3] proposed an Approximated ML (AML) frequency weighting by estimating $\mathrm{var}[\theta_k]$ by assuming that at each microphone the phase variance is inversely proportional to the *a posteriori* SNR $|X_l[k]|^2/|V_l[k]|^2$, for $l = 1, 2$ and that $\mathrm{var}[\theta_k]$ is a sum of independently estimated phase variances

$$\psi_{\mathrm{AML}}[k] = \frac{|X_1[k]||X_2[k]|}{|V_1[k]|^2|X_2[k]|^2 + |V_2[k]|^2|X_1[k]|^2}, \tag{14}$$

which is shown to be more robust than the original ML weighting [2] and outperforms the PHAT weighting at low SNR [3]. Equation (14) is often referred to as the ML weighting instead of the original ML weighting of Eq. (12) in the literature [18]. Note that in order to apply the AML weighting we still need to have the noise spectra available. This knowledge is seldom available in practice, and is one of the main reasons that the PHAT method, requiring no such knowledge, is popular.

Recently, Lee et al. [12] proposed a method for estimating the ML frequency weighting in Eq. (13) by blindly estimating the phase variance. In particular, based on the complex Gaussian model in Eq. (7), they showed that

$$\mathrm{var}[\theta_k] \approx \sqrt{\log|\bar{\Sigma}_k|^{-2}}, \tag{15}$$

where $\bar{\Sigma}_k$ is the mean of the observed complex phase $e^{j\theta_k}$, which gives the following ML frequency weighting

$$\psi_{\mathrm{PML}}[k] = \frac{1}{|X_1[k]X_2^*[k]|\sqrt{\log|\bar{\Sigma}_k|^{-2}}}. \tag{16}$$

This frequency weighting has been shown to be more robust than PHAT weighting in Eq. (6) without requiring knowledge of signal and noise statistics [12].

3. MULTIMODAL FUSION AND OBJECT LOCALIZATION

In Section 2, we described a probabilistic model for acoustic object localization via ML-TDOA and tracking via MAP-TDOA methods. In this section we present a method to fuse the acoustic modality to the depth information in order to find the posterior probability of object location. For simplicity we assume that the interval T for the audio frame is equivalent to the duration of each depth information corresponding to the refresh rate of the depth sensor. Also we consider that we know the locations of the depth sensor and microphones.

3.1 Multimodal MAP Localization

For multimodal object localization, our purpose is to find the object location $\mathbf{L} = (i, j, k)$ in a three-dimensional space. We denote a set of acoustic observations as

$$\mathcal{S}_A = \left\{ X_1[k], X_2[k] \middle| k = 1, 2, \cdots, N \right\}$$

consisting of $2N$ complex variables, and a set of depth observations as

$$\mathcal{S}_D = \left\{ \mathbf{l}_p \middle| p = 1, 2, \cdots, P \right\},$$

where $\mathbf{l}_p = (i_p, j_p, k_p, I_p)$ is the depth sensor reading at the p^{th} pixel with i_p, j_p, k_p denoting the object location and I_p denoting the corresponding signal intensity where P denotes the total number of pixels of the depth sensor.

Given these observations, we formulate the MAP estimation of the object location as

$$\begin{aligned} \hat{\mathbf{L}}_{MAP} &= \arg\max_{\mathbf{L}} p(\mathbf{L} \big| \mathcal{S}_A, \mathcal{S}_D) \\ &= \arg\max_{i,j,k} p(i, j, k \big| \mathcal{S}_A, \mathcal{S}_D) \end{aligned} \tag{17}$$

We assume that acoustic and depth observations are independent with each other and use the Bayes' rule to rewrite Eq. (17) as

$$\begin{aligned} \hat{\mathbf{L}}_{MAP} &= \arg\max_{i,j,k} p(\mathcal{S}_A, \mathcal{S}_D \big| i, j, k) p(i, j, k) \\ &= \arg\max_{i,j,k} p(\mathcal{S}_A \big| i, j, k) p(\mathcal{S}_D \big| i, j, k) p(i, j, k) \\ &= \arg\max_{i,j,k} p(\mathcal{S}_A \big| i, j, k) p(i, j, k \big| \mathcal{S}_D) \end{aligned} \tag{18}$$

We see that Eq. (18) consists of two components, $p(\mathcal{S}_A \big| i, j, k)$ a likelihood of acoustic observation for object located at (i, j, k) and $p(i, j, k \big| \mathcal{S}_D)$ a posterior PDF of object location given depth information. They will be described in more detail in the following sections.

3.2 Likelihood of Acoustic Observation Given Object Location

In Section 2, we described the likelihood of a set of acoustic observations $\mathcal{S}_A$ conditioned upon $\Delta = f_s\tau$. In order to solve Eq. (18), we need to find a likelihood conditioned upon (i, j, k) instead of Δ.

We can show that any points on a surface of a hyperboloid can be candidates of any given TDOA. In other words, for microphones positioned along the i axis with their center located at $i = 0$, any points satisfying the following condition share the same Δ

$$\frac{i^2}{b^2} - \frac{j^2}{a^2 - b^2} - \frac{k^2}{a^2 - b^2} = 1, \tag{19}$$

where $b = c\Delta/2f_s$, c is propagation speed of acoustic wavefronts, and a is a half of the distance between two microphones. In other words, any given Δ corresponds to a hyperboloid in 3D space. An example can be seen in Fig. 1 (c).

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6235
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6235

Since we know the likelihood of observing $\mathcal{S}_A$ given Δ from Eq. (9) we can compute likelihoods of all possible object locations by finding corresponding Δ using Eq. (19).

3.3 Posterior PDF Given Depth Information

The posterior PDF $p(i,j,k|\mathbf{l}_p)$ given the depth sensor reading from the p^{th} pixel is modeled as a Gaussian PDF with an assumption that i, j, and k are independent. Then we find

$$p(i,j,k|\mathbf{l}_p) = \frac{1}{\left(\sqrt{2\pi}\sigma_p\right)^3} \times \exp\left\{-\frac{(i-i_p)^2+(j-j_p)^2+(k-k_p)^2}{2(\sigma_p)^2}\right\}, \tag{20}$$

where $(\sigma_p)^2$ is the variance modeled as inversely proportional to the signal strength I_p. Once we find the PDF given the each pixel reading of the depth sensor with Eq. (20), we model the entire posterior PDF $p(i,j,k|\mathcal{S}_D)$ as a Gaussian Mixture Model (GMM) with equal weight for each mixture, i.e.,

$$p(i,j,k|\mathcal{S}_D) = \frac{1}{P}\sum_{p=1}^{P} p(i,j,k|\mathbf{l}_p). \tag{21}$$

4. EXPERIMENTS

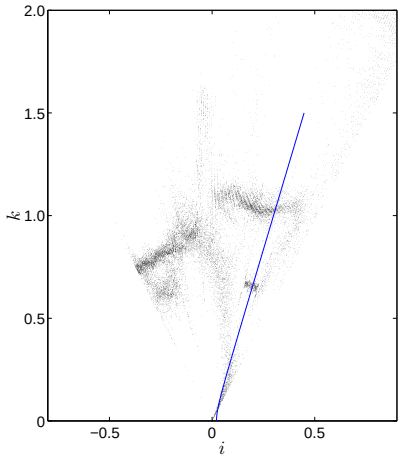
In order to validate our method, we used a TOF sensor from Canesta [1] and a pair of omnidirectional microphones. We placed the depth sensor between the two microphones for them to share the same origin along the i axis. Microphone spacing has been chosen to be 15 centimeters and sampling rate of 48 kHz was used for experiments. We recorded the audio data and depth data simultaneously. The raw depth sensor output $\mathcal{S}_D = \{i,j,k,I\}$ is a set of (i,j,k) points with corresponding signal intensity I in 3D.

Fig. 1(b) shows the raw depth sensor output projected onto the $i-k$ plane, where the i-axis passes through the two microphones and the k-axis is parallel to the depth sensor's optical axis. For synchronization purposes we also record a third audio channel which has short acoustic pulses generated as each depth sensor frame is captured. We found that the sensor setup did not need to be precisely calibrated. We confirmed this in Fig. 1(b), where the depth sensor output is plotted along with the hyperboloid corresponding to the maximum-likelihood TDOA (shown as a blue curve). Clearly, the hyperboloid passes through the location of the portable white noise generator, confirming that the multimodal sensor data are well-aligned. Figure 1(c) shows a slice of the 3D joint PDF along the maximum-likelihood TDOA hyperboloid. Figure 2 shows the 3D PDFs of the depth sensor, microphones, and the joint multimodal PDF. By fusing information from the two different sensor modalities, the 3D localization solution can be found.

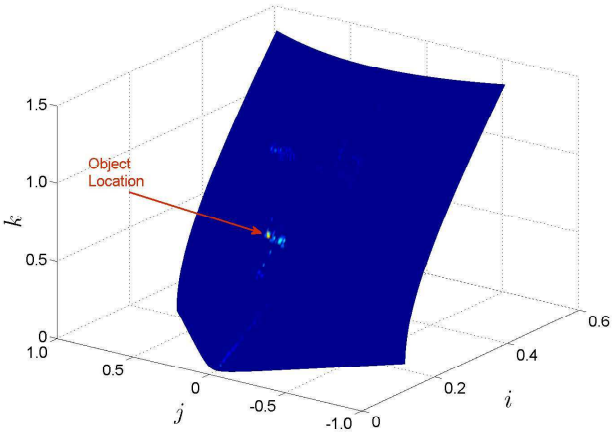
5. CONCLUSION AND FUTURE WORK

We have presented a method for object localization using a pair of microphones and a TOF depth sensor. Our method offers a principled method for fusing multimodal data and we have experimentally shown that it is able to located objects reliably. We also found that the method works without

(a)



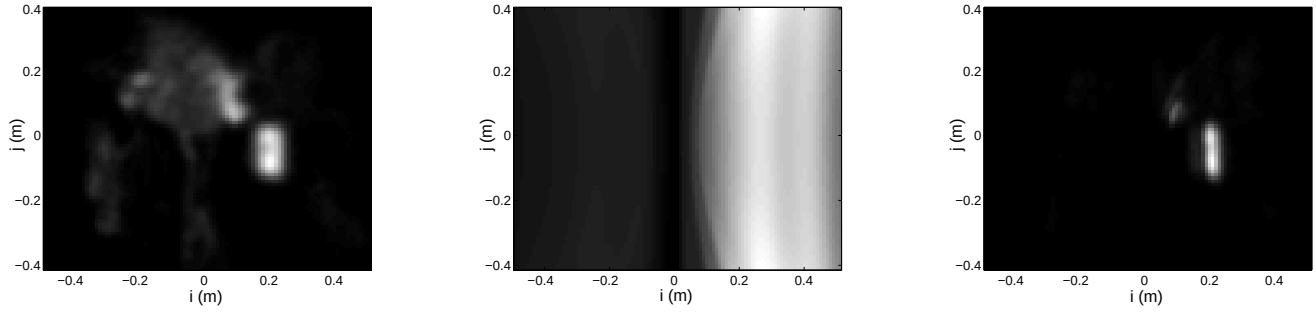
(b)



(c)

Figure 1: Multimodal object localization result. (a) RGB image of the scene. (b) Localization: the hyperboloid corresponding to the maximum-likelihood TDOA is shown as a blue curve. Clearly the curve passes through the location of the noise generator. (c) A slice of the 3D joint PDF along the maximum-likelihood TDOA hyperboloid. The maximum *a posteriori* localization solution is indicated with an arrow.

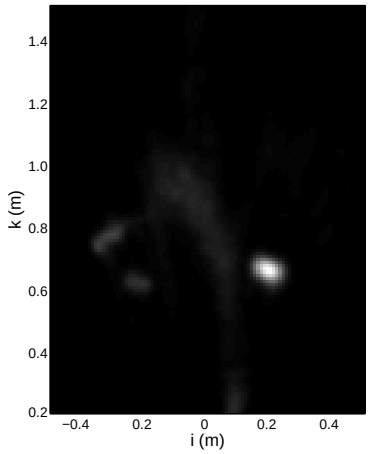
Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6235
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6235



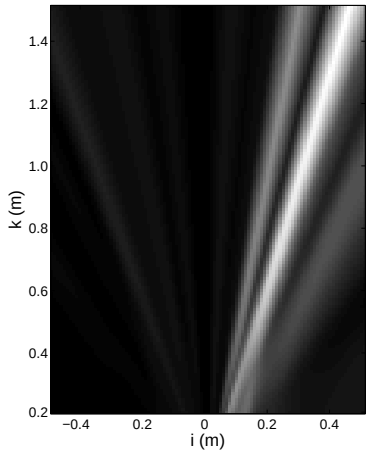
(a) Depth PDF Front View

(c) Acoustic PDF Front View

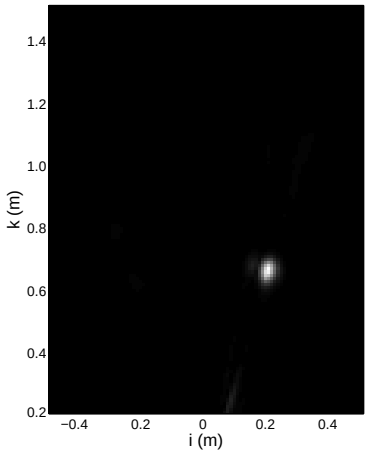
(e) Multimodal PDF Front View



(b) Depth PDF Top View



(d) Acoustic PDF Top View



(f) Multimodal PDF Top View

Figure 2: Multimodal joint 3D PDF in the spatial domain. (a), (b) Front and top view of the 3D PDF from the depth sensor. (c), (d) Front and top views of the 3D PDF from the pair of microphones. (e), (f) Joint 3D PDF. Frontal views correspond to the view shown in Fig. 1(a) and top views correspond to the view shown in Fig. 1(b). Clearly object localization is much easier in the joint PDF than the separate depth and acoustic PDFs which are fairly noisy.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6235
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6235

requiring careful calibration of the sensor setup. We believe the mathematical framework we developed can be naturally extended to solve the tracking problem, and in future work, we would like to create a method for real-time object tracking.

6. REFERENCES

[1] Canestavision^{TM} electronic perception development kit, canesta inc. http://www.canesta.com/html/development_kits.htm.

[2] M. Brandstein. Time-delay estimation of reverberated speech exploiting harmonic structure. *J. Acoust. Soc. Am.*, 105(5):2914–2919, 1999.

[3] M. Brandstein, J. Adcock, and H. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer Speech and Language*, 9:153–169, 1995.

[4] M. S. Brandstein and D. B. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, Berlin, Germany, 2001.

[5] C. Busso, S. Hernanz, C.-W. Chu, S. Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan. Smart room: participant and speaker localization and identification. In *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, volume 2, pages 1117–1120, 2005.

[6] J. DiBiase. *A high-accuracy, low-latency technique for talker localization in reverberant environments*. PhD thesis, Brown University, Providence, RI, May 2000.

[7] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, and Signal Process.*, 32(6):1109–1121, December 1984.

[8] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(2):601–616, Feb. 2007.

[9] T. Gehrig, K. Nickel, H. Ekenel, U. Klee, and J. McDonough. Kalman filters for audio-video source localization. pages 118–121, Oct. 2005.

[10] G. M. Jenkins and D. G. Watts. *Spectral Analysis and Its Applications*. Holden-Day, San Francisco, CA, 1968.

[11] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time-delay. *IEEE Trans. Acoust., Speech and Audio Process.*, ASSP-24(4):320–327, 1976.

[12] B. Lee, A. Said, T. Kalker, and R. W. Schafer. Maximum likelihood time delay estimation with phase domain analysis in the generalized cross correlation framework. In *Proc. HSCMA*, pages 89–92, 2008.

[13] H. Maganti, D. Gatica-Perez, and I. McCowan. Speech enhancement and recognition in meetings with an audio-visual sensor array. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2257–2269, Nov. 2007.

[14] A. O'Donovan, R. Duraiswami, and J. Neumann. Microphone arrays as generalized cameras for integrated audio visual processing. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[15] N. Strobel, S. Spors, and R. Rabenstein. Joint audio-video object localization and tracking. *Signal Processing Magazine, IEEE*, 18(1):22–31, Jan 2001.

[16] H. L. V. Trees. *Detection, Estimation and Modulation Theory, Part I*. Wiley, New York, 1968.

[17] J. Vermaak, M. Gangnet, A. Blake, and P. Perez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proc. Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 741–746, 2001.

[18] C. Zhang, Z. Zhang, and D. Florêncio. Maximum likelihood sound source localization for multiple directional microphones. In *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, volume I, pages 125–128, 2007.

[19] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

Digital Object Identifier: 10.4108/ICST.IMMERSCOM2009.6235
http://dx.doi.org/10.4108/ICST.IMMERSCOM2009.6235