HARDWARE AND ALGORITHMS FOR ULTRASONIC DEPTH IMAGING

Ivan Dokmanić[*]

School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015, Switzerland

Ivan Tashev

Microsoft Research
One Microsoft Way
Redmond, WA, USA

ABSTRACT

Depth imaging is commonly based on light. For example, LIDAR and Kinect use infrared light, while stereo cameras use visible light. These systems require hardware operating at high sampling frequencies, precise calibration, and they dissipate significant power. In this paper, we investigate the potential of ultrasound for image and depth acquisition, with applications to human-computer interaction and skeletal tracking in mind. We use a loudspeaker array and a microphone array to sense the scene. We discuss a technique for offline loudspeaker beamforming (commonly used for microphone beamforming) which enables us to significantly increase the frame rate. Further, we propose a sound-source-localization-based method for computing the depth image, giving a substantial improvement over the naïve time-of-flight approach. We designed inexpensive hardware with eight elements per array to obtain both the depth and the intensity images. Even with this limited number of transducers we obtain promising experimental results.

***Index Terms*—** Ultrasound, depth imaging, beamforming, sound source localization, skeletal tracking, array processing

1. INTRODUCTION

Domains of application of depth imaging include automotive industry, surveying, computer vision, robotics, and lately with the advent of Kinect, human-computer interaction (HCI).

Typically, light is used for depth imaging—either visible as in stereo cameras, or infrared as in Kinect. On the other hand, depth imaging by sonic or ultrasonic means is attractive because of the relatively low power consumption, and simpler, low-rate hardware. Additionally, it could complement light in scenarios where light fails, for example mirrors, windows and glass walls, imaging through thin fabric, or spaces filled with smoke.

Two major issues are related to ultrasonic imaging in air. The first one is that the sound reflects mostly specularly from typical surfaces. It is not reasonable to aim to detect the times of arrivals of diffuse reflections, especially in noise, or when strong specular reflections are present. But for some applications, specular reflections alone may suffice if the objects have uneven surfaces, as the surface details may give rise to usable echoes. Fortunately, this is the case with the human figure. The second issue is the comparatively low speed of sound in air—around 340 m/s. Naïve scanning results in unacceptably low frame rates. For example, to acquire 900 depth pixels with the maximum range of 4 meters, raster scanning requires at least 20 seconds for a single frame. This issue is commonly addressed by employing only a single transmit element, thus avoiding the creation of physical beams in transmission (microphone beams are created offline in all cases).

1.1. Prior Art

A common technique in computer vision is to estimate depth from two or more offset photos of the same scene [1] (stereo or multi-view matching). Another group of methods are time-of-flight and phase difference methods, most often using light. These methods offer many advantages over stereo and multi-view matching, but require specialized, relatively expensive and power hungry equipment [2, 3].

A particularly interesting approach is proposed by Kirmani et al [4]. They use a single omnidirectional light source, and a single photo diode, and rely on temporal information to obtain the depth image. The scene is sampled by a small number of patterned light emissions: The light is patterned by passing through random binary masks. A central assumption in their approach is that the scene is composed of planar surfaces. Their results are impressive, but the assumption is too strict for general purposes. Also, we are constrained to very simple processing in order to achieve reasonable sampling rates on embedded hardware.

Moebus and Zoubir [5] studied ultrasound imaging in air, and discussed its suitability for biometric applications [6]. Their system is based on beamforming with a synthetic 2D array of 400 acoustic receivers. In [7], the authors demonstrate an acoustic range imaging system based on audible sound. Sound is used to develop an obstacle detecting aid for visually impaired people in [8]. The system described by the authors operates at 18.4 kHz, and uses beamforming with an array of 64 microphones. In [9], the authors also propose a mobility aid for the blind based on ultrasound. They use a system with six transmitting and four receiving elements, organized in linear arrays for imaging in the horizontal plane.

1.2. Main Contributions

In the present paper, we describe algorithms and hardware for ultrasonic imaging in air. We propose to use a method that enables fast frame acquisition while still creating the loudspeaker beams. Furthermore, we propose an algorithm based on sound source localization (SSL) that addresses imperfect beamforming and the effect of strong specular reflectors. We designed a simple ultrasonic device (Fig. 1) with eight microphones and eight piezo transducers, for acquiring images in both azimuth and elevation. The number of array elements that we use is considerably smaller than in other devices reported in the literature. Using the proposed algorithms and the designed device, we obtain promising experimental results, especially considering the small number of elements in the arrays.

[*]This investigation was carried out during his internship at Microsoft Research, Redmond.

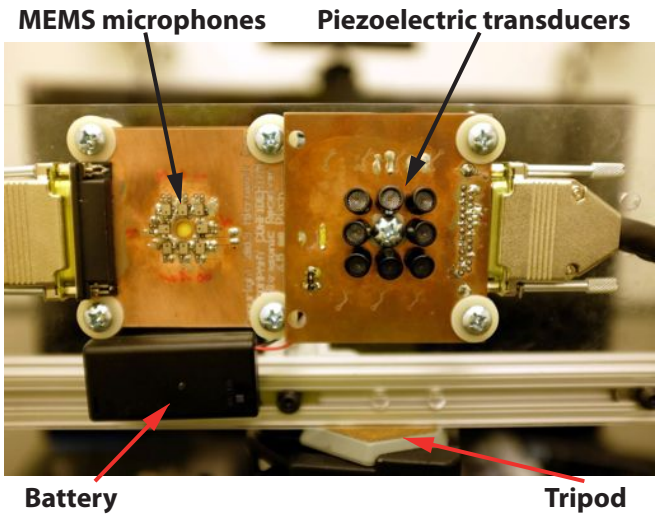

Fig. 1. Hardware design. The ultrasonic camera consists of two arrays: an array of piezoelectric transducers (source), and an array of MEMS microphones.

2. ARRAY GEOMETRIES AND BEAMFORMING

Given a set of M transducers at positions $\{\mathbf{p}_m\}_{m=1}^M$, $\mathbf{p}_m \in \mathbb{R}^3$, we can derive the weights of a beamformer optimal in the noise suppression sense,

$$\mathbf{W}_{\Omega_c}(f) = \frac{\mathbf{\Phi}_{NN}^{-1}(f)\mathbf{D}_{\Omega_c}(f)}{\mathbf{D}_{\Omega_c}^*(f)\mathbf{\Phi}_{NN}^{-1}(f)\mathbf{D}_{\Omega_c}(f)}. \quad (1)$$

This is the well-known minimum-variance-distortionless-response (MVDR) beamformer [10]. In (1), $\Omega_c = (\theta_c, \phi_c)$ indexes the elevation and the azimuth of the look-up point, $\mathbf{D}_{\Omega_c}(f)$ is the vector of transfer functions from the look-up point to the microphones, and $\mathbf{\Phi}_{NN}(f)$ is the noise cross-power spectrum, typically diagonally loaded with the microphone self-noise and the representation of the manufacturing tolerances (Chapter 5, [11]). We assume the far-field regime, so that the look-up points lie on sufficiently large sphere. The array's output is computed as $R_{\Omega_c}(f) = \mathbf{W}_{\Omega_c}^*(f)\mathbf{X}(f)$. This is equivalent to the output of a single microphone with the directivity pattern $B(f, \Omega) = \mathbf{W}_{\Omega_c}^*(f)\mathbf{D}_{\Omega}(f)$, located at the center of the array. We define the directivity index for the direction Ω_T as [11]

$$\mathsf{di}(f) \stackrel{\text{def}}{=} \frac{|B(f, \Omega_T)|^2}{\frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} |B(f,\Omega)|^2 \, \mathrm{d}\Omega}. \quad (2)$$

Usually we express it in dB, $\mathsf{DI}(f) \stackrel{\text{def}}{=} 10\log_{10}[\mathsf{di}(f)]$. We can get the total directivity index that summarizes the directivity over all frequencies of interest as

$$\mathsf{DI}_{tot} = 10\log_{10}\int_{f_1}^{f_2} \mathsf{di}(f)\, \mathrm{d}f. \quad (3)$$

In the *scanning* mode, we emit pulses towards each angular "pixel", and measure the time it takes for an echo from that direction to arrive. To maximize the resolution, and reduce the effects of finite beam width, we want the beam to be as narrow as possible, at all the frequencies of interest. The beam *narrowness* is quantified through the directivity index. We varied $\{\mathbf{p}_m\}_{m=1}^M$ to get the narrowest beam in the sense of DI_{tot}. The optimization is constrained by the number of transducers, by the realistic design constraints (transducer size) and by the common sense (symmetric geometries). Searching over several parametric geometry classes (cross, circle, square, double square), and varying the distance between the microphones, allowed us to find realizable geometries with good beam directivity.

For the microphone array, the square geometry as shown in Fig. 1, with the spacing of 6.5 mm between the microphones yields the highest DI_{tot}. For the piezoelectric transducer array, we used the smallest mechanically achievable spacing. Since the transducers themselves are directive, we further optimized the tilt from the main axis, in order to achieve uniform beam profile over the target range of angles. We found the optimal tilt to be 20° with respect to the main response axis. Figures showing the DI_{tot} for different geometries are omitted in the interest of space.

3. BASIC IMAGING AND SINGLE-SHOT ACQUISITION

Similarly to previous approaches, we follow the time-of-flight paradigm. That is, we want to direct beams of sound (by emitting sound directionally, and by listening to the echoes directionally) towards every *angular pixel*, and to measure the time it takes for the sound to travel to a reflector and back. Previous approaches addressing this problem often use only a single transmit element to avoid the transmit beamforming in the interest of frame rate. Clearly, the receive beamforming can be performed offline. We observe that the same goes for transmit beamforming, thus eliminating the need to raster scan the scene, even when using multiple sources of ultrasound.

This elementary observation seems to be overlooked in the literature. The reasons are probably twofold. First, in audio, loudspeaker arrays are meant to be listened to by a person. This means that necessarily the array must be used at once. Second, in LIDAR or Kinect depth imaging, the laser is itself highly directional—there is no question of transmit beamforming, be it online or offline. We do note that similar techniques were used with ultrasound for medical and non-destructive evaluation purposes [12], but not in air.

Denote the signal emitted by the ith transmitter by $s_i(t)$, the signal received by jth receiver by $r_j(t)$, and their Fourier transforms by $S_i(f)$ and $R_j(f)$. Let the total number of microphones be M, and the number of sources K. Signals emitted by the transmitters are all filtered versions of the same template pulse, $u(t)$, $s_i(t) = [w_i * u](t)$, where $w_i(t)$ is the impulse response of the beamforming filter corresponding to the ith transducer. For the delay-and-sum beamformer, the filters $w_i(t)$ are simply delays. Additionally, they may include the calibration filters that compensate for non-ideal source characteristics.

If h_{ij} is the impulse response of the acoustic channel between the ith source and the jth microphone, then the signal picked up by the jth microphone can be expressed as

$$r_j(t) = \sum_{i=1}^{K}[h_{ij} * s_i](t) = \sum_{i=1}^{K}[h_{ij} * w_i * u](t), \quad (4)$$

or in the frequency domain,

$$R_j(f) = \sum_{i=1}^{K} H_{ij}(f)W_i(f)U(f). \quad (5)$$

Denote further by $r_j^i(t)$ the signal recorded by jth microphone, if all the sources except the ith one remain silent, and the ith source emits $u(t)$ without passing it through the beamforming filter. Concretely,

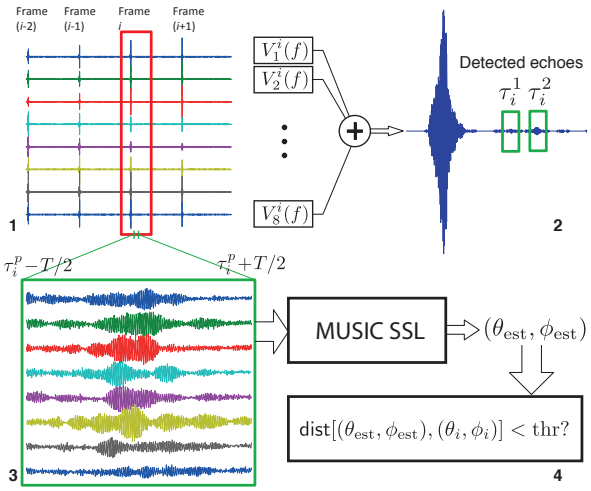

Fig. 2. Illustration of the SSL enhanced imaging process.

$r_j^i(t) = [h_{ij} * u](t)$. Then, we can rewrite (5) as

$$R_j(f) = \sum_{i=1}^{K} W_i(f) R_j^i(f). \tag{6}$$

But this means that if we know $R_j^i(f)$, we can do the transmit beamforming computationally at the receive end, without raster scanning. Conveniently, in order to obtain the set of KM responses $\left\{R_j^i, j \in [\![1, M]\!], i \in [\![1, K]\!]\right\}$ we need to emit only K pulses, where K is the number of transducers. To compare with the naïve approach, for a resolution of 30×30, we need 900 beams in transmission. With 8 transducer elements, by exploiting the properties of linear, time-invariant (LTI) systems as described above, we need to emit only 8 pulses. We obtain a reduction of several orders of magnitude in the time needed to acquire one frame, while still doing the transmit beamforming. Importantly, the scanning time does not scale with the resolution—whatever the target resolution is, we always need to emit only K pulses per frame to reconstruct all the transmit beams. Thus, we get the benefit of having both the transmit and the receive beamformers, at a fixed constant scanning time. We still get all the benefits of loudspeaker beamforming, such as noise suppresion and reduced reverberation, by the fact that the effective received signal must be the same when beamforming in post-processing by the properties of LTI systems.

4. SSL FOR DEPTH IMAGING

Transducer arrays with a small number of elements (and without precise calibration) produce beams that are far from perfect. A considerable amount of energy is transmitted into the side lobes, and the main lobe is relatively wide. If there is a particularly strong reflector in the scene, it will always appear that the reflection is coming form that strong reflector, as there will always be some amount of acoustic energy radiated towards it.

To deal with imperfect beamforming, we propose to combine the beamformer with sound source localization algorithms. For each beam direction, we detect the moments of the multiple returned echoes (note that this signal includes both the transmit and the receive beamforming), as illustrated in Fig. 2. After determining the delays of returned pulses in the current beam, we go back to the raw, unbeamformed microphone signals. These signals did not pass through the receive beamforming filters, but they are still concentrated on a part of the scene through loudspeaker beamforming. We select segments that correspond to the detected peaks, and we feed them into the sound source localizer. Most of the SSL algorithms can be applied—we use the MUSIC [13] algorithm adapted to search in azimuth and elevation, but other SSL algorithms could be used in place of MUSIC. The key step is that we create a depth pixel only if the output of the source localizer agrees (within some prescribed tolerance) with the direction where we are pointing the beam. More precisely, for each direction in the image, we create a list of candidate distances, and after going through all the beams, we select the *smallest* distance for every direction.

The described procedure is summarized in Algorithm 1. The total beamformed signal corresponding to the ith beam is given as

$$X_i(f) = \sum_{m=1}^{M} V_m^i(f) R_m^i(f), \tag{7}$$

where $V_m^i(f)$ is the beamforming filter for the mth microphone and the ith beam. The signal $x_i(t)$ is first processed to find the returned peaks by cross-correlating it with the pulse template, and a list of time delays corresponding to P strongest peaks is created $\left\{\tau_p^i\right\}_{p=1}^{P}$. Then we go back to the microphone signals $\left\{r_m^i(t)\right\}_{m=1}^{M}$, and for every detected peak in the list, we extract the segments of microphone signals around the detected echo return times τ_p^i. We feed these segments into a sound source localization algorithm. We used an implementation of the MUSIC algorithm that outputs an estimate of the direction of arrival. If the estimated direction agrees with the beam direction, we add the depth of the current echo to the list of candidate depths for the current direction.

Algorithm 1 SSL enhancement

Input: ▷ Set of transmit-beamformed signals received by the microphones, $\{R_m(f)\}_{m=1}^{M}$
▷ Number of echoes per beam to consider, P
▷ Threshold for angular distance, d_{thr}
▷ SSL window duration, T
Output: ▷ SSL enhanced depth image

For every receive beam:

(i) Do the microphone beamforming, $X_i(f) = \sum_{m=1}^{M} V_m^i(f) R_m(f)$.
(ii) $\left\{\tau_p^i\right\}_{p=1}^{P} \leftarrow \mathsf{GetEchoTimes}[x_i(t)]$
(iii) **For** p **from** 1 **to** P**:**
▷ $[\theta_{\text{est}},\ \phi_{\text{est}}] = \mathsf{SSL}\left[\left\{r_m^i(\tau_p^i - T/2 : \tau_p^i + T/2)\right\}_{m=1}^{M}\right]$
▷ If $\mathsf{dist}([\theta_{\text{est}}, \phi_{\text{est}}], [\theta, \phi]) \le d_{\text{thr}}$, then $D_i \leftarrow D_i \cup \left\{\tau_p^i / c\right\}$

For all D_i**:**

(i) Set the pixel value $I_{\text{SSL}}^i \leftarrow \min D_i$
(ii) If the depth is outside of $[d_{\min}, d_{\max}]$, set it to ∞.

5. EXPERIMENTAL RESULTS

We designed simple hardware to investigate the ultrasonic imaging in air, shown in Fig. 1. To transmit the ultrasonic pulses, we use 8 Senscomp 40LT10 piezoelectric buzzers. In comparison with similar devices, they have a slightly smaller diameter of 10 mm, which allows for more freedom in geometric design. The transducers are operating at 40kHz with a relatively small bandwidth—at 38kHz,

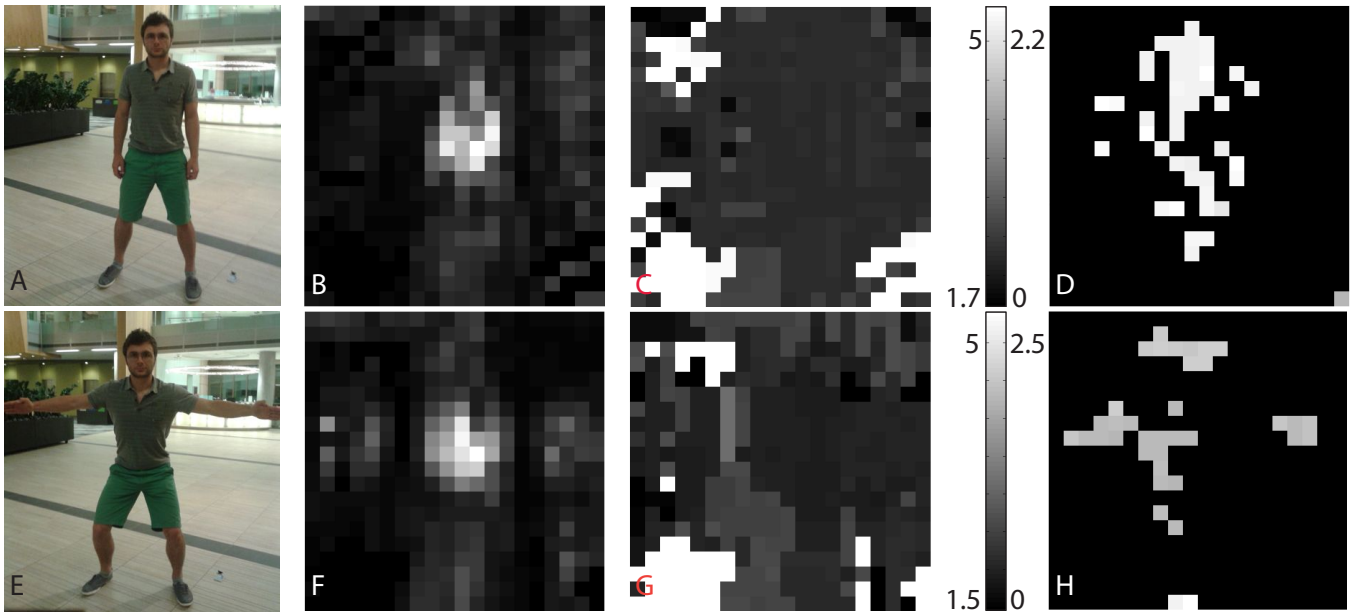

Fig. 3. Experimental results: The photograph of a subject (A, E), intensity images (B, F), naïve depth images (C, G) and SSL-enhanced depth images (D, H).

the signal is attenuated by 10 dB relative to the maximum value. On the receive side, we use the top-port MEMS microphones Knowles SPM0406HE3H, typically used in cellphones. Conveniently, they have a preamplified differential output, so they are easily integrated in our system design. Microphone signals are fed into Presonus DigiMax D8 preamplifier and then into MOTU 828mk3 Hybrid unit operating at the sampling frequency of 192 kHz.

A major advantage of ultrasonic devices is their low power dissipation. Taking into account the power consumption of piezo buzzers, and the target duty cycle, we can achieve powers of less than 100 mW, maybe even below 10 mW. In the prototype design we used off-the-shelf operational amplifiers with a large quiescent current, resulting in the power dissipation of around 3 W. The signal power was set low enough so that no audible artifacts were present.

Experimental results are shown in Fig. 3. We acquired images of a person standing approximately 2.5 meters from the ultrasonic device, first with arms relaxed along the sides of the body, and then with arms outstretched. The aim of the experiments was to understand if the pose change is clearly observed in the reconstructed images.

Figs. 3B and 3F show the *intensity image*, where pixel value is proportional to the the energy of the strongest echo. In particular, the pixel value for the ith beam in this image is computed as $I^i_{\text{int}} = \int_{t_{\max}-T/2}^{t_{\max}+T/2} |x_i(t)|^2 \, \mathrm{d}t$, where $t_{\max}$ is the time of the largest returned peak, and T is the window size for energy computation. The field of view is $60^\circ \times 60^\circ$, and the angular resolution is 3°.

We see in Fig. 3B that the body indeed gives reflections corresponding to larger intensity values. A more important finding is that the spread-arms pose is clearly observed in the ultrasonic intensity image. This suggests that ultrasound could be used for skeletal tracking, or more generally, HCI.

Figs. 3C and 3G show the corresponding *naïve* depth images created by finding the time delay of arrival of the largest returned pulse for every beam. As explained previously, this largest pulse will always seem to be coming from strong reflectors in the scene, thus the information it provides is not reliable. This is clear from the corresponding images, as they bear no resemblance to the intensity image—most of the pixels are at the distance of the person. On the other hand, Figs. 3D and 3H show the depth images created by Algorithm 1. A major improvement is observed (black encodes no echo received): Depth pixels are places only where there actually is an object in the scene, and the distances are correct. In particular, the spread-arms pose is clearly distinguishable. We remind the reader that these images were obtained by an eight element array, to be compared with an eight pixel camera (8 + 8).

6. CONCLUSION

In this paper, we explored the use of ultrasonic transducer arrays with a small number of transmitters and receivers for creating depth and intensity images. Our aim was to assess the potential of ultrasound in the design of HCI interfaces.

We addressed two issues. First, we demonstrated how to do the transmit beamforming offline when using loudspeaker beamforming, by leveraging the basic properties of LTI systems. This solves the frame rate problem due to the low speed of sound, and enables frame rates suitable for real-time applications. Second, using the proposed SSL-based imaging algorithm, we obtained depth images that reveal the pose of the human subject, unlike with the naïve approach. Due to the highly specular nature of sound reflections, it is not possible to estimate the depth image naively. However, intensity images clearly indicate the pose of the subject, and so do SSL-enhanced depth images, suggesting that there is enough information for HCI. Skeletal tracking in Kinect is based on large scale frame-by-frame machine learning. In this context, and given the experimental results, we can conclude that ultrasound is a cheap, low-power alternative technology for HCI.

Future work includes increasing the number of transducers, improving the calibration, and designing skeletal trackers and gesture recognizers using depth and intensity features.

7. REFERENCES

[1] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, 2006, pp. 519–528.

[2] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, 2008, pp. 1–7.

[3] B. Schwarz, "Mapping the world in 3D," *Nature Photonics*, 2010.

[4] A. Kirmani, A. Colaço, F. Wong, and V. K. Goyal, "Exploiting sparsity in time-of-flight range acquisition using a single time-resolved sensor," *Opt. Express*, 2011.

[5] M. Moebus and A. M. Zoubir, "Three-Dimensional Ultrasound Imaging in Air using a 2D Array on a Fixed Platform," in *IEEE Int. Conf. Acoust. Speech, and Signal Process.* 2007, pp. 961–964, IEEE.

[6] M. Moebus, A. M. Zoubir, and M. Viberg, "Parametrization of acoustic images for the detection of human presence by mobile platforms," in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Process.* 2010, pp. 3538–3541, IEEE.

[7] R. Miyake, K. Hayashida, M. Nakayama, and T. Nishiura, "A study on acoustic imaging based on beamformer to range spectra in the phase interference method," in *Proceedings of Meetings on Acoustics*. June 2013, pp. 055041–055041, Acoustical Society of America.

[8] M. R. Strakowski, B. B. Kosmowski, R. Kowalik, and P. Wierzba, "An ultrasonic obstacle detector based on phase beamforming principles," *IEEE Sensors J.*, vol. 6, no. 1, pp. 179–186, Feb. 2006.

[9] S. Harput and A. Bozkurt, "Ultrasonic Phased Array Device for Acoustic Imaging in Air," *IEEE Sensors J.*, vol. 8, no. 11, pp. 1755–1762, Nov. 2008.

[10] H. L. Van Trees, *Optimum Array Processing (Part IV of Detection, Estimation, and Modulation Theory)*, John Wiley & Sons, New York, USA, 2002.

[11] I. J. Tashev, *Sound Capture and Processing*, Practical Approaches. John Wiley & Sons, Chichester, UK, 2009.

[12] C. Holmes, B. W. Drinkwater, and P. D. Wilcox, "Post-processing of the full matrix of ultrasonic transmit–receive array data for non-destructive evaluation," *NDT&E International*, vol. 38, no. 8, pp. 701–711, Dec. 2005.

[13] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.