Exploring the Resolution Limit for In-Air Synthetic-Aperture Audio Imaging

Hisham Bedri, Micha Feigin
MIT Media Lab
Cambridge MA
hbedri@media.mit.edu, michaf@mit.edu

Petros T. Boufounos
MERL
Cambridge MA
petrosb@merl.com

Ramesh Raskar
MIT Media Lab
Cambridge MA
raskar@media.mit.edu

Abstract

SONAR imaging can detect reflecting objects in the dark and around corners, however many SONAR systems require large phased-arrays and immobile equipment. In order to enable sound imaging with a mobile device, one can move a microphone and speaker in the air to form a large synthetic aperture. We demonstrate resolution limited audio images using a moving microphone and speaker of a mannequin in free-space and a mannequin located around a corner. This paper also explores the 2D resolution limit due to aperture size as well as the time resolution limit due to bandwidth, and proposes Continuous Basis Pursuits (CBP) to super-resolve.

1. Introduction

Mobile phones are as of yet not capable of seeing in the dark, through smoke, and around occluding objects. While there is a race to commercialize time-of-flight sensors [11, 7, 6], mobile phones have readily available audio hardware which can perform audio imaging tasks. This would be useful for rescue situations and indoor mapping. [5, 10, 2]

In order to generate a sensing aperture to perform scene reconstructions, a user moves their mobile phone to a set of static positions which allows the user to sample a plane. At each location the phone transmits and receives a signal, similar to synthetic aperture sonar/radar. The received signals are then processed (pulse compression) and the data inverted using backprojection to generate a 3D image.

Figure 3 shows an example of data acquisition. Each row of the data is a range measurement taken from a microphone and speaker pair which is moved to another location. The range measurements are then backprojected to form an image. An ideal reflecting point lies on an ellipsoid in space whose foci are the speaker and microphone. The distance traveled by the sound is the major axis of the ellipsoid. Through tomography, it is possible to reconstruct the location of all reflectors.

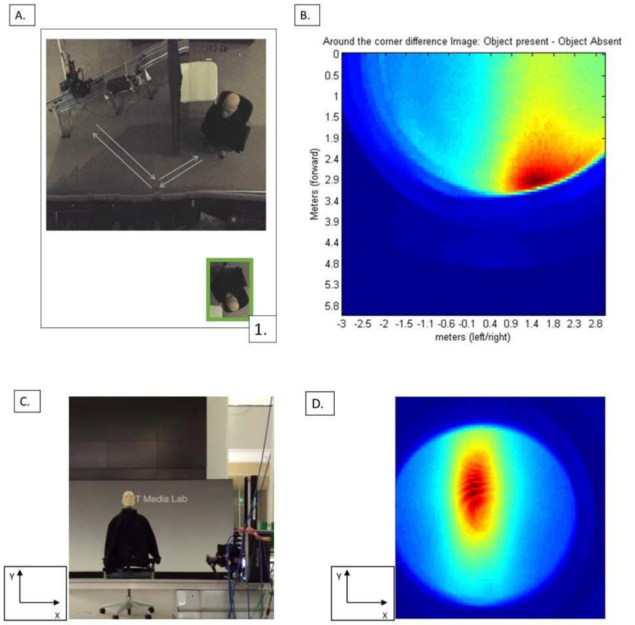

Figure 1. In A, a mannequin is placed around the corner from our system. In B, an audio image returned by the system. One can see that the hot spot of the audio image is located near where the reflection of the mannequin should be. In C, a mannequin is placed directly in front of the system, and in D, the corresponding audio image.

2. Experiment and Results

We implement mobile-audio imaging by moving a speaker and microphone on an x-y stage, transmitting a chirp between 20KHz and 30KHz. While today's phones cannot reliably produce sound in this range, microphones and speakers exist which can easily operate in these ranges. We move a microphone and speaker pair to 216 positions (18 columns by 12 rows) in a 1m x 1m 2D plane and perform a range measurement at each location (transmit and receive). We demonstrate two imaging results, one with a mannequin in front of the setup, and another with a mannequin located around a corner. The result with the mannequin in front of the setup is shown in Figure 1 in the

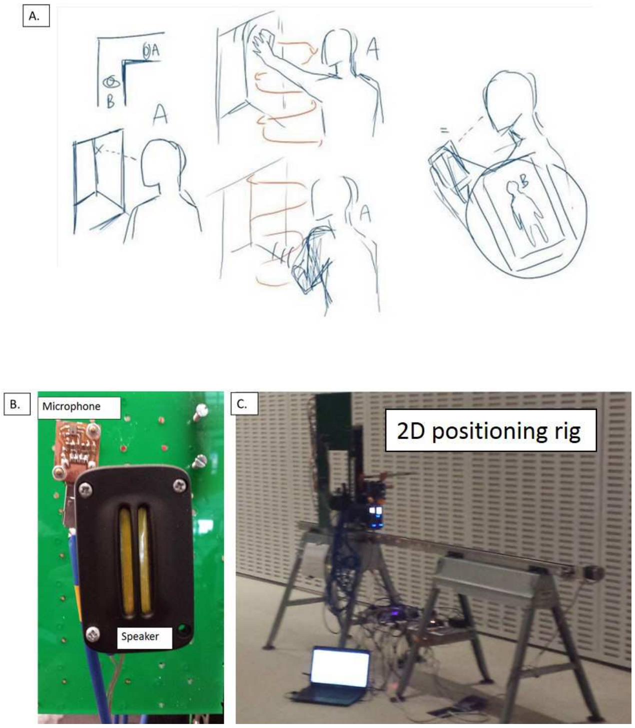

Figure 2. Top, an artist's vision of synthetic aperture audio imaging, where a user moves a mobile phone in the air and reconstructs an object around the corner. Graphic Credit Dina Bedri. Left, a photo of our microphone and speaker pair used (20KHz - 30KHz bandwidth). Right, 2D positioning rig used in the experiments.

upper-right. The result for imaging of a mannequin behind a corner is shows in Figure 2. The system is able to see around the corner since sound bounces specularly off the wall, causing a virtual image of the mannequin to be in view.

3. Angular Resolution Limit

The diffraction-limited angular resolution of a camera is determined by its aperture and wavelength of illumination: $\theta = 1.22 \cdot \lambda / D$[4]. The same equation applies for the synthetic aperture covered by the motion of the microphone and speaker, thus a larger area covered leads to more angular acuity. When using the backprojection method, this defines the lateral PSF of our system. By covering an aperture of 1m x 1m with $\lambda/2$ density, our experiment has a theoretical rayleigh-limited angular resolution of 0.8cm.

4. Time Domain Resolution Limit

Hardware and physical limitations prevent the pulse compression step from acheiving a delta response. This means that an ideal reflection of a sound ping cannot be localized to a single point in time and is usually spread over a range. A more complete treatment of the model of how sound travels in air will start with boundary conditions and differential equations. In this paper, we will begin with a simple intuitive model for how a pressure signal in the air medium travels between a transmitter and a receiver. Here we assume an ideal reflector k of known position, a speaker s of known position, an omni-directional microphone m of known location. The received signal due to propogation in air is proportional to:

$$y\left(t\right) = \sum\nolimits_{k=0}^{K-1} a_k \phi\left(t - \tau_k\right) \tag{1}$$

where,

- $\mathbf{p}^{(m)}$, $\mathbf{p}^{(s)}$ and $\mathbf{p}_k^{(r)}$, respectively, are the 3D locations of microphone, speaker, and k^{th} reflector.
- $\tau_k = \frac{\left\|\mathbf{p}^{(m)} - \mathbf{p}_k^{(r)}\right\| + \left\|\mathbf{p}^{(s)} - \mathbf{p}_k^{(r)}\right\|}{\nu}$ is the propagation time delay due to transmission and reflection.
- ϕ is the time-domain function transmitted by the speaker
- $\{a_k\}_{k=0}^{K-1}$ are the amplitudes of each reflection (based on material and illumination properties).

In this case, we use a chirp signal instead of an impulse, since it enables our system to transmit more energy over a longer time period. The time-shifted signal f is shown below:

$$\phi\left(t\right) = \cos\left(\xi t^2 + \omega t\right) \mathbb{1}_{[0,T]}\left(t\right) \tag{2}$$

where ξ relates to the slope of the chirp such that $\xi = \frac{2\pi}{T}\left(f_2 - f_1\right)$ and $\omega = 2\pi f_1$ is the initial frequency. Furthermore, in (2), we use the indicator function defined on domain $\mathcal{D}$ by,

$$\mathbb{1}_{\mathcal{D}}\left(t\right) = \begin{cases} 1 & t \in \mathcal{D} \\ 0 & t \notin \mathcal{D} \end{cases}.$$

We assume that $\{\tau_k\}_{k=0}^{K-1} \in [0, T]$.

We utilize FMCW processing (multiplication of the transmitted signal with the received signal and performing a Fourier transform and a low-pass filter) to recover the shift τ.

$$\begin{aligned} \phi(t) &= \cos(\psi(t)) \mathbb{1}_{[0,T]}(t) \\ \phi(t-\tau) &= \cos(\psi(t) + \theta(t)) \mathbb{1}_{[0,T]}(t) \end{aligned}$$

where

$$\begin{aligned} \psi(t) &= \frac{\xi}{2} t^2 + wt \\ \theta(t) &= -\xi \tau t + \xi \tau^2 - 2\pi \tau f1 \end{aligned}$$

therefore:

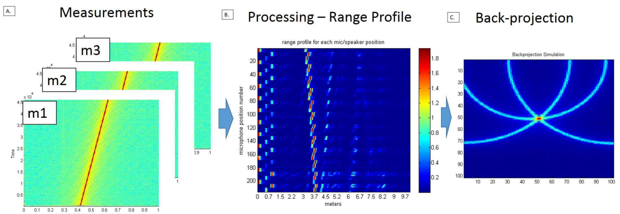

Figure 3. A flow-chart of the rayleigh-limited reconstruction. In A, the system transmits acoustic chirps. In B, thos chirps are compressed through FMCW processing [9], and in C, the signals are reconstructed through backprojection.

$$\phi(t-\tau)\cdot\phi(t)=\frac{1}{2}[\cos(\phi(t))+\cos(\psi(t))]\mathbb{1}_{[0,T]}(t)$$

and by examining the low-frequency portion by low-passing above $2f_1-\frac{\xi}{2\pi}\tau$

$$\cos(-\xi\tau t+\xi\tau^2-2\pi\tau f_1)\mathbb{1}_{[0,T]}(t)$$

Thus τ is encoded into the frequency of the processed cosine, and can be recovered via a fourier transform. Increasing the duration of the transmission window T has the effect of reducing ξ, thus the limit in resolution depends only on f_1, f_2.

4.1. Physical Bandwidth Limit

There is a hardware limit to the bandwidth of sound which can be transmitted in-air. This limit is caused by absorption in the air of high-frequency signals. The attenuation has been characterized by the following equation, which means the attenuation of a signal in air is nearly 100dB at 50Khz at over 1m [1]:

Attenuation $=\alpha\cdot L\cdot f$, where:

- α is the attenuation coefficient of air in units $\frac{db}{MHz\cdot cm}$
- L is the distance propagated
- f is the frequency of the sound

The limitations on the frequencies which can be reliably transmitted through the medium impose a bandwidth limit on the signal which is transmitted. This bandwidth limit causes an ambiguity in the estimation of the signal delays and thus the estimations for the distances to the target. This puts the rayleigh-limited time-domain resolution to 10cm for the 20KHz - 30KHz range of the audio spectrum.

5. Model-based Super-resolution

Due to physical hardware limitations for in-air imaging, we turn to computational methods to address the imaging problem. In the following sections, we describe how sparse reconstruction can theoretically extend the resolution of the system, and show simulations which achieve higher theoretical resolution.

5.1. Backprojection Matrix Coherence

One can use model based methods for time-domain super-resolution by creating a dictionary of shifted signals and searching for the support within the dictionary and the corresponding coefficients that describe the received signal. One method for recovering the sparse vector x is through Basis Pursuit denoising (BPDN) which uses the ℓ_1 penalty instead of the expensive-to-compute ℓ_0 penalty. BPDN has advantages over greedy methods since it is guaranteed to converge to the global minimum solution[8],

$$\mathbf{x}^\star=\arg\min_{\mathbf{x}}\|\mathbf{b}-\mathbf{D}_\Delta\mathbf{x}\|_2^2+\lambda\|\mathbf{x}\|_1. \tag{3}$$

Using sparse recovery algorithms, it is possible to find the best sparse representation of the signal within resolution of the shifts of the dictionary. However, this approach is not guaranteed to work since the signal may lie in an off-grid location. In order to reduce the chance of this happening, one can form finer and finer sampling grids, however, this leads to an increase in the coherence of the dictionary. The algorithms can have a hard time choosing which atom is the correct one, especially in the presence of signal or quantization noise. The resolution of the estimation of the parameter τ is again limited by the coherence of the matrices using these methods.

The coherence of a matrix is defined as follows [8]:

$$\mu\left(\mathbf{D}_\Delta\right)=\max_{1\le k,j\le m,k\ne j}\frac{\mathbf{d}_k^\top\mathbf{d}_j}{\|\mathbf{d}_k\|\,\|\mathbf{d}_j\|}. \tag{4}$$

5.2. Continuous Basis Pursuits

Continuous Basis Pursuit (CBP) overtakes on-grid methods by introducing a bilinear model which finds the atoms of the dictionary which approximates the signal the best and then improves that approximation by finding a coefficient for a corresponding dictionary which perturbs the approximation closer to the original signal. The result is a recovery which is more accurate, and for sparse signals, results in a more sparse solution than BPDN for off-grid elements. In order to understand CBP, one can think of any N-dimensional signal as a point in N-dimensional space. The set of all time-domain shifts of the signal form a manifold in N-dimensions. On-grid solutions approximate this manifold by uniformly sampling it. CBP improves upon this by forming an approximation of the shape of the manifold (either Taylor or polar) and perturbing the sample points to more accurately represent a shifted signal.

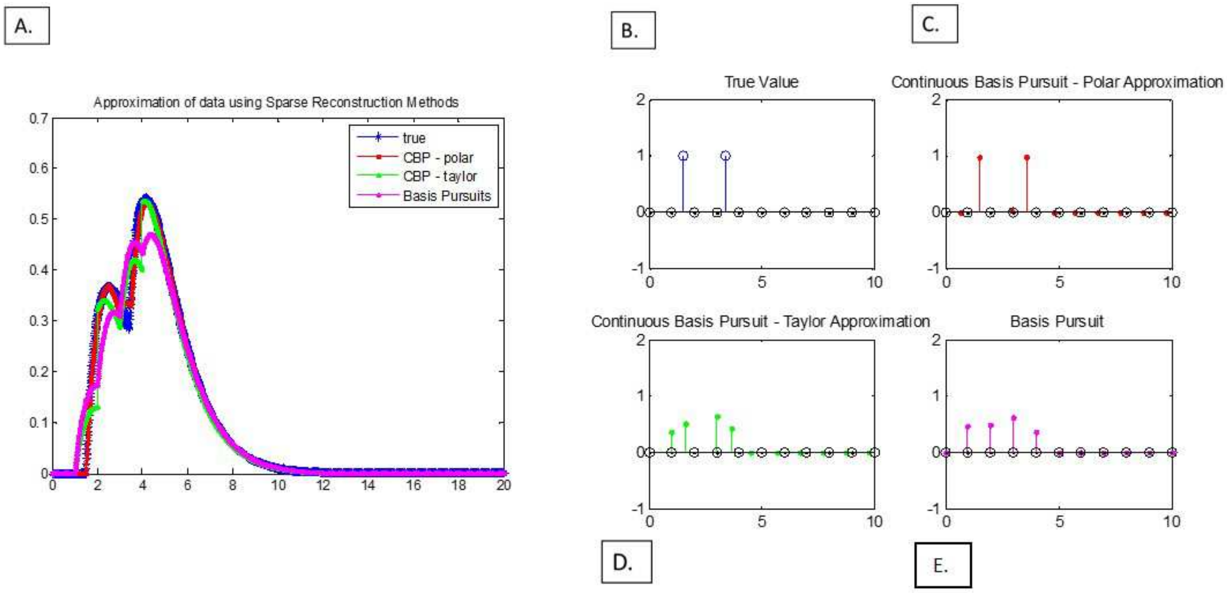

Figure 4. Results of super-resolution in 1 dimensions. In A, approximations of the sum of shifted versions of ground-truth decaying exponential signals (blue) and the result of sparse recovery techniques (Basis Pursuit DeNoising [BPDN] magenta, Continuous Basis Pursuit Taylor [CBPT] green, Continuous Basis Pursuit Polar [CBPP] red). In B-E, recovered time shifts and their coefficients. From upper left clockwise: True shifts in blue, upper-right, CBPP in red, most accurate, Lower Right, BPDN in magenta, Lower Left, CBPT in green. Grid points are in black.

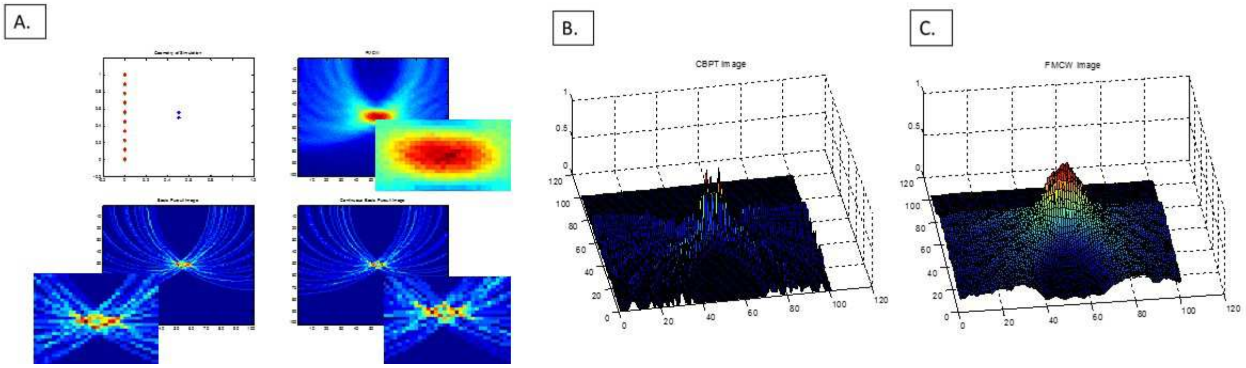

Figure 5. Result of 2D super-resolution. In A, simulation set-up, microphones in green, speakers in red, and reflectors in blue. On the right, reconstruction using Continuous Basis Pursuit Taylor (B) and FMCW image (C).

6. Comparing Algorithms for Image Reconstruction

In order to evaluate super-resolution, backrpojection (linear inversion), BPDN, and CBP were compared for reconstruction. There is a challenge in comparing the performance of algorithms, especially since they behave differently with different reconstruction parameters (λ, Δ). Furthermore, comparing reconstructions is difficult since continuous basis-pursuit returns vectors which are not on-grid, thus recovered vectors cant be directly compared by discretely by taking the norm of the difference of the signals. In order to measure accurate signal reconstruction, a simple error term is defined:

$$\mathsf{E}\left(\widetilde{\mathbf{\Theta}}, \mathbf{\Theta}\right) = \sum_{k=0}^{K-1} \left(|\widetilde{\tau}_k - \tau_k|\, \widetilde{a}_k\right)^2 + \|\mathbf{b} - \widetilde{\mathbf{b}}\|_2^2, \quad (5)$$

where $\widetilde{\mathbf{\Theta}} = \{\widetilde{a}_k, \widetilde{\tau}_k\}_{k=0}^{K-1}$ denotes the set of estimated parameters and $\mathbf{\Theta}$ denotes the ground truth. $\widetilde{\tau}_k$ are ordered such that $\sum_{k=0}^{K-1} ||\widetilde{\tau}_k - \tau_k||$ is minimized. $\mathbf{b}$ is the ground-truth measured signal, and $\widetilde{\mathbf{b}}$ is the signal produced by the recovered parameters $\widetilde{\mathbf{\Theta}}$.

The error function is a trade-off between accurate time-shifts and accurate data-matching. If the recovered time shifts are inaccurate and the data is matched perfectly, then the left part of the error will grow high. If the time shifts and coefficients are accurate, then both the left and right parts of the error will be zero. If the time shifts are accurate but the coefficients are zero due to lambda being too strong, then the right part of the error function will grow. The following signal was used as base-function in the 1D simulation (for ease of visualization):

$$y\left(t\right) = \mathbb{1}_{\left[\tau_1,\infty\right)}\left(t\right)\left(t-\tau_1\right)e^{-\alpha\left(t-\tau_1\right)} \\ + \mathbb{1}_{\left[\tau_2,\infty\right)}\left(t\right)\left(t-\tau_2\right)e^{-\alpha\left(t-\tau_2\right)}.$$

6.1. Comparing 1D recovery

1D recovery is shown in Figure 4 for BP, BPDN, CBPT, and CBPP. One can see that CBPP performs the best and most accurately detects the time shifts in the most sparse manner. BPDN approximates the signal with two coefficients on the grid points closest to each signal point. CBPT also approximates the signal with two coefficients, however the coefficients are closer to the signal. Figure 6 shows the performance of each of the algorithms with different delta (dictionary spacings) and lambdas (regularization parameters). One can see that as the dictionary spacing increases, the error increases. Similarly, for each delta, there is only one lambda which is optimal. CBP-Taylor and CBP-Polar both outperform BPDN such that for each delta, there is a corresponding lambda in each of the other two algorithms with lower error.

6.2. Comparing Image recovery

In order to simulate the audio-imaging process, a 2D room was modelled with a set of 10 microphone/speaker pairs and a two reflectors located 10cm apart. Each microphone/speaker measures the reflected signal from the system separate from the other pairs. The algorithms are applied to find range profiles at each measurement and a backprojection algorithm is used to reconstruct the image. One can see in Figure 6 that the model-based algorithms (BPDN, CBP) perform much better than FMCW processing and it is possible to discern two peaks. Furthermore, the off-grid CBPT reconstruction performs better than the on-grid BPDN reconstruction. One can see better defined peaks at the reflectors.

7. Discussion

There are many uses for mobile audio imaging, including searching for persons trapped in rubble, room reconstruction while your phone is in your pocket, mapping out caves, automated vehicles, depth imaging, and kaleidoscopic reconstructions of objects. However, there are many resolution limits due to bandwidth. On the high end, high frequency sound does not carry in air, and on the low-end, the system must operate outside of the human hearing range (so it is not disturbing).

CBP is effective in finding a more accurate representation of off-grid signals, however the approximations necessary to perform CBP limit the types of signals one can use. The CBP Taylor approximation assumes the function f is differentiable on all points t, thus a discontinuity can ruin the reconstruction. Furthermore Ekanadham [3] highlights that the polar approximation deteriorates as the signal increases in bandwidth.

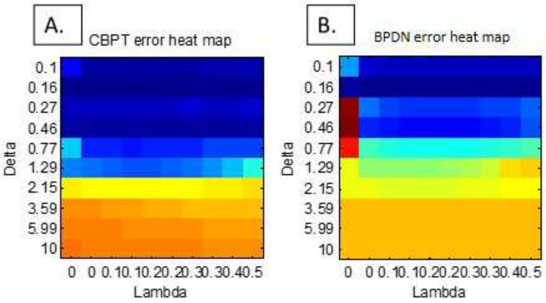

Figure 6. Heat maps showing error associated with varying deltas (grid spacing) (Red = large error), lambdas (regularization parameters) and algorithms. In A, Continuous Basis Pursuit Taylor, In B, Basis Pursuit Denoising

A limitation to the CBP approach to audio imaging is the failure of the assumption of sparsity in the scene. In the super-resolution reconstructions, the scene was assumed to consist of sparse-reflectors. Real-life scenes, however, are composed of complex objects with curvatures and shapes. The ideal omnidirectional reflector assumption breaks down if you have a reflecting plane or surface. In-air scenes usually consist of a small number of objects, thus future work will explore modelling the scene geometry as a set of primitive shapes of varying sizes and locations, and use bilinear or trilinear sparse recovery to estimate the shapes.

Acknowledgements: We would like to thank M. Everett, I. Filho, and G. Charvat for contributing to the design and engineering of the imaging setup. We would like to thank G. Satat, for conversations. P. Boufounos was exclusively supported by Mitsubishi Electric Research Laboratories.

References

[1] Acoustics – attenuation of sound during propagation outdoors – part 1: Calculation of the absorption of sound by the atmosphere. ISO 9613-1, 1993. 3

[2] A. O. Donovan, R. Duraiswami, and D. Zotkin. Imaging concert hall acoustics using visual and audio cameras. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 5284–5287. IEEE, 2008. 1

[3] C. Ekanadham. *Continuous basis pursuit and its applications*. PhD thesis, New York University, 2012. 5

[4] J. W. Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005. 2

[5] S. Gupta, D. Morris, S. Patel, and D. Tan. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1911–1914. ACM, 2012. 1

[6] A. KADAMBI, H. ZHAO, B. SHI, and R. RASKAR. Occluded imaging with time of flight sensors. 1

[7] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar. Looking around the corner using transient imaging. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 159–166. IEEE, 2009. 1

[8] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993. 3

[9] A. Meta, P. Hoogeboom, and L. P. Ligthart. Signal processing for fmcw sar. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(11):3519–3532, 2007. 3

[10] O. Oçal, I. Dokmanic, and M. Vetterli. Source localization and tracking in non-convex rooms. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1429–1433. Ieee, 2014. 1

[11] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Communications*, 3:745, 2012. 1