Digital Ventriloquism: Giving Voice to Everyday Objects

Yasha Iravantchi
University of Michigan
Ann Arbor, MI
yiravan@umich.edu

Mayank Goel
Carnegie Mellon University
Pittsburgh, PA
mayank@cs.cmu.edu

Chris Harrison
Carnegie Mellon University
Pittsburgh, PA
chris.harrison@cs.cmu.edu

ABSTRACT
Smart speakers with voice agents are becoming increasingly common. However, the agent's voice always emanates from the device, even when that information is contextually and spatially relevant elsewhere. Digital Ventriloquism allows smart speakers to render sound onto everyday objects, such that it appears they are speaking and are interactive. This can be achieved without any modification of objects or the environment. For this, we used a highly directional pan-tilt ultrasonic array. By modulating a 40 kHz ultrasonic signal, we can emit sound that is inaudible "in flight" and demodulates to audible frequencies when impacting a surface through acoustic parametric interaction. This makes it appear as though the sound originates from an object and not the speaker. We ran a study in which we projected speech onto five objects in three environments, and found that participants were able to correctly identify the source object 92% of the time and correctly repeat the spoken message 100% of the time, demonstrating our digital ventriloquy is both directional and intelligible.

Author Keywords
Ultrasound; Smart Speakers; IoT; Interaction; VR/AR.

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI);** *Interaction devices;* Sound-based input / output;

INTRODUCTION
Smart speakers with voice agents have seen rapid adoption in recent years, with 41% of U.S. consumers owning one by the end of 2018 [30]. These devices use traditional speaker coils, which means the agent's voice always emanates from the device itself, even when that information might be more contextually and spatially relevant elsewhere.

Predating smart speakers by almost three decades, Mark Weiser described a future with ever-present voice agents: even a paper instruction manual can speak and is interactive [39]. One option is to instrument everything or to have multiple speakers in each room, but this comes with installation, maintenance, and aesthetic downsides. In this paper, we describe our work on Digital Ventriloquism, which allows a single smart speaker to render sounds onto many passive objects in the environment. Not only can these items speak, but also make other sounds, such as notification chimes. Importantly, objects need not be modified in any way: the only requirement is line of sight to our speaker. As smart speaker microphones are omnidirectional, it is possible to have interactive conversations with totally passive objects, such as doors and plants.

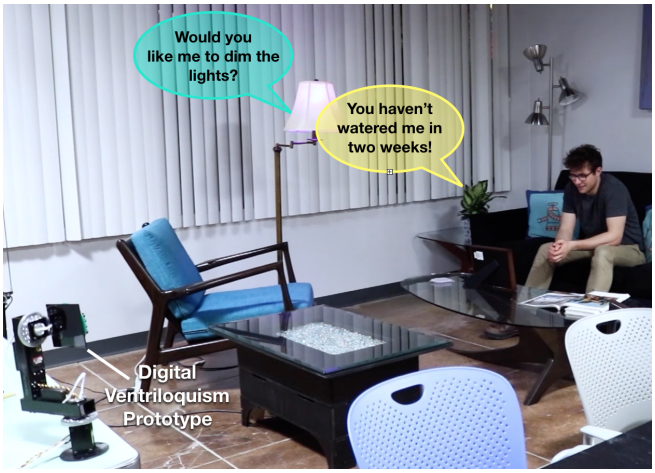

Figure 1. Our system uses a parametric speaker to project ultrasonic beams and render sounds onto passive, everyday objects, appropriating them as interactive voice agents and distributing intelligence throughout environments.

To achieve this effect, we use a dense, 2D array of ultrasonic transducers. This produces a highly directional emission due to the Huygens-Fresnel principle [22], which is critical for rendering sounds onto specific objects in the environment. We amplitude modulate a 40 kHz ultrasonic signal, which is inaudible "in flight" prior to collision with an object's surface. Upon collision, it demodulates to audible frequencies through parametric interaction [31]. Thus, the object becomes the origin of the audible sound. Humans can then localize this "digital" ventriloquism as they would with any other sound (i.e., though binaural localization and their head-related transfer function).

We started our development with a series of physical studies, characterizing the parametric effect and audible response

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376503

across different materials and surface geometries, before moving to capturing data from real-world objects. We conclude with a user study, which tests the perceived directionality and intelligibility of speech across three settings and 15 commonplace items. Over 225 trials, participants were able to localize the source of sound 92.0% of the time, and were able to accurately hear the spoken phrase 100% of the time, demonstrating imminent feasibility.

RELATED WORK

We review three key literatures that intersect with our work on Digital Ventriloquism. First is a brief review of ultrasound in HCI systems. We then review prior work that overlaps with Digital Ventriloquism in enabling localized audio for augmented sound environments. Finally, we cover work that more closely aligns with our technical approach using parametric interaction.

Ultrasound in HCI

Ultrasound can be generated using low cost components: transducers can be found for as little as $0.25/pc [1], making ultrasound popular in many sensor-driven systems in the Human-Computer Interaction literature. For example, ultrasound is used as a rangefinder [16, 4] and for Doppler sensing [13, 32]. Beyond these two more common uses of ultrasound, interference effects [9] and beamforming [8] have been utilized for face and hand gesture recognition. Finally, the closest implementation of ultrasound to Digital Ventriloquism is for in-air haptics, which uses a similar array of coplanar transducers working in concert to create focused ultrasonic energy [3, 12, 19, 41]. See [32] for a survey of uses of ultrasound in HCI.

Localized Audio

Sound localization plays a strong role in how users perceive their environment, especially in conjunction with visual input to generate immersive spatialized audio environments [26]. In applications where acoustics are tied to coordinated visual feedback, the directionality of the sound is attained through Interaural Time Difference (ITD) in which uses two speakers (often headphones) use phase shifts to give the perception of directionality. This is particularly important for 360° videos [17], enhanced mobile apps [18], and interactions with augmented virtual objects [28].

To create a soundscape for more than one observer, individual objects can be augmented with speakers to create a multitude of audio sources that multiple observers can experience [7]. In particular, there is significant work in creating audible overlays in museum environments [6, 15, 14, 29]. Museums may embed paintings with transducers using the canvas as a diaphragm, allowing multiple viewers to experience sound emanating from the painting itself [2]. However, each painting would require instrumentation in order to "speak".

Another approach to localize audio is directional audio, which most commonly uses either a parabolic reflector or parabolic speaker array to produce a directed acoustic beam towards a chosen target [31]. Parabolic speakers do not significantly modify their input signal; the sound is completely audible in flight and have a useable listening range of a few meters [31].

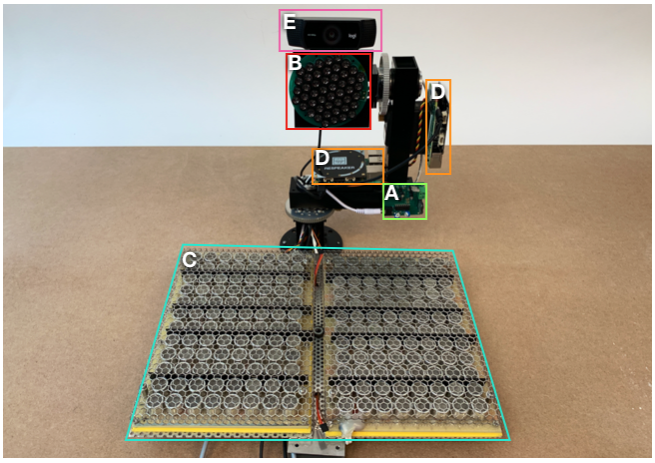

Figure 2. Our proof of concept apparatus. A) Driver Board. B) Small Speaker Array. C) Large Speaker Array. D) Raspberry Pi Microphone Array. E) Webcam.

Such directional speakers have been widely used in museum environments [15, 14] and spherical loudspeakers have been used to reproduce the directivity of musical instruments [21]. With these types of speakers, users still localize the speaker itself as the origin of the sound, and not the targeted object.

Parametric Interaction

Unlike parabolic speakers, parametric speakers do not emit any audible frequencies [31]. Instead, parametric speakers use an array of ultrasonic transducers to create aimed ultrasound waves with narrow beam width compared to audible frequencies [31]. These ultrasonic beams can be modulated with an input signal and the nonlinearities in air and surface interactions create a heterodyning effect, resulting in the modulation signal separating from the ultrasound carrier upon striking a solid surface [31]. To an observer, this effect makes it appear as if the sound emanates from the targeted surface itself, and unlike parabolic speakers, the beam is completely inaudible in flight until the signal is demodulated.

These speakers have been used previously in HCI applications to render audio-enhanced spots through reverse ray tracing [28] and as a localized communication channel [38], including with handheld systems [27]. There is also significant commercial interest in taking advantage of parametric speakers for directional audio, though these systems (e.g., [20]) do not offer implementation details, nor physical or user studies to aid the HCI community. Finally, we note that prior research has explored using parametric speakers on pan-tilt platforms [11, 10], exactly like our setup, but it is used to simply direct audio at listeners (i.e., directional audio) and does not explore the notion of ventriloquism, where other objects are given voice, nor using this ventriloquy to enabled distributed intelligence.

IMPLEMENTATION

Digital Ventriloquism consists of four key components: the ultrasonic array, signal generation board, pan/tilt platform, and

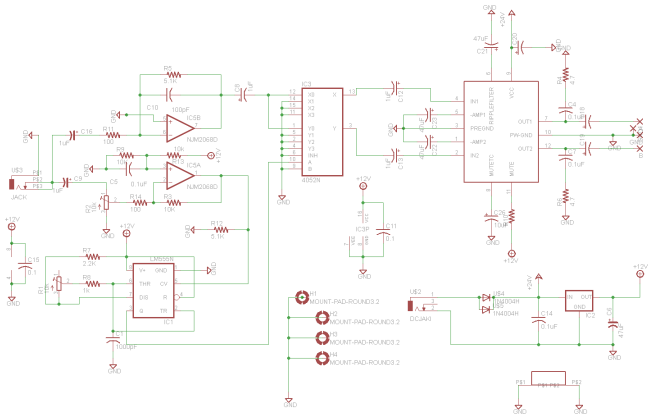

Figure 3. The schematic of the driver board. Image courtesy of Miura [23]

control software. We now describe these elements in further detail.

Ultrasonic Array

The "loudness" of the ventriloquy effect is chiefly a function of the number, size, drive voltage, and quality of the ultrasonic transducers [31]. For this reason, we present two different form factors (Figure 2): a compact array for indoors and a larger one for louder and outdoor applications. Both units use 40 kHz ultrasonic transducers, which are common and low cost ($0.25/pc), connected in parallel and in phase. Our smaller unit is octagonal, measuring 86 mm across and uses a honeycomb arrangement of 48, 10 mm transducers and can be purchased from [23]. Our larger unit is a custom piece: a 330 mm x 230 mm rectangular array of 16 mm transducers. We sourced our larger transducers from [1]. The smaller transducers cannot be driven above $35V_{pp}$ without damage, but the larger transducers can go to $100V_{pp}$. Both transducers draw very little current due to their high impedance. The arrays operate independently as standalone devices.

Signal Generation

To produce a parametric interaction effect, a 40 kHz carrier wave is pulse width modulated (PWM) with an input signal [31]. A Class-D amplifier (without the final low-pass filter stage) can be used for this purpose [40, 24]. We adopted a MOSFET-based design from [24] that used PWM to create our modulated signal. The schematic for the driver board (Figure 3) and links to purchase can be found [23]. We used a custom driver board, that produces a similarly modulated signal, to drive our larger parametric speaker array. Our parametric speakers are connected to their driver hardware via electrically shielded cables, as our modulated signal can interfere with the servo motors. Our driver hardware scales its output relative to its input voltage (i.e., the greater input voltage, the louder the effect).

Pan-Tilt Platform

For our pan-tilt platform, we use an off-the-shelf Servo City PT785-S [37], which consists of two PWM controlled servos,

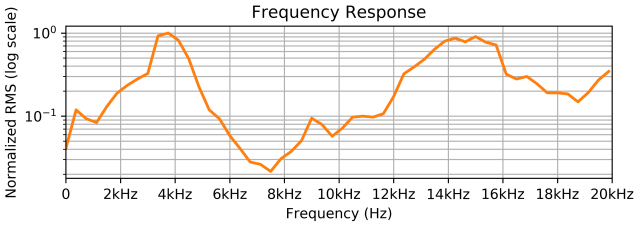

Figure 4. The pre-adjusted frequency response of our parametric speaker from 0 to 20 kHz.

permitting full 360° movement in two axes. We control these servos with an Arduino UNO, connected to a laptop and wall power.

Control Software

Our hardware is driven by a 2017 MacBook Pro 15". The 3.5mm audio jack on the MacBook provides the audible component to our signal generator board and a USB-Serial connection sends commands to an Arduino for servo control. For speech synthesis, we use MacOS's built-in Text-to-Speech engine, to which we apply a custom Equalizer (EQ), discussed next.

Sound Adjustment, Equalization, & Volume Control

Due to non-linearities in the parametric interaction effect, there are corresponding non-linearities in the frequencies that we are able to render onto objects. In our studies, we found a sharply decreased frequency response starting from 5 kHz to 12 kHz (Figure 4). In the frequencies outside of that region, there is significant "peakiness". In response, we took the following approaches to address this limitations. We can strategically select sounds that are outside of this "hole", such as choosing a synthesized male voice, which spans from 100 Hz to 8 kHz (including harmonics), compared to a female voice which spans 350 Hz to 17 kHz (including harmonics) [36]. We can also pitch shift sounds into the two ideal frequency regions. Finally, we can use an equalizer to flatten the non-linear response. In our implementation, we use a male voice ("Tom" in MacOS Speech) and use equalizer software (Boom 2 [5]) to flatten the response.

It is also important to control volume, as Digital Ventriloquism could be distracting to others in the environment. To control

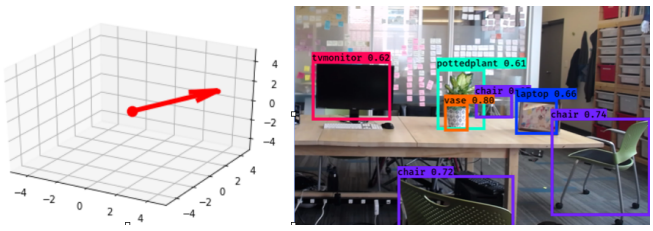

Figure 5. We used a finger snapping interaction and two microphone arrays to capture a 3D vector to an object (left). YOLOv3 provides computer vision recognition and bounding boxes of everyday objects (right).

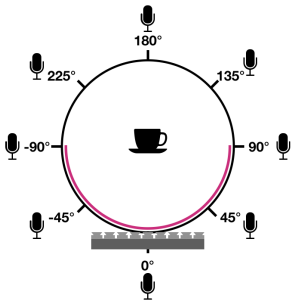

Figure 6. For our physical studies, we placed the object in the center and the parametric speaker at 0°. We then recorded frequency sweeps at eight angles. For our "real world" studies, where room layout prevented 360° access, we only recorded at angles denoted in pink.

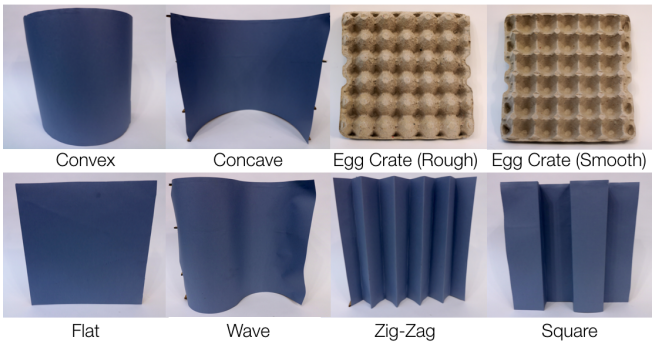

Figure 7. Our eight surface geometry exemplars.

the volume, we can either 1) modify the volume of the input signal or 2) modify the voltage of input power. We control the volume using the former, as it is significantly easier to programmatically control the input volume rather than power supply voltage. In the Video Figure, we drive the small array as loud as possible to demonstrate how pronounced the ventriloquy effect can be, but similar to traditional speakers, a user can easily select a desired volume.

Object Discovery

Before sound can be rendered onto an object, our system must know where it is located in space relative to the smart speaker (in our case, the phi and theta of our pan-tilt platform). The simplest approach would be for users to manually add the locations and labels for each object they wished to expose to the system, though obviously this is tedious. Thus, we implemented two other methods that are more automated.

In our first method, we ask the user to state the name of the object (e.g. "coffee machine"), and then repeatedly snap their fingers or clap their hands directly above (or next to) the object. We use these sounds and acoustic Directional of Arrival (DoA) to automatically determine the 3D vector of the object with respect to the smart speaker. As a proof of concept (Figure 5, left), we used two perpendicular, four-channel microphone arrays (Respeaker's XYZs [34]) connected to a Raspberry Pi, and the Respeaker's DoA API [35].

Our second method is fully automated. We use a camera mounted to our pan-tilt rig to raster scan the environment (Figure 5, right), identifying objects for ventriloquism augmentation with computer vision. As a proof of concept, we step in 10° increments in both axes: vertically from -40° below the horizon to straight up (+90°), and the full 360° horizontally. This yields 504 images, on which we run YOLOv3 object detection [33]. Any object bounding box that intersects with the center of the image is recorded with its label, along with the servo values for later use.

PHYSICAL STUDIES

To gain a better understanding of the capabilities of our system, we conducted a series of physical studies evaluating the frequency response and reflected power of different materials and surface geometries. Additionally, we performed a "real world" study looking at common objects in office, domestic, workshop, and outdoor environments. We used the smaller array for all physical studies with the exception of the outdoor real objects study, in which we used the larger array.

Materials Study

Procedure

In this study, we created 40 cm x 40 cm square sheets of nine common materials ranging from 1/16" to 1/4" in thickness: acrylic, concrete, drywall, foamboard, glass, foam insulation, paper, steel, and wood. We placed the square of material in the center of a circle (radius = 2 m) and placed the parametric speaker directly perpendicular. We placed a microphone at 8 different angles around the object (see Figure 6). We then performed a linear frequency sweep from 0-20 kHz and captured the audio reflected from the object. Note that our frequency sweep is of the input signal and not the carrier frequency, which is always 40 kHz. This was in order to 1) measure the frequency response of our speaker (Figure 4) and 2) to calculate the RMS for each of the materials/geometries (Figures 8 and 9). We did not want to allow for a scenario where a selected frequency would by happenstance work better for one geometry/material/item than the others, and thus a sweep was decided to be most fair. As a method of quantifying the performance of the ventriloquy effect, we performed a dynamic FFT bandpass filter on the captured reflection (such that only the sweep frequency remains) and calculated the RMS of the signal.

Results

We found that most materials behaved relatively similarly in "off" angles (i.e., not head on or directly opposite). There are differences at 0° and 180°, which seem to roughly correlate to material stiffness. Figure 8, top, shows the reflective power.

Geometries Study

Procedure

In our geometries studies, we constructed paper artifacts with 40 mm x 40 mm cross sections (identical to our previous study). We evaluated nine different geometries: Concave, Convex, Egg Crate (rough side), Egg Crate (smooth side), Flat, Wave, Zig-Zag, and Square (Figure 7). We utilized the same recording and frequency sweep procedure as the previous study (radius = 2 m, 8 angles).

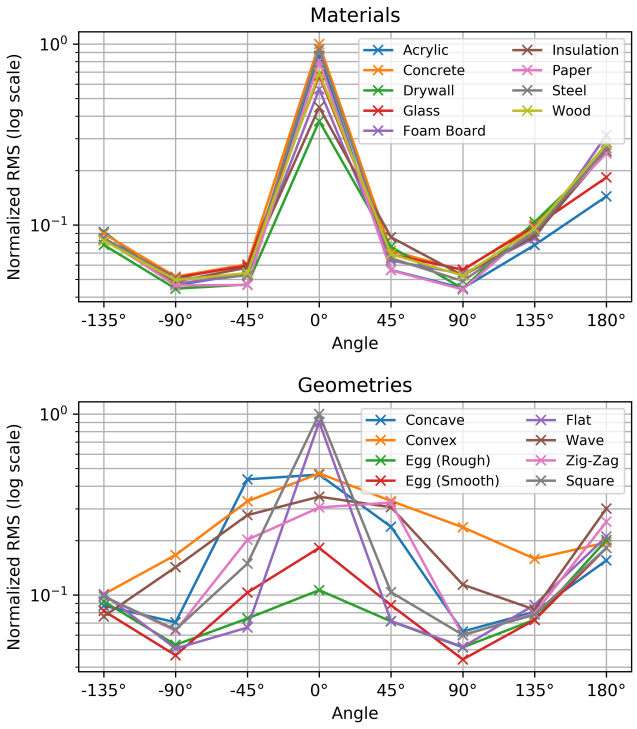

Figure 8. The reflected acoustic power at different angles across our various test materials (top) and geometries (bottom).

Results

We found that geometry had a particularly strong influence on reflected emission. Flat and orthogonal surfaces, such as flat and square, performed particularly strongly head on and more complex geometries, such as convex and concave, performed better at off angles. Unsurprisingly, egg crate (both textures), which is commonly used in sound deadening, performed poorly at all angles. Figure 8, bottom, shows the reflected acoustic power.

Real Objects Study

Procedure

In our real world objects study, we evaluated 24 objects across four environments (office, domestic, workshop, and outdoors). In the office environment, we looked at objects that would be typically found on an office desk. We placed these objects in the center of an office desk and placed the speaker at 0° and 1 m away with the microphone being placed at 5 angles (Figure 6, in pink). We then performed a similar procedure in a domestic environment, except placing both the microphone and speaker 2 m away. For the workshop and outdoors environment, the microphone and speaker were placed 3 m away. For all of these studies, we did not artificially control background sound (e.g., distant conversation, HVAC hum, general office noise), but we also did not turn on especially noisy equipment.

Results

In the office environment, we found that large objects (e.g., computer monitors) and those with complex geometries (e.g., plants) produced a stronger ventriloquy effect at non-frontal angles. Smaller objects performed similar to the desk itself, as it contributes most of the reflective surface. In our domestic, workshop, and outdoors environments, we found that object material and geometry played a significant role in reflected power. For example, the dishwasher, laser cutter, and parking sign (which have a large flat metal surfaces) have significant head-on reflections, but are worse off axis. Conversely, objects with more complex geometries, including asymmetries, had better performance at off angles. It is important to note that while the ventriloquy effect is significantly weaker at off angles, the sound is still intelligible. The reflected power for objects in each of our four contexts can be seen in Figure 9.

USER STUDIES

We also performed a human perception study to better quantify the performance of the ventriloquy effect: both in localization ability and intelligibility of the rendered audio. We recruited 5 participants (1 female, mean age = 25) and performed a perception task across three generic environments: office, domestic, and workshop. The study took approximately 15 minutes per environment and participants were paid $20 USD for their time.

Procedure

For each environment, participants were asked to sit or stand at a specified location. Five objects characteristic of that environment were distributed in front of the participant. Depending on the environment, the parametric speaker was either placed adjacent or above the participant, but in a manner in which the participant could not observe the speaker or be in its line line of sight. We did not control for environmental factors such as distant people talking, HVAC noise, etc. Object order was randomized, and our speaker re-aimed itself each trial.

To test intelligibility, we wanted a controlled, random and unbiased set of words that could pair with randomly selected objects. For this, our study software randomly generated a number (from 0 to 9) and a color (black, blue, green, purple, red, white, or yellow) and used speech-to-text to announce the number-color pair.

In each trial, participants were tasked with pointing out the object from which the sound emanated, and repeating out loud the number-color pair that they heard. Each object was repeated 3 times, for a total of 15 instances per environment. In all environments, the input voltage was set to 50% and the input was set to 50% of our MacBook's 3.5mm output volume. We felt that this provided adequate volume in all of our contexts.

Test Environments

We selected three physical contexts – office, domestic, and workshop – to evaluate our system's resolution (i.e., can objects placed closely together be distinguished?) but also its range (i.e., can objects far away speak and be intelligible?). The office environment represents close-range/high-density, the domestic medium-range/medium-density, and the workshop represents long-range/low-density.

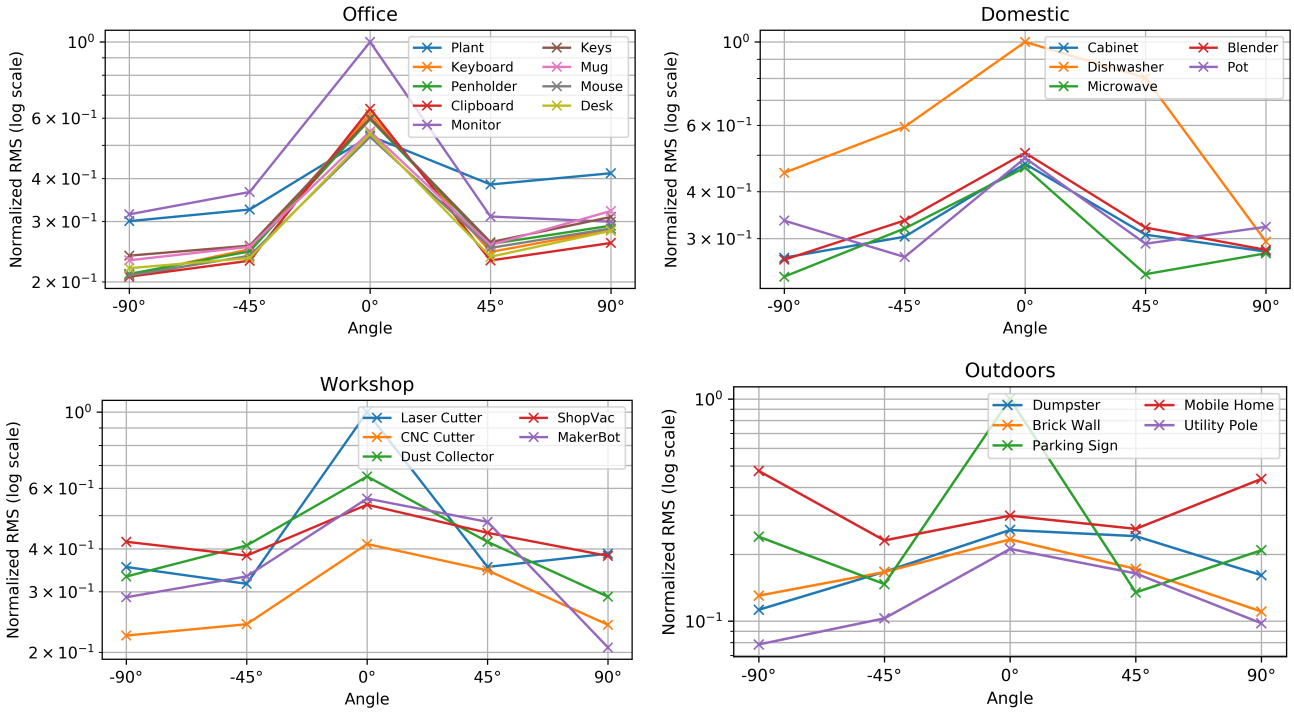

Figure 9. The reflected acoustic power at different angles for objects in each of our "real world" contexts.

For the office environment, five objects (printer, paper tray, monitor, plant, picture frame) were placed creating a typical desk setup (Figure 10, left). Participants were asked to sit facing the desk. For the domestic environment, five objects (dishwasher, cabinet, coffee maker, microwave, refrigerator) were selected (Figure 10, center. Participants were asked to stand centered in front of the kitchen countertop. For the workshop environment, five objects (CNC, fire extinguisher, laser cutter, storage cabinet, and trash can) were selected (Figure 10, right). In this context, participants were asked to stand in the center of the room. Figure 10 presents top-down, schematic views of our test objects relative to participants (P), denoted in green.

Results

Across all three environments, we found that participants were able to localize the object correctly 92.0% (SD = 6.8%, chance = 20%) of the time. Of the 8% error, 81.1% were off-by-one errors (i.e., the object identified was immediately adjacent to the target object). Participants were able to correctly identify both the number and the color 100% of the time (SD = 0.0%, chance = 1.4%). Individual confusion matrices for office, domestic, and workshop environments can be found in Figure 11.

LIMITATIONS & FUTURE WORK

There are important limitations of Digital Ventriloquism, which we discuss in this section along with potential avenues for future work. The most immediate limitation for our approach is that it requires line of sight in order to operate. While ultrasound can pass through thin fabrics and materials, it cannot pass through walls or large objects. In a similar vein, the illusion of Digital Ventriloquism breaks if a a user walks in the path of the ultrasound beam. In this case, the signal demodulates on them, rather than on the target object. Prior work has looked into "bouncing" ultrasound off of surfaces in the environment, such that it arrives to the user using ray tracing [25]. However, in our experience, the sound appears to be coming from the first surface of contact rather than the target object. It might be possible to use second-order modulated signals (i.e., a second carrier signal is used such that the first impact demodulates into a 40 kHz modulated signal). More immediately, the line of sight limitation could be overcome by using several arrays distributed strategically in an environment to provide good coverage and alternative viewpoints should users occlude line of sight.

There are also some noteworthy surface and geometry limitations, such as absorptive surfaces and materials (as we found with egg-crate and foam in our first study) that do not reflect back sufficient demodulated signal or scatter the demodulated signal in ways that do not reach the user. A possible solution is to use a more powerful system.

Another limitation of Digital Ventriloquism is the limited frequency response in the audible range. As so, the generated audio does not sound entirely natural and cannot be used as a drop-in replacement for traditional speakers. While adjusting EQ improves the audio quality, a more complete solution would involve using higher frequency transducers and modulating the signal using true AM modulation rather than a PWM

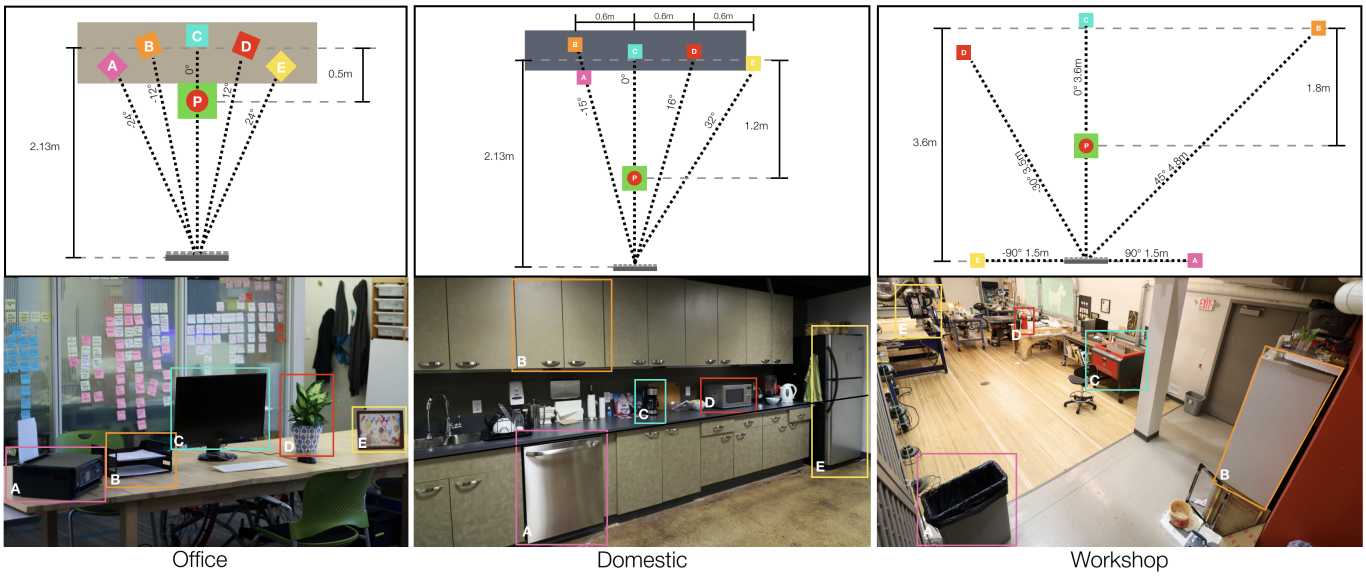

Figure 10. The arrangement of objects for each of our experimental contexts.

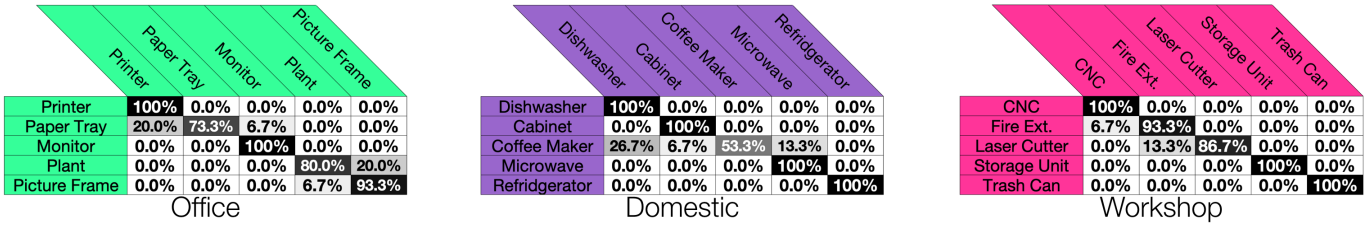

Figure 11. Localization confusion matrices for each of our contexts.

approximation. Pompei was able to achieve 1% distortion in audio quality using 80 kHz transducers, which are significantly more expensive than the common 40 kHz transducers that we employed.

While our current implementation uses an impractical pan/tilt apparatus to direct the ultrasonic beam, previous work has shown that ultrasound can be beamformed in both direction and focal length using arrays. Thus we envision a future implementation of Digital Ventriloquism utilizing a skin of transducers to steer a directed Digital Ventriloquism beam without any moving parts: our mockup presents one such form factor (Figure 12).

For practical reasons (inclement weather and a noisy outdoor environment), we were unable to perform a controlled user study outdoors. However, we do not see our system being solely limited to indoor use. Ultimately, how well a person can hear the ventriloquy effect is chiefly a function of the Signal-to-Noise Ratio (SNR). In this regard, Digital Ventriloquism is no different than conventional speakers and can be about as "loud" as traditional speakers. In future work, larger arrays specifically designed for outdoor applications could be created to explore novel streetscape interactions, such as talking cross walks, mailboxes, parking meters and storefronts.

DISCUSSION

Digital Ventriloquism intends to illuminate a new and interesting use of audio. It is not intended to be a better speaker, but rather a different type of speaker. In particular, we believe a ventriloquism approach has unique benefits with respect to embodiment and immersion that traditional speakers cannot offer. Apart from our formal studies, we often found that the simplest use cases turn out to be the most compelling. One item that created an amazing amount of surprise and delight was our test plant, which reminded people to water it. Another was our picture frame, which would "replay" brief stories. Anecdotally, we found the more "stupid" and analog the item (i.e., not digital and not powered), the stronger the user reaction. Colleagues who experienced Digital Ventriloquism during development would ask, "Can you make this [item] talk too?"

Beyond granting the ability for inanimate objects to talk, there is a deeper significance to Digital Ventriloquism. The distribution of intelligence to everyday objects improves the interaction with assistive voice agents. While we can ask Alexa "Do I need to water the plant?", it is a much more natural

Figure 12. Rather than using a mechanical pan/tilt platform, a future Digital Ventriloquism implementation could utilize a skin of ultrasonic transducers to beamform, as seen in this mockup.

interaction to ask the plant directly "Do you need watering?". Without Digital Ventriloquism, the user's expectation for the plant to reply would be left unfulfilled. When *all* the objects in the home have this ability, the intelligent agent can easily shift between the user's attention and the periphery and not restricted to its embodiment as a smart speaker.

CONCLUSION

We have presented Digital Ventriloquism, a method to allow smart speakers to render audio onto everyday objects; enabling them to become smart voice assistants. Digital Ventriloquism uses directed ultrasonic beams that, when modulated with an input signal, are inaudible in flight and demodulate when striking a surface, allowing the sound to emanate from the target rather than the speaker itself. We evaluated this method with a series of physical studies, characterizing the parametric effect and audible response in controlled and real world environments. We then performed a user study, which evaluated localization and intelligibility of speech, and found promising results. While parametric audio does not aim to replace traditional speakers for music and entertainment, a ventrilioquism approach has unique benefits with respect to embodiment and immersion that traditional speakers cannot offer. We hope this paper spurs future work in Digital Ventriloquism and uses of acoustic parametric interaction in HCI.

REFERENCES

[1] AliExpress. 2019. Ultrasonic Transducers 40kHz. Website. (20 September 2019). Retrieved September 20, 2019 from https://www.aliexpress.com/item/100PCS-50-Pairs-16MM-Ultrasonic-Sensor-Probe-Transceiver-Receiver/32814891738.html?spm=a2g0s.9042311.0.0.7ed54c4dmffnNW.

[2] Benjamin B. Bederson. 1995. Audio Augmented Reality: A Prototype Automated Tour Guide. In *Conference Companion on Human Factors in Computing Systems (CHI '95)*. ACM, New York, NY, USA, 210–211. DOI: http://dx.doi.org/10.1145/223355.223526

[3] Tom Carter, Sue Ann Seah, Benjamin Long, Bruce Drinkwater, and Sriram Subramanian. 2013. UltraHaptics: Multi-point Mid-air Haptic Feedback for Touch Surfaces. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 505–514. DOI: http://dx.doi.org/10.1145/2501988.2502018

[4] Anthony Carton and Lucy E. Dunne. 2013. Tactile Distance Feedback for Firefighters: Design and Preliminary Evaluation of a Sensory Augmentation Glove. In *Proceedings of the 4th Augmented Human International Conference (AH '13)*. ACM, New York, NY, USA, 58–64. DOI: http://dx.doi.org/10.1145/2459236.2459247

[5] Global Delight. 2019. Boom Mac Audio Equalizer. Website. (20 September 2019). Retrieved September 20, 2019 from https://www.globaldelight.com/boom/index.php.

[6] Dimitris Grammenos, Xenophon Zabulis, Damien Michel, Thomas Sarmis, Giannis Georgalis, Konstantinos Tzevanidis, Antonis Argyros, and Constantine Stephanidis. 2011. Design and development of four prototype interactive edutainment exhibits for museums. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 173–182.

[7] Kaj Grønbæk, Karen Johanne Kortbek, Claus Møller, Jesper Nielsen, and Liselott Stenfeldt. 2012. Designing playful interactive installations for urban environments–the swingscape experience. In *International Conference on Advances in Computer Entertainment Technology*. Springer, 230–245.

[8] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019a. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 15, 10 pages. DOI: http://dx.doi.org/10.1145/3290605.3300245

[9] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019b. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 276, 13 pages. DOI: http://dx.doi.org/10.1145/3290605.3300506

[10] Kentaro Ishii, Yuji Taniguchi, Hirotaka Osawa, Kazuhiro Nakadai, and Michita Imai. 2013. Merging Viewpoints of User and Avatar in Automatic Control of Avatar-Mediated Communication. In *1st International Conference on Human-Agent Interaction, iHAI*.

[11] Kentaro Ishii, Yukiko Yamamoto, Michita Imai, and Kazuhiro Nakadai. 2007. A Navigation System Using Ultrasonic Directional Speaker with Rotating Base. In *Human Interface and the Management of Information. Interacting in Information Environments*, Michael J. Smith and Gavriel Salvendy (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 526–535.

[12] Louis Jackowski-Ashley, Gianluca Memoli, Mihai Caleap, Nicolas Slack, Bruce W. Drinkwater, and Sriram Subramanian. 2017. Haptics and Directional Audio Using Acoustic Metasurfaces. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces (ISS '17)*. ACM, New York, NY, USA, 429–433. DOI: http://dx.doi.org/10.1145/3132272.3132285

[13] Kaustubh Kalgaonkar and Bhiksha Raj. 2009. One-handed gesture recognition using ultrasonic Doppler sonar. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1889–1892.

[14] Karen Johanne Kortbek and Kaj Grønbæk. 2008a. Communicating Art Through Interactive Technology: New Approaches for Interaction Design in Art Museums. In *Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building Bridges (NordiCHI '08)*. ACM, New York, NY, USA, 229–238. DOI:http://dx.doi.org/10.1145/1463160.1463185

[15] Karen Johanne Kortbek and Kaj Grønbæk. 2008b. Interactive Spatial Multimedia for Communication of Art in the Physical Museum Space. In *Proceedings of the 16th ACM International Conference on Multimedia (MM '08)*. ACM, New York, NY, USA, 609–618. DOI: http://dx.doi.org/10.1145/1459359.1459441

[16] Bogdan Kreczmer. 2011. Gestures recognition by using ultrasonic range-finders. In *2011 16th International Conference on Methods & Models in Automation & Robotics*. IEEE, 363–368.

[17] Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng. 2018. Scene-aware Audio for 360&Deg; Videos. *ACM Trans. Graph.* 37, 4, Article 111 (July 2018), 12 pages. DOI:http://dx.doi.org/10.1145/3197517.3201391

[18] Mats Liljedahl and Nigel Papworth. 2012. Using Sound to Enhance Users' Experiences of Mobile Applications. In *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound (AM '12)*. ACM, New York, NY, USA, 24–31. DOI: http://dx.doi.org/10.1145/2371456.2371460

[19] Benjamin Long, Sue Ann Seah, Tom Carter, and Sriram Subramanian. 2014. Rendering Volumetric Haptic Shapes in Mid-air Using Ultrasound. *ACM Trans. Graph.* 33, 6, Article 181 (Nov. 2014), 10 pages. DOI: http://dx.doi.org/10.1145/2661229.2661257

[20] Loyola. 2019. Sennheiser AudioBeam). Website. (20 September 2019). Retrieved September 20, 2019 from https://av.loyola.com/products/audio/sennheiser-speaker-audiobeam.html.

[21] Katuhiro Maki, Toshiyuki Kimura, and Michiaki Katsumoto. 2010. Reproduction of Sound Radiation Directivities of Musical Instruments by a Spherical Loudspeaker with Multiple Transducers. In *Proceedings of the 9th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry (VRCAI '10)*. ACM, New York, NY, USA, 85–88. DOI:http://dx.doi.org/10.1145/1900179.1900197

[22] David AB Miller. 1991. Huygens's wave propagation principle corrected. *Optics letters* 16, 18 (1991), 1370–1372.

[23] Kazunori Miura. 2009. Ultrasonic Directive Speaker. Website. (12 August 2009). Retrieved September 20, 2019 from http://zao.jp/radio/parametric/index_e.php.

[24] Kazunori Miura. 2011. Ultrasonic Directive Speaker. Magazine. (01 March 2011). Retrieved September 20, 2019 from https://www.elektormagazine.com/magazine/elektor-201103/19559.

[25] Naoya Muramatsu, Kazuki Ohshima, Ryota Kawamura, Ooi Chun Wei, Yuta Sato, and Yoichi Ochiai. 2017. Sonoliards: Rendering Audible Sound Spots by Reflecting the Ultrasound Beams. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 57–59. DOI: http://dx.doi.org/10.1145/3131785.3131807

[26] Martin Naef, Oliver Staadt, and Markus Gross. 2002. Spatialized Audio Rendering for Immersive Virtual Environments. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST '02)*. ACM, New York, NY, USA, 65–72. DOI: http://dx.doi.org/10.1145/585740.585752

[27] Ken Nakagaki and Yasuaki Kakehi. 2011. SonalShooter: A Spatial Augmented Reality System Using Handheld Directional Speaker with Camera. In *ACM SIGGRAPH 2011 Posters (SIGGRAPH '11)*. ACM, New York, NY, USA, Article 82, 1 pages. DOI: http://dx.doi.org/10.1145/2037715.2037807

[28] Steven Neale, Winyu Chinthammit, Christopher Lueg, and Paddy Nixon. 2011. Natural Interactions Between Augmented Virtual Objects. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference (OzCHI '11)*. ACM, New York, NY, USA, 229–232. DOI:http://dx.doi.org/10.1145/2071536.2071573

[29] Nikolaos Partarakis, Emmanouil Zidianakis, Margherita Antona, and Constantine Stephanidis. 2015. Art and Coffee in the Museum. In *Distributed, Ambient, and Pervasive Interactions*, Norbert Streitz and Panos Markopoulos (Eds.). Springer International Publishing, Cham, 370–381.

[30] Sarah Perez. 2018. Smart speakers hit critical mass in 2018. Website. (28 December 2018). Retrieved September 20, 2019 from https://techcrunch.com/2018/12/28/smart-speakers-hit-critical-mass-in-2018/.

[31] F Joseph Pompei. 2002. *Sound from ultrasound: The parametric array as an audible sound source*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[32] Bhiksha Raj, Kaustubh Kalgaonkar, Chris Harrison, and Paul Dietz. 2012. Ultrasonic doppler sensing in hci. *IEEE Pervasive Computing* 11, 2 (2012), 24–29.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[34] Respeaker. 2019a. Respeaker 4-Mic Array for Raspberry Pi. Website. (20 September 2019). Retrieved September 20, 2019 from http://wiki.seeedstudio.com/ReSpeaker_4_Mic_Array_for_Raspberry_Pi/.

[35] Respeaker. 2019b. Respeaker Mic Array v2.0 - DOA (Direction of Arrival). Website. (20 September 2019). Retrieved September 20, 2019 from http://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/#doa-direction-of-arrival.

[36] SEAIndia. 2019. Human Voice Frequency Range). Website. (20 September 2019). Retrieved September 20, 2019 from http://www.seaindia.in/blog/human-voice-frequency-range/.

[37] ServoCity. 2019. PT785-S Pan Tilt System). Website. (20 September 2019). Retrieved September 20, 2019 from http://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/#doa-direction-of-arrival.

[38] Hitomi Tanaka and Yasuaki Kakehi. 2013. SteganoSonic: A Locally Information Overlay System Using Parametric Speakers. In *ACM SIGGRAPH 2013 Posters (SIGGRAPH '13)*. ACM, New York, NY, USA, Article 95, 1 pages. DOI: http://dx.doi.org/10.1145/2503385.2503489

[39] Mark Weiser. 1999. The Computer for the 21st Century. *SIGMOBILE Mob. Comput. Commun. Rev.* 3, 3 (July 1999), 3–11. DOI: http://dx.doi.org/10.1145/329124.329126

[40] Wikipedia. 2019. Class-D Amplifier. Website. (20 September 2019). Retrieved September 20, 2019 from https://en.wikipedia.org/wiki/Class-D_amplifier.

[41] Graham Wilson, Thomas Carter, Sriram Subramanian, and Stephen A. Brewster. 2014. Perception of Ultrasonic Haptic Feedback on the Hand: Localisation and Apparent Motion. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1133–1142. DOI: http://dx.doi.org/10.1145/2556288.2557033