Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Device Ecosystems

Karan Ahuja **Andy Kong** **Mayank Goel** **Chris Harrison**
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
{kahuja, akong, mayank, chris.harrison}@cs.cmu.edu

ABSTRACT
Future homes and offices will feature increasingly dense ecosystems of IoT devices, such as smart lighting, speakers, and domestic appliances. Voice input is a natural candidate for interacting with out-of-reach and often small devices that lack full-sized physical interfaces. However, at present, voice agents generally require wake-words and device names in order to specify the target of a spoken command (*e.g.*, "Hey Alexa, kitchen lights to full brightness"). In this research, we explore whether speech alone can be used as a directional communication channel, in much the same way visual gaze specifies a focus. Instead of a device's microphones simply receiving and processing spoken commands, we suggest they also infer the Direction of Voice (DoV). Our approach innately enables voice commands with addressability (*i.e.*, devices know if a command was directed *at them*) in a natural and rapid manner. We quantify the accuracy of our implementation across users, rooms, spoken phrases, and other key factors that affect performance and usability. Taken together, we believe our DoV approach demonstrates feasibility and the promise of making distributed voice interactions much more intuitive and fluid.

Author Keywords
Speaker orientation, addressability, voice interfaces.

CSS Concepts
• **Human-centered computing~Human computer interaction (HCI)**; User studies;

INTRODUCTION
Where a person is looking is an important social cue in human-human interaction, allowing someone to address a particular person in conversation or denote an area of interest. For several decades, human-computer interaction researchers have looked at using gaze data to ease and enhance interactions with computing systems, ranging from social robots [56] to smart environments [8]. However, to capture gaze direction, special sensors must either be worn on the head [15][57] (unlikely for consumer adoption) or external cameras are used [3][37] (which can be privacy invasive).

UIST '20, October 20–23, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7514-6/20/10.
https://doi.org/10.1145/3379337.3415588

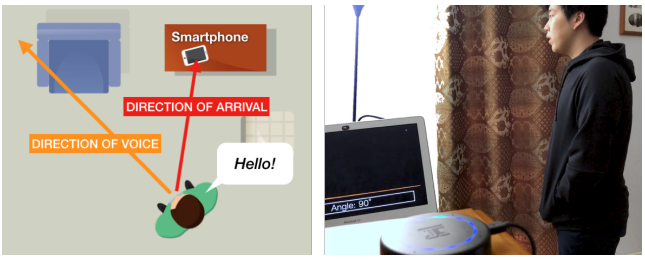

Figure 1. Left: an illustration of Direction of Voice (DoV) vs. Direction of Arrival (DoA). Right: our approach allows devices ("smartspeaker" in foreground) to compute DoV (debug output shown on laptop) without any external sensors, which could enable more robust and natural voice interactions, especially in contexts with many devices.

In this research, we explored the use of speech as a directional communication channel. In addition to receiving and processing spoken content, we propose that devices also infer the *Direction of Voice* (DoV). Note this is different from Direction of Arrival (DoA) algorithms (Figure 1, red), which calculate *from where* a voice originated. In contrast, DoV calculates the direction *along which* a voice was projected (Figure 1, orange).

Such DoV estimation innately enables voice commands with addressability, in a similar way to gaze, but without the need for cameras. This allows users to easily and naturally interact with diverse ecosystems of voice-enabled devices, whereas today's voice interactions suffer from multi-device confusion (illustrated in Figure 2A). With DoV estimation providing a disambiguation mechanism, a user can speak to a particular device and have it respond; *e.g.*, a user could ask their smartphone for the time (Figure 2B), laptop to play music (Figure 2C), smartspeaker for the weather, and TV to play a show. Another benefit of DoV estimation is the potential to dispense with wakewords (*e.g.*, "Hey Siri", "OK Google") if devices are confident that they are the intended target for a command. This would also enable general commands – *e.g.*, "up" – to be innately device-context specific (*e.g.*, window blinds, thermostat, television).

As we will discuss in greater detail, our approach relies on fundamental acoustic properties of both human speech and multipath effects in human environments. Our machine learning model leverages features derived from these phenomena to predict both angular direction of voice, and more coarsely, if a user is facing or not facing a device. Our software is lightweight, able to run on a wide variety of consumer devices without having to send audio to the cloud for processing (helping to preserve privacy).

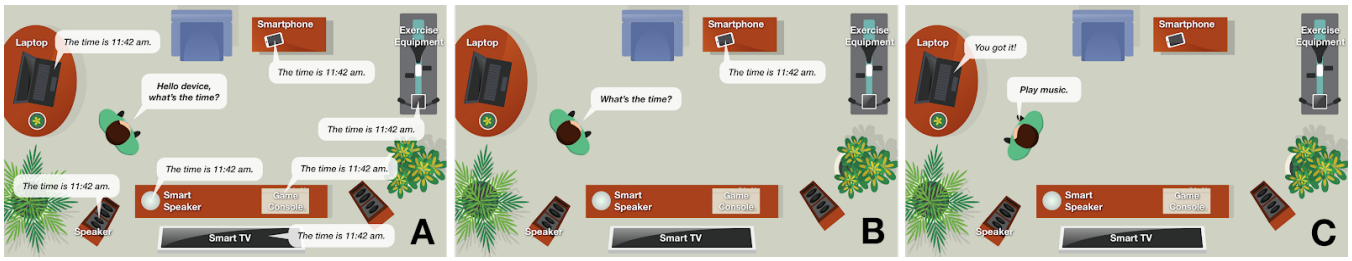

Figure 2. Future smart homes and offices are envisioned to contain many "smart" devices able to respond to voice commands. However, without device-specific wakewords, multiple devices may try to respond to generic queries (left). Ideally, users would be able to face and speak *to* a device, more akin to human-human interaction (center & right). Thus, there is a need for Direction-of-Voice estimation approaches, especially those than can run locally on self-contained devices, without having to install extra sensors in the environment or rely on multi-device interoperability, which does not appear to be forthcoming in the near future.

RELATED WORK

Our approach intersects with several different literature, including voice interfaces generally, as well as direction of arrival estimation and speaker orientation estimation.

Voice Interfaces

HCI researchers have long studied voice interfaces, dating back to seminal works such as "Put-that-there" [8]. For a comprehensive review, we recommend [41][42][43][52]. In recent years, voice interfaces have entered the mainstream in the form of smartphone assistants (*e.g.*, Apple Siri, Google Assistant) and smart speakers (Google Home, Amazon Alexa, and Apple Homepod).

Before understanding and executing a user's command, these voice interfaces need to infer whether the user is speaking to them. To help minimize false positive interactions and resolve target ambiguity, researchers have explored mechanisms such as wakewords [23] and approaches that localize speakers and estimate the direction of arrival of sounds, which we discuss in greater detail next.

Direction of Arrival and Speaker Localization

Estimating the direction of arrival of sounds dates as far back as World War I for detecting and tracking aircraft (with human observers listening via two or more large sound horns). Major approach categories include angle of arrival (AoA), time difference of arrival (TDOA), and frequency difference of arrival (FDOA), which work across radio, optical and acoustic signals (see [18][26] for an in-depth review). Most techniques use multiple receivers and cross-correlation algorithms [5][27][39], though non-heuristic deep learning techniques are also possible [14][21][54]. When sensors can be distributed in an environment (*i.e.*, not limited to the confines of one small device), techniques such as bi-channel sound-source localization (SSL) [30][31], multilateration [51][53] and multi-channel SSL [29][35] can be employed to localize the position of a signal source.

More related to our present work are Direction of Arrival (DoA) algorithms. These generally use many microphones operating together as an array, but contained in a single device (*e.g.*, a smartspeaker). DoA is inherently different from Direction of Voice – one way to conceptualize the difference is with a user standing in one position several meters away from a device but rotating their body to speak at different angles. In this case, the DoA will always be the same value, but the DoV will change.

DoA and speaker localization unlock many interesting applications. For example, they can enable social robots to turn and reply to different users [32][44]. In multi-user collaborative scenarios, such as meetings, they can enable external devices such cameras to face the active speaker [59][61] and also enhance the quality of speech audio [24]. DoA is also intrinsic to solving source separation [4][40] and speaker diarization [38] that enable a plethora of applications on their own (see *e.g.*, [10][17][22]).

Speaker Orientation

Most similar to this research are prior systems that infer speaker orientation, most often as additional metadata for speaker localization systems [36]. Almost all prior work uses multiple microphones *distributed around a room*. These include works such as [28][34][46] that make use of large microphone arrays (scores of microphones per room) to estimate a speaker's yaw orientation. Works such as [2] and [49] show the number of microphones can be reduced by using a few small, T-shaped microphone arrays distributed in a room of known geometry. The latter two systems achieved an accuracy of 37.3% and 76.8%, respectively, for predicting speaker head orientation in 1 of 8 classes (*i.e.*, 45° segments).

Most closely related to our approach are [33][54][60], which use a single, small microphone array placed in a room. Of note, [33] and [54] used loudspeakers to generate sounds (not humans) and data collection and testing are limited to a single room, limiting generalizable insights. On the other hand, [60] is trained/tested across rooms and real humans (as is our system) offering a better estimation of real-world accuracy. The latter system serves as an excellent complement to our work in that the technical approaches are orthogonal. While [60] leverages deep learning (CNN-LSTM), we pursue a featurization approach informed by the physics of sound propagation and the human vocal tract.

THEORY OF OPERATION AND ML FEATURES

Our development and investigations followed a principled approach, requiring an understanding of the fundamental physics involved (as opposed to a more "black box" machine

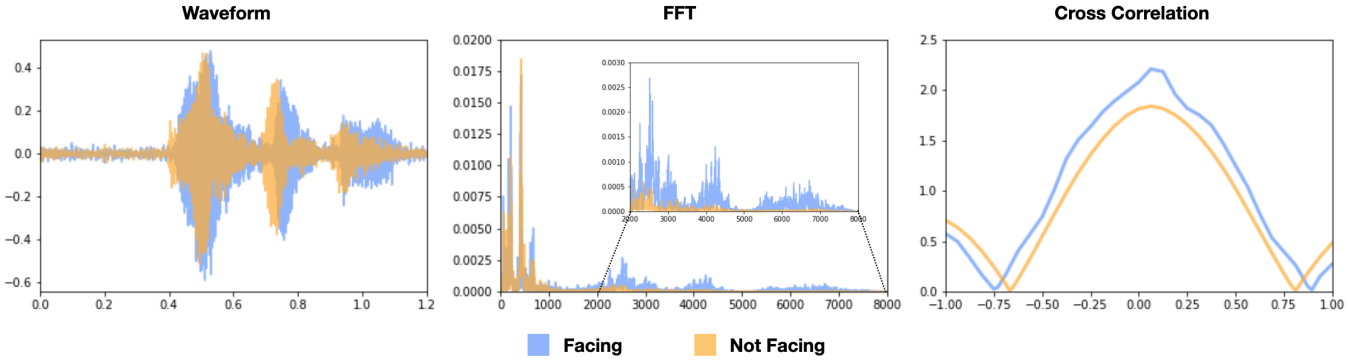

Figure 3. Example waveform, FFT, and cross correlation for an example "facing" and "not facing" trial (all other factors same).

learning approach). This offers the benefit of interoperability of results, and more importantly, offers generalizable insights for future researchers and practitioners to apply in their own systems.

Voice Frequency Distribution

The first property we leverage is that the distribution of human speech frequencies varies by spoken angle. There are two key effects:

Foremost, due to the complex operation and geometry of the human vocal tract, voice frequencies are not uniformly distributed across the acoustic "field of view" even just outside the mouth [16]. Specifically, higher frequencies are more rapidly attenuated off-axis. Secondly, and closely related to the previous effect, is the fact that once voice has left a speaker's mouth, higher frequencies are more directional [7], carrying with greatest amplitude in their emitted direction, while lower frequencies spread out in a more omnidirectional fashion.

Both of these effects come together and manifest as a characteristic imbalance between high- and low-frequency voice bands, ("high-low band ratio", or HLBR, in the literature [1][48]). Put simply, if a voice is directed at a microphone (*i.e.* facing), high and low voice frequencies are present. However, if we receive a sound when a user was facing another direction, or if the sound has had to echo to reach the microphone, we typically see reduced high frequencies compared to low frequencies. An illustrative example can be found in Figure 3.

More specifically, we calculate the following features to capture aspects of HLBR: sum power of low frequencies (< 7 kHz), sum power of high frequencies (> 7 kHz), ratio of the latter two values, the 4 coefficients of a three-degree polynomial fit to a 128-bin FFT, and the 2 coefficients of a linear regression fit to a 128-bin FFT. These features alone are not sufficient for robust DoV estimation, as HLBR effects are most apparent when comparing two signals when all other factors are held constant (such as speaker, distance and utterance). Nonetheless, it offers useful and complementary information when used in concert with the other machine learning features we employ, described next.

Crispness of First Wavefront

Built environments introduce characteristic multipath effects [7]. When a user speaks towards a device, the first, loudest, and least-distorted signal to arrive is the original sound, which took a direct path (*i.e.*, shortest path and time). All subsequent signals are echoes (assuming single speaker, a limitation we discuss later), having scattered off of various surfaces in the environment – these signals are delayed, quieter, and more distorted. This effect is apparent to the naked ear, and we encourage the reader to play our Video Figure with headphones. Figure 3 offers a visual example.

To take advantage of this effect, we need metrics that capture the "crispness" (*i.e.*, "reverblessness") of a signal. First, and most straightforward, is to run autocorrelation on the sound. If we receive a "facing" sound that came first and directly to the microphone, there should be no echoes on which to correlate against in the first few milliseconds of audio. However, if the sound has bounced and scattered off of other surfaces before reaching the microphone, we typically see duplicated overlapping waveforms, which manifest as small peaks in the autocorrelation. To capture this for machine learning, we use the ratio of the max peak and the average of all other peaks within ±10ms, the ratio between the max peak and average of the next highest nine peaks, the standard deviation and area under the curve of the autocorrelation, and the standard deviation and area under the curve of the absolute first derivate autocorrelation. As an additional measure of reverb, we calculate the speech-to-reverberation modulation energy ratio (SRMR) [20].

Our smartspeaker-esque test device contains four microphones, offering six pairings on which to run more advanced correlations. For this we use Generalized Cross-Correlation with Phase Transform (GCC-PHAT [9][25]), which has been shown to be more robust to DoA, reflections, reverberations and noise [5][48][49]. We compute all six GCC-PHAT correlations (*i.e.*, all pairs of our four microphones) and crop ±0.236ms (which represents the maximum theoretical delay based on the distance between orthogonal microphones) around the max peak. We then take the raw GCC output as features, as well as the max peak value, max peak index (*i.e.*, delay between channels), and area under the curve. For the latter three values, we compute

the standard deviation, range, and mean across all six GCC-PHAT correlations. We also use GCC-PHAT to calculate TDOA as another feature.

IMPLEMENTATION

Our test hardware consists of a ReSpeaker USB 4-channel microphone [45] made by Seeedstudio (visible in foreground of Figure 1). We use a MacBook Pro with 16 GB of RAM and a dual-core Intel i5 processor @ 3.1G Hz for audio processing and classification. We configured the ReSpeaker to transmit five audio streams at 48 kHz sampling rate to our laptop over USB. Four of the channels are raw microphone audio, with the fifth channel containing processed audio for Automatic Speech Recognition (ASR).

We use a Python backend for data collection, signal processing, and machine learning. For our machine learning algorithm, we use an Extra-Trees Classifier (sklearn implementation with 1000 estimators). We train our classifier across all eight angles of the direction of voice to mitigate class imbalance and bin the output predictions based on the facing definition we are testing. This approach performed better than training with class weights for class imbalances and also provided us with a global classifier rather than individual classifiers, which would be prone to overfitting to the test condition Our ML model had an average latency of 107 ms, which included signal processing and prediction.

We tested several other classifiers, such as bagging classifiers, SVM, and neural networks and found that ensemble-based decision trees work the best. Although neural networks with deeper layers achieved similar performance on the same feature set, we chose Extra-Trees due to their computational efficiency and interpretability.

DEFINITION OF "FACING"

There is no universally accepted definition of what constitutes "facing" a direction. Clearly someone oriented perfectly towards a person or object (*i.e.*, 0° off-axis) would be considered facing, but what about angled ±10° or ±45°? Rather than select an arbitrary definition, we decided to run a small investigation to explore possible definitions. This pilot study also offered some preliminary insights into how well humans can estimate direction of voice (even at unambiguous speech angles, such as 0°).

To collect data, we recruited two pairs of participants (3 male, 1 female; mean age of 26). A "listening" participant stood in the middle of a 3x3 polar grid (distances = 1, 3 and 5m; radial angles = 0, 45 and 90°), illustrated in Figure 4. At each grid intersection (pink circles), 8 angles were marked on the floor (blue arrows), spanning 360° in 45° intervals (0° = facing straight towards the listener, depicted by black arrows in Figure 4).

Twenty random grid positions and speech angles were requested per "speaking" participant. Once situated at each location and angle, they spoke the phrase "the quick brown fox jumped over the lazy sheep". The "listening" participant then stated aloud whether they thought the speaker was "facing" or "not facing" them. (In an even earlier pilot experiment, we had participants guess the direction of the voice vector by pointing with their fingers, but we found this procedure to be very confusing and challenging for users, and so we simplified to the binary choice of facing/not facing).

Spoken Angle	*Reported "facing"*	*Reported "not facing"*
0°	94.1 %	5.9 %
45°	63.2 %	36.8 %
-45°	61.5 %	38.5 %
90°	9.1 %	90.9 %
-90°	0.0 %	100.0 %
135°	9.1 %	90.9 %
-135°	0.0 %	100.0 %
180°	0.0 %	100.0 %

Table 1. Participant responses for a pilot study looking at human perception of what spoken angles constitute "facing".

To mitigate the listener picking up on cues other than the spoken phrase, they were blindfolded throughout the data collection period, and wore noise-canceling headphones between trials (*e.g.*, to hide footfalls and other noises generated by the speaking participant). Only when the speaking participant was fully situated and ready to speak the utterance were the headphones removed.

When all 20 trials were completed, the participants swapped "speaking" and "listening" roles and the experiment repeated. To collect more data, we had our participant pairs repeat the same procedure in another larger room with different acoustics. In total, this process yielded 160 trials.

The results of this data collection are shown in Table 1. Even when the speaking participant was directly facing the listener (0°), the listener incorrectly reported the phrase as "not facing" in 5.9% of trials. At spoken angles of ±45° relative to the listener, our participants reported this as "facing" 61.9% of the time. This result suggests that facing, at least acoustically, has some angular variance in human perception. At ±90°, just 4.6% of trials are reported as facing, a precipitous drop that suggests something distinct is happening acoustically (in line with our theory of operation). At ±135°, 4.6% of trials are reported as facing, and at 180°, no trials are reported as facing.

From these findings, we derived three working definitions of "facing":

- **Directly Facing**: 0° is facing; all other angles are not facing.
- **Forward Facing**: +45, 0 and -45° are considered facing; all other angles are not facing.
- **Mouth Line-of-Sight**: +90, +45, 0, -45, -90° are considered facing, all other angles are not facing.

To provide the most generalizable results, we report our system's accuracy under all three of these definitions in our later evaluation. We also plot system performance alongside

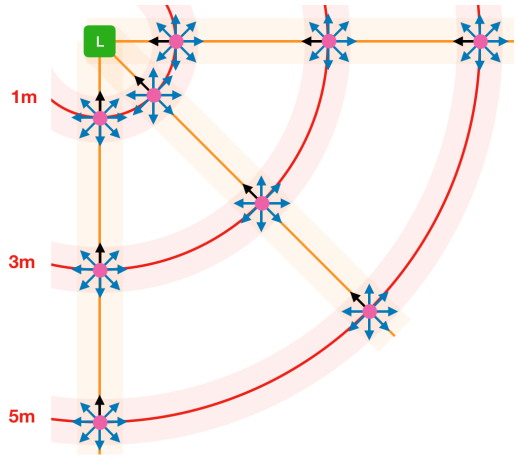

Figure 4. Illustration of data collection layout. The "listening" user or device lies in the center (green). From this point extends a polar grid at 1, 3 and 5 meters (red lines) and 0, 45 and 90 degrees (orange rays). Intersections denote the 9 positions where "speaking" participants stood (pink circles). At each location, data was collected at 8 angles (blue and black arrows). 0° is directly facing the device (black arrow), with +45, +90 +135 and 180° proceeding clockwise, and -45, -90 and -135° proceeding counterclockwise from 0°.

human accuracy under the same definitions as a useful, but obviously preliminary benchmark given our small number of participants.

EVALUATION

There are many variables that could affect the accuracy of our Direction-of-Voice estimation approach, including robustness across people, time, utterances, rooms, device placement, user position and spoken angle. In order to analyze our performance across these factors, we collected data across:

- 10 participants (4 male, 6 female, mean age 20)
- 2 sessions (run back-to-back, but otherwise independent in time)
- 2 utterances ("hey assistant" and "the quick brown fox jumped over the lazy sheep")
- 2 rooms (24.3×9.1×4.0m classroom and 13.7×6.1×3.6m office)
- 2 devices placements (<50cm and >2m away from wall)
- 3 user distances (1, 3 and 5 meters)
- 3 user polar positions (0, 45 and 90°)
- 8 spoken angles (0, +45, -45, +90, -90, +135, -135 180°)

This full factorial study design (10×2×2×2×2×3×3×8) yielded 11,520 multi-channel audio recordings (350+ mins of audio). The 3 speaker distances, 3 speaker polar positions, and 8 spoken angles followed the same design as our study in the Definitions of Facing section (Figure 4). Rather than have a human stand in the center of the polar grid, we placed our ReSpeaker microphone, connected over USB to a laptop where all recording and processing occurred.

To streamline data collection over so many combinations, the order of polar distances and positions were randomized, but then once the participant was standing at that spot, data for all 8 angles and both utterances were collected. This constituted one session of data collection, which was then repeated to provide two equivalent, but independent sessions. Finally, the above procedure was repeated for the 2 rooms, within which there were 2 device placements. In total, this study took 2 hours per participant, who were compensated $20 for their time.

OPEN SOURCE DATA

To enable other researchers to explore this domain, we have made our study data freely available at https://github.com/FIGLAB/DirectionOfVoice with the gracious permission of our participants.

RESULTS AND DISCUSSION

We designed our study procedure in order to systematically isolate and analyze different factors. We now describe our main findings, broken out into sections, using the aforementioned definitions of facing.

Machine Learning and Feature Importance

We calculated the relative importance of our machine learning features via their information entropy. We found that those responsible for characterizing the crispness of the first wavefront had the highest information entropy, followed by voice frequency distribution, and finally the echoes of sound. However, all features contribute to our global model and provide value in different physical settings.

Overall Accuracy

To estimate the overall accuracy of our system, we performed a leave-one-out cross validation by training on all data (across people, utterances, rooms, device placements, user distance/position and spoken angle) from one session and testing on the second (hold-out) session.

Using our "directly facing" definition, our system had an accuracy of 93.1% (F1 score = 0.83), which is comparable to our prior human accuracy result. Under "forward facing" and "mouth line-of-sight" definitions, our system was 92.0% (F1 score = 0.91) and 87.3% (F1 score = 0.86) accurate, respectively. Note that the test set classes are imbalanced, so it is also important to consider zero classifier accuracies, which are 87.5% (F1 score = 0.47), 62.5% (F1 score = 0.38; *i.e.*, it always chooses the majority class of not facing) and 62.5% (F1 score = 0.38, *i.e.*, it always chooses the majority class of facing) across the facing definitions respectively.

For both brevity and clarity in all subsequent sections, we only report accuracies under our "forward facing" definition. We selected this as the accuracy metric as it contains a more balanced distribution of facing vs. not facing classes and offers a middle ground metric between "directly facing" and "mouth line-of-sight" definitions. Note that full results (with all three facing definitions) can be found in Figure 5.

Cross Utterance Accuracy

To test robustness across utterances, we train our system on all data for one utterance and test on the other utterance, a similar leave-one-out protocol as used previously. Such a procedure gives us a preliminary understanding of how DoV could generalize across many phrases. We found that our

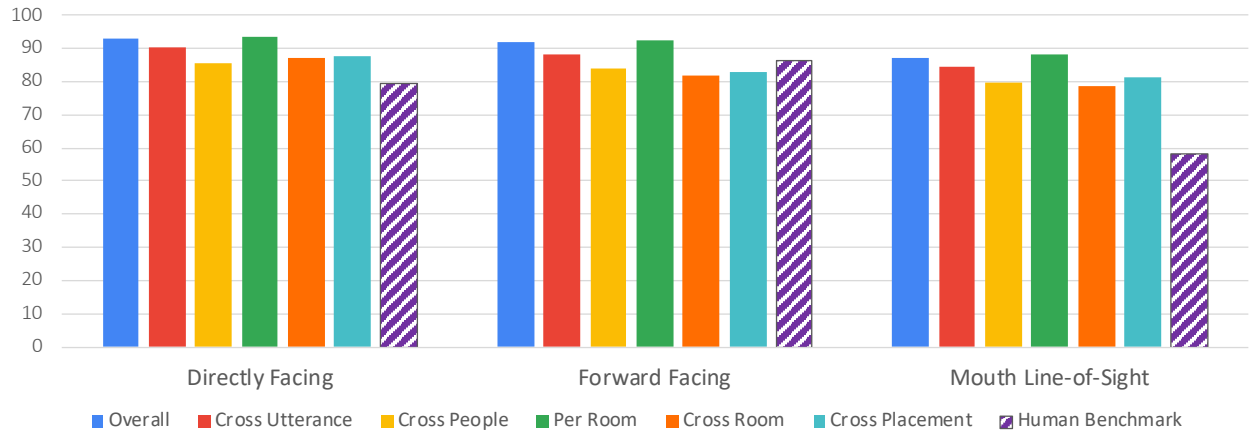

Figure 5. Accuracy (%) across three facing *definitions* and six study *factors*. We also include a human accuracy benchmark.

system achieved a facing / not facing accuracy of 88.4% (F1 score = 0.87).

As discussed in Theory of Operation, we purposely engineered our machine learning features to be utterance independent. To explore how successful we were, we ran a Linear Discriminant Analysis (LDA) for both test utterances (Figure 6). As can be seen, there is little difference in the distribution (and thus separation boundaries would be similar), a positive result suggesting generalizability.

Cross People Accuracy

We also wished to test our system's ability to generalize across people, without any prior per-user calibration. This is generally a high bar, and even highly engineered commercial systems, such as the Google Assistant, recommend providing spoken examples. To assess this, we performed a leave-one-participant-out cross validation, combining data from all other factors. Facing / not facing accuracy stands at 83.9% (SD=0.03; F1 score = 0.82). This is just 4.5% lower than our cross-utterance accuracy, where the model was able to train on a user's data (though a different utterance, which is more like the Google Assistant setup process).

Cross Room and Cross Device Placement Accuracy

To investigate accuracy across rooms and device placement, we follow a similar procedure as above: a leave-one-room-out and leave-one-placement-out cross validation, such that the model has no prior data about the room or placement it is tested in. However, we note that while smartphone assistants are mobile, smart speakers are generally stationary, and thus would have ample opportunity to build a per room model (which we discuss in the next section).

Our system achieved a cross-room accuracy of 82.1% (F1 score = 0.79) and cross-placement accuracy is 82.8% (F1 score = 0.80). We also ran an LDA (Figure 7) to help visualize how different rooms and device placements manifest in our machine learning feature set. While there are subtle differences, there is no systematic pattern, and helps explain why the accuracy results are within 1% of each other.

Per Room Accuracy

In this analysis, we trained on all session one data for a room and tested on all session two data for that room (and vice versa, averaging the results). This simulates a "calibrated" model that has in situ training data. We note that commercial devices, such as the Apple HomePod, run a calibration step to capture the impulse response of a room to improve audio processing. It is thus unsurprising that in this train/test procedure, our system achieves its highest accuracy: 92.6% (F1 score of 0.89).

Per Device Placement Accuracy

We hypothesized that placing a device near to a wall would have a deleterious effect on accuracy, introducing excessive multipath interference. Our study data let us test this theory, and the results show no significant effect. Specifically, we trained on all session one data for a device placement and tested on all session two data for that placement (and vice versa, averaging the results). We found a mean "against wall" accuracy of 91.7%, and a mean "away from wall" accuracy of 92.6%. This result matches our Cross Rooms & Device Placement Accuracy results, and also the LDA in Figure 7.

Speaker Distance and Polar Position

We found that speaker polar position had no effect on facing / not facing accuracy. This is not the case for distance, which

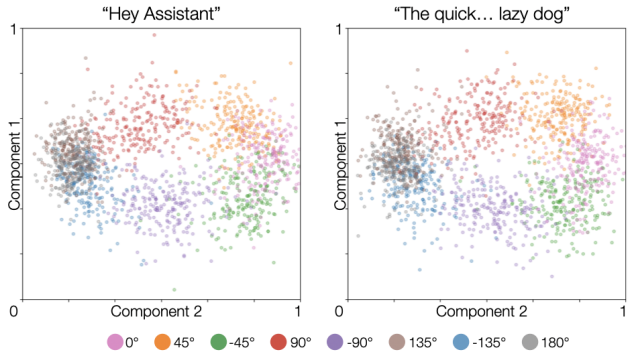

Figure 6. LDA of two utterances at 1m.

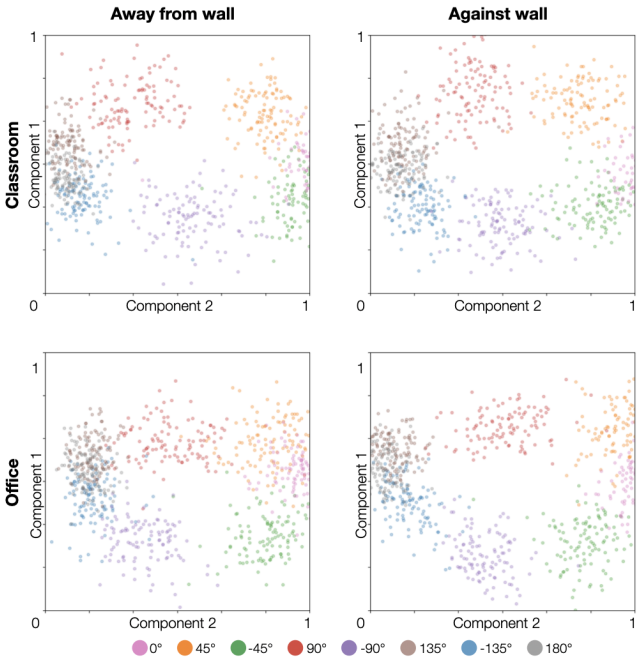

Figure 7. LDA of room × device placement at 1m.

has a prominent effect, with accuracy degrading as distance increases. To formally test this, we trained a model on all session one data for a distance, and then tested on all session two data for that distance (and vice versa, averaging the result). Results show accuracy at 1 meter away is 96.8%, dropping to 93.0% at 3m, and finally to 88.5% at 5m. Looking at information loss, we find that features derived from voice frequency distribution drop the most at longer distances. As with our other analyses, we ran an LDA (Figure 9). This time, the effect of speaker distance is pronounced, with separation between DoV classes almost non-existent at 5m (and presumably beyond).

This also underscores the potential to improve DoV accuracy by training distance-aware classifiers that make use of state-of-the-art techniques for sound source localization [58].

Overall

	0	45	-45	90	-90	135	-135	180
0	0.73	0.1	0.11	0.01	0.02	0.01	0.01	0.01
45	0.1	0.68	0.02	0.09	0.01	0.04	0.03	0.03
-45	0.16	0.03	0.65	0.01	0.08	0.01	0.02	0.03
90	0.01	0.05	0	0.62	0.01	0.22	0.04	0.05
-90	0.01	0.01	0.09	0.01	0.59	0.03	0.2	0.06
135	0	0.01	0	0.07	0.01	0.73	0.06	0.14
-135	0	0	0.01	0.01	0.07	0.1	0.65	0.16
180	0.01	0.01	0.01	0.02	0.03	0.21	0.15	0.58

Cross Utterance

	0	45	-45	90	-90	135	-135	180
0	0.65	0.1	0.16	0.01	0.03	0.01	0.01	0.03
45	0.17	0.56	0.04	0.08	0.02	0.04	0.04	0.05
-45	0.2	0.03	0.56	0.01	0.11	0.01	0.04	0.03
90	0.02	0.08	0	0.53	0.01	0.23	0.05	0.07
-90	0.04	0.01	0.11	0	0.53	0.04	0.2	0.07
135	0.01	0.02	0	0.07	0	0.66	0.07	0.18
-135	0.01	0	0.02	0	0.09	0.11	0.59	0.17
180	0.02	0.03	0.01	0.02	0.03	0.24	0.16	0.51

Cross People

	0	45	-45	90	-90	135	-135	180
0	0.63	0.11	0.14	0.03	0.03	0.01	0.02	0.03
45	0.26	0.4	0.05	0.12	0.02	0.06	0.04	0.05
-45	0.31	0.05	0.39	0.02	0.12	0.02	0.04	0.05
90	0.05	0.13	0.01	0.39	0.01	0.26	0.06	0.1
-90	0.08	0.02	0.15	0.02	0.37	0.07	0.18	0.11
135	0.01	0.02	0.01	0.1	0.01	0.57	0.08	0.19
-135	0.02	0.01	0.03	0.02	0.11	0.19	0.39	0.22
180	0.03	0.04	0.02	0.04	0.04	0.32	0.16	0.35

Per Room

	0	45	-45	90	-90	135	-135	180
0	0.76	0.08	0.1	0.01	0.02	0.01	0.01	0.01
45	0.08	0.72	0.02	0.09	0.01	0.04	0.02	0.03
-45	0.15	0.03	0.68	0.01	0.08	0.01	0.02	0.03
90	0.01	0.06	0	0.66	0	0.2	0.02	0.05
-90	0.01	0.01	0.08	0.01	0.62	0.03	0.19	0.05
135	0	0.01	0	0.06	0.01	0.74	0.04	0.15
-135	0	0	0.01	0.01	0.08	0.08	0.65	0.17
180	0	0.01	0.01	0.02	0.02	0.2	0.14	0.61

Figure 8. Confusion matrices for per angle classification across different study factors.

Per Angle Classifier

Our study data also permitted us to evaluate a model that predicts DoV angle (8 classes), and not just binary facing vs. not facing. This is a considerably harder classification task, and as noted earlier, when we attempted to give this task to participants in a pilot test, we were met with confusion and frustration. To create such a model, we trained a model on all session one data for an angle, and then tested on all session two data for that angle (and vice versa). Note that "all" means data across all rooms, people, utterances, device placement distances, and positions.

Overall, our system is able to predict DoV angle with an accuracy 65.4% (prior probability is 12.5%), exceeding our expectations, though not yet accurate enough for user-facing applications. We also ran equivalent analyses for cross utterance (57.6%), cross people (43.5%) and per room (67.9%). Confusion matrices for these results can be found in Figure 8.

We find that most of the confusion is caused by angle classes adjacent to one another (*e.g.*, 45° and 90°), but less so with symmetric angles (*e.g.*, 45° and -45°). The extreme "not facing" angles of -135°, +135° and 180° also have a lot

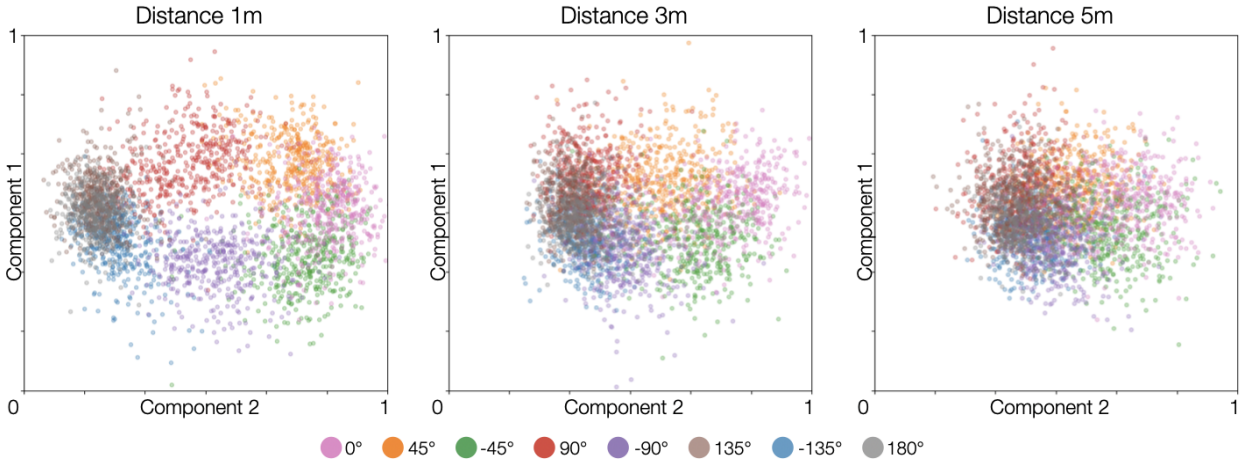

Figure 9. LDA at 1, 3 and 5m speaker distances.

of confusion, which is also visible in the LDAs in Figures 6, 7 and 9. It is likely there is not much discernable difference in the latter signals, as voice frequency features drop in value past ±90° and reverb effects from the environment will be roughly equivalent, having to bounce off opposing walls and objects before reaching the device.

In comparison to prior work, [49] achieves an accuracy of 76.8% on the same eight-angle classification task, however it makes use of six T-shaped microphone arrays (each with 4 mics) distributed across a room of known geometry. [60] uses a single compact microphone array similar to our work and reports an average orientation error of ±40° when trained and tested in the same room (cross user), and ±57° when tested in different rooms (cross user).

LIMITATIONS AND FUTURE WORK

It is important to note that while our system has promising results, there are several key limitations that will need to be overcome before consumer use. First is the accuracy of the system itself. Our model can distinguish between speakers facing and not facing a device at ~90% accuracy. If we assume a device can calibrate on data from its room and owner, accuracy rises to ~93% While approaching feasibility, it does fall short of the 99%+ accuracies consumers have come to expect, though they are perhaps more forgiving with voice interfaces. Nonetheless, 93% accuracy is the best result in the present literature and constitutes an important step towards feasibility.

Another potential use of DoV and facing / not facing recognition (if it is not used directly as a waking trigger) is to combine its confidence with wakeword confidence to increase robustness to false positive triggers. Thus if wakeword confidence is low (*e.g.*, "Hey Sarah", but not "Hey Siri"), but it was spoken with high confidence towards a device, the voice agent might still decide to activate.

We also make the basic assumption that facing a device signals intent for interaction via voice. However, this is not always the case, and infeed users can interact with voice-first interfaces while focused on another task (*e.g.*, cooking) or even from an entirely different room. That said, it is less common for users to speak directly towards a voice-first device if there is no intention to interact with it.

We also acknowledge that our current implementation does not attempt multiple speaker DoV, which would preclude its use in highly social settings (restaurant, birthday party). It may be that speaker diarization algorithms, able to isolate and extract speech at varying DoAs, could be brought to bear on this problem. Additionally, robustness in noisy environments remains a future challenge.

In the future, we hope to explore applications beyond device addressability and disambiguation. For instance, DoV could be used in social extended reality to facilitate richer communication between participants. It could also be used to enhance hearing aids by selectively amplifying speech directed at the user. We also envision DoV being used in conjunction with existing body and head pose systems (optical, RF, etc.) to boost their accuracies.

CONCLUSION

We have presented an angular direction of voice estimation system using features that leverage human voice frequency distributions and characteristic multipath effects in human environments. We comprehensively analyze the efficacy of our system across various factors such as people, time, utterance, room, device placement, user position and spoken angle. Our approach is lightweight, software-only and could run on a plethora of consumer devices without having to transmit audio data to the cloud. We also make our data open to researchers and practitioners, which we hope spurs future algorithmic efforts.

REFERENCES

[1] Alberto Abad, Dusan Macho, Carlos Segura, Javier Hernando, and Climent Nadeu. "Effect of head orientation on the speaker localization performance in smart-room environment." In *Ninth European Conference on Speech Communication and Technology*. 2005.

[2] Alberto Abad, Carlos Segura, Climent Nadeu, and Javier Hernando. "Audio-based approaches to head orientation estimation in a smart-room." In *Eighth Annual Conference of the International Speech Communication Association*. 2007.

[3] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical Classroom Sensing at Scale. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 71 (September 2019), 26 pages. DOI: https://doi.org/10.1145/3351229

[4] Dan Barry, Bob Lawlor, and Eugene Coyle. "Sound source separation: Azimuth discrimination and resynthesis." In *7th International Conference on Digital Audio Effects*, DAFX 04. 2004.

[5] Dirk Bechler, and Kristian Kroschel. "Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array." In *Proceedings of the International Workshop on Acoustic Echo and Noise Control*, pp. 315-318. 2003.

[6] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. "Understanding the long-term use of smart speaker assistants." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, no. 3 (2018)*: 1-24. DOI:https://doi.org/10.1145/3264901

[7] Leo L. Beranek. 1986. Acoustics. *American Institute of Physics*, Woodbury, NY, USA.

[8] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques (SIGGRAPH '80).* Association

for Computing Machinery, New York, NY, USA, 262–270. DOI: https://doi.org/10.1145/800250.807503

[9] M. S. Brandstein and H. F. Silverman. 1997, A robust method for speech signal time-delay estimation in reverberant rooms, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany. DOI: https://doi.org/10.1109/ICASSP.1997.599651

[10] Hervé Bredin and Grégory Gelly. 2016. Improving Speaker Diarization of TV Series using Talking-Face Detection and Clustering. In *Proceedings of the 24th ACM international conference on Multimedia (MM '16)*. Association for Computing Machinery, New York, NY, USA, 157–161. DOI: https://doi.org/10.1145/2964284.2967202

[11] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer. "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays." In *Ninth European Conference on Speech Communication and Technology*. 2005.

[12] Cristian Canton-Ferrer, Carlos Segura, Montse Pardas, Josep R. Casas, and Javier Hernando. "Multimodal real-time focus of attention estimation in smartrooms." In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-8. IEEE, 2008. DOI: https://doi.org/10.1109/CVPRW.2008.4563180

[13] Rastislav Cervenak, and Pavel Masek. "ARKit as indoor positioning system." In *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 1-5. IEEE, 2019. DOI: https://doi.org/10.1109/ICUMT48472.2019.8970761

[14] Soumitro Chakrabarty, and Emanuël AP Habets. "Multi-speaker localization using convolutional neural network trained with noise." *arXiv preprint arXiv:1712.04276* (2017).

[15] Craig A. Chin, Armando Barreto, Gualberto Cremades, and Malek Adjouadi. 2007. Performance analysis of an integrated eye gaze tracking / electromyogram cursor control system. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility (Assets '07).* Association for Computing Machinery, New York, NY, USA, 233–234. DOI: https://doi.org/10.1145/1296843.1296888

[16] W.T. Chu and A.C.C. Warnock. Detailed directivity of sound fields around human talkers, *Tech. Rep. RR-104*, National Research Council Canada, 2002.

[17] Antoine Deleforge and Radu Horaud. 2012. The cocktail party robot: sound source separation and localisation with an active binaural head. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12).* Association for Computing Machinery, New York, NY, USA, 431–438. DOI: https://doi.org/10.1145/2157689.2157834

[18] Nilanjan Dey, and Amira S. Ashour. Direction of arrival estimation and localization of multi-speech sources. *Springer International Publishing,* 2018.

[19] Nilanjan Dey, and Amira S. Ashour. "Challenges and future perspectives in speech-sources direction of arrival estimation and localization." In *Direction of arrival estimation and localization of multi-speech sources*, pp. 49-52. Springer, Cham, 2018. DOI: https://doi.org/10.1007/978-3-319-73059-2_5

[20] Tiago H. Falk, Chenxi Zheng, and Wai-Yip Chan. "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech." *IEEE Transactions on Audio, Speech, and Language Processing 18, no. 7* (2010): 1766-1774. DOI: https://doi.org/10.1109/TASL.2010.2052247

[21] Eric L. Ferguson, Stefan B. Williams, and Craig T. Jin. "Sound source localization in a multipath environment using convolutional neural networks." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2386-2390. IEEE, 2018. DOI: https://doi.org/10.1109/ICASSP.2018.8462024

[22] Yangyang Huang, Takuma Otsuka, and Hiroshi G. Okuno. "A speaker diarization system with robust speaker localization and voice activity detection." In *Contemporary Challenges and Solutions in Applied Artificial Intelligence*, pp. 77-82. Springer, Heidelberg, 2013. DOI: https://doi.org/10.1007/978-3-319-00651-2_11

[23] V. Z. Këpuska, and T. B. Klein. "A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation*." Nonlinear Analysis: Theory, Methods & Applications 71*, no. 12 (2009): e2772-e2789. DOI: https://doi.org/10.1016/j.na.2009.06.089

[24] Seon Man Kim, and Hong Kook Kim. "Direction-of-arrival based SNR estimation for dual-microphone speech enhancement*." IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, no. 12 (2014): 2207-2217. DOI: https://doi.org/10.1109/TASLP.2014.2360646

[25] C. H. Knapp, and G. C. Carter: 1976, The generalized correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech and Signal Processing* ASSP-24(4), 320-327. DOI: https://doi.org/10.1109/TASSP.1976.1162830

[26] V. Krishnaveni, T. Kesavamurthy, and B. Aparna. "Beamforming for direction-of-arrival (DOA) estimation-a survey." *International Journal of Computer Applications* 61, no. 11 (2013).

[27] Byoungho Kwon, Youngjin Park, and Youn-sik Park. "Multiple sound sources localization using the spatially mapped GCC functions." In 2009 I*CCAS-SICE*, pp. 1773-1776. IEEE, 2009.

[28] Avram Levi, and Harvey Silverman. "A robust method to extract talker azimuth orientation using a large-aperture microphone array." *IEEE transactions on audio, speech, and language processing* 18, no. 2 (2009): 277-285. DOI: https://doi.org/10.1109/TASL.2009.2025793

[29] Jun-seok Lim, and Hee-Suk Pang. "Time delay estimation method based on canonical correlation analysis." *Circuits, Systems, and Signal Processing* 32, no. 5 (2013): 2527-2538. DOI: https://doi.org/10.1007/s00034-013-9578-3

[30] Rong Liu, and Yongxuan Wang. "Azimuthal source localization using interaural coherence in a robotic dog: modeling and application." *Robotica* 28, no. 7 (2010): 1013-1020. DOI: https://doi.org/10.1017/S0263574709990865

[31] Michael I. Mandel, Daniel P. Ellis, and Tony Jebara. "An EM algorithm for localizing multiple sound sources in reverberant environments." In *Advances in neural information processing systems*, pp. 953-960. 2007. DOI: https://doi.org/10.7916/D84176FK

[32] Ivan Meza, Caleb Rascon, Gibran Fuentes, and Luis A. Pineda. "On indexicality, direction of arrival of sound sources, and human-robot interaction." *Journal of robotics* 2016 (2016). DOI: https://doi.org/10.1155/2016/3081048

[33] Menno Müller, Steven van de Par, and Joerg Bitzer. "Head-Orientation-Based Device Selection: Are You Talking to Me?." In *Speech Communication*; 12. ITG Symposium, pp. 1-5. VDE, 2016.

[34] Hirofumi Nakajima, Keiko Kikuchi, Toru Daigo, Yutaka Kaneda, Kazuhiro Nakadai, and Yuji Hasegawa. "Real-time sound source orientation estimation using a 96 channel microphone array." In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 676-683. IEEE, 2009. DOI: https://doi.org/10.1109/IROS.2009.5354285

[35] Keisuke Nakamura, Kazuhiro Nakadai, Futoshi Asano, and Gökhan Ince. "Intelligent sound source localization and its application to multimodal human tracking." In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 143-148. IEEE, 2011. DOI: https://doi.org/10.1109/IROS.2011.6094558

[36] Alberto Yoshihiro Nakano, Kazumasa Yamamoto, and Seiichi Nakagawa. "Directional acoustic source's position and orientation estimation approach by a microphone array network." In *2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, pp. 606-611. IEEE, 2009. DOI: https://doi.org/10.1109/DSP.2009.4785995

[37] Aanand Nayyar, Utkarsh Dwivedi, Karan Ahuja, Nitendra Rajput, Seema Nagar, and Kuntal Dey. 2017. OptiDwell: Intelligent Adjustment of Dwell Click Time. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. Association for Computing Machinery, New York, NY, USA, 193–204. DOI: https://doi.org/10.1145/3025171.3025202

[38] Kazuhiro Otsuka, Shoko Araki, Kentaro Ishizuka, Masakiyo Fujimoto, Martin Heinrich, and Junji Yamato. "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization." In *Proceedings of the 10th international conference on Multimodal interfaces*, pp. 257-264. 2008. DOI:https://doi.org/10.1145/1452392.1452446

[39] Beom-Chul Park, Kyu-Dae Ban, Keun-Chang Kwak, and Ho-Sup Yoon. "Performance analysis of GCC-PHAT-based sound source localization for intelligent robots." *Journal of Korea Robotics Society 2*, no. 3 (2007): 270-274.

[40] Despoina Pavlidi, Symeon Delikaris-Manias, Ville Pulkki, and Athanasias Mouchtaris. "3D DOA estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 96-100. IEEE, 2016. DOI: https://doi.org/10.1109/ICASSP.2016.7471644

[41] Cathy Pearl. Designing voice user interfaces: principles of conversational experiences. " *O'Reilly Media*, Inc.", 2016.

[42] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Paper 640, 1–12. DOI: https://doi.org/10.1145/3173574.3174214

[43] Om Prakash Prabhakar, and Navneet Kumar Sahu. A survey on: Voice command recognition technique. *International Journal of Advanced Research in Computer Science and Software Engineering* 3, no. 5 (2013).

[44] Caleb Rascón, Héctor Avilés, and Luis A. Pineda. "Robotic orientation towards speaker for human-robot interaction." In *Ibero-American Conference on Artificial Intelligence*, pp. 10-19. Springer, Berlin, Heidelberg, 2010. DOI: https://doi.org/10.1007/978-3-642-16952-6_2

[45] ReSpeaker. URL: https://wiki.seeedstudio.com/ReSpeaker-USB-Mic-Array

[46] Joshua M. Sachar, and Harvey F. Silverman. "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array." In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. iv-iv. IEEE, 2004. DOI: https://doi.org/10.1109/ICASSP.2004.1326764

[47] Andreas Schwarz, and Walter Kellermann. "Coherent-to-diffuse power ratio estimation for dereverberation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing 23*, no. 6 (2015): 1006-1018. DOI: https://doi.org/10.1109/TASLP.2015.2418571

[48] Carlos Segura, Alberto Abad, Javier Hernando, and Climent Nadeu. "Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR." In *Ninth Annual Conference of the International Speech Communication Association*. 2008.

[49] Carlos Segura, and Francisco Javier Hernando Pericás. "GCC-PHAT based head orientation estimation." In *13th Annual Conference of International Speech Communication Association*, pp. 1-4. 2012.

[50] Carlos Segura, Cristian Canton-Ferrer, Alberto Abad, Josep R. Casas, and Javier Hernando. "Multimodal head orientation towards attention tracking in smartrooms." In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2, pp. II-681. IEEE, 2007. DOI: https://doi.org/10.1109/ICASSP.2007.366327

[51] Hing Cheung So, Yiu Tong Chan, and Frankie Kit Wing Chan. "Closed-form formulae for time-difference-of-arrival estimation." *IEEE Transactions on Signal Processing 56*, no. 6 (2008): 2614-2620. DOI: https://doi.org/10.1109/TSP.2007.914342

[52] Hannu Soronen, Markku Turunen, and Jaakko Hakulinen. "Voice commands in home environment-a consumer survey." In *Ninth Annual Conference of the International Speech Communication Association*. 2008.

[53] Norbert Strobel, and Rudolf Rabenstein. "Classification of time delay estimates for robust speaker localization." In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*. ICASSP99 (Cat. No. 99CH36258), vol. 6, pp. 3081-3084. IEEE, 1999. DOI: https://doi.org/10.1109/ICASSP.1999.757492

[54] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. "Estimation of talker's head orientation based on discrimination of the shape of cross-power spectrum phase coefficients." In *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[55] Ryu Takeda, and Kazunori Komatani. "Discriminative multiple sound source localization based on deep neural networks using independent location model." In *2016 IEEE Spoken Language Technology Workshop (SLT),* pp. 603-609. IEEE, 2016. DOI: https://doi.org/10.1109/SLT.2016.7846325

[56] Yunus Terzioğlu, Bilge Mutlu, and Erol Şahin. 2020. Designing Social Cues for Collaborative Robots: The Role of Gaze and Breathing in Human-Robot Collaboration. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 343–357. DOI: https://doi.org/10.1145/3319502.3374829

[57] Tobii, Pro Glasses 2. URL: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

[58] Jose Velasco, Daniel Pizarro, and Javier Macias-Guarasa. "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints." *Sensors* 12, no. 10 (2012): 13781-13812. DOI: https://doi.org/10.3390/s121013781

[59] Hong Wang, and Peter Chu. "Voice source localization for automatic camera pointing system in videoconferencing." In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 187-190. IEEE, 1997. DOI: https://doi.org/10.1109/ASPAA.1997.625639

[60] Jackie (Junrui) Yang, Gaurab Banerjee, Vishesh Gupta, Monica S. Lam, and James A. Landay. 2020. Soundr: Head Position and Orientation Prediction Using a Microphone Array. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/3313831.3376427

[61] Cha Zhang, Dinei Florêncio, Demba E. Ba, and Zhengyou Zhang. "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings." *IEEE Transactions on Multimedia* 10, no. 3 (2008): 538-548. DOI: https://doi.org/10.1109/TMM.2008.917406