# A high accuracy, low-latency, scalable microphone-array system for conversation analysis

**David Sun and John Canny**
Berkeley Institute of Design and Computer Science Division
University of California, Berkeley
{davidsun, jfc}@eecs.berkeley.edu

## ABSTRACT

Understanding and facilitating real-life social interaction is a high-impact goal for Ubicomp research. Microphone arrays offer the unique capability to provide continuous, calm capture of verbal interaction in large physical spaces, such as homes and (especially open-plan) offices. Most microphone array work has focused on arrays of custom sensors in small spaces, and a few recent works have tested small arrays of commodity sensors in single rooms. This paper describes the first working scalable and cost-effective array that offers high-precision localization of conversational speech, and hence enables ongoing studies of verbal interactions in large semi-structured spaces. This work represents significant improvements over prior work in three dimensions – cost, scale and accuracy. It also achieves high throughput for real-time updates of tens of active sources using off-the-shelf components. We describe the system design, key localization algorithms, and a systematic performance evaluation. We then show how source location data can be usefully aggregated to reveal interesting patterns in group conversations, such as dominance and engagement.

### Author Keywords

Microphone array, conversation analysis, localization

### ACM Classification Keywords

H.5.2 Information interfaces and presentation (e.g., HCI): Miscellaneous.

### General Terms

Design, Performance, Experimentation

## INTRODUCTION

Location sensing of individuals has been an active and fruitful area of research for ubiquitous computing over the past two decades. Real-time location information provides rich contextual information which has become a key enabler for a myriad of novel location-aware applications and services. With the advent of wearable and mobile computing, there has been significant interest in outdoor location sensing solutions, where individuals are tracked via actively transmitting devices using GPS [2] and Radio Frequency [14]. For indoor localization, while RFID [1] and WiFi [16] have been explored, passive location sensing systems have shown increasing promise to provide calm capture of context [11]. To this end, much research so far has focused on vision-based camera sensing and tracking of individuals [10], while less attention has been given to the uses of speech.

In this work, we explore the use of microphone arrays to extract locations of individuals and conversation patterns of small groups in a large semi-structured space in real-time. We achieve this through the use of SLAAM, a Scalable Large Aperture Array of Microphones built entirely using off-the-shelf hardware and covers a physical space of some 1000sq feet. We describe the localization algorithms implemented in SLAAM which provide low-latency and high accuracy location information of multiple conversations. We also present a systematic performance evaluation and demonstrate its potential for a multitude of *conversation analysis* tasks. We highlight the following key characteristics of SLAAM that contribute to its capability of delivering scalable speech localization services:

- High accuracy: SLAAM achieves improvements in precision over previous large-array realizations and quantification of localization with natural, conversational speech.
- Modularity: SLAAM adopted a modular design approach using an array of cells, currently 5 x 5 covering approximately 1000 square feet, which is scalable by replication to arbitrary areas.
- Cost effectiveness: the SLAAM system costs about $10/square foot installed, similar to modular carpeting, and has required no maintenance in 4 years.
- Simple, efficient, easy to use: the SLAAM API allows applications to easily connect to, and use the service. The current hardware uses a single server CPU to provide an efficient service, which can simultaneously track up to 10s of targets.

## RELATED WORK

A microphone array system provides multiple tracks of redundant recordings in both temporal and spatial domains. This redundancy has been exploited in many algorithms to improve audio quality and reduce the impact of noise for


*UbiComp '12*, Sep 5-Sep 8, 2012, Pittsburgh, USA.

applications such as tele-conferencing and automatic speech recognition. A comprehensive review of the techniques and algorithms for microphone arrays can be found in [3] and [5].

A large number of different array designs have been explored. In the Huge Microphone Array (HMA) project [23] an array of 512 microphones was deployed in a lab space of 690 sq feet. The system included a fully custom hardware-software stack, comprised of custom-made DSP chips and operating systems, to handle a high (4GFlops) computation rate. The Large acOUstic Data Array (LOUD) project experimented with an array consisting of 1020 microphones [26], laid out uniformly on a rectangular panel of approximately 180cm wide and 50cm high. The high data rate of the array also required the researchers to design and fabricate custom hardware to handle the computation load. The CHIL project [17] experimented with a number of microphone array designs operating in unison, including a 4-element T-shaped array and a 64-element array developed by NIST.

As microphone array technologies become increasingly accessible, there has been a growing interest in classifying the activities of occupants in various indoor settings through sound localization [4, 11, 24]. Scott [22] designed a 6-element microphone-array for localizing implosive sounds (finger clicks or hand claps) using only off-the-shelf components. The system reported a localization accuracy of up to 27cm (at 90th percentile). Bian et al. [4] instrumented a home of 410sq feet with an array of 16 microphones to detect significant sound events (cooking, footsteps, conversations). The system reported high-latency updates at 1-5 second intervals, a modest localization accuracy of 68cm at 95th percentile and a high rejection rate of 60-80%. Guo et al. [11] designed a system which performed simultaneous classification and localization of human-generated acoustic events in a lab environment. Their system reported high accuracy for a four-class classification task (speech, clasp, footstep, and mug placed on table). For speech, location estimates were accurate up to 30cm at the 70th percentile. However, as most prior work quantified system performance using either artificially generated plosive sounds [4], or use a single frames of recorded speech [11], it's unclear if these performance figures also generalize to natural, continuous speech.

Multiple tracks of simultaneously recorded speech have also been utilized for behavior modeling. Karahalios et al. [13] created real-time visualizations of on-going social conversations to provide feedback as a reflective mechanism to help maintaining social norms. Roy et al. [21] instrumented homes with eleven video cameras and fourteen-element microphone arrays to obtain unobtrusive continuous recording of all activities to understand the language acquisition process of children. Choudhury et al. [6] explored the use of conversation dynamics extracted from multiple streams of recorded audio as contextual information to model and understand the formation of social networks.

## DESIGN GOALS AND CHALLENGES

Our work draws from and extends existing techniques in microphone array processing. We first motivate the desired characteristics of our system based on specific goals of our intended applications.

1. **Cost-effectiveness**. Cost is an important consideration in our design. Many indoor environments such as smart homes and offices cannot afford expensive setups. Thus, a clear understanding of the expected cost will be crucial in assessing feasibility and extensibility. Given the total hardware and installation costs for a specific space, the following cost measure can be computed:

$$\text{cost-per-unit-area} = \frac{\text{cost of hardware + installation}}{\text{area of sensitive region}}$$

2. **Off-the-shelf hardware** We wanted to examine the feasibility of building a robust array by using only off-the-shelf components. These components include: microphones, input/output modules, computing devices, communication protocols and software stacks.

3. **Low-latency localization** Latency requirements for a location sensing service are generally dependent on the applications to be designed and supported. For instance, update rates of 200-300ms are very acceptable for automatic camera steering in video-conferencing systems. However, to support the analysis of conversations where dynamics such as turn-taking need to be tracked in near real-time, the system will need to update estimates as quickly as 20-50ms [6, 8].

4. **Multi-source localization**Simultaneous speech can play an important role in analyzing the turn-taking patterns of conversations. Turn-stealing or interruption is an event that causes momentary simultaneous speech. The system will need to be able to locate two or more simultaneously active sources within close spatial proximity.

5. **Spatial precision** For conversation analysis, spatially separating the speakers is an important requirement. It has been well established in social psychology that the "social distance" between individuals has high correlation with physical distance, which can range from 0.5m to 3.7m, depending on the degree of intimacy [12]. Hence the physical distance between speakers in a social conversations can be lower-bounded by 50cm ( 20 inches). This implies that a source location estimate which deviates by more than 25cm from the true location will likely be incorrectly classified. The system will need to produce location estimates well below this error margin.

Low-latency, robust location sensing in the presence of multiple simultaneous sources is a highly challenging problem for microphone arrays. In this paper we describe how to address this and other related challenges in SLAAM.

## SYSTEM DESIGN

### Hardware

SLAAM is constructed entirely using off-the-shelf modules (Figure 1). Thirty-six Shure Easyflex omnidirectional boundary microphones (Figure 1a) were laid out in a grid pattern over the unistrut ceiling, at roughly 5 feet separation in each direction. The mics are mounted at the same height to simplify and more accurately calibrate the system. Each bank of six mics are feed into a single SM Pro Audio Preamp which provides a maximum gain of 60dB for each discrete input channel (Figure 1d). The preamp outputs to a MOTU MKII digital mixer and streaming device (Figure 1b). To avoid phase distortion in processing all thirty-six channels, adjacent MKIIs are connected by ADAT light-pipes and processed as a single virtual unit. The virtual units are clock synchronized to a single master module via their firewire interfaces. The grid based layout can easily scale up – for example, the array density can be doubled by a duplicate set up where the components run orthogonal to the current array.

All the audio signal processing, delay estimation and source computation take place on a quad-core 2.8GHz workstation. The MKIIs are configured with the lowest available sampling rate setting of 44.1kHz at 21-bit precision. The MKIIs use internal circular buffers of 1024 samples, which imply that a new data frame is generated by each channel at 23ms intervals. This places a hard time-constraint on the host-side algorithms – taking longer than 23ms to process all 36 audio frames will result in the dropping of future data frames without additional host-level buffering. This is a challenging constraint for robust localization system design since it rules out many algorithmic choices due to their high degree of computational complexity.

At the time of its construction (48 months ago), equipment cost of SLAAM was approximately $1270 per bank of mics. Including the workstation, the total equipment cost amounts to $8920. The quoted installation price averages to $30 per mic, due to retrofitting to an existing ceiling. Under our cost model we have:

$$\text{cost-per-unit-area} = \frac{\$8920 + 1080}{1000 \text{ sq ft}} = 10$$

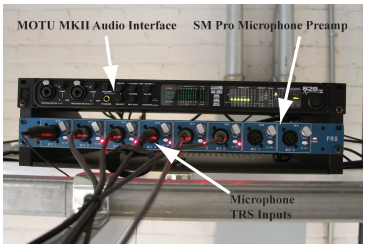
which is comparable to cost of installing office carpeting the same area. We believe this is a highly feasible pricing range to work with.

A one-time physical calibration took place post-installation and no calibration had been needed since. The hardware has also been very resilient to failure – no mic or streaming device failure has been detected since installation.

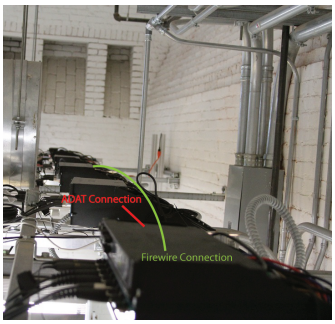
### Software Stack

A custom software stack was implemented which includes a streaming client, a localization module, and a network service listener module. The streaming client communicates with the MKIIs directly via the ASIO 2.0 audio interface. The key advantage in using the ASIO interface is that the client is able to obtain audio streams with minimal latency directly from hardware (by bypassing the OS kernel). The signal processing and localization modules are implemented in C++ and Intel Math Kernel Library (MKL) to leverage

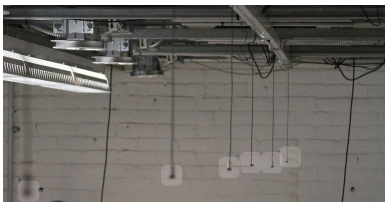

(a) Shure Boundary Microphone




(b) A single MOTU MKII and SM Pro Preamp setup




(c) Ceiling mount setup



(d) A single bank of microphones

Figure 1: Hardware configuration

its highly tuned linear algebra routines. The outputs of the localization module are published by the network service listener as a streaming service with an open IP address. Applications, such as online visualization, can easily tap into the service via a simple API to receive streams of source location updates.

## SOURCE LOCALIZATION

Acoustic source localization methods can be classified into three types: steered-beamforming, high-resolution spectral-estimation, and time-delay of arrival estimation [9]. Time-delay based estimators can be efficiently realized by computing delays between pairs of microphones outputs which are then used to parametrize a set constraints on the location of the target source.

### Array Partitioning

SLAAM consists of 36 microphones laid out in a 5-by-5 rectangular grid. We statically group each adjacent set of four microphones as a single unit, creating 25 such unique units. The cell defined by each such unit of four microphones is called a quad. This procedure effectively partitions the aperture area of SLAAM into non-overlapping subregions. We runs the localization algorithm described in this section independently for each quad, accepting only source estimates which fall within its perimeter.

### Source Detection

To detect the presence of speech sources, we make use of an efficient voice activity detector (VAD) algorithm similar to [20, 19]. The detector produces a binary classification of a data frame $x_t(n)$ at time $t$ as either speech or

non-speech. The metric is based on the Kullback-Leibler (KL) divergence between the current estimate of the noise spectrum probability distribution $p_n(\omega)$ and the estimated spectrum of the current data frame $p_d(\omega)$, as $H(p_n \| p_d) = \int p_n(x) \log\left(\frac{p_n(x)}{p_d(x)}\right) dx$. The noise spectrum probability distribution is an exponentially weighted average all non-speech frames. The spectrum of the current data frame is temporally smoothed via a double-exponential filter running at the same rates as EXP-GCC-PHAT (see next section). The VAD classifies each mic stream independently and a quad-level majority voting (i.e. more than 2 channels are classified as speech) is used to decide if delay estimates and triangulation should be attempted within the quad.

#### Time-delay Estimation

The General Cross-Correlation (GCC) with Phase-transform (PHAT) is a popular array technique for estimating time-delays [15]. The GCC-PHAT procedure takes as input two signals received at the microphone doublet $(m_1, m_2)$: $x_1(t) = x(t-\tau_1) + n_1(t)$ and $x_2(t) = x(t-\tau_2) + n_2(t)$, which are time-shifted, and noise corrupted replicas of a source $x(t)$. The GCC function between $x_1(t)$ and $x_2(t)$, given by $c(\tau) = \int_{-\infty}^{\infty} x_1(t)x_2(t+\tau)\, dt$, is maximized at the lag $\tau = \tau_1 - \tau_2$. A computationally efficient implementation of the GCC function is by working in the frequency domain, as $R(\tau) = IDFT\Big(\Psi_{12}(\omega)X_1(\omega)X_2^*(\omega)\Big)$ where $X_1(\omega)$ and $X_2(\omega)$ are the spectra of the input signals, $*$ is the complex conjugation operator, and $\Psi_{12}$ is a suitably chosen weighting function so that the cross-correlation function peaks at the true delay. Finally, the estimated delay or Time-Delay-of-Arrival (TDOA) is obtained via $\hat{\tau} = \arg\max_{\tau \in D} R(\tau)$ where $D$ is the plausible interval of delays governed by a doublet separation. The GCC-PHAT weighting function is defined as $\Psi_{12} = \frac{1}{|X_1(\omega)X_2^*(\omega)|}$

GCC-PHAT is well known for its simplicity and efficient implementation, and has been shown to provide fair localization accuracy under a range of acoustic conditions. For instance, Guo et al.[11] applied the GCC-PHAT procedure to the first voiced 204.3ms frame, or 8376 samples, of a continuous speech segment to obtain an absolute TDOA accuracy of 60cm at $70\%$ of the time, though errors can be as high as 2.5meters. While these results are encouraging, it is unclear how well GCC-PHAT can consistently and accurately produce TDOA estimates for continuous speech in a low-latency setting (e.g. data frames of 1024 samples). In fact our experiments suggests that when applied over an *entire segment* of multiple frames of speech, one might be much less optimistic about GCC-PHAT in its basic form.

We conducted delay-estimation experiments using the TIMIT test dataset (see Section Playback speech). Given a quad, the GCC-PHAT function is computed for each doublet on frame-by-frame basis over the entire speech segment, producing six time-series of delay estimates. A hamming window was applied to each data frame of 1024 samples (with a $50\%$ overlap). As shown in Figure 3, GCC-PHAT only achieved an absolute error of 60cm about $40\%$ of the time.

GCC-PHAT is a frequency domain technique, and when applied to *isolated* frames of speech data yields suboptimal TDOA estimates. In our work, we observed that the relative time-delays change only slowly over time for both stationary and even slowly moving sources. To track temporally slow varying TDOAs, we designed a double-exponential filter which is applied over the short-term GCC-PHAT. This procedure, which we call EXP-GCC-PHAT, takes the conventional PHAT-weighted cross-spectrum at time $t$: $\text{GCC}^t$ and recursively smooths the estimate via a weighted moving average:

$$\begin{aligned} \text{GCC}_\alpha^t &= (\alpha \text{GCC}_\alpha^{t-1}) + (1-\alpha)\text{GCC}^t \\ \text{GCC}_\beta^t &= (\beta \text{GCC}_\beta^{t-1}) + (1-\beta)\text{GCC}^t \\ \text{GCC}^t &= \text{GCC}_\alpha^t - \text{GCC}_\beta^t \end{aligned}$$

At $t = 0$, we have $\text{GCC}_\alpha^0 = \text{GCC}^0, \text{GCC}_\beta^0 = \text{GCC}^0$. Note $\alpha$ and $\beta$ are filter parameters which trades the relative emphasis between present and past cross-spectra. Taking the difference of two single-exponential averages effectively implements a non-causal IIR-filter. We empirically designed the filter coefficients with $\alpha = 0.5$ and $\beta = 0.3$, which provided good temporal smoothing while introducing a minimum of 1-frame delay in the output.

The result of applying EXP-GCC-PHAT to the TIMIT dataset is shown in Figure 3. An absolute error of $60cm$ and $75cm$ are obtained over $70\%$ and $80\%$ of the time. These are significant improvements in the cumulative distribution function of the TDOA errors. In Figure 2, the same frame of data is processed by EXP-GCC-PHAT and GCC-PHAT; compared to GCC-PHAT, EXP-GCC-PHAT significantly boosts the primary peak relative to secondary peaks. To quantify this effect, consider the ratio between the secondary peak to the primary peak in the cross-correlation under each processor, as shown in Figure 4. Under EXP-GCC-PHAT, the primary peaks are at least 2 times stronger than secondary peaks $30\%$ of the time and at least 1.5 times stronger in $60\%$ of the time. In comparison, under GCC-PHAT, only $10\%$ of the primary peaks are twice stronger and $30\%$ are 1.5 times strong. In general, EXP-GCC-PHAT achieves significantly stronger primary peaks which contributes to much lower TDOA estimation errors.

#### Peak Selection

While the EXP-GCC-PHAT procedure obtains encouraging TDOA estimate results, the estimation errors could still run as high as 1 meter over $15\%$ of the time. Furthermore, since TDOAs are key parameters in defining the triangulation problem, these errors would result in suboptimal or inconsistent models for solvers, which lead to poor localization estimates. While residual-norms from the triangulation solver could be used to evaluate the quality of a set of localization estimates, solving for the model in the first place makes it an expensive peak quality evaluation procedure.

In most indoor residential or office enclosures, hard con-

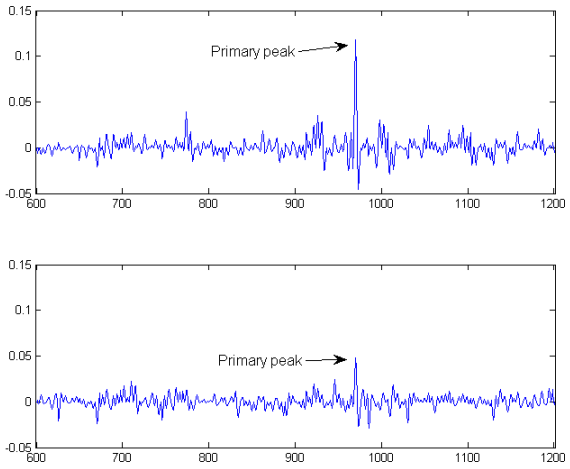



Figure 2: EXP-GCC-PHAT accentuate the desired peak

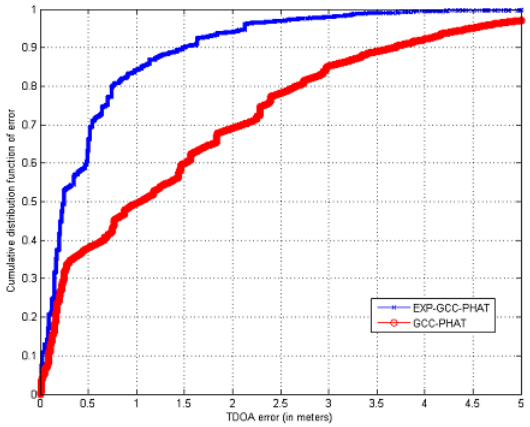



Figure 3: TDOA estimation accuracy

structs such as walls, windows, and furniture will cause reflections and reverberations. These lead to multipath propagation of the signal and will manifest as *false* peaks in the GCC. To account for possible strong reflections, we consider simultaneously a number of candidate peaks. This can be accomplished by softening the requirement to return the lag which maximizes the GCC function but rather consider the top $k$ peaks in the GCC (which is slower than $\max$ by a factor of only $\log n$)

### *Fast Peak Matching*

With up to $k$ candidate peaks returned per GCC function (for a single doublet), the resulting problem is a matching problem – for $d$ pairs of doublets involved in the localization procedure, determining which $d$ peaks out of $k^d$ possible candidate combinations best match the possible location of the source. While in general such problems are combinatorial and would cause scalability concerns, good heuristics exist to simplify our particular problem.

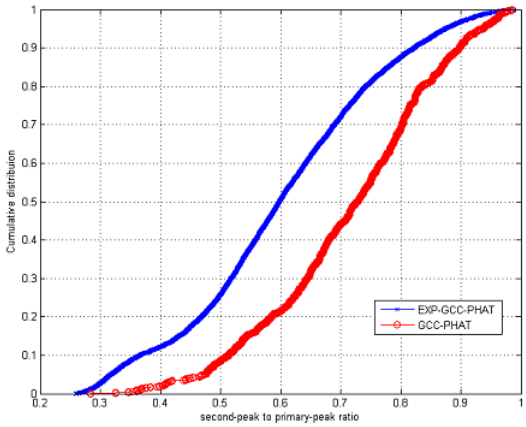



Figure 4: Secondary peak to primary peak ratio

We observe that if the source location is known, then the TDOAs can be described by a set of deterministic relations. Consider Figure 5, letting $d_i$ denote the distance between the source to mic $i$ and $\tau_i = d_i/c$, we have the following relations: $\tau_{12} = \tau_1 - \tau_2, \tau_{23} = \tau_2 - \tau_3, \tau_{13} = \tau_1 - \tau_3 = \tau_{12} + \tau_{23}$, where the last relation reveals that $\tau_{13}$ is fixed by fixing the two other delays $\tau_{12}$ are $\tau_{23}$. We can in fact write six such independent constraints to form a set of necessary consistency conditions for which a given set of TDOAs must satisfy. Furthermore, since each consistency constraint only requires knowing three, rather than all six delays, we can quickly prune the candidate space; without needing to generate and evaluate all $k^6$ combinations. In actual realization, since all the delays are estimated, rather than requiring equality, we consider $\hat{\tau}_{13} = \hat{\tau}_{12} + \hat{\tau}_{23} + \epsilon$ for some suitably chosen threshold $\epsilon$. A lower bound on $\epsilon$ obtained by considering the best-case delay estimator, where the estimation error $\epsilon$ is zero.Since the actual delays are continuous time quantities, discretization introduces an uncertainty bounded below by the sampling period $1/f$.

We evaluate the effectiveness of this matching algorithm with respect to error CDF by varying the threshold $\epsilon$ in multiples of $1/f$. We decision performed thresholding in an OR-fashion, i.e. the entire frame of array data is rejected if any of the $d$ peaks (time-delays) fails the consistency check. We see an absolute error of 3cm is achieved at the 99th percentile (for $\epsilon = 1/f$) at a surprisingly reasonable rejection rate of around 20%. At 5% rejection rate ($\epsilon = 3/f$), 99% of errors are below $7cm$. We note that in theory only three pairs of time-delay estimates are require for localizing targets in 3D – however the inclusion redundant constraints will generally improve accuracy in the presence of noisy estimates. Hence one can imagine a less aggressive thresholding scheme where the array data frame is only rejected when the number of consistent TDOA estimates are less than a constant between 3 and 6.

## Multiple Sources

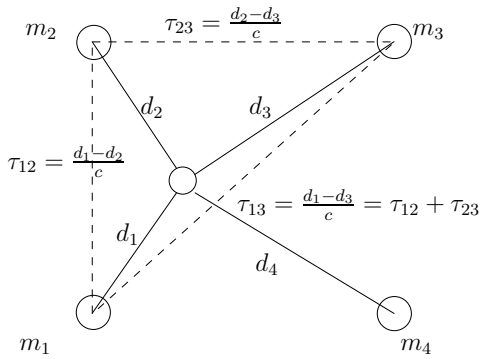



Figure 5: TDOA consistency constraints

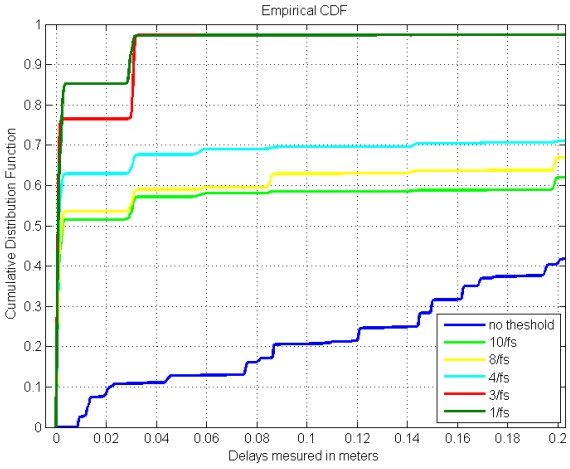



Figure 6: EXP-GCC-PHAT with consistency check

Robust multi-source localization is still a challenging problem in microphone array processing. Recently proposed techniques based on blind-source-separation (BSS) have shown great promise [7, 18]. These techniques compute a factorizations of the signal coherent matrix and/or computes likelihoods over a 3D space, and thus are are not easily adapted for real-time processing.

Within the time-delay based localization framework, the existence of multiple sources implies the GCC-PHAT function may produce multiple significant peaks. Unlike peaks introduced by strong reflections, these peaks are not only likely to be internally consistent, but should also exhibit longer term persistence in the temporal domain. Our multi-source localization strategy is thus based on the fast peak matching algorithm described previously. The scheme will return $S \times k$ numbers of peaks for each doublet, where $S$ is the number potential sources and $k$ is the number of candidate peaks per source. We consider the performance of this scheme in the Evaluation section.

### Triangulation

Triangulation of the target source location based on delay estimates is a geometric problem. Given a source located at $x$, the time-delay relative to the $i$th doublet is: $\tau = \frac{\|x - m_1\| - \|x - m_2\|}{c}$ where $m_1$, $m_2$ are the known locations of the doublet. Together with the sign of $\tau$, the expression defines a locus of candidate source locations that form a one-half of a hyperboloid of two sheets in 3D, centered at $m_1$ and $m_2$, with a semi-major axis $\alpha$ and a semi-minor axis $\beta$ related by $\alpha^2 + \beta^2 = (\frac{d}{2})^2$, where $d$ is the distance between $m_1$ and $m_2$. A hyperboloid of two sheets in 3D whose foci lie on the x-axis is given by the quadratic form: $x^T A x = 0$, where $A = \text{diag}[\alpha^{-2}, -\beta^{-2}, -\beta^{-2}]$ is a diagonal coefficient matrix.

Given a set of six time-delay estimates in a quad, the source location estimate can then be defined as the point that lies at the intersection of all the hyperboloids. However, such a point may not exist since TDOAs are estimated and may not be exact. As such, we seek to find the point that is closest to all the hyperbolic constraints, which is effectively a least squares problem. Letting $f_i(x) = x^T A_i x$, the least squares problem can be stated as:

$$\hat{x} = \arg\max_x \sum_{i=1}^{6} w_i f_i(x)^2 \quad (1)$$

The optimization problem (1) involves solving a weighted system of nonlinear equations. We developed an iterative solver using gradient descent. At each iteration, the solver forms a linear approximation to the original problem and obtains an approximate solution via weighted linear least squares. In general, descent methods can take many iterations to converge. To ensure our solver converges under the low-latency requirement, we developed a number of heuristics to speed up convergence, which we describe below.

1. *Initialization* Descent methods are often sensitive to the initialization point. For our solver we leveraged the geometric model to produce a good (probable) initial guess of the source location. For a hyperboloid given by $f_i(x) = x^T A_i x = 0$, one can fit a plane tangential to the surface at a given point $x_i$ by evaluating the partial derivatives at that point $\bigtriangledown f_i(x_i) = A_i x_i$ and forming the equation $\bigtriangledown f_i(x_i)^T x = \bigtriangledown f_i(x_i)^T x_i$. One such tangential planes is obtained for each of the six hyperboloid constraints, and together, form the linear approximation to the problem, i.e., let $J = [\bigtriangledown f_1(x_1)^T, \ldots, \bigtriangledown f_6(x_6)^T]^T$ and $b = [\bigtriangledown f_1(x_1)x_1, \ldots, \bigtriangledown f_6(x_6)x_6]^T$, solve $Jx = b$. Provided $J$ is full-rank, the linear system can be solved via the normal equations: $x = (J^T J)^{-1} J^T b$. We initialize each $x_i$ to be the vertex of the corresponding hyperboloid. Furthermore, the normalized weights indicate the relative importance of each constraint $f_i(x)$. We encode our belief of the accuracy of the model into $w_i$ by setting them to the normalized GCC function values corresponding to the TDOA peaks.

2. *Z-dim constraint* The microphones in SLAAM are ceiling mounted at the same height. Consequently, the linear approximation to the quadratic problem is under-constrained and the Jacobian is singular. To mitigate this issue, we add

a soft height-constraint $z = c$ to the solver. To locate both sitting and standing speakers, this height-constraint is set at 1m, roughly the height of a sitting person.

3. *Planar reduction* When an active speaker is positioned at roughly the half way point between a doublet, the TDOA $\tau$ approaches zero and the coefficient matrix $A$ is undefined. However, we observe that in this degenerate case, the hyperboloid constraint effectively reduces to a planar constraint – a plane that bisects the line joining the two microphones. The solver replaces the quadratic constraint with a corresponding plane constraint, by consulting the known geometry of the microphone array layout.

### Benchmarking

We benchmarked our solver against the Matlab optimization toolbox implementation of Nonlinear least squares using Levenberg-Marquardt. For localizing a single source, the Levenberg-Marquardt solver took a frame-average of 30ms cpu-time for a single source computation. Our solver took an average of 1ms cpu-time to obtain a solution at the same precision, which represents a factor of 30 increase in performance at the same accuracy. This suggests that, for real-time localization (23ms), our solver is theoretically able to simultaneously track 10s of targets.

## PERFORMANCE EVALUATION

In this section, we discuss evaluation of SLAAM for continuous speech, using both prerecorded and authentic conversation data. To provide an honest appraisal of our system, we adopted a multi-step evaluation approach in order to understand its performance characteristics.

### Testing Environment

The receiving region of SLAAM overlays one-half of an open semi-structured lab space (Figure 7) which hosts regular seminar series, small group research meetings, and graduate student and adviser meetings, teaching-assistant office hours and student socials. The space was retrofitted with a unistrut ceiling top at a height of 10 feet from ground over the original vaulted ceiling of 30 feet. HVACs are installed 3 feet away, parallel to the unistrut from which the mics hang. The space under the mic array is connected to an open cubicle area (of 30 by 32 feet) which hosts 20 graduate students with 15 workstations. Rudimentary analysis of the room acoustics revealed a fairly complex condition for acoustic localization: an average ambient noise of 42dB is detected across all microphones (about the same as a house in the middle of a major city). The RT60 of the room averages to 3.6 seconds and was measured as high as 5 seconds at some locations.

### Playback speech

Evaluation of localization performance for naturally occurring speech is complicated by the fact that the human upper-body will not remain perfectly rigid even without active movement. We start therefore by examining the performance characteristics of the system for a completely stationary sound source. A stationary sound source is generated by playing back 50 clips selected from the TIMIT test dataset through a

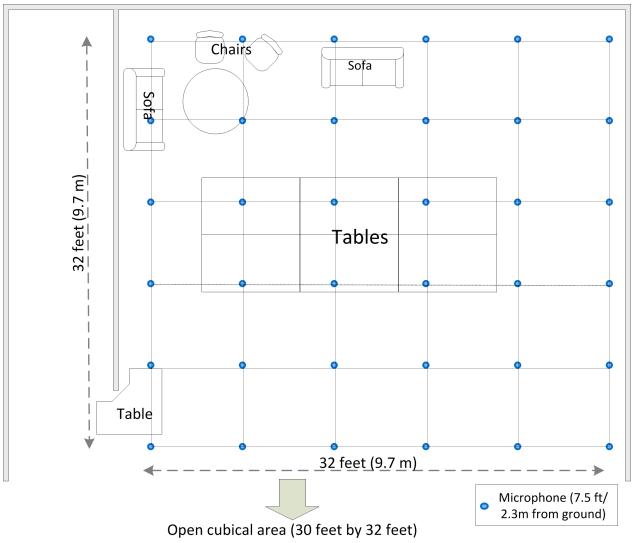



Figure 7: Layout of lab space and array placement

quality loud speaker, at 25 calibrated locations under SLAAM (one location per quad) and at two different heights: a "standing height" (of 6 feet) and a "sitting height" (of 4.5 feet). The loudspeaker was adjusted to produce a maximum of volume of 70dB at 3 feet away which is roughly the loudness of adult speech. A total of 750 audio clips were collected at each height level. TDOAs are then estimated using EXP-GCC-PHAT and triangulation is performed using our custom least-squares solver, as described in previous sections. We summarize both the 3D localization errors and rejection ratios in Table 1.

Table 1: Localization performance for prerecorded speech

| | RMSE (standing) | RMSE(sitting) |
|---|---|---|
| $< 25cm$ | 100 % | 99% |
| $< 10cm$ | 98.20% | 83.90% |
| $< 8cm$ | 95.91% | 73.34% |
| $< 5cm$ | 92.62% | 65.71% |
| Reject % | 1.01% | 2.14% |

High localization accuracy was obtained by the system: 100% of the standing data obtained an absolute error of less than $25cm$, while 92.6% of the time, the error was less than $5cm$. For the sitting corpus, we obtain an absolute error of less than $25cm$ at 99% but only $< 5cm$ at 65% of the time. The rejection rates were at 1% for the standing corpus and 2% for the sitting corpora.

### Human speakers

We consider the case of a single, stationary human speaker. For this we recruited seven subjects and each subject was assigned to 3 different locations under the coverage of SLAAM. At each location, a subject read aloud the content of a paper written in English. This was repeated for both standing as well as sitting heights. We asked the subjects to speak naturally at a normal conversational volume and pace, and did

not ask them to remain consciously rigid during the recording process. Under each condition, a multichannel track recording of 2.5-3 minutes was obtained. Localization was performed in real-time on the fly and the location estimates and residuals were also recorded. This corpus contained an average of 18 minutes of audio recording for each of the seven subjects.

For each condition, we consider the RMS deviation of the estimated source location from the calibrated locations. The results are summarized in Table 2. For the standing corpus, $95.2\%$ of the estimates were less than 10cm of error. No error exceeded the permissible margin of 25cm. Localization performance is slightly worse for the sitting corpus; the mean deviation from the calibrated locations was 8.8cm and $92.5\%$ of estimates were below 25cm.

Given the ceiling-mount construction of the array, the distance of a sound source to the microphones will increase at sitting height. Thus we expect the signal strength to degrade in proportion to inverse distance squared, which contributes to the increase in RMSE. Furthermore, some of the error could be attributed to slight movements of a subject's upper-body position (e.g. head movement).

Table 2: Localization performance for human speakers

| | RMSE(standing) | RMSE(sitting) |
|---|---|---|
| $< 25cm$ | 100% | 92.50% |
| $< 10cm$ | 95.20% | 61.90% |
| $< 8cm$ | 80.90% | 52.30% |
| $< 5cm$ | 47.60% | 33.30% |
| Reject % | 3.91% | 4.52% |

### Human conversations

We evaluate SLAAM's localization performance for human conversations by considering dyadic conversations (2 interlocutors). We recruited six subjects, forming three dyads. We calibrated three pairs of fixed locations, where each pair of locations are separated by a maximum of 70cm. Each dyad was asked to carry a normal conversation, first standing and then sitting down, at each location; all other requirements were identical to the single-speaker condition. In total we recorded between 30-40 minute of conversational speech at each calibrated location. As shown in Table 3, for standing conversations, $97\%$ of the errors were under $25cm$ while for the sitting corpus, $89\%$ of the estimates were under $25cm$ of absolute error. These results are strong indications that it is feasible to spatially separate and implicitly tag stationary speaker directly from the localization results.

## CONVERSATION ANALYSIS

While our system has the potentially for the analysis of larger group conversations, to keep issues tractable we set out with a modest initial goal of analyzing conversations for small group interactions. Furthermore, we wanted to perform our analysis using natural authentic conversations. For this, we consented and recorded three group conversations. We briefly summarize the demographics of the groups and the topics of their conversations in Table 4 .

Table 3: Localization performance for dyadic conversations

| | RMSE(standing) | | RMSE(sitting) | |
|---|---|---|---|---|
| | A | B | A | B |
| $< 25cm$ | 97.3% | 97.2% | 90.03% | 89.3% |
| $< 10cm$ | 93.10% | 94.23% | 60.10% | 59.4% |
| $< 8cm$ | 78.10% | 77.93% | 50.40% | 51.5% |
| $< 5cm$ | 45.33% | 41.30% | 31.3% | 30.1% |
| Reject % | 4.22% | 4.17 % | 5.20% | 5.41% |

Table 4: Authentic conversations

| | Demographics | Topic |
|---|---|---|
| Group 1 | 3 graduate students | weekend plan |
| Group 2 | 3 undergraduate students | football |
| Group 3 | 1 graduate student and 2 undergraduate researchers | research |

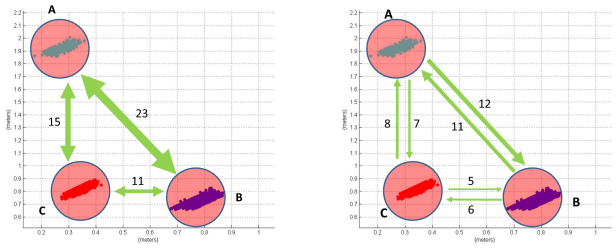
In order to not influence their interaction, we asked each group to take up whatever location was natural and appropriate. However, we did ask all the groups to avoid making significant movements (e.g. swapping locations). To obtain ground truth, each subject was wired with a close-talking microphone just below the chin. To perform conversation analysis, we first used the standard $k$-means algorithm to cluster our localization data, which consisted of coordinate vectors into meaningful constructs: sources. The number of clusters was chosen via a five-fold cross-validation. The cost function was the sum of the euclidean distances from the coordinate vectors to the centroids of the clusters. Next, we perform a temporal mapping of the sources by assigning the source-id to the frame in which it is active. From this we compute a set of high level summary statistics for each conversation (Table 5).

#### *Conversation dynamics*

From the summary data it is clear that Group 1 and Group 2 had more balanced conversations than Group 3. Each member in Group 1 and Group 2 spoke for approximate the same total time as well as holding the same number of floors. This is unsurprising given the informal and casual nature of these two conversations. For Group 3, allocation of speaking times and turn-takes were more skewed and a pattern of dominance and engagement reveals itself – member A, the graduate student, was clearly the dominant speaker; member C held the floor the highest number of times (19), which seems to suggest that she was highly engaged in the conversation. To better understand this dynamics, we break-down the turn-taking counts to examine the number of turn-takes between each pair of members in the group and the directionality of the turns-takes, i.e. whether the turn was passed from member X to Y or Y to X. As visualized in Figure 8, we overlay a directed graph over the localization visualization, with the following mapping: (1) nodes represented members in the group (2) color of the nodes is proportional to the mean speech energy (3) directed arrows indicate the exchange of floors and the thickness of the arrow lines is proportional to the number of exchanges.

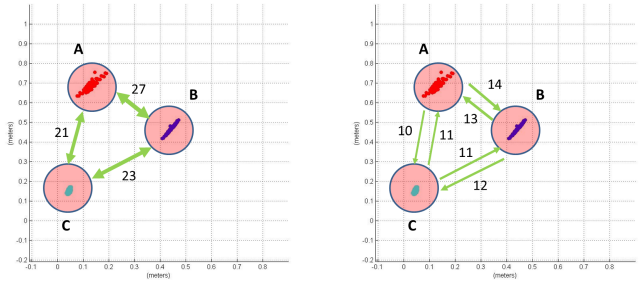
Table 5: Summary statistics of three-way conversations

| | Group 1 | | | Group 2 | | | Group 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Member | A | B | C | A | B | C | A | B | C |
| Total Speaking Time (minutes) | 15.2 | 19.3 | 17.1 | 10.3 | 8.2 | 9.2 | 25 | 2.4 | 10.8 |
| Percentage of total conversation time | 29% | 37 % | 33% | 37% | 30% | 33% | 65% | 6% | 28% |
| Total # turns | 24 | 25 | 22 | 16 | 18 | 14 | 17 | 6 | 19 |
| Time per turn (sec) | 38 | 48 | 49 | 39 | 27 | 39 | 88 | 24 | 34 |
| Mean energy (db) | 59 | 58 | 59 | 56 | 57 | 56 | 57 | 55.6 | 55 |




(a) pair-wise turn-takes (b) directional turn-takes

Figure 8: Turn-taking dynamics of Group 3

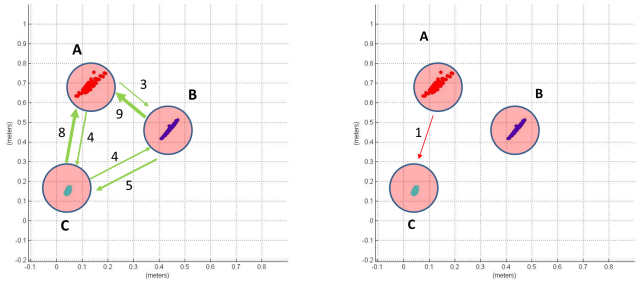



(a) pair-wise turn-takes (b) directional turn-takes

Figure 9: Turn-taking dynamics of Group 1

Even without sophisticated methods, we can begin gaining insight into the dynamics of the groups. Members A and B was the more engaged pair, as indicated by the high number of floor exchanges they shared. Furthermore, the smaller number of interchanges between members B and C suggests that the communication pattern was more of a two 2-way conversation (A to B and A to C). This is shown more prominently in Figure 8b, where the directionality of the turn-exchanges show a roughly balanced number of floor exchanges between each dyad. Indeed, from the audio transcripts we were able to ascertain that the graduate student (A) was getting a weekly progress update from the two undergraduate researchers (B and C) who worked independently on parts of a project. Applying the same analysis to Group 1, we see in Figure 9 that turn-taking behavior was more balanced, suggesting equal participation from each member. This is consistent with their speaking times (Table 3).

We have remarked that both interruption and cooperative-overlap can produce frames in which multiple sources are active. As explained in [25] cooperative-overlaps are devices used by the interlocutors to "chime in" to show rapport and do not interrupt the flow of the conversation. Examples of cooperative overlap devices include utterances such as "oh" or "oh yeah". To differentiate multi-source frames that correspond to true turn-stealing interruptions versus those that are cooperative overlaps, we use the following heuristic: if the number of consecutive overlapping frames is greater than 5 frames (i.e. 2 seconds) and the overlapping frames are followed by a turn-exchange, then the multi-source frames are classified as interruptions; otherwise, they are classified as cooperative overlap. This heuristic is consistent with the observation that cooperative overlap constructs are short utterances and that they do not typically cause a floor change. We visualize the detected interruptions and cooperative overlaps for Group 1 and Group 3 in Figures 10 and 11. Again, we show the directionality of the overlap. It is of interest to observe that in the formal meeting case (Group 3), few cooperative overlaps were detected and the majority of the overlaps were actual interruptions and $50\%$ of those interruptions were caused by member A (graduate student). For the more casual conversation (Group 1) the opposite is true – most of the overlaps were cooperative rather than interruptions. These are consistent with [25] in that cooperative overlaps happen more in conversations between acquaintances.

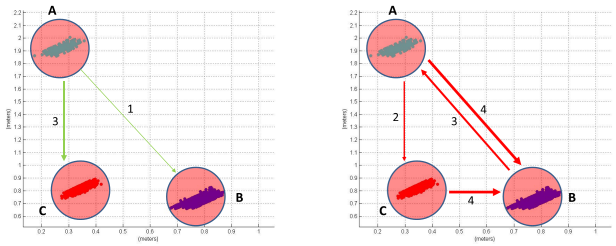



(a) Cooperative overlaps (b) Interruptions

Figure 10: Overlapping dynamics of Group 1

## CONCLUSION AND FUTURE WORK

In this work, we have examined the design, implementation, and evaluation of a scalable microphone array system that is capable of achieving high accuracy, low latency human speech localization and small-group conversation analysis. We discussed system design requirements for capturing conversation dynamics, and algorithms and techniques capable




(a) Cooperative overlaps (b) Interruptions

Figure 11: Overlapping dynamics of Group 3

of meeting these requirements. While this work is built on a known localization strategy using time delays, we achieved significant performance gains over previously reported results in both precision and recall for localization in a real-world acoustic environment. Furthermore, we proposed an efficient multi-source localization scheme under this framework based on peak matching, which was the key enabler to pick up the nuances in human to human conversations. This working platform has enabled a number of initial conversation analysis work, revealing interesting patterns of rapport, interruption, and dominance within different social structures.

There are a number of interesting directions which can be built on this work. We are currently exploring machine learning techniques, such as Bayesian networks, for automated analysis and classification of meeting types, speaker roles, and other nuances of social interaction (e.g. mimicry) on a large scale. We hope that this line of work will inform models to better understand the formation and evolution of physical social networks.

We have also been exploring uses of this infrastructure for emotion detection, classification, and feedback for collaborating teams. It is well known that emotional and social feedback can be powerful incentives to enable teams to function more effectively. We have since collected a significant amount of authentic conversational and meeting data for model training and prediction which are being studied and will be reported in future papers. Lastly, microphone arrays have long been envisioned as ubiquitous input devices and to this end we are extending the SLAAM API as an array front-end to drive novel speech-based ubicomp applications.

## ACKNOWLEDGMENTS

We would like to thank the reviewers and members of the BiD lab for many helpful suggestions and comments.

## REFERENCES

1. G.D. Abowd, A. Battestini, and T. OConnell. The location service: A framework for handling multiple location sensing technologies. *College of Computing and GVU Center, Georgia Institute for Technology, Atlanta,. Georgia, USA*, 2002.
2. M. Agrawal and K. Konolige. Real-time localization in outdoor environments using stereo vision and inexpensive gps. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1063–1068. IEEE, 2006.
3. J. Benesty, J. Chen, and Y. Huang. *Microphone array signal processing*, volume 1. Springer Verlag, 2008.
4. X. Bian, G.D. Abowd, and J.M. Rehg. Using sound source localization in a home environment. *Pervasive Computing*, pages 19–36, 2005.
5. M. Brandstein and D. Ward. *Microphone arrays*. Springer, 2001.
6. T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In *Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWC03)*, volume 1530, pages 17–00, 2003.
7. E.D. Di Claudio, R. Parisi, and G. Orlandi. Multi-source localization in reverberant environments by root-music and clustering. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages II921–II924. IEEE, 2000.
8. J.H. DiBiase. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Brown University, 2000.
9. J.H. DiBiase, H.F. Silverman, and M.S. Brandstein. Robust localization in reverberant rooms. *Microphone arrays: signal processing techniques and applications*, pages 157–180, 2001.
10. D.M. Gavrila. The visual analysis of human movement: A survey* 1. *Computer vision and image understanding*, 73(1):82–98, 1999.
11. Y. Guo and M. Hazas. Localising speech, footsteps and other sounds using resource-constrained devices. In *Proc. of the 10th International Conference on Information Processing in Sensor Networks*, number 2, pages 330–341, 2011.
12. E.T. Hall. A system for the notation of proxemic behavior1. *American anthropologist*, 65(5):1003–1026, 1963.
13. K. Karahalios and T. Bergstrom. Social mirrors as social signals: Transforming audio into graphics. *IEEE computer graphics and applications*, 29(5):22–32, 2009.
14. D.H. Kim, J. Hightower, R. Govindan, and D. Estrin. Discovering semantically meaningful places from pervasive rf-beacons. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 21–30. ACM, 2009.
15. C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4):320–327, 1976.
16. J. Koo and H. Cha. Autonomous construction of a wifi access point map using multidimensional scaling. *Pervasive Computing*, pages 115–132, 2011.
17. R. Malkin, D. Macho, A. Temko, and C. Nadeu. First evaluation of acoustic event classification systems in chil project. In *HSCMA'05 Workshop*, 2005.
18. F. Nesta and M. Omologo. Generalized state coherence transform for multidimensional localization of multiple sources. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 237–240. IEEE, 2009.
19. J. Ramırez, J.C. Segura, C. Benıtez, A. De La Torre, and A. Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech communication*, 42(3):271–287, 2004.
20. J. Ramírez, J.C. Segura, C. Benítez, Á. de la Torre, and A.J. Rubio. A new kullback-leibler vad for speech recognition in noise. *Signal Processing Letters, IEEE*, 11(2):266–269, 2004.
21. D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guinness, et al. The human speechome project. *Symbol Grounding and Beyond*, pages 192–196, 2006.
22. J. Scott and B. Dragovic. Audio location: Accurate low-cost location sensing. *Pervasive Computing*, pages 307–311, 2005.
23. H.F. Silverman, W.R. Patterson III, and J.L. Flanagan. The huge microphone array. *Concurrency, IEEE*, 6(4):36–46, 1998.
24. Z. Sun, A. Purohit, K. Yang, N. Pattan, D. Siewiorek, A. Smailagic, I. Lane, and P. Zhang. Coughloc: Location-aware indoor acoustic sensing for non-intrusive cough detection. In *International Workshop on Emerging Mobile Sensing Technologies, Systems, and Applications*, 2011.

25. D. Tannen. *Conversational style: Analyzing talk among friends*. Oxford University Press, USA, 1984.

26. E. Weinstein, K. Steele, A. Agarwal, and J. Glass. Loud: A 1020-node modular microphone array and beamformer for intelligent computing spaces. Technical report, Citeseer, 2004.