Relative Sound Localization for Sources in a Haphazard Speaker Array

Neal Andersen
Department of Music and Arts Technology,
Indiana University-Purdue University Indianapolis
andersne@iupui.edu

Benjamin D. Smith
Department of Music and Arts Technology,
Indiana University-Purdue University Indianapolis
bds6@iupui.edu

ABSTRACT

A rapidly deployable, easy to use method of automatically configuring multi-channel audio systems is described. Compensating for non-ideal speaker positioning is a problem seen in immersive audio-visual art installations, home theater surround sound setups, and live concerts. Manual configuration requires expertise and time, while automatic methods promise to reduce these costs, enabling quick and easy setup and operation. Ideally the system should outperform a human in aural sound source localization. A naïve method is proposed and paired software is evaluated aiming to cut down on setup time, use readily available hardware, and enable satisfactory multi-channel spatialization and sound-source localization.

1. HAPHAZARD ARRAYS

A haphazard speaker array involves any number of speakers (more than 2), placed in a space with little regard to precise alignment, orientation, or positioning. Unlike speaker grids or uniform array setups, the haphazard array is created at the whims of the user, potentially responding to constraints of the environment to guide placement (such as limitations in mounting, positioning, and cable lengths), or to take advantage of unique acoustics of a given installation space. Further, the haphazard array may use any mix of speakers with significantly different acoustic characteristics. While a conventional, uniform array focuses on pristine, reproducible audio, the haphazard model seeks to exploit unique elements of a given installation, equipment, and space.

The haphazard array presents a complex system with potential acoustic richness unique to each setup. The array also works within each environment it is setup in, providing a further layer of acoustic interaction that makes each configuration unique. A primary goal of haphazard arrays is a quick and inexpensive setup, using equipment that is on hand and spending a minimum of time calibrating the system.

The goal of this project is to research and define methods of working with haphazard arrays that make their complex nature transparent to the user. Ideally the capabilities of the system should be easy to use, leveraging current live mixing practices. The user should not be burdened with learning the particulars of the array's configuration, rather they should be able to use a uniform panning interface (sec. 3) which hides the complexities of the array. Similarly the setup and configuration of the system should support rapid deployment and minimal time from connection to use.

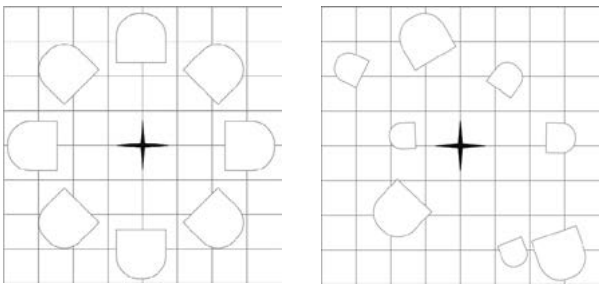

Figure 1. Fixed-Speaker Array (Left); Haphazard-Speaker Array (Right)

2. BACKGROUND/CONCEPT

Researchers in both acoustics and robotics address automatic identification of speaker array characteristics, such as sound-source/speaker location and frequency responses. The goal of this project is to provide a single point of interaction for a user to mix one or more tracks of audio within the array's acoustic space. Given a fixed uniform array (Fig. 1, left), the controls typically take the form of a panning potentiometer or digital dial to mix the source audio between output channels. This same model can be extended to work across non-uniform arrays (Fig. 1, right) if the characteristics of the setup can be accurately mapped.

2.1 Auditory Localization Issues

Describing the speaker locations and characteristics is closely related to research in robotic audition which looks at building systems to isolate and locate sound sources to inform robot functionality. Popular robotic approaches are based on models of human hearing, and typically start with two or more microphones mounted in opposed directions, performing calculations based on inter-aural intensity difference [6] and time difference of arrival [3] (i.e. the difference in time between a sound's arrival at each 'ear'). The accuracy of these systems (typically

within centimeters for nearby sounds) greatly improves with the employ of more than two microphones, allowing the robot to assess sounds in a 3-dimensional field [8].

Tests conducted with human subjects show a wide range of error in localizing depending on the frequency and angle in which the sound source is played. In one study [1], test subjects displayed horizontal angle accuracy between 8.5–13 degrees in testing audio along the horizontal plane without visual cues.

The minimum audible angle of humans for the horizontal plane has improved accuracy if the sound is in front of the listener and the test tone is brief [5]. This optimized scenario displayed accuracy between 2–3.5 degrees. However as the sounds moved to the side and behind the head, the error reached up to 20 degrees.

Measuring the perceived distance from human listeners is almost incalculable, as distance is considered to be lateralized, or processed internally as opposed to localized from an external cue. [8] To accurately localize distance from the arrival time of a sound source in a human, there needs to be some kind of non-auditory sensory feedback. [9]

2.2 Speaker Systems

Another similar problem is the automatic calibration of home surround sound systems, which are commonly setup in a less than ideal fashion. Using a microphone array these approaches play test tones through all the speakers in the setup in order to identify the particulars of the setup, acoustic characteristics of the room, and listener's sitting location [2, 7]. Time difference of arrival is the primary approach taken for speaker identification. They anecdotally report speaker location accuracy to several centimeters, in spaces no larger than 9 feet square.

The performance requirements of a haphazard array are based on the discriminatory ability of the people who will be experiencing it. Thus human auditory accuracy defines operating success of a calibration system.

Evaluation of panning algorithms with human participants showed a consistent average accuracy across all models of 10 degrees [4]. However every test showed many individual errors of up to 45 degrees, regardless of panning algorithm employed.

3. LOCALIZATION FACTORS

In order to create a panning interface the characteristics of the array have to be measured and analyzed to build a virtual map of the array. This can be accomplished manually, with a user entering data for each speaker into the system, but this is cumbersome and expensive (in terms of time), requires expertise, and works against the goals of having a quickly usable system. Automating the configuration of the system is the preferred solution and involves analyzing the acoustic space for the following information:

- The position of each speaker,
- The relative loudness of each speaker,
- The relative frequency response of each speaker.

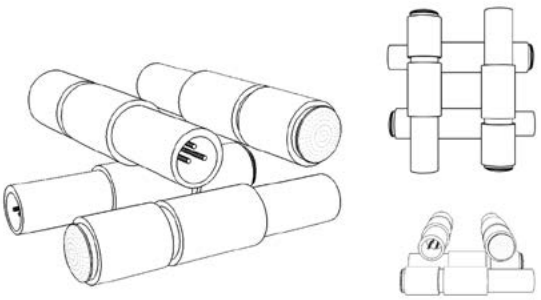

Figure 2. Microphone X-Y grid

The accuracy of this system needs to be more accurate than human listeners in order to convincingly spatialize sounds. With the final aim of informing a real-time panning system, such as a 360° dial, for live use we prefer a simpler, naïve approach.

The proposed system analyzes the speaker array using a 4-way X grid of microphones (see Fig. 2) setup in the nominal center of the space (Fig. 3). A frequency rich[1] test tone is played through each speaker in turn and recorded through the four microphones. These recordings are then analyzed and the source location estimation is performed. This information is then used to inform a panning interface.

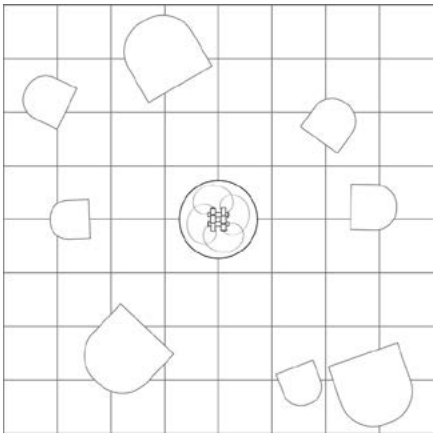

Figure 3. Diagram of microphone within array

4. MODEL AND MAXFORLIVE OBJECT

Before other aspects can be analyzed the overall latency of the audio system must be measured (i.e. the time from sound output to the return of that sound through a microphone). We accomplish this by holding a microphone on the grill of a speaker, playing a test tone through it, and calculating the interval between onsets by looking at signal threshold crossings. This latency time is used as a

[1] Tests with a straight sine tone and tests with various colors of noise resulted in widely anomalous estimations across different speaker positions. Tones with many frequencies (such as those of an acoustic instrument or voice) were found to be more consistent.

baseline to estimate speaker distances based on tone times of arrival.

Estimation of speaker position is performed using brute force loudness estimations rather than inter-aural timing differences. Given the priorities of speed and robustness this method is able to take advantage of the pickup patterns of commonly available unidirectional microphones (such as the Shure SM57, see fig. 4).

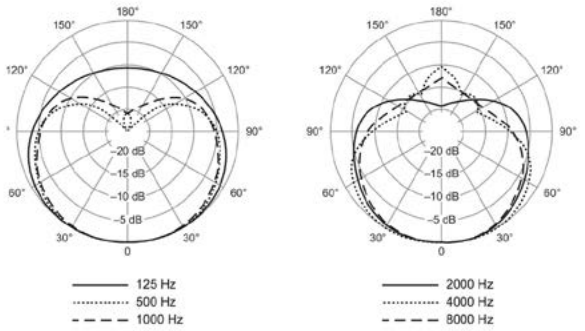

Figure 4. Shure SM57 Polar Pattern

Theoretically simple triangulation of the speaker positions (from the decibel level captured over the four mic grid) would be possible with ideally isolated microphones with precise pickup patterns. Commonly available microphones pickup much more than 90° and have non-linear input responses (i.e. discontinuous around the polar pattern of the microphone). However, given a set of four identical microphones (within the specifications of the manufacturer) it is possible to deduce position through cancellation. That is, as a sound source moves along the axis of two opposed microphones the change in measured intensity will vary in a consistent fashion. The decibel level (D) is calculated as the root mean square of one-second of audio samples ($x_{1:N}$) from one microphone.

$$D = \sqrt{\frac{(x_1^2 + x_2^2 + \ldots + x_n^2)}{n}} \quad (1)$$

With two matched uni-polar microphones facing in opposite directions the location of a sound source along the axis correlates with the signed difference between the measured input levels. This is repeated for the same sound source along the laterally perpendicular axis giving a Cartesian estimation of the source (speaker) location (axis x, y with input levels 1, 2). This allows an estimation of the angle to the sound source from the center microphone grid:

$$\theta = \tan^{-1}\left(\frac{y_1 - y_2}{x_1 - x_2}\right) \quad (2)$$

The distance that can be calculated from these measurements will be highly influenced by the characteristics of the microphones (for example, hyper-cardioid microphones pickup effectively at 90° off-axis and this would make sound sources seem closer than they are). Using the amplitude (and the theoretical reduction in decibels over distance) recorded by the microphones as an estimation of distance is similarly influenced by reflections and resonances of the environment. We found a simple time of arrival measurement performs consistently across speakers and with minimal environmental sensitivity.

The distance to each speaker is estimated based on the latency between the initiation of the test tone and the time of arrival (τ) of the same tone at the microphone grid. Removing the known system latency (z) gives a time indicating distance (d) to the speaker, using the known speed of sound at sea level (C) of 1,126 ft./second.

$$d = C(\tau - z) \quad (3)$$

With an estimation of all of the speaker locations, panning between speakers is accomplished with a software interface, implemented for Ableton Live as a MaxforLive device (see Fig. 5).

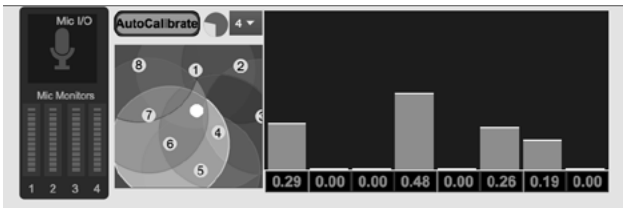

Figure 5. Multichannel panning interface prototype.

There are special configuration requirements for the Ableton Live session file to work properly with the device. The latest version of the device is implemented to Live by corresponding the estimated distance and angle data for each speaker to set the level of each Return Track, which is determined by how many external outputs (speakers) you are using.

After calibration, the method of control takes place on each individual track or group. The current control interface utilizes a node object as the main point of control. Upon dragging the unnumbered node closer or further from the numbered nodes (speakers) the levels will rise and fall accordingly in real time (Fig. 5).

5. TESTING

To evaluate the proposed system 800+ data points were recorded at around 240 different speaker positions. Four Shure SM57 microphones were used for the test grid and the same hardware was used for all data measurements. The tests were performed in a large (20x30 ft.), acoustically treated room with a minimum of sound-reflective surfaces and background noise. At each speaker position the location in the room was measured relative to the center of the microphone grid, and a 4-channel recording was captured of the test tone playback. This recording was processed as described above and the angle and distance to the speaker was estimated. The performance is characterized in Table 1 by error in estimated distance, error in estimated angle, and the magnitude of the distance error (i.e. error divided by measured distance to show scale of error). Figure 6 shows the error in angle in degrees across all data points.

The error in angle measurement appears to be independent of actual distance to the speaker (within the tested 2-30 foot range), and does not correlate with the distance to the speaker, as shown in figure 7 (error in angle, in degrees, graphed over distance to speaker).

Error:	Mean	Standard Dev.	Max.
Angle to speaker	4.95°	4.45°	23.09°
Distance to speaker	1.15 ft.	1.60 ft.	9.25 ft.
Magnitude of distance error	10.75%	13.07%	135.25%

Table 1. Estimation error.

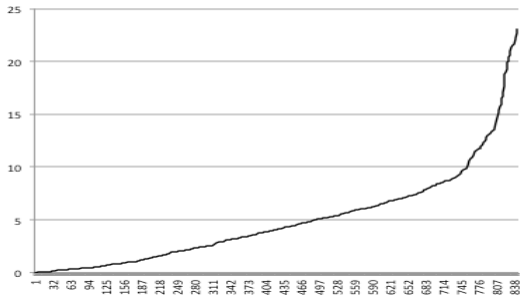

Figure 6. Error in angle (in degrees) across all test data.

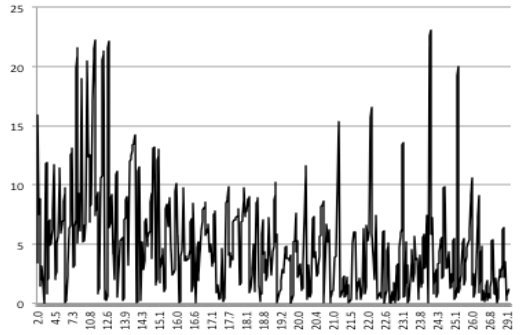

Figure 7. Error in angle over distance (in feet).

This data shows that the system can estimate the angle to a speaker within 4.45 degrees, and the distance to the speaker within 1.6 feet. The error in distance does not strongly correlate with actual distance, (i.e. the error does not increase with actual distance). Likewise the error in angle does not significantly correlate with distance (i.e. the system performs independent of actual distance).

While solutions such as [2, 3, 6, 7, 8] are able to locate sound sources with an accuracy of centimeters given smaller spaces, our system works with reasonable accuracy at a larger scale.

6. CONCLUSIONS

Considering the goal of supporting rapid deployment of speakers and minimal setup time, the software achieves a 2 second calculation time for each speaker could theoretically detect the speaker location of an 8-source array within 16 – 32 seconds depending on the level of desired accuracy. The accuracy of the estimation performs roughly twice as well as human audition, suggesting that the resulting panning system is accurate enough to satisfy a listener's discriminatory ability. Future studies with human participants will determine if practical application is satisfactory for real world use.

Frequency response of each speaker is an important characteristic in building an accurate system. However the current model does not address this aspect. Performing spectral analyses of the test tone playing through each speaker, de-convolved with the tone, should identify the frequency response of each individual speaker. This can then be used to inform an EQ calibration to ensure a uniform audio image across the entire array.

Future goals include putting this software into practice in a full 8-speaker setup. In this environment the accuracy of human listeners standing in the same position as our microphone array can be tested. Further, use tests can be conducted to compare panning algorithms with different practical sound material.

Extending the current model to enable speaker elevation detection could be accomplished through reconfiguration to a tetrahedral microphone grid (i.e. 4 microphones facing out in a pyramid formation).

7. REFERENCES

[1] Carlile, Simon. "Auditory space." *Virtual auditory space: Generation and applications*. Springer Berlin Heidelberg, 1996. 1-25.

[2] Fejzo, Zoran, and James D Johnston. 2011. "DTS Multichannel Audio Playback System: Characterization and Correction." In Audio Engineering Society.

[3] Hu, Jwu-Sheng, Chen-Yu Chan, Cheng-Kang Wang, Ming-Tang Lee, and Ching-Yi Kuo. "Simultaneous localization of a mobile robot and multiple sound sources using a microphone array." *Advanced Robotics* 25, no. 1-2 (2011): 135-152.

[4] Kostadinov, Dimitar, Joshua D Reiss, and Valeri Mladenov. 2010. "Evaluation of Distance Based Amplitude Panning for Spatial Audio." In *ICASSP*, pp. 285-288..

[5] Makous, James C., and John C. Middlebrooks. "Two‐dimensional sound localization by human listeners." *The journal of the Acoustical Society of America* 87.5 (1990): 2188-2200.

[6] Nakadai, Kazuhiro, Hiroshi G. Okuno, and Hiroaki Kitano. "Real-time sound source localization and separation for robot audition." In *INTERSPEECH*. 2002.

[7] Shi, Guangji, Martin Walsh, and Edward Stein. 2014. "Spatial Calibration of Surround Sound Systems Including Listener Position Estimation." In Audio Engineering Society.

[8] Valin, Jean-Marc, François Michaud, Jean Rouat, and Dominic Létourneau. "Robust sound source localization using a microphone array on a mobile robot." In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, vol. 2, pp. 1228-1233. IEEE, 2003.

[9] Zwiers, M., Al Van Opstal, and J. Cruysberg. "Two-dimensional sound-localization behavior of early-blind humans." *Experimental Brain Research*140.2 (2001): 206-222.