Proceedings of the 2001 Workshop on
Perceptual User Interfaces (PUI'01),
Orlando, FL, May 2001.

Audio-Video Array Source Separation for Perceptual User Interfaces

Kevin Wilson, Neal Checka, David Demirdjian and Trevor Darrell

MIT Artificial Intelligence Laboratory
{kwilson, nealc, demirdj, trevor}@ai.mit.edu

Abstract. Steerable microphone arrays provide a flexible infrastructure for audio source separation. In order for them to be used effectively in perceptual user interfaces, there must be a mechanism in place for steering the focus of the array to the sound source. Audio-only steering techniques often perform poorly in the presence of multiple sound sources or strong reverberation. Video-only techniques can achieve high spatial precision but require that the audio and video subsystems be accurately calibrated to preserve this precision. We present an audio-video localization technique that combines the benefits of the two modalities. We implement our technique in a test environment containing multiple stereo cameras and a room-sized microphone array. Our technique achieves an 8.9 dB improvement over a single far-field microphone and a 6.7 dB improvement over source separation based on video-only localization.

1 Introduction

Many current perceptual user interface applications require high-quality audio signals for acceptable performance. Examples include automated speech recognition (ASR) and smart teleconferencing. When hands-free operation is required, the most common ways to obtain audio signals for these applications are to use close-talking microphones that are attached to the speakers of interest or to use single-element directional microphones pointed at the speakers of interest.

However, both of these techniques leave much to be desired. Close-talking microphones require that each user be equipped with his own microphone, while directional microphones are often bulky and are limited to a fixed beam pattern, thus restricting their ability to track multiple users.

An alternate technique that has become more attractive with the decreasing cost of computation and digital communication is the microphone array. A microphone array consists of several microphones in fixed locations relative to each other. The microphones' audio signals can be filtered and summed to perform spatial filtering of the audio sources in the room. By altering the filters applied to the individual microphones' signals, sounds coming from different regions of the room can be selectively amplified or attenuated.

Microphone arrays address many of the problems inherent in more passive audio capture techniques. Unlike close-talking microphone systems, microphone arrays do not require users to remember to wear special equipment when they anticipate that they will interact with the environment. Instead, microphone arrays have, as a fundamental property, an explicit notion of the spatial relationships among sound sources.

This association between sound and location makes a microphone array a powerful tool in the context of perceptive environments. In combination with additional sensors and contextual information from the environment, a microphone array can effectively amplify and separate sounds of interest from complex background noise.

To focus a microphone array, the location of the speaker(s) of interest must be known in order for the microphone array to modify its filter response to amplify the selected speakers. A number of techniques exist for localizing sound sources using the array data itself [12], but the performance of these localization techniques tends to degrade significantly in the presence of reverberation and/or multiple sound sources. Unfortunately, most common office and meeting room environments are highly reverberant, with reflective wall and table surfaces, and will normally contain multiple speakers.

In our application, we can take advantage of other sensors in the perceptive environment domain to perform multimodal localization of multiple speakers despite reverberation. Because the wavelength of visible light is much smaller than the wavelength of audible sound, cameras can be much more precise in their localization, and multiple users can be more easily segmented in space.

Cameras, however, are not perfect for steering a microphone array. It may be difficult to obtain a precise joint calibration between the cameras and the microphone array. In addition, the features that a camera-based system can most easily track, such as extremities of the body, are not directly relevant to the microphone array; the microphone array requires information about the location of the speaker's mouth, which is difficult to obtain from wide-angle camera views of the environment.

Because of these issues, a microphone array aimed using only information from a set of cameras will likely be incorrectly aimed, resulting in a loss of several decibels of performance and an undesirable spectral coloration of the signal of interest. In spite of these problems, video localization information is accurate enough to restrict the range of possible acoustic source locations to a region small enough to allow for acoustic localization techniques to operate without severe problems with reverberation and multiple speakers.

As far as we are aware, our system is the first visually guided large-aperature microphone array. This paper demonstrates the use of 3-D visual localization in combination with acoustic localization to acquire high-quality audio speech signals from moving users in a perceptually enabled environment.

2 Background

This work brings together techniques from array signal processing with techniques from vision-based person tracking to implement a system that can selectively amplify audio from a selected speaker as he moves through the room. Much work has been done in both of these areas. The relevant background is summarized below.

2.1 Microphone arrays

Microphone arrays are a special case of the more general problem of sensor arrays, which have been studied extensively in the context of applications such as radar and sonar [11]. The Huge Microphone Array project[10] is investigating the use of very large arrays containing hundreds of microphones. Their work concentrates on audio-only solutions to array processing. Another related project is Wang and Brandstein's audio-guided active camera[13], which uses audio localization to steer a camera on a pan/tilt base.

Many problems can be addressed through array processing. The two array processing problems that are relevant to our system are beamforming and source localization.

Beamforming is a type of spatial filtering in which the signals from individual array elements are filtered and added together to produce an output that amplifies signals coming from selected regions of space and attenuates sounds from other regions of space. In the simplest form of beamforming, delay-and-sum beamforming, each channel's filter is a pure delay. The delay for each channel is chosen such that signals from a chosen "target location" are aligned in the array output. Signals from other locations will tend to be combined incoherently. For example, if a three element array consists of elements that are 2, 4, and 7 meters away from a target location, the elements' signals should be delayed by the time that it takes for sound to travel 5, 3, and 0 meters, respectively. This type of beamforming is simple and robust to small uncertainties in microphone and target locations.

Source localization is a complementary problem to beamforming whose goal is to estimate the location of a signal source. One way to do this is to beamform to all candidate locations and to pick the location that yields the strongest response. This method works well, but the amount of computation required to do a full search of a room is prohibitively large. Another method for source localization consists of estimating relative delays among channels and using these delays to calculate the location of the source. Delay-estimation techniques are computationally efficient but tend to perform poorly in the presence of multiple sources and/or reverberation.

A number of projects [2–4] have used vision to steer a microphone array, but because they use a single camera to steer a far-field array, they cannot obtain or make use of full 3-D position information; they can only select sound coming from a certain direction.

For microphone arrays that are small in size compared to the distance to the sources of interest, incoming wavefronts are approximately planar. Because of this, only source direction can be determined; source distance remains ambiguous. When the array is large compared to the source distance, the sphericity of the incoming wavefronts is detectable, and both direction and distance can be determined. These effects of array size apply both to localization and to beamforming, so if sources at different distances in the same direction must be separated, a large array must be used.

As a result, with large arrays the signal-to-noise ratio (for a given source) at different sensors will vary with source location. Because of this, signals with better signal-to-noise ratios should be weighted more heavily in the output of the array. Our formulation of the steering algorithm presented below takes this into account.

2.2 Person tracking

Tracking people in known environments has recently become an active area of research in computer vision. Several person-tracking systems have been developed to detect the number of people present as well as their 3D position over time. These systems use a combination of foreground/background classification, clustering of novel points, and trajectory estimation over time in one or more camera views [6, 9].

Color-based approaches to background modeling have difficulty with illumination variation due to changing lighting and/or video projection. To overcome this problem, several researchers have supported the use of background models based on stereo range data [6, 8]. Unfortunately, most of these systems are based on computationally intense, exhaustive stereo disparity search.

We have developed a system that can perform dense, fast range-based tracking with modest computational complexity. We apply ordered disparity search techniques to prune most of the disparity search computation during foreground detection and disparity estimation, yielding a fast, illumination-insensitive 3D tracking system. Details of our system are presented in [5]; here we review the details of our visual tracking system which are relevant to the integration with audio processing in our microphone array.

When tracking multiple people, we have found that rendering an orthographic vertical projection of detected foreground pixels is a useful representation (see also [1, 9]). A ”plan view” image facilitates correspondence in time since only 2D search is required. Previous systems would segment foreground data into regions prior to projecting into a plan-view, followed by region-level tracking and integration, potentially leading to sub-optimal segmentation and/or object fragmentation. Instead, we develop a technique that altogether avoids any early segmentation of foreground data. We merge the plan-view images from each view and estimate over time a set of trajectories that best represents the integrated foreground density. Trajectory estimation is performed by finding connected components in a spatio-temporal filtered volume.

To estimate the trajectory of objects over time, we combine information from multiple stereo views. The true extent of an individual object in a given image is generally difficult to identify. An optimal trajectory segmentation should consider the assignment of an individual pixel to all possible trajectories estimated over time. Systems which perform an early segmentation and grouping of foreground data before trajectory estimation preclude this possibility.

We adopt a late-segmentation strategy that finds the best trajectory in an integrated spatio-temporal representation by combining foreground pixels from each view. By assuming that objects move on a ground plane, a “plan-view assumption” allows us to completely model instantaneous foreground information as a 2-D orthographic density

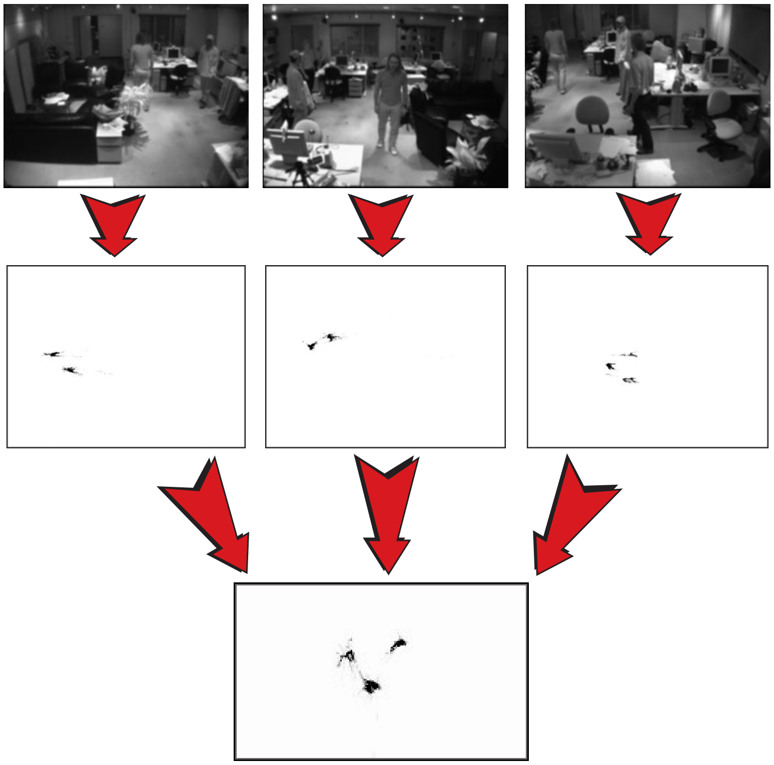

Fig. 1. Detecting locations of users in a room using multiple views and plan-view integration. Three people are standing in a room, though not all are visible to each camera. Foreground points are projected onto a ground plane. Ground plane points from all cameras are then superimposed into a single data set before clustering the points to find person locations.

projection. Over time, we compute a 3-D spatio-temporal plan-view volume.

We project (x_j, y_j, d_j) from each foreground point $\boldsymbol{p}_j$ into world coordinates (U_j, V_j, W_j). (See Figure 4.) U, V are chosen to be orthogonal axes on the ground plane, and W normal to the ground plane. We then compute the spatio-temporal plan view volume (Figure 1), with

$$P(u, v, t) = \sum_{\{\boldsymbol{p}_j | U_j = u, V_j = v, t_j = t\}} 1$$

Each independently moving object in the scene generates a continuous volume in the spatio-temporal plan view volume $P(u, v, t)$. When the trajectories of moving objects do not overlap, the trajectory estimation is easy and consists in running a connected-component analysis in $P(u, v, t)$ (each component is then a trajectory).

When the trajectories of moving objects overlap (*e.g.* crossing of two people), the volume associated with these trajectories in $P(u, v, t)$ also overlap and make the extraction of trajectories more difficult. In order to overcome this, a graph is built from a piece-wise connected-component analysis of $P(u, v, t)$. Nodes correspond here to trajectory crossing and branches to non-ambiguous trajectories between two crossing. A color histogram is then estimated for each branch of the graph (using all images associated with this branch). Trajectories are estimated by finding in the graph the paths consisting of branches having the most similar color histograms. This may be done instantaneously using a greedy search strategy or using the slower but optimal dynamic programming technique described in [5].

3 Large-array volume selection

Our system performs both audio localization and beamforming with a large, ceiling-mounted microphone array. Localization uses information from both audio and video, while beamforming uses only the audio data and the results of the localization processing. A large array gives the ability to select a *volume* of 3-D space, rather than simply form a 2-D beam of enhanced response as anticipated by the standard array localication algorithms. However, the usual assumption that of constant target signal-to-noise ratio (SNR) across the array does not hold when the array geometery is large (array width on same scale as target distance.) As described below, we need to model the SNR term in the array localization algorithm.

3.1 Localization

Our system uses the location estimate from the vision tracker as the initial guess from which to begin a gradient ascent search for a local maximum in beam output power. Beam power is defined as the integral over a half-second window of the square of the output amplitude.

It is difficult to characterize the error in the tracker's estimate because this error depends on the person's position in the room, the person's appearance, and a number of other characteristics of the situation. However, experience leads us to believe that the vision tracker is accurate to within less than one meter. Gradient ascent to the nearest local maximum can therefore be expected to converge to the location of the speaker of interest when no other speakers are very close by.

Gradient ascent is complicated by the fact that there are many high-spatial-frequency ripples superimposed on the large-scale peak whose maximum we wish to find. These small ripples in the response result in many undesirable local maxima that must be avoided. Because speech is a broadband signal, it is possible to start the gradient ascent using a low-passed version of the speech signal. As the peak is approached, the cut-off frequency of the filter can be raised, thus incorporating more of the signal into the location estimate. This technique is similar to one used in [7] as part of an exhaustive search for a power maximum.

3.2 Source separation

For small microphone arrays, the relative SNRs of the individual channels do not vary significantly as a function of source location. This is, however, not true for larger microphone arrays. For our array, which is roughly 4 meters across, we must take into account the fact that some elements will have better signals than others. Specifically, if we assume that we have signals x_1 and x_2 which are versions of the unit-variance desired signal, s, that have been contaminated by unit-variance uncorrelated noise, we can analyze the problem as follows:

$$x_1 = a_1 s + n_1$$

$$x_2 = a_2 s + n_2$$

In this model, the signal to noise ratios of x_1 and x_2 will be a_1^2 and a_2^2, respectively. Their optimal linear combination will be of the form $y = bx_1 + x_2$. Because of the uncorrelated noise assumption, the SNR of this combination will be

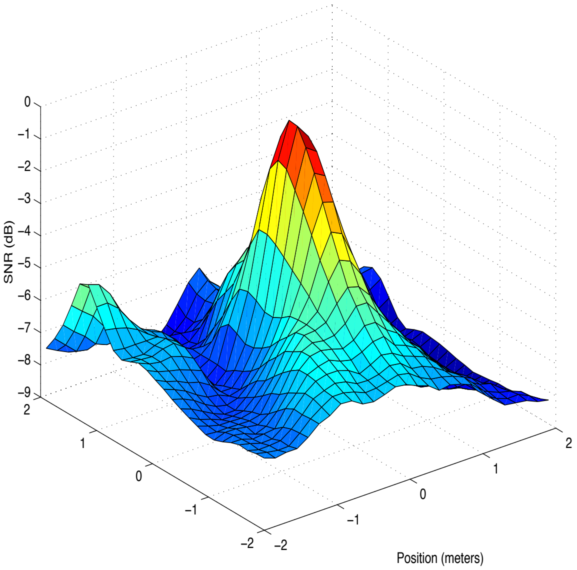

Fig. 2. Array power response as a function of position (single speaker close-up). This plot shows the array output power as the array's focus is scanned through a plane centered on a speaker.

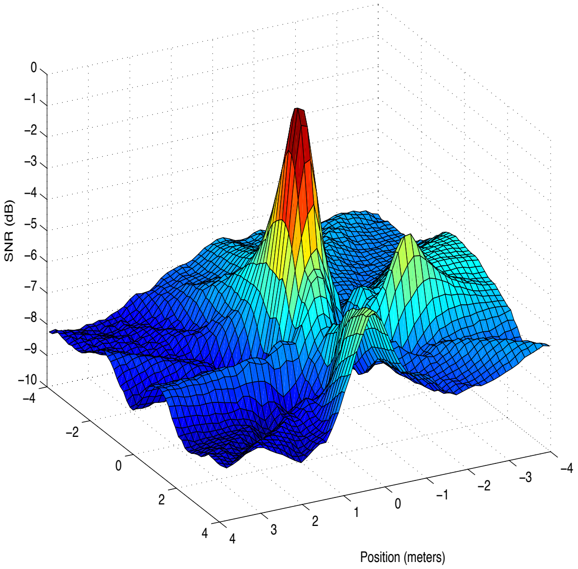

Fig. 3. Array power response as a function of position (two speakers). This plot shows the array output power as the array's focus is scanned through a plane centered on one speaker while another speaker is nearby. The central speaker is easily discernible in the plot, but the peak corresponding to the weaker speaker is difficult to distinguish among the sidelobe peaks.

$$SNR(y) = \frac{(ba_1 + a_2)^2}{b^2 + 1}$$

By taking the derivative of this expression with respect to b and setting the result equal to zero, one finds that the optimal value of b is:

$$b = \frac{a_1}{a_2} = \frac{SNR(x_1)}{SNR(x_2)}$$

Because of the symmetry of the signals, this result implies that, in general, individual elements' signals should be scaled by a constant proportional to the square root of their SNRs.

Ideally, we would like to have complete knowledge of the strengths and statistical relationships among the noise signals at the individual sensors. This information is not easy to obtain, but because of our large array and multiple stereo cameras, it is easy for us to use our location estimate to weight individual channels assuming a $1/r$ attenuation due to the spherical spreading of the source. Assuming $1/r$ attenuation from a source to each microphone, we have $a_n = 1/r_n$ in the above equations, so the optimal weighting factor for channel n is $1/r_n$. This is intuitively appealing since it means that microphones far from the source contribute relatively little to the array output.

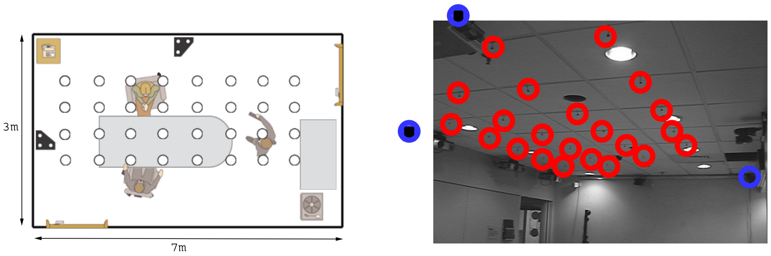

Fig. 4. The test environment. On the left is a schematic view of the environment with stereo cameras represented by black triangles and microphones represented by empty circles. On the right is a photograph of the environment with microphones and camera locations highlighted.

4 Results

Our test environment, depicted in Figure 4, is a conference room equipped with 32 omnidirectional microphones spread across the ceiling and 2 stereo cameras on adjacent walls.

The audio and video subsystems were calibrated independently, and for our experiments, we performed a joint calibration by finding the least-squares best-fit alignment between the two coordinate systems.

Table 1. Audio-video localization performance comparison

Localization Technique	SNR (dB)
Single microphone	−6.6
Video only	−4.4
Audio only	2.0
Audio-Video	2.3

Figure 3 is an example of what happens when multiple speakers are present in the room. Audio-only gradient ascent could easily find one of the undesirable local maxima. Because our vision-based tracker is accurate to within one meter, we can safely assume that we will find the correct local maximum even in the presence of interferers.

To validate our localization and source separation techniques, we ran an experiment in which two speakers spoke simultaneously while one of them moved through the room. We tracked the moving speaker with the stereo tracker and processed the corresponding audio stream using three different localization techniques. For each, we used a reference signal collected with a close-talking microphone to calculate both a time-averaged SNR (Table 1) and a sequence of short-time SNRs (Figure 5).

As a reference for performance comparison, we use the signal from a single microphone near the center of the room. This provides no spatial selectivity, but for our scenario it tends to receive the desired speech more strongly than the interfering speech. The SNR for this case is negative because of a combination of the interfering speaker and diffuse noise from the room's ventilation system.

To evaluate the video-only approach, we steer the microphone array directly to the location returned by the stereo tracker. If the stereo tracker could reliably return the location of the speaker's mouth, this method would work quite well. For our system, this technique improves

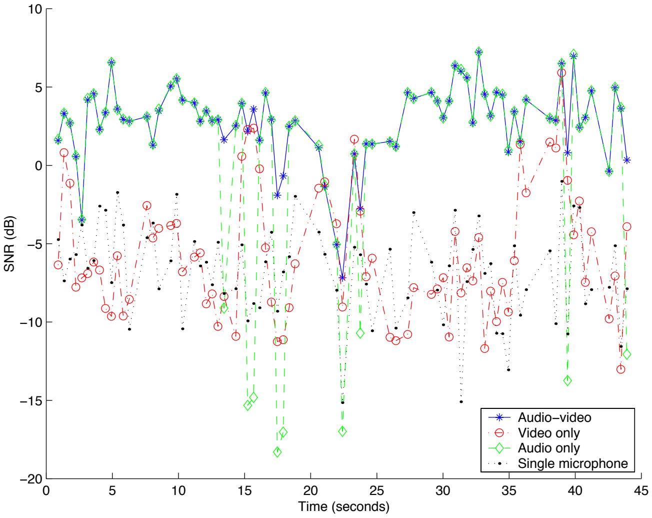

Fig. 5. Short-time signal-to-noise ratio comparison. We calculate a sequence of SNRs for non-overlapping half-second windows of audio. Much of the variation in the SNR of audio-video result is due to variations in speech energy.

the SNR by 2.2 dB, which, while noticeable, is not close to the theoretical performance of a 32 element array (15 dB in uncorrelated noise). Figure 5 shows large fluctuations in SNR for this and other methods. For some t, all three curves are low, corresponding to times when the speaker pauses between words. Other fluctuations for this technique, however, are due to stereo tracking errors and other biases of the stereo system or microphone array.

To evaluate the audio-only approach, we search the room for the location of maximum acoustic power and steer the array to that location. For our test scenario, this worked quite well when tracking the louder speaker. Even so, there are several points in time where the array locks onto the interfering speaker. When attempting to track the quieter speaker, this method fails completely.

The fourth entry in Table 1 and Figure 5 uses the stereo tracker's location estimate as the initial guess from which to perform gradient ascent in the signal output power. This technique's average SNR is well above that of either the single-microphone or video-only methods, and its short-time SNRs are consistently highest or nearly the highest of any of the four techniques.

These experiments demonstrate that audio-video localization is superior to video alone in our environment. We believe our approach improves upon audio-only localization in cases where there are multiple simultaneous speakers and the reverberant energy is nearly equal or greater than the direct path energy. The initial position estimate provided by video localization reduces the amount of computation required compared to an unconstrained audio-only search.

5 Conclusion

We have implemented a computationally efficient hybrid sound source localization scheme. This scheme makes use of the complementary information available in the audio and video streams available in our test environment and is suitable for use as part of perceptive environments requiring high-quality audio signals for higher-level applications such as automated speech recognition.

In the future, we plan to incorporate more sophisticated beamforming techniques into our system to further improve the SNR of the output. In addition, we hope to be able to feed the final results of the audio-video localization back to the vision subsystem to allow it to refine its location and trajectory estimates.

6 Acknowledgments

We would like to thank the NIST Smart Spaces Laboratory for the use of their microphone array infrastructure, upon which our microphone array is based. We would also like to thank Corinna Sherman for her help in conducting the experiments.

References

1. D. J. Beymer and K. Konolige. Real-time tracking of multiple people using stereo. In *Frame-Rate Workshop*, 1999.
2. U. Bub, M. Hunke, and A. Waibel. Knowing who to listen to in speech recognition: Visually guided beamforming. In *1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995.
3. M. Casey, W. Gardner, and S. Basu. Vision steered beam-forming and transaural rendering for the artificial life interactive video environment, (alive). In *99th Convention of the Audio Engineering Society*, 1995.
4. M. Collobert, R. Feraud, G. LeTourneur, O. Bernier, J. E. Viallet, Y. Mahieux, and D. Collobert. Listen: a system for locating and tracking individual speakers. In *2nd International Conference on Face and Gesture Recognition*, 1996.
5. T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb. Plan-view trajectory estimation with dense stereo background models. In *2001 International Conference on Computer Vision*, 2001.
6. T. Darrell, G. G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *IJCV*, (37(2)):199–207, June 2000.
7. R. Duraiswami, D. Zotkin, and L. S. Davis. Active speech source localization by a dual course-to-fine search. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.
8. Y. A. Ivanov, A. F. Bobick, and J. Liu. Fast lighting independent background subtraction. *IJCV*, 2000.
9. J. Krumm, S. Harris, B. Meyers, B. Brummit, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. In *3rd IEEE Workshop on Visual Surveillance*, 2000.
10. H. F. Silverman, W. R. Patterson, and J. L. Flanagan. The huge microphone array. *IEEE Concurrency*, pages 36–46, October 1998.
11. Barry D. Van Veen and Kevin M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, April 1988.
12. M. Viberg and H. Krim. Two decades of statistical array processing. In *31st Asilomar Conference on Signals, Systems, and Computers*, 1997.
13. C. Wang and M. Brandstein. Multi-source face tracking with audio and visual data. In *IEEE International Workshop on Multimedia Signal Processing*, 1999.