Audiovisual arrays for untethered spoken interfaces

Kevin Wilson, Vibhav Rangarajan, Neal Checka, Trevor Darrell
Artificial Intelligence Lab, M.I.T., Cambridge, MA, 02139 USA
trevor@ai.mit.edu

Abstract

When faced with a distant speaker at a known location in a noisy environment, a microphone array can provide a significantly improved audio signal for speech recognition. Estimating the location of a speaker in a reverberant environment from audio information alone can be quite difficult, so we use an array of video cameras to aid localization. Stereo processing techniques are used on pairs of cameras, and foreground 3-D points are grouped to estimate the trajectory of people as they move in an environment. These trajectories are used to guide a microphone array beamformer. Initial results using this system for speech recognition demonstrate increased recognition rates compared to non-array processing techniques.

1 Introduction

Most existing conversational speech systems require *tethered* interaction, and work primarily for a single user. Users must wear an attached microphone or speak into a telephone handset, and do so one at a time. This limits the range of use of dialog systems, since in many applications users might expect to freely approach and interact with a device. We are interested in building a system for *untethered* spoken interface, where multiple users can move about an environment and speak to a computer system.

With only a single sensing modality disambiguating the audio from multiple speakers can be a challenge. But with multiple modalities, and possibly multiple sets of sensors, segmentation can become feasible. In this paper we describe a audio-visual array approach to tracking speakers in a noisy environment.

We have developed our system in a “smart environment” or “smart room” enabled with multiple stereo cameras and a ceiling mounted large-aperture microphone array grid. Users can move arbitrarily in the room or environment while focused audiovisual streams are generated from their appearance and utterance. In our system multi-view image correspondence and tracking methods are combined with acoustic beamforming techniques to focus a virtual microphone on each speaker. Our multimodal approach can track sources even in acoustically reverberant environments with dynamic illumination, conditions that are tough for audio or video processing alone.

First we review related work, and then present our method for geometric source separation and vision-guided microphone array processing. We show results integrating our technique with a conversational speech system, and describe avenues for future work combining other types of audiovisual multimodal information.

2 Related Work

Several authors have explored geometric approaches to audiovisual segmentation using array processing techniques. 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. A number of projects [1, 2, 3] 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.

3 Audiovisual array processing

To focus a microphone array, the location of the speaker(s) of interest must be known. A number of techniques exist for localizing sound sources using only acoustic cues [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. However, in a multimodal setting we can take advantage of other sensors in the environment to perform localization of multiple speakers despite reverberation. We use a set of cameras to track the position of speakers in the environment, and report the relative geometry of speakers, cameras, and microphones.

The vision modality is not affected by acoustic reverberation, but its accuracy will depend on the the calibration and segmentation procedures. In practice we use video information 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.

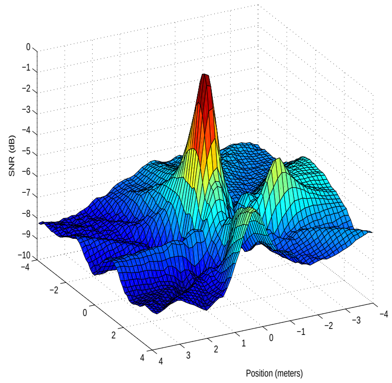

Figure 1. Array power response as a function of position (two speakers). This plot shows the array output power as the array's focus is scanned through a plane centered on one speaker while another speaker is nearby. The central speaker is easily discernible in the plot, but the peak corresponding to the weaker speaker is difficult to distinguish among the sidelobe peaks. Using vision-based person tracking cues can disambiguate this case.

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.

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 may have dif£culty in the presence of multiple sources and/or reverberation.

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.

3.1 Person tracking with multiple stereo views

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 [5, 9].

Color-based approaches to background modelling 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 [5, 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 [4]. Our system reports the 3-D position of people moving about an environment equipped with an array of stereo cameras.

3.2 Vision-guided acoustic volume selection

We perform 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 localization algorithms. However, the usual assumption that of constant target signal-to-noise ratio (SNR) across the array does not hold when the array geometry is large (array width on same scale as target distance.)

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. 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

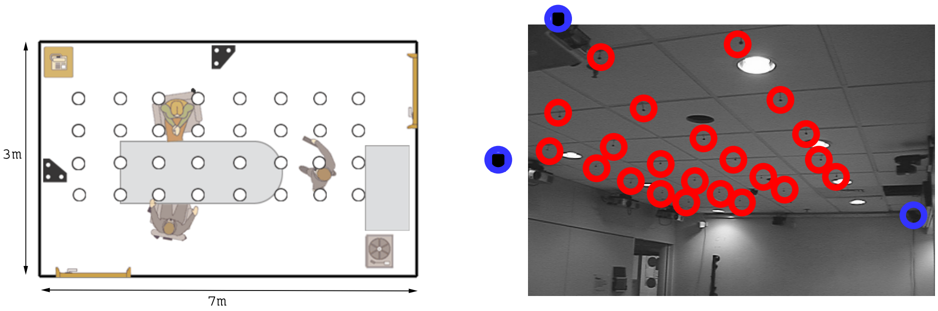

Figure 2. The test environment. On the left is a schematic view of the environment with stereo cameras represented by black triangles and microphones represented by empty circles. On the right is a photograph of the environment with microphones and camera locations highlighted.

are very close by.

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

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

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:

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

	SNR (dB)
Distant microphone	-6.6
Video only	-4.4
Audio only (dominant speaker)	2.0
Audio-Video	2.3

Table 1. Audio-video localization performance.

Individual elements' signals should be scaled by a constant proportional to the square root of their SNRs. We use the location estimate to weight individual channels assuming a $1/r$ attenuation due to the spherical spreading of the source: $a_n = 1/r_n$.

4 Results

Our test environment, depicted in Figure 2, 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.

Figure 1 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-

-12db Interferer	Male	Female
Close-talking microphone	95	95
Microphone array	64	13
Distant microphone	51	7
-24db Interferer	Male	Female
Close-talking microphone	95	96
Microphone array	80	41
Distant microphone	74	28
No Interferer	Male	Female
Close-talking microphone	95	96
Microphone array	82	43
Distant microphone	73	35

Table 2. Word recognition rates (percent correct) calculated in each condition from 5 male and 3 female speakers. The close-talking microphone was clipped to the lapel of the speaker. The microphone array is as described above. The distant microphone is one array element from near the center of the room.

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 a time-averaged SNR (Table 1). For performance comparison we use the signal from a single distant 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 the single microphone case is negative because of a combination of the interfering speaker and diffuse noise from the room's ventilation system.

To evaluate the microphone array's effects on recognition rates for automated speech recognition (ASR), we connected our system to the MIT Spoken Language Systems (SLS) Group's JUPITER weather information system [14]. We had two male speakers issue each of nine weather-related queries from two different locations in the room. As collected, the data contains quiet but audible noise from the ventilation system in the room. To evaluate the results under noisier conditions, additional noise was added to these signals. The results are shown in Table 2. The interferer used in these experiments was a male speaker located one to two meters from the target speaker. The interferer level (-24 dB or -12 dB) is a rough measure relative to the male test subjects. The same absolute interferer level was used for all test subjects. The beamformed signal from the microphone array was in all cases superior to the single distant microphone. The distant microphone, which was approximately 1.5m from the speaker, yielded recognition rates that were too low to be useful in our current environment. While recognition rates in low noise case were high enough to be useful, the -12 dB interferer significantly degraded the performance of the array. We are currently working on adaptive null-steering algorithms that should improve performance in the presence of stronger interferers such as this.

5 Conclusion and Future Work

We have shown how an audiovisual array approach can help enable untethered conversational interaction. Our approach was primarily geometric, using 3-D tracking and array processing. In our current architecture vision processing was used to for coarse-scale localization and audio cues for fine-scale localization; we are exploring a symmetric localization architecture where both audio and video cues can influence coarse or fine tracking.

In other work, we have demonstrated how modelling joint appearance variation can also be used to reduce environmental noise and identify corresponding audio and video from individual speakers [6, 7]. The statistical approach used a mutual information analysis of appearance and spectral variation and ignored 3-D geometry. In contrast to the geometric approach presented above, it worked with just a single microphone and camera. While each approach is valuable in its intended domain, it is clear that they are orthogonal and would benefit from combination. We are thus exploring an integrated approach that combines geometric and statistical insights in a common source separation algorithm.

Acknowledgements

This research has been supported in part by a National Science Foundation Graduate Research Fellowship, NTT, and MIT Project Oxygen.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Trevor Darrell, John Fisher, Paul Viola, and Freeman Bill. Audio-visual segmentation and the cocktail party effect. In *Proceedings of the International Conference on Multimodal Interfaces*, Oct 2000.

[7] John W. Fisher III and Trevor. Darrell. Probabalistic models and informative subspaces for audiovisual correspondence. In *Proceedings ECCV 2002.*, 2002.

[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.

[14] V. Zue, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington. Jupiter: a telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96, 2000.