A Multi-Modal Approach for Determining Speaker Location and Focus

Michael Siracusa siracusa@ai.mit.edu
Louis-Philippe Morency lmorency@ai.mit.edu
Kevin Wilson kwilson@ai.mit.edu
John Fisher fisher@ai.mit.edu
Trevor Darrell trevor@ai.mit.edu

Computer Science and Artificial Intelligence Laboratory at MIT. Cambridge, MA 02139.

ABSTRACT

This paper presents a multi-modal approach to locate a speaker in a scene and determine to whom he or she is speaking. We present a simple probabilistic framework that combines multiple cues derived from both audio and video information. A purely visual cue is obtained using a head tracker to identify possible speakers in a scene and provide both their 3-D positions and orientation. In addition, estimates of the audio signal's direction of arrival are obtained with the help of a two-element microphone array. A third cue measures the association between the audio and the tracked regions in the video. Integrating these cues provides a more robust solution than using any single cue alone. The usefulness of our approach is shown in our results for video sequences with two or more people in a prototype interactive kiosk environment.

1. INTRODUCTION

Currently, most speech interfaces are designed for a single speaker. Designing systems that can understand conversational dynamics between multiple people is a difficult task. We would like to have a system that can understand not only who is speaking but to whom they are speaking. This should also be done without burdening the user with devices such as microphone headsets. The use of multiple modalities and multiple sensors makes the problem more tractable.

Visual cues alone can be used to determine who is in the field of view, where they are facing and if their lips are moving. However, these cues cannot tell if a subject's lip movement is caused by speaking or by some other process such as a change in expression. Audio cues can tell us when someone is speaking, and even where the sound is coming from. However, audio alone cannot identify who the speaker is speaking to and whether or not that speaker can be seen by the camera. With audio and video cues from multiple cameras and microphones, we can estimate where people are, toward whom (or what) they are facing, whether audio is coming from their direction, and whether their lips are moving synchronously with the audio.

In this paper, we discuss a prototype system which statistically fuses these cues, serving as a front end for a multi-person conversational kiosk. Our system combines modules for stereo-based head tracking, direction of arrival (DOA) estimation with a microphone array, and audio/visual synchrony processing. We show these three components allow for robust front-end processing in a multi-person conversational interface.

Faces have been used as salient cues for a conversational interface; [13] demonstrated a state-of-the-art machine learning approach to detecting face patterns. The use of face detection and tracking in an interactive kiosk was demonstrated in [1], and [2] showed how fine-grained pose tracking can estimate the conversational target of a user in an interactive environment when the speakers wore an attached microphone.

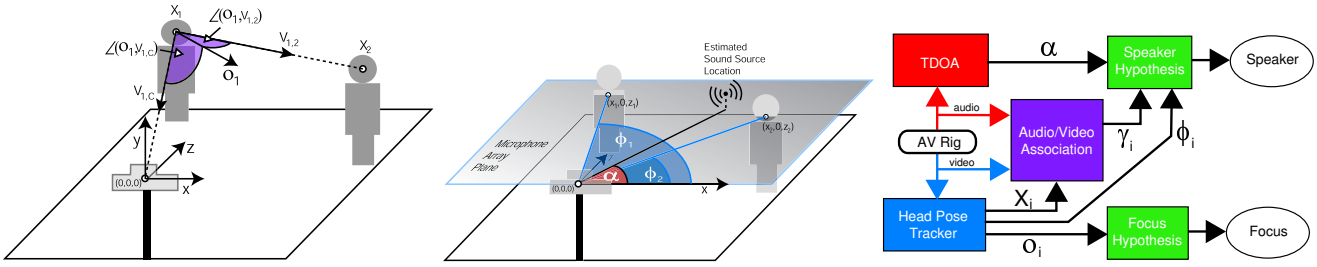
User pose, proximity and visual speech activity were combined with simple fusion rules in [6] to determine whether a person is speaking to the camera and to enable automatic control of a speech recognition system in a traditional desktop environment. Several systems for speaker detection using visual cues have been proposed using Bayesian Networks [9, 11]. These systems exploit a sophisticated statistical model for fusion, but only limited audio and visual cues from a single microphone and single camera, and were primarily designed for a single speaker.

There are many approaches to speaker localization from one or more audio sensors. Microphone array processing can be used to estimate the DOA from one or more sources [7]. In [10] a microphone array and a single camera was used to locate a speaker in a scene; they used a time-difference-of-arrival (TDOA)/cross correlation technique to locate the direction of speakers relative to the interface. They do not consider to whom the person is speaking, nor handle speakers who are not in the camera view but possibly in the same direction as a visible person (relative to the array).

With just a single microphone and video information, audio/visual synchrony can associate an utterance with two or more possible visual targets. Hershey and Movellan treat the audio and video signals as two separate random variables and measure the correlation between them [4] with the implicit assumption that the audio and video are individually and jointly Gaussian over a small window of time. Others have considered projections or mappings into low dimensional subspaces. Canonical correlation analysis is considered in [12]. The non-Gaussian case is considered in [5].

None of the above systems can simultaneously and ro-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.
Copyright 2003 ACM 1-58113-621-8/03/0011 ...$5.00.



(a) Focus Model (b) Audio DOA Model (c) System Block Diagram

Figure 1: Models for determining speaker intent and audio source localization. The camera and array coordinate systems were aligned and their X,Z plane is parallel to the ground. By positioning the microphone array at approximately mouth level, we can make the simplifying assumption that speech sources emanate from within the microphone array plane.

bustly discern the conversational target of two or more users and identify when an off screen audio source in the direction of a visible user is actually responsible for a received utterance. We believe we are the first to combine these three necessary components - head pose tracking, microphone array processing, and audiovisual synchrony detection - to achieve this goal.

2. VISUAL AND AUDIO CUES

Our first cue is obtained using a six degree of freedom head pose tracker. We estimate the head orientation and position of each visible speaker independently using adaptive view-based appearance models created online [8] from a two-frame registration algorithm which combines the robustness of ICP (Iterative Closest Point) and the precision of the normal flow constraint. Our tracking technique takes advantage of depth information available from a stereo camera [3] which makes it less sensitive to lighting variations.

For each person i tracked we obtain a 3D position, X_i, which represents their head's (x, y, z) location in the camera's coordinate system. In addition, we are given parameters for the person's head orientation, O_i, which is shown as a vector in Figure 1(a)

Our second cue is derived purely from the audio. We determine the DOA of the audio source based on estimates of the TDOA between the microphones in the array. To estimate TDOA, we use the standard technique of finding the maximum in the microphone signals' time-domain cross-correlation function. In general, the uncertainty in this TDOA estimate will depend on the signal-to-noise ratios (SNRs) of the microphone signals and the bandwidth of the audio source. In our prototype system, we make the simplifying assumption that the SNRs and bandwidths are fixed, and we test on scenarios where these assumptions are approximately true.

Given our estimated TDOA and knowing the geometry of our microphone array, we estimate a DOA, α. To incorporate the uncertainty of the DOA estimate into our hypothesis framework, we calculated the sample variance, σ_α^2, of a training dataset from a speaker at a known location. Knowing the array geometry allows us to define a function, $\beta(\theta; \sigma_\alpha)$, which expresses the uncertainty at any arbitrary DOA, θ, parameterized by σ_α^2. In particular, our A/V rig has lower DOA uncertainty directly in front of the camera than it does to the sides.

Our third cue combines information from both the audio and video streams. We use the information from the head tracker to obtain a stabilized region of interest (ROI) in the video, $V_i(x, y)$, for each person i. To measure the synchrony of each region of interest with the audio signal we calculate a correlation/association score in a style similar to [4].

$$C_t(x, y) = \frac{1}{2} \log \left(\frac{|\Sigma_A||\Sigma_{V_i(x,y)}|}{|\Sigma_{A,V_i(x,y)}|} \right) \quad (1)$$

where Σ_A , $\Sigma_{V_i(x,y)}$ and $\Sigma_{A,V_i(x,y)}$ are estimates of the covariance for marginal and joint Gaussian distributions of A, $V_i(x, y)$, and $A * V_i(x, y)$ respectively. These estimates are made over a one second window centered at time t. For single dimensional audio and video representations $C_t(x, y)$ is a function of the Pearson's correlation coefficient ρ [4].

The association calculation in Equation 1 is performed independently for each pixel location (x, y) in $V_i(x, y)$. We use the output of a horizontal edge filter on the region of interest for our video feature, and the energy in the 1kHz to 6kHz band as our audio representation. This energy is calculated over a window associated with each video frame.

In an attempt to smooth out any spuriously high correlation scores we define a variable γ_i to give us a sense of the overall association of the audio and video at time t by treating each pixel and frame independently and calculating the average of value $C_t(x, y)$ over all x and y. This average, γ_i, is computed over a window length of one second.

3. INTEGRATION

In order to facilitate the integration of different modalities into our system we adopt a statistical framework. We enumerate a set of hypothesis and associated statistical measurement models. While more sophisticated measurement models might be available, we purposely choose simple models which satisfy the requirements of a hypothesis test, agree with our intuition, and do not exceed our computational capacity.

We decompose the hypotheses into two subsets. The first set, via the head pose tracker, infers the focus of each subject asking "Who is looking at who/what?". We refer to this as our focus hypothesis. The second set integrates DOA information with the measure of audio/visual association to form our speaker hypothesis which asks the question "Which subject is speaking?". We assume a higher level process, such

as a speech recognizer, has told our system when someone is speaking.

3.1 Focus Hypothesis

The hypotheses used to describe the focus, Hf, of each subject i out of the possible M that are tracked, are enumerated as follows:

$\mathsf{Hf}_{i,\emptyset}$	:	person i is looking in a random direction, or at someone who is not being tracked.
$\mathsf{Hf}_{i,j}$	:	person i is looking at person j.
$\mathsf{Hf}_{i,c}$	:	person i is looking at the camera

with the associated statistical models:

$$p\left(O_i|\mathsf{Hf}_{i,\emptyset}\right) = \text{unif}\left(O_i\right) \quad (2)$$

$$p\left(O_i|\mathsf{Hf}_{i,j}\right) = \mathcal{N}\left(\angle(O_i, V_{i,j}); 0, \sigma_p\right) \quad (3)$$

$$p\left(O_i|\mathsf{Hf}_{i,C}\right) = \mathcal{N}\left(\angle(O_i, V_{i,c}); 0, \sigma_c\right) \quad (4)$$

where $\mathcal{N}(\cdot; \mu, \sigma)$ indicates a Gaussian density with mean μ and standard deviation σ and $i, j \epsilon \{1...M\}$. Equation 2 assumes a uniform distribution over the possible directions a subject can be looking toward when he or she is conversing with neither the camera nor another tracked individual.

As shown in Figure 1(a), $\angle(O_i, V_{i,j})$ is defined as the angle between subject i's orientation, and the vector from subject i to j. Similarly $\angle(O_i, V_{i,c})$ is the angle between O_i and the vector pointing toward the camera from location X_i.

In our experiments σ_p is approximately 10° while σ_c is approximately 2°. The difference arises from the observation that when subjects look at the camera their gaze (as measured from the head pose tracker) is much less variable than when looking at other subjects in the scene.

The prior probabilities of each hypothesis are set to be equal, and the hypothesis with the highest posterior probability is picked.

3.2 Speaker Hypothesis

The hypothesis for associating the audio signal with a subject in the scene, Hs, we consider are:

$\mathsf{Hs}_\emptyset$	:	a person who is not tracked is speaking
Hs_i	:	person i is speaking

with associated statistical models

$$p\left(\alpha, \Gamma|\mathsf{Hs}_\emptyset\right) = \text{unif}\left(\alpha, \Gamma\right) \quad (5)$$

$$p\left(\alpha, \Gamma|\mathsf{Hs}_i\right) = \mathcal{N}(\alpha; \phi_i, \beta(\phi_i; \sigma_\alpha)) * p\left(\Gamma|\mathsf{Hs}_i\right) \quad (6)$$

where α is the estimated audio DOA, and Γ represents the set $\{\gamma_1, ..., \gamma_M\}$.Again we assume a simple uniform distribution over all possible values of α, and Γ when the speaker is someone who is not tracked. ϕ_i is the angle between the microphone array and person i as shown in Figure 1(b). We define:

$$p\left(\Gamma|\mathsf{Hs}_i\right) \propto \frac{\gamma_i}{\sum_{j=1}^{M} \gamma_j} \quad (7)$$

The intuition is that as the association measure γ_i for one subject is significantly higher than the others, $p\left(\Gamma|\mathsf{Hs}_i\right)$ approaches a maximum. In the other extreme it approaches zero. Additionally, if the association measures are nearly equal then $p\left(\Gamma|\mathsf{Hs}_i\right)$ approach an equal value for every i. The prior probabilities for our speaker hypotheses are set such that the probability of the speaker being someone who is tracked is equally as likely as being someone who is not tracked.

4. EMPIRICAL RESULTS

We used a prototype audio/video rig consisting of a firewire stereo camera head [3] and a linear microphone array to record our sequences. The stereo camera provided us with two 320x240 pixel rectified stereo image pairs at 25 frames per second, and each omnidirectional microphone in the array provided a separate audio stream sampled at 44.1kHz. We chose to use only two microphones, 9" apart, in our array so that our system could be interfaced to any computer with a standard stereo audio jack. Our recording environment was a 20' x 20' open room with a small amount of stationary background noise. In our implementation our DOA estimate was computed over 125 ms windows to enable us to detect rapid speaker changes. Figure 1(c) describes the information flow between the components of our system

Each of our experimental sequences involved two or three individuals conversing with each other or the audio/video rig. Two individuals were tracked and remained in the field of view of the camera at all times. Their conversation was restricted to a turn-taking dialog and there was no simultaneous speech. Each sequence consisted of more than 2000 frames.

For each sequence, ground truth was established for the frames in which someone was speaking. These frames were labelled with who was speaking and toward whom or what he or she was facing. Focus information was only labelled and compared for the tracked individuals. The results of the two hypothesis tests were independently compared to the ground truth. Table 1 shows the result of the focus hypothesis for each of the sequences. We see that our head tracker is close to 100% accurate in estimating toward whom, or what a person is looking. However, during the first sequence there were a few times when the first person simultaneously gestured with his head toward the camera and spoke to the second person. This resulted in a period of ambiguity. Head tracking alone cannot predict such complex conversational dynamics.

We ran three versions of our system. The first version calculated the speaker hypothesis using only the DOA measurements by ignoring Γ, i.e. it sets $p\left(\Gamma|\mathsf{Hs}_i\right)$ in Equation 6 to 1. The second version only used the audio/video association measure to determine which of the two tracked subjects were speaking. The resulting hypotheses was therefore limited to Hs_i and the speaker was determined by which γ_i was greatest. The final version of the system performed the full hypothesis test outlined in section 3.2.

A summary of the speaker hypothesis using each version of our system is shown in Table 2. These results show that using DOA alone gives approximately 90% accuracy in the two person sequence, and 80% accuracy in the three person sequence. Using only the audio/video association performs slightly worse. The combined system reduces the error by only 1% in the two person sequence, but by 45% in the three person sequence.

Table 3 shows confusion matrices for these experiments. In the second sequence involving three people talking, the first person spoke 34% of the time, the second 53% of the time, and the third individual who was off camera spoke 13% of the time. Using DOA only on this sequence never predicted that an off camera source was speaking. This can be explained by the fact that the second and third person were standing close to each other and toward the side of the microphone array where the variance was greater. This allowed

hypothesis Hs_2 to out-weight the null hypothesis. By combining the audio/video association measure we gain 62% accuracy when the third person is speaking. The audio/video association can be thought of as a confidence weighting factor. In the situation where the third person was talking, if the second person's motion was not associated with the audio, it would lower Hs_2, thus allowing for the correct hypothesis. This situation is shown on the right-hand example of Figure 2.

Sequence	Person 1	Person 2
Two Person	87.91 %	97.17 %
Three Person	99.79 %	100.00 %

Table 1: Subject Focus/Intent (% Accuracy)

Sequence	DOA Only	A/V Assoc. Only	Combined
Two Person	89.36 %	76.17 %	89.49 %
Three Person	80.06 %	76.76 %	89.01 %

Table 2: Speaker Association (% Accuracy)

	Person 1	Person 2	Person 3 (off camera)
Person 1	100.0 %	0.00 %	0.00 %
Person 2	12.31 %	87.69 %	0.00 %
Not Tracked	28.72 %	71.28 %	0.00 %

(a) Speak Association Confusion Matrix, DOA Only

	Person 1	Person 2	Person 3 (off camera)
Person 1	100.0 %	0.00 %	0.00 %
Person 2	10.53 %	88.78 %	0.69 %
Not Tracked	24.47 %	12.76 %	62.77 %

(b) Speak Association Confusion Matrix, Combined

Table 3: Confusion matrices for results for the three person sequence using DOA only (a) and combined (b) speaker hypothesis testing. The "y-axis" is the ground truth, and the "x-axis" is our system's hypothesis.

5. DISCUSSION AND CONCLUSIONS

This paper presents a simple probabilistic framework for combining multiple visual and audio cues in order to locate a speaker in a scene and determine to whom he or she is speaking. We have shown that the visual cues obtained by our head pose tracker are suitable for determining a speaker's focus. The DOA measurements found using the two-element microphone array were satisfactory for locating the speaker in a scene in which the subjects were sufficiently separated. In the cases where the subjects were at similar angles from the array, it was shown that using audio/visual association measurements could help disambiguate them.

Future work will examine more sophisticated joint audio/video measurements. Here we considered simple pixel based features when measuring audio/video synchrony. However, it is expected that domain specific features (e.g. facial models) would improve results. In addition, we wish to explore incorporating a speech recognizer and integrating a higher level conversation model for determining to whom each person is speaking.

6. REFERENCES

[1] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *CVPR*, 1998.

[2] T. Darrell, K. Tollmar, F. Bentley, N. Checka, L.-P. Morency, A. Rahimi, and A. Oh. Face-responsive interfaces: from direct manipulation to perceptive presence. In *International Conference of Ubiquitous Computing*, 2002.

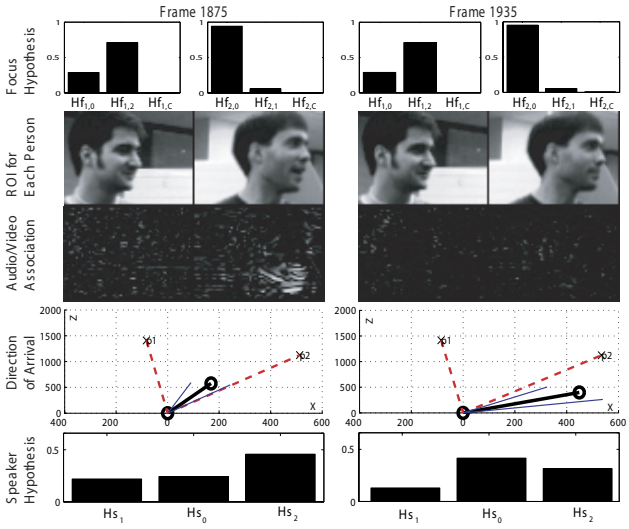


Figure 2: Results for two frames of the three person video sequence. The top row shows the posterior probability of each focus hypothesis (see Section 3.1). The third row shows the audio/video association measurements for each pixel in the ROI. A plan-view plot of the estimated configuration, is shown in the fourth row. The position of each person is marked with an 'x'. The DOA is represented by the solid black line with circle end point markers, and the blue lines surrounding it represent the standard deviation. The last row shows the posterior probabilities of each speaker hypothesis (see Section 3.2).

[3] Videre Design. *MEGA-D Megapixel Digital Stereo Head*. http://www.ai.sri.com/~konolige/svs/, 2000.

[4] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. In *NIPS*, pages 813–819, 1999.

[5] J.W. Fisher III, T.Darrell, W.T. Freeman, and P.A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NIPS*, pages 772–778, 2000.

[6] G. Iyengar and C. Neti. A vision-based microphone switch for speech intent detection. In *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 2001.

[7] D.H. Johnson and D.E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice Hall, 1993.

[8] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *CVPR*, 2003.

[9] V. Pavlovic, A. Garg, J. Rehg, and T. Huang. Multimodal speaker detection using error feedback dynamic bayesian networks. In *CVPR*, 2000.

[10] G. Pingali, G. Tunali, and I. Carlbom. Audio-visual tracking for natural interactivity. In *Proceedings of the Seventh ACM International Conference on Multimedia*, pages 373–382, 1999.

[11] J. M. Rehg, K.P. Murphy, and P. W. Fieguth. Vison-based speaker detection using bayesian networks. In *CVPR*, 1999.

[12] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *NIPS*, 2000.

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.