AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking

Guillaume Lathoud[1,2], Jean-Marc Odobez[1] and Daniel Gatica-Perez[1]

[1] IDIAP Research Institute, CH-1920 Martigny, Switzerland
[2] EPFL, CH-1015 Lausanne, Switzerland
{lathoud,odobez,gatica}@idiap.ch

Abstract. Assessing the quality of a speaker localization or tracking algorithm on a few short examples is difficult, especially when the ground-truth is absent or not well defined. One step towards systematic performance evaluation of such algorithms is to provide time-continuous speaker location annotation over a series of real recordings, covering various test cases. Areas of interest include audio, video and audio-visual speaker localization and tracking. The desired location annotation can be either 2-dimensional (image plane) or 3-dimensional (physical space). This paper motivates and describes a corpus of audio-visual data called "AV16.3", along with a method for 3-D location annotation based on calibrated cameras. "16.3" stands for 16 microphones and 3 cameras, recorded in a fully synchronized manner, in a meeting room. Part of this corpus has already been successfully used to report research results.

1 Introduction

This paper describes a corpus of audio-visual data called "AV16.3", recorded in a meeting room context. "16.3" stands for 16 microphones and 3 cameras, recorded in a fully synchronized manner. The central idea is to use calibrated cameras to provide continuous 3-dimensional (3-D) speaker location annotation for testing audio localization and tracking algorithms. Particular attention is given to overlapped speech, i.e. when several speakers are simultaneously speaking. Overlap is indeed an important issue in multi-party spontaneous speech, as found in meetings [9]. Since visual recordings are available, video and audio-visual tracking algorithms can also be tested. We therefore defined and recorded a series of scenarios so as to cover a variety of research areas, namely audio, video and audio-visual localization and tracking of people in a meeting room. Possible applications range from automatic analysis of meetings to robust speech acquisition and video surveillance, to name a few.
In order to allow for such a broad range of research topics, "meeting room context" is defined here in a wide way. This includes a high variety of situations, from "meeting situations" where speakers are seated most of the time, to "motion situations" where speakers are moving most of the time. This departs from existing, related databases: for example the ICSI database [4] contains audio-only recordings of natural meetings, the CUAVE database [7] does contain audio-visual recordings (close-ups) but

focuses on multimodal speech recognition. The CIPIC [1] database focuses on Head-Related Transfer Functions. Instead of focusing the entire database on one research topic, we chose to have a single, generic setup, allowing very different scenarios for different recordings. The goal is to provide annotation both in terms of "true" 3-D speaker location in the microphone arrays' referent, and "true" 2-D head/face location in the image plane of each camera. Such annotation permits systematic evaluation of localization and tracking algorithms, as opposed to subjective evaluation on a few short examples without annotation. To the best of our knowledge, there is no such audio-visual database publicly available. The dataset we present here has begun to be used: two recordings with static speakers have already been successfully used to report results on real multi-source speech recordings [5].
While investigating for existing solutions for speaker location annotation, we found various solutions with devices to be worn by each person and a base device that locates each personal device. However, these solutions were either very costly and extremely performant (high precision and sampling rate, no tether between the base and the personal devices), or cheap but with poor precision and/or high constraints (e.g. personal devices tethered to the base). We therefore opted for using calibrated cameras for reconstructing 3-D location of the speakers. It is important to note that this solution is potentially non-intrusive, which is indeed the case on part of the corpus presented here: on some recordings no particular marker is worn by the actors.
In the design of the corpus, two contradicting constraints needed to be fulfilled: 1) the area occupied by speakers should be large enough to cover both "meeting situations" and "motion situations", 2) this area should be entirely visible by all cameras. The latter allows systematic optimization of the camera placement. It also leads to robust reconstruction of 3-D location information, since information from all cameras can be used.
The rest of this paper is organized as follows: Section 2 describes the physical setup and the camera calibration process used to provide 3-D mouth location annotation. Section 3 describes and motivates a set of sequences publicly available via Internet. Section 4 discusses the annotation protocol, and reports the current status of the annotation effort.

2 Physical Setup and Camera Calibration

For possible speakers' locations, we selected a L-shaped area around the tables in a meeting room, as depicted in Fig. 1. A general description of the meeting room can be found in [6]. The L-shaped area is a 3 m-long and 2 m-wide rectangle, minus a 0.6 m-wide portion taken by the tables. This choice is a compromise to fulfill the two constraints mentioned in the Introduction. Views taken with the different cameras can be seen in Fig. 2. The data itself is described in Sect. 3.
The choice of hardware is described and motivated in Sect. 2.1. We adopted a 2-step strategy for placing the cameras and calibrating them. First, camera placement (location, orientation, zoom) is optimized, using a looping process including sub-optimal calibration of the cameras with

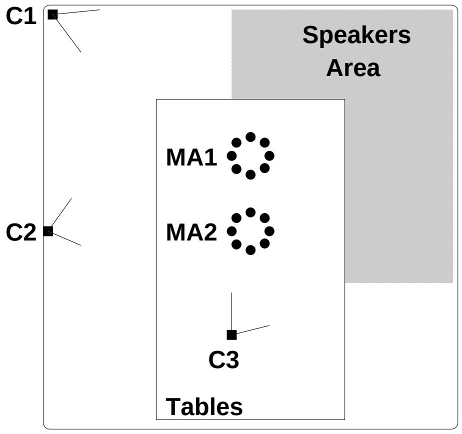

Fig. 1. Physical setup: three cameras C1, C2 and C3 and two 8-microphone circular arrays MA1 and MA2. The gray area is in the field of view of all three cameras. The L-shaped area is a 3 m-long by 2 m-wide rectangle, minus a 0.6 m-wide portion taken by the tables.

2-D information only (Sect. 2.2). Second, each camera is calibrated in a precise manner, using both 2-D measurements and 3-D measurements in the referent of the microphone arrays (Sect. 2.3).
The idea behind this process is that if we can track the mouth of a person in each camera's image plane, then we can reconstruct the 3-D trajectory of the mouth using the cameras' calibration parameters. This can be useful as audio annotation, provided the 3-D trajectory is defined in the referent of the microphone arrays. We show that the 3-D reconstruction error is within a very acceptable range.

2.1 Hardware

We used 3 cameras and two 10 cm-radius, 8-microphone arrays from an instrumented meeting room [6]. The two microphone arrays are placed 0.8 m apart. The motivation behind this choice is threefold:

– Recordings made with two microphone arrays provide test cases for 3-D audio source localization and tracking, as each microphone array can be used to provide an (azimuth, elevation) location estimate of each audio source.
– Recordings made with several cameras generate many interesting, realistic cases of visual occlusion, viewing each person from several viewpoints.
– At least two cameras are necessary for computing the 3-D coordinates of an object from the 2-D coordinates in cameras' image planes. The use of three cameras allows to reconstruct the 3-D coordinates

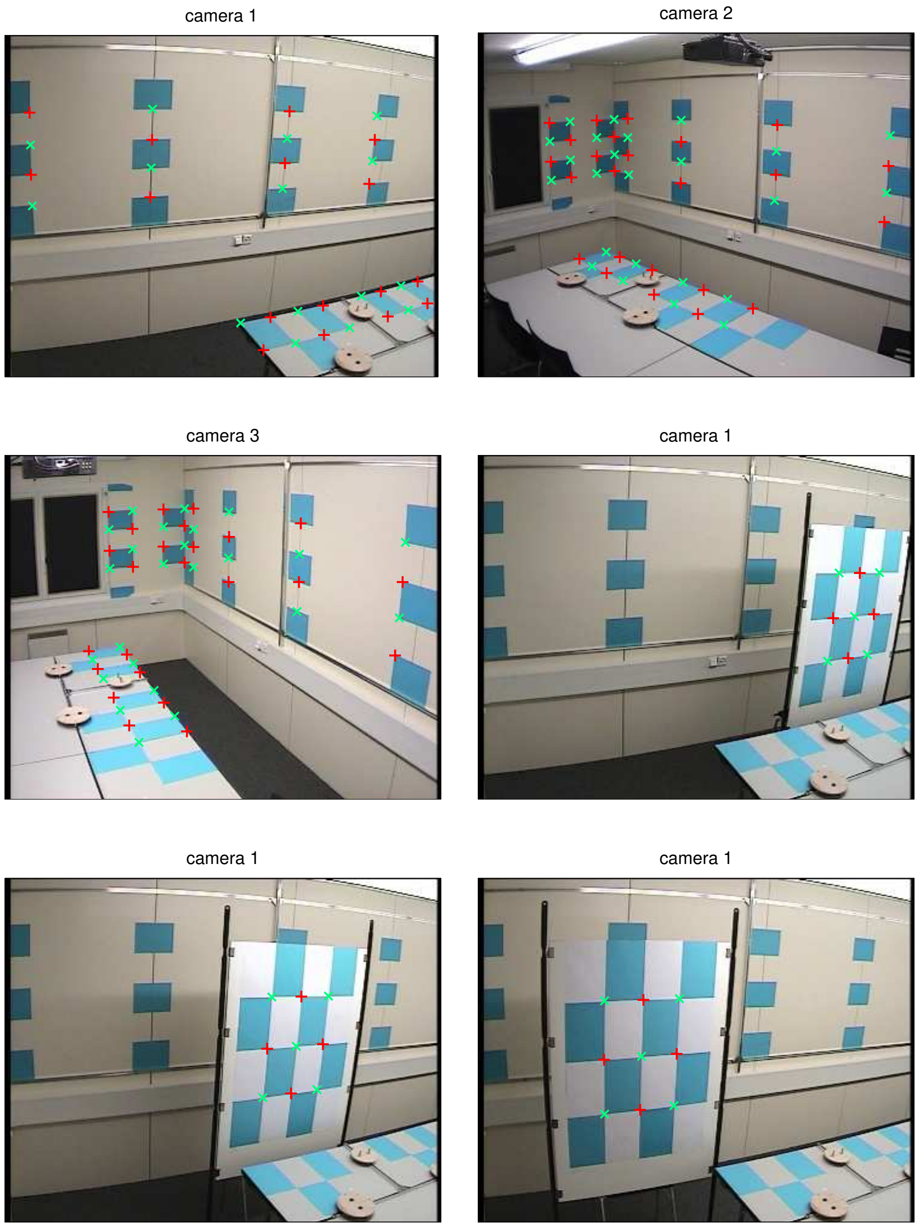

Fig. 2. Snapshots from the cameras at their final positions. Red "+" designate points in the calibration training set Ω_{train}, green "x" designate points in the calibration test set Ω_{test}.

of an object in a robust manner. Indeed, in most cases, visual occlusion occurs in one camera only; the head of the person remains visible from the two other cameras.

2.2 Step One: Camera Placement

This Section describes the looping process used to optimize cameras placement (location, orientation, zoom) using 2-D information only. We used a freely available Multi-Camera Self-Calibration (MultiCamSelfCal) software [10]. "Self-calibration" means that 3-D locations of the calibration points are unknown. The MultiCamSelfCal uses only the 2-D coordinates in the image plane of each camera. It *jointly* produces a set of calibration parameters[3] for each camera and 3-D location estimates of the calibration points, by optimizing the **"2-D reprojection error"**. For each camera, **"2-D reprojection error"** is defined as the distance in pixels between the recorded 2-D points and the projection of their 3-D location estimates back onto the camera image plane, using the estimated camera calibration parameters. Although we used the software with the strict minimum number of cameras (three), the obtained 2-D reprojection error was decent: its upper bound was estimated as less than 0.17 pixels.
The camera placement procedure consists in an iterative process with three steps: Place, Record and Calibrate:

1. *Place* the three cameras (location, orientation, zoom) based on experience in prior iterations. In practice the various cameras should give views that are as different as possible.
2. *Record* synchronously with the 3 cameras a set of calibration points, i.e. 2-D coordinates in the image plane of each camera. As explained in [10], waving a laser beamer in darkness is sufficient.
3. *Calibrate* the 3 cameras by running MultiCamSelfCal on the calibration points. MultiCamSelfCal optimizes the 2-D reprojection error.
4. To try decreasing the 2-D reprojection error, loop to 1. Else go to 5. In practice, a 2-D reprojection error below 0.2 pixels is reasonable.
5. Select the camera placement that gave the smallest 2-D reprojection error.

Multi-camera self-calibration is generally known to provide less precision than manual calibration using an object with known 3-D coordinates. The motivation for using it was ease of use: the calibration points can be quickly recorded with a laser beamer. One iteration of the Place/Record/Calibrate loop thus takes about 1h30. This process converged to the positioning of the camera depicted in Fig. 1.
For detailed information, including the multi-camera self-calibration problem statement, the reader is invited to refer to the documentation in [10].

2.3 Step Two: Camera Calibration

This Section describes precise calibration of each camera, assuming the cameras' placement fixed (location, orientation, zoom). This is done by

[3] For a description of camera calibration parameters see [2].

selecting and optimizing the calibration parameters for each camera, on a calibration object. For each point of the calibration object, both true 3-D coordinates **in the microphone arrays' referent** and true 2-D coordinates in each camera's image plane are known. 3-D coordinates were obtained on-site with a measuring tape (measurement error estimated below 0.005 m). Crosses in Fig. 2 show the 3-D calibration points. These points were split in two sets: Ω_{train} (36 points) and Ω_{test} (39 points).
Particular mention must be made of the model selection issue, i.e. how we chose to model non-linear distortions produced by each camera's optics. An iterative process that evaluates adequacy of the calibration parameters of all three cameras in terms of **"3-D reconstruction error"** was adopted: the Euclidean distance between 3-D location estimates of points visible from at least 2 cameras, and their true 3-D location. The camera calibration procedure can be detailed as follows:

1. *Model selection*: for each camera, select the set of calibration parameters based on experience in prior iterations.
2. *Model training*: for each camera, estimate the selected calibration parameters on Ω_{train} using the software available in [2].
3. *3-D error*: for each point in Ω_{train}, compute the Euclidean distance between true 3-D coordinates and 3-D coordinates reconstructed using the trained calibration parameters and the 2-D coordinates in each camera's image plane.
4. *Evaluation*: estimate the "training" maximum 3-D reconstruction error as $\mu + 3\sigma$, where μ and σ respectively stand for mean and standard deviation of the 3-D error, across all points in Ω_{train}.
5. To try decreasing the maximum 3-D reconstruction error, loop to 1. Else go to 6.
6. Select the set of calibration parameters and their estimated values, that gave the smallest maximum 3-D reconstruction error.

The result of this process is a set of calibration parameters and their values for each camera. For all cameras the best set of parameters were focal center, focal lengths, r^2 radial and tangential distortion coefficients. Once the training was over, we evaluated the 3-D error on the unseen test set Ω_{test}. The maximum 3-D reconstruction error on this set was 0.012 m. This maximum error was deemed decent, as compared to the diameter of an open mouth (about 0.05 m).

3 Online Corpus

This Section first motivates and describes the variety of sequences recorded, and then describes in more details the annotated sequences. "Sequence" means:

- 3 video DIVX AVI files (resolution 288x360), one for each camera, sampled at 25 Hz. It includes also one audio signal.
- 16 audio WAV files recorded from the two circular 8-microphone arrays, sampled at 16 kHz.
- When possible, more audio WAV files recorded from lapels worn by the speakers, sampled at 16 kHz.

All files were recorded in a synchronous manner: video files carry a time-stamp embedded in the upper rows of each image, and audio files always start at video time stamp `00:00:10.00`. Complete details about the hardware implementation of a unique clock across all sensors can be found in [6]. Although only 8 sequences have been annotated, many other sequences are also available. The whole corpus, along with annotation files, camera calibration parameters and additional documentation is accessible[4] at: `http://glat.info/ma/av16.3 v6`. It was recorded over a period of 5 days, and includes 42 sequences overall, with sequence duration ranging from 14 seconds to 9 minutes (total 1h25). 12 different actors were recorded. Although the authors of the present paper were recorded, many of the actors don't have any particular expertise in the fields of audio and video localization and tracking.

3.1 Motivations

The main objective is to study several localization/tracking phenomena. A non-limiting list includes:

- Overlapped speech.
- Close and far locations, small and large angular separations.
- Object initialization.
- Variable number of objects.
- Partial and total occlusion.
- "Natural" changes of illumination.

Accordingly, we defined and recorded a set of sequences that contains a high variety of test cases: from short, very constrained, specific cases (e.g. visual occlusion), for each modality (audio or video), to natural spontaneous speech and/or motion in much less constrained context.

Each sequence is useful for at least one of three fields of research: analysis of audio, video or audio-visual data. Up to three people are allowed in each sequence. Human motion can be static (e.g. seated persons), dynamic (e.g. walking persons) or a mix of both across persons (some seated, some walking) and time (e.g. meeting preceded and followed by people standing and moving).

3.2 Contents

As mentioned above, the online corpus comprises of 8 annotated sequences plus many more unannotated sequences. These 8 sequences were selected for the initial annotation effort. This choice is a compromise between having a small number of sequences for annotation, and covering a large variety of situations to fulfill interests from various areas of research. It constitutes a minimal set of sequences covering as much variety as possible across modalities and speaker behaviors. The process of annotation is described in Sect. 4.

The name of each sequence is unique. Table 1 gives a synthetic overview. A more detailed description of each sequence follows.

[4] both HTTP or FTP protocols can be used to browse and download the data.

Table 1. List of the annotated sequences. Tags mean: [A]udio, [V]ideo, predominant [ov]erlapped speech, at least one visual [occ]lusion, [S]tatic speakers, [D]ynamic speakers, [U]nconstrained motion, [M]outh, [F]ace, [H]ead, speech/silence [seg]mentation.

Sequence name	Duration (seconds)	Modalities of interest	Nb. of speakers	Speaker(s) behavior	Desired annotation
seq01-1p-0000	217	A	1	S	M, seg
seq11-1p-0100	30	A, V, AV	1	D	M, F, seg
seq15-1p-0100	35	AV	1	S,D(U)	M, F, seg
seq18-2p-0101	56	A(ov)	2	S,D	M, seg
seq24-2p-0111	48	A(ov), V(occ)	2	D	M, F
seq37-3p-0001	511	A(ov)	3	S	M, seg
seq40-3p-0111	50	A(ov), AV	3	S,D	M, F
seq45-3p-1111	43	A(ov), V(occ), AV	3	D(U)	H

seq01-1p-0000 A single speaker, static while speaking, at each of 16 locations covering the shaded area in Fig. 1. The speaker is facing the microphone arrays. The purpose of this sequence is to evaluate audio source localization on a single speaker case.

seq11-1p-0100 One speaker, mostly moving while speaking. The only constraint on the speaker's motion is to face the microphone arrays. The motivation is to test audio, video or audio-visual (AV) speaker tracking on difficult motion cases. The speaker is talking most of the time.

seq15-1p-0100 One moving speaker, walking around while alternating speech and long silences. The purpose of this sequence is to 1) show that audio tracking alone cannot recover from unpredictable trajectories during silence, 2) provide an initial test case for AV tracking.

seq18-2p-0101 Two speakers, speaking and facing the microphone arrays all the time, slowly getting as close as possible to each other, then slowly parting. The purpose is to test multi-source localization, tracking and separation algorithms.

seq24-2p-0111 Two moving speakers, crossing the field of view twice and occluding each other twice. The two speakers are talking most of the time. The motivation is to test both audio and video occlusions.

seq37-3p-0001 Three speakers, static while speaking. Two speakers remain seated all the time and the third one is standing. Overall five locations are covered. Most of the time 2 or 3 speakers are speaking concurrently. (For this particular sequence only snapshot image files are available, no AVI files.) The purpose of this sequence is to evaluate multi-source localization and beamforming algorithms.

seq40-3p-0111 Three speakers, two seated and one standing, all speaking continuously, facing the arrays, the standing speaker walks back and forth once behind the seated speakers. The motivation is both to test multi-source localization, tracking and separation algorithms, and to highlight complementarity between audio and video modalities.

seq45-3p-1111 Three moving speakers, entering and leaving the scene, all speaking continuously, occluding each other many times. Speakers' motion is unconstrained. This is a very difficult case of overlapped speech and visual occlusions. Its goal is to highlight the complementarity between audio and video modalities.

3.3 Sequence Names

A systematic coding was defined, such that the name of each sequence (1) is unique, and (2) contains a compact description of its content. For example "seq40-3p-0111" has three parts:
- "seq40" is the unique identifier of this sequence.
- "3p" means that overall 3 different persons were recorded – but not necessarily all visible simultaneously.
- "0111" are four binary flags giving a quick overview of the content of this recording. From left to right:
  bit 1: 0 means "very constrained", 1 means "mostly unconstrained" (general behavior: although most recordings follow some sort of scenario, some include very strong constraints such as the speaker facing the microphone arrays at all times).
  bit 2: 0 means "static motion" (e.g. mostly seated), 1 means "dynamic motion". (e.g. continuous motion).
  bit 3: 0 means "minor occlusion(s)", 1 means "at least one major occlusion", involving at least one array or camera: whenever somebody passes in front of or behind somebody else.
  bit 4: 0 means "little overlap", 1 means "significant overlap". This involves audio only: it indicates whether there is a significant proportion of overlap between speakers and/or noise sources.

4 Annotation

Two types of annotations can be created: in space (e.g. speaker trajectory) or time (e.g. speech/silence segmentation). The definition of annotation intrinsically defines the performance metrics that will be used to evaluate localization and tracking algorithms. How annotation should be defined is therefore debatable. Moreover, we note that different modalities (audio, video) might require very different annotations (e.g. 3-D mouth location vs 2-D head bounding box). Sections 4.1 and 4.2 report the initial annotation effort done on the AV16.3 corpus. Sections 4.3, 4.4 and 4.5 detail some examples of application of the available annotation. Section 4.6 discusses future directions for annotation.

4.1 Initial Effort

The two sequences with static speakers only have already been fully annotated: "seq01-1p-0000" and "seq37-3p-0001". The annotation includes, for each speaker, 3-D mouth location and speech/silence segmentation. 3-D mouth location is defined relative to the microphone arrays' referent.

The origin of this referent is in the middle of the two microphone arrays. This annotation is also accessible online. It has already been successfully used to evaluate recent work [5]. Moreover, a simple example of use of this annotation is available whithin the online corpus, as described in Sect. 4.3.
As for sequences with moving speakers and occlusion cases, three Matlab graphical interfaces were written and used to annotate location of the head, of the mouth and of an optional marker (colored ball) on the persons' heads:

BAI: the Ball Annotation Interface, to mark the location of a colored ball on the head of a person, as an ellipse. Occlusions can be marked, i.e. when the ball is not visible. The BAI includes a simple tracker to interpolate between manual measurements.

HAI: the Head Annotation Interface, to mark the location of the head of a person, as a rectangular bounding box. Partial or complete occlusions can be marked.

MAI: the Mouth Annotation Interface, to mark the location of the mouth of a person as a point. Occlusions can be marked, i.e. when the mouth is not visible.

All three interfaces share very similar features, including two windows: one for the interface itself, and a second one for the image currently being annotated. An example of snapshot of the HAI can be seen in Fig. 3. All annotation files are simple matrices stored in ASCII format.
All three interfaces are available and documented online, within the corpus itself. We have already used them to produce continuous 3-D mouth location annotation from sparse manual measurements, as described in Sect. 4.5.

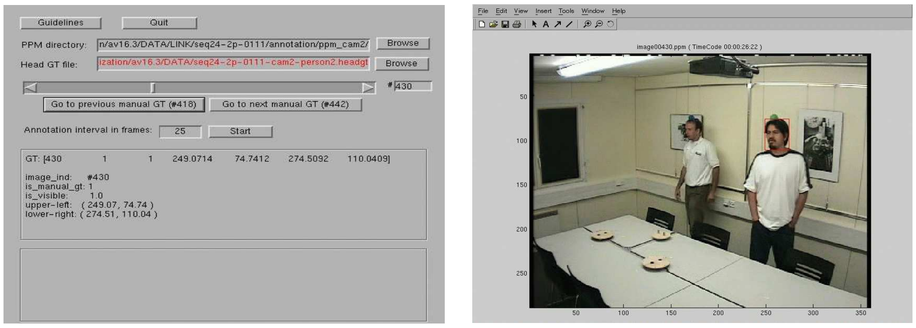

Fig. 3. Snapshots of the two windows of the Head Annotation Interface.

4.2 Current State

The annotation effort is constantly progressing over time, and Table 2 details what is already available online as of August 31st, 2004.

Table 2. Annotation available online as of August 31st, 2004. "C" means continuous annotation, i.e. all frames of the 25 Hz video are annotated. "S" means sparse annotation, i.e. the annotation is done at a rate less than 25 Hz (given in parenthesis).

Sequence	ball		mouth		head	speech/silence
	2-D	3-D	2-D	3-D	2-D	segmentation
seq01-1p-0000			C	C		precise
seq11-1p-0100	C	C	C	C		
seq15-1p-0100			S(2 Hz)	S(2 Hz)		
seq18-2p-0101	C	C	C	C		
seq24-2p-0111	C	C	C	C	S(2 Hz)	
seq37-3p-0001			C	C		undersegmented
seq40-3p-0111			S(2 Hz)	S(2 Hz)		
seq45-3p-1111			S(2 Hz)	S(2 Hz)	S(2 Hz)	

4.3 Example 1: Audio Source Localization Evaluation

The online corpus includes a complete example (Matlab files) of single source localization followed by comparison with the annotation, for "seq01-1p-0000". It is based on a parametric method called SRP-PHAT [3]. All necessary Matlab code to run the example is available online[5]. The comparison shows that the SRP-PHAT localization method provides a precision between -5 and +5 degrees in azimuth.

4.4 Example 2: Multi-Object Video Tracking

As an example, the results of applying three independent, appearance-based particle filters on 200 frames of the "seq45-3p-1111" sequence, using only one of the cameras, are shown in Fig. 4, and in a video[6]. The sequence depicts three people moving around the room while speaking, and includes multiple instances of object occlusion. Each tracker has been initialized by hand, and uses 500 particles. Object appearance is modeled by a color distribution [8] in RGB space.
In this particular example we have not done any performance evaluation yet. We plan to define precision and recall based on the intersecting surface between the annotation bounding box and the result bounding box.

[5] `http://glat.info/ma/av16.3 v6/EXAMPLES/AUDIO/README`
[6] `http://glat.info/ma/av16.3 v6/EXAMPLES/VIDEO/av-video.mpeg`

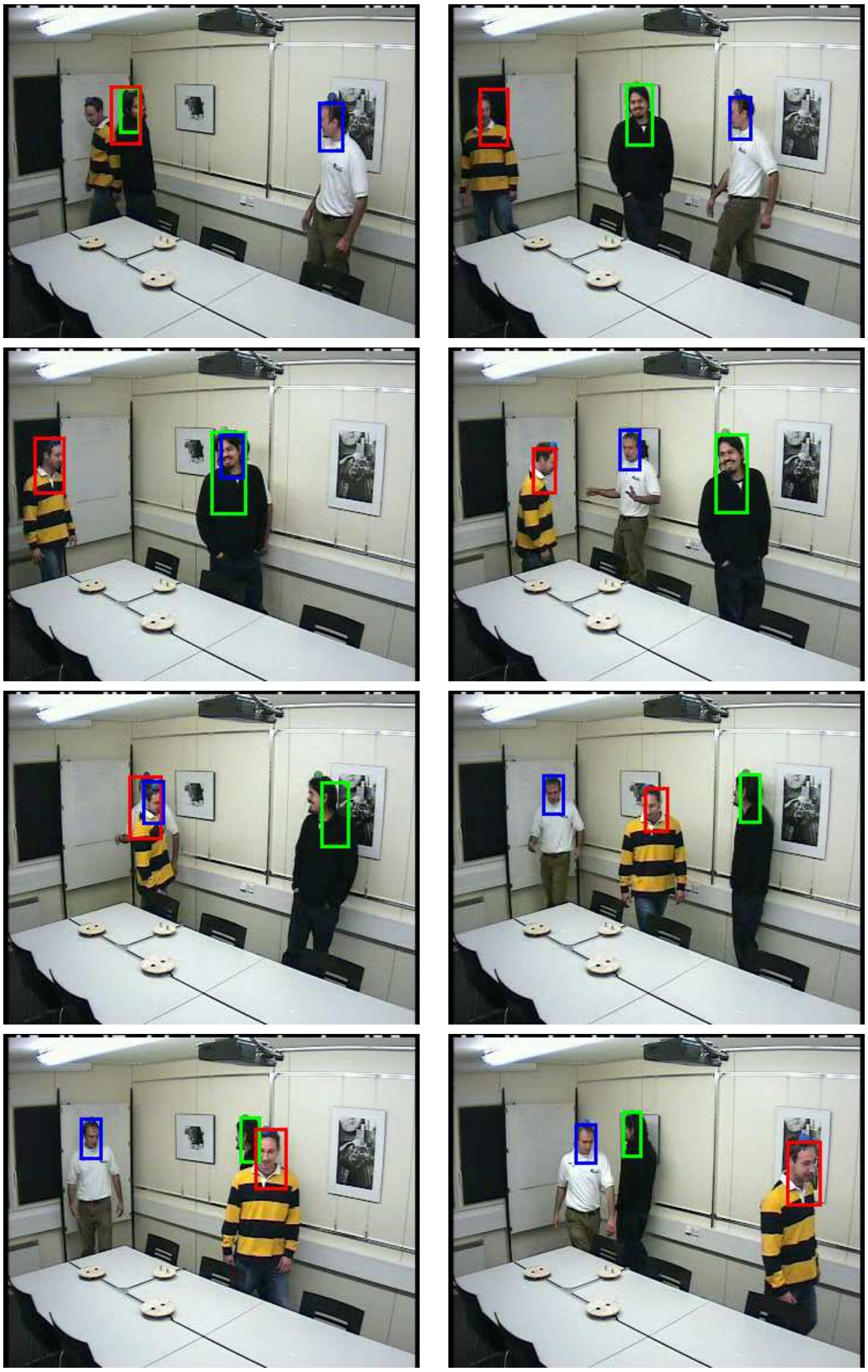

Fig. 4. Snapshots from visual tracking on 200 frames of "seq45-3p-1111". 200 frames (initial timecode: 00:00:41.17). Tracking results are shown every 25 frames.

4.5 Example 3: 3-D Mouth Annotation

From sparse 2-D mouth annotation on each camera we propose to (1) reconstruct 3-D mouth location using camera calibration parameters estimated as explained in Sect. 2.3, (2) interpolate 3-D mouth location using the ball location as origin of the 3-D referent. The 3-D ball location itself is provided by the 2-D tracker in the BAI interface (see Sect. 4.1) and 3-D reconstruction. The motivation of this choice was twofold: first of all, using simple (e.g. polynomial) interpolation on mouth measurements was not enough in practice, since human motion contains many complex nonlinearities (sharp turns and accelerations). Second, visual tracking of the mouth is a hard task in itself. We found that interpolating measurements in the moving referent of an automatically tracked ball marker is effective even at low annotation rates (e.g 2 Hz = 1 video frame out of 12), which is particularly important since the goal is to save on time spent doing manual measurements. A complete example with all necessary Matlab implementation can be found online[7]. This implementation was used to create all 3-D files available within the corpus.

4.6 Future Directions

Difficulties arise mostly in two cases: 1) predominance of overlapped speech, and 2) highly dynamic situations, in terms of motions and occlusions. 1) can be addressed by undersegmenting the speech and defining proper metrics for evaluation. By "undersegmenting" we mean that less segments are defined, each segment comprising some silence and speech which is too weak to be localized. An example is given in [5].
2) is more difficult to address. It is intrinsically linked to the minimum interval at which annotation measurements are taken, and therefore the interval at which performance will be evaluated. Considering the fact that location between two measurements can be interpolated, two attitudes can be envisaged:

1. On short sequences, with very specific test cases, the interval can be chosen very small, in order to obtain fine-grained, precise spatial annotation. Even with interpolation, this would require independent observer(s) to give many true location measurements.
2. On long sequences, the interval can be chosen larger. If the interpolated annotation is used for performance evaluation, slight imprecision can be tolerated, as compensated by the size of the data ("continuous" annotation). If the manual annotation measurements only are used for performance evaluation ("sparse" annotation), the evaluation will be more precise, and the relatively large number of such measurements may still lead to significant results. By "significant" we mean that the standard deviation of the error is small enough for the average error to be meaningful.

[7] `http://glat.info/ma/av16.3 v6/EXAMPLES/3D-RECONSTRUCTION/README`

5 Conclusion

This paper presented the AV16.3 corpus for speaker localization and tracking. AV16.3 focuses mostly on the context of meeting room data, acquired synchronously by 3 cameras, 16 far-distance microphones, and lapels. It targets various areas of research: audio, visual and audio-visual speaker tracking. In order to provide audio annotation, camera calibration is used to generate “true” 3-D speaker mouth location, using freely available software. To the best of our knowledge, this is the first attempt to provide synchronized audio-visual data for extensive testing on a variety of test cases, along with spatial annotation. AV16.3 is intended as a step towards systematic evaluation of localization and tracking algorithms on real recordings. Future work includes completion of the annotation process, and possibly data acquisition with different setups.

6 Acknowledgments

The authors acknowledge the support of the European Union through the AMI, M4, HOARSE and IM2.SA.MUCATAR projects. The authors wish to thank all actors recorded in this corpus, Olivier Masson for help with the physical setup, and Mathew Magimai.-Doss for valuable comments.

References

1. Algazi, V., Duda, R., Thompson, D.: The CIPIC HRTF Database. Proceedings of WASPAA (2001).
2. Bouguet, J.Y.: Camera Calibration Toolbox for Matlab (01/2004). `http://www.vision.caltech.edu/bouguetj/calib_doc/`
3. DiBiase, J., Silverman, H., Brandstein, M.: Robust Localization in Reverberant Rooms. Microphone Arrays, Eds. Brandstein, M., Ward, D., Springer (2001), 157-180.
4. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI Meeting Corpus. Proceedings of ICASSP (2003).
5. Lathoud, G., McCowan, I.A.: A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays. To appear in Proceedings of SAPA (2004).
6. Moore, D.: The IDIAP Smart Meeting Room. IDIAP Communication COM-02-07 (2002).
7. Patterson, E., Gurbuz, S., Tufekci, Z., Gowdy, J.: Moving Talker, Speaker-Independent Feature Study and Baseline Results Using the CUAVE Multimodal Speech Corpus. Eurasip Journal on Applied Signal Processing (2002), vol. 11, 1189–1201.
8. Perez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based Probabilistic Tracking. Proceedings of ECCV (2002).
9. Shriberg, E., Stolcke, A., Baron, D.: Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation. Proceedings of Eurospeech (2001) vol. 2, 1359–1362.
10. Svoboda, T.: Multi-Camera Self-Calibration (08/2003). `http://cmp.felk.cvut.cz/ svoboda/SelfCal/index.html`