Multiple Active Speaker Localization based on Audio-visual Fusion in two Stages*

Zhao Li[1], *Student Member, IEEE*, Thorsten Herfet[1], *Senior Member, IEEE*,
Martin Grochulla[2], and Thorsten Thormählen[2], *Member, IEEE*

***Abstract*— Localization of multiple active speakers in natural environments with only two microphones is a challenging problem. Reverberation degrades performance of speaker localization based exclusively on directional cues. The audio modality alone has problems with localization accuracy while the video modality alone has problems with false speaker activity detections. This paper presents an approach based on audio-visual fusion in two stages. In the first stage, speaker activity is detected based on the audio-visual fusion which can handle false lip movements. In the second stage, a Gaussian fusion method is proposed to integrate the estimates of both modalities. As a consequence, the localization accuracy and robustness compared to the audio/video modality alone is significantly increased. Experimental results in various scenarios confirmed the improved performance of the proposed system.**

I. INTRODUCTION

The problem of localizing the active speakers in natural environments arises in a series of human computing applications, e.g. human-robot interaction, video conference systems where cameras are turned towards the persons that are speaking [1], or autonomous recording systems [2] where only the camera streams with the best view of speakers are recorded. Localization of one or more speakers is fundamental to auditory perception and signal processing strategies that seek to enhance a source signal by spatial filtering. Because of the potentially large number of subjects moving and speaking in such cluttered environments the problem of robust speaker localization is challenging.

In many systems that handle speaker localization, audio and video data are treated separately. Such systems usually have subsystems that are specialized for the different modalities and are optimized for each modality separately [3], [4]. With increasing computing capabilities, both auditory and visual modalities of the speech signal may be used to improve active speaker detection and lead to major improvements in the perceived quality of man-machine interaction, where each modality may compensate for weaknesses of the other one.

The problem of multimodal multiple speaker localization poses various challenges. For audio, the signal propagating from the speaker is usually corrupted by reverberation and multipath effects and by background noise, making it difficult to identify the time delay. For video, the camera view may be cluttered by objects other than the speaker, often causing a tracker to lose the subjects. And video alone cannot deal with false lip movements and leads to false detection of lip activity (called speaker activity in this paper). Another problem that needs to be addressed is the audio-visual data fusion that makes use of the modalities' complementarity. Audio-visual correlations cannot always be observed and the fusion approach needs to be robust against missing correlations.

Among the different methods that perform speaker localization, only a few are performing the fusion of both audio and video modalities. Some of them just select the active face among all detected faces based on the distance between the peak of audio cross correlation and the position of the detected faces in the azimuth domain [2], [5]. A few of the existing approaches perform the fusion directly at the feature level, which relies on explicit or implicit use of mutual information [1], [6], [7]. Most of them address the detection of active speaker among a few face candidates, where it is assumed that all the faces of speakers can be successfully detected by the video modality. However, this assumption does not always hold in practice, especially in cluttered environments.

In this paper the audio modality performs the audio source localization based on advanced audio processing algorithms and extracts audio signal for each visual face to help the video modality with speaker activity detection. The video modality localizes the detected faces and computes the number of pixels with low intensities in the mouth region of speakers, where the latter one indicates lip movements and will be used to detect speaker activity. Then we propose an approach based on audio-visual fusion in two stages. In the first stage, speaker activity is detected based on the audio-visual fusion which can handle false lip movements. In the second stage, a Gaussian fusion method is proposed to integrate the estimates from both modalities in a way that video results can compensate for the localization deviation of the audio modality while audio results can still contribute to the final results when video modality have occluded faces in view.

This approach is applied in a human-machine interaction scenario, where a motorized human dummy head – called Bob – with three degrees of freedom is used (shown in Fig. 1). Bob resides in a normal office meeting room and is able to turn its head to investigate the surrounding auditory scene, which in our case consists of multiple speaking subjects. The auditory scene is recorded via two microphones

*Research carried out in the Excellence Cluster Multimodal Computing and Interaction (MMCI) and is funded by the German National Science Foundation DFG.

[1]Zhao Li and Thorsten Herfet, Telecommunications Lab, Saarland University, Saarbrücken, Germany, li@nt.uni-saarland.de, herfet@nt.uni-saarland.de

[2]Martin Grochulla and Thorsten Thormählen, Max-Planck-Institut für Informatik, Saarbrücken, Germany, mgrochul@mpi-inf.mpg.de, thormae@mpi-inf.mpg.de

in Bob's ears. Bob has also two eyes (cameras) which have a horizontal field of view of approximately 43 degrees and can move approximately from -15 to 15 degrees in the horizontal direction.

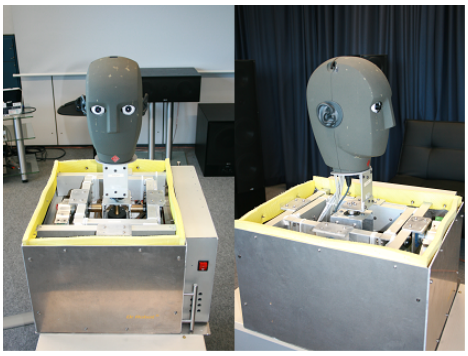

Fig. 1. *Bob – the movable human dummy head.*

In summary, this paper presents the following contributions. Firstly, we propose a robust system for speaker localization that is based on the combination of advanced audio and video processing algorithms. Secondly, in contrast to [2], [5], our approach requires only two microphones and two cameras. Our approach can handle the difficult scenario where multiple speakers are talking at the same time. Thirdly, the proposed fusion approach allows the simultaneous improvement of the estimation accuracy and robustness. If both modalities are available, the estimation accuracy is improved due to the accurate video localization. Nevertheless, the approach is also robust if only a single modality contributes information.

The rest of this paper is organized as follows. Section II present the overall system architecture, audio modality, and video modality. The fusion in two different stages is described in Section III. In Section IV, we will show experimental results of both fusions. The last section provides conclusion and future work.

II. LOCALIZATION SYSTEM

In this section, we first introduce the architecture of the proposed localization system and then describe the audio and video modalities separately. We compute the position of the subjects relative to the robot (relative coordinates).

A. System Architecture

The architecture of the proposed localization system is shown in Fig. 2. The audio modality (blue arrows) contributes to two different modules of the system. Firstly, it performs the audio source localization based on advanced audio processing algorithms. This information is directly fed into the late-stage integration module. Secondly, the audio signal is extracted for each visual face and the data is used in the speaker activity detection.

The video modality (green arrows) detects faces and the corresponding mouth regions in both left and right images. Besides calculating the location of each face, the video modality also calculates the number of pixels with low intensities for each mouth region, which will be used in the first stage of the audio-visual fusion (speaker activity detection, see Section III-A).

In the second stage, the result is further improved based on the integration of the estimates from the audio modality and the speaker activity detection. The details about the audio and video modalities will be described in the following subsections and the fusion in two stages will be presented in the next section.

B. Audio Modality

It is widely acknowledged that for human audition, Interaural Time Differences (*ITD*) are the main localization cues used at low frequencies (< 1.5 kHz), whereas in the high frequency range both Interaural Level Differences (*ILD*) and *ITD* between the envelopes of the signals are used [8]. The resolution of the binaural cues has implications for both localization and recognition tasks.

Human cochlear filtering can be modeled by a bank of bandpass filters. The filterbank employed here consist of 128 fourth-order gammatone filters [9], the output of which is half-wave rectified in order to simulate firing rates of the auditory nerve. Saturation effects are modeled by taking the square root of the rectified signal.

1) Azimuth Localization and the Skeleton *Method:* Current models of azimuth localization almost invariably employ cross correlation, which provides excellent time delay estimation for broadband stimuli, and for narrow band stimuli in the low-frequency range. For high frequency narrow band signals it produces multiple ambiguous peaks and the support of *ILD* is needed to estimate the time delay. *ITD* is estimated by computing the cross-correlation between the outputs of the precedence processed auditory filter response at the two ears. Given the output of the precedence effect model for the left and right ear in channel i, $l_i(n)$ and $r_i(n)$, the cross correlation for delay τ and time frame j is given by

$$C(i,j,\tau) = \sum_{n=0}^{M-1} l_i(jT-n) r_i(jT-n-\tau)\, win(n), \quad (1)$$

where win is a window of width M time steps and T is the frame period (10 ms, or 441 samples with a sampling rate of 44 100). Currently, we use a Hann window with $M = 441$, corresponding to a duration of 10 ms, and consider values of τ between -1 and 1 ms.

Ideally, the cross-correlogram should exhibit a 'spine' at the delay τ corresponding to the *ITD* of a sound source. This feature can be emphasized by summing the channel cross-correlation functions, giving a pooled cross-correlogram

$$P(j,\tau) = \sum_{i=0}^{N} C(i,j,\tau). \quad (2)$$

Each cross-correlation function is warped into the azimuthal axis, giving a modified cross-correlogram of the form $C(i,j,\phi)$, where ϕ is azimuth in degrees. The azimuth is quantized to a resolution of one degree, giving 181 points between -90 and $+90$ degree. Warping is achieved by a table look-up, which relates the azimuth in degrees

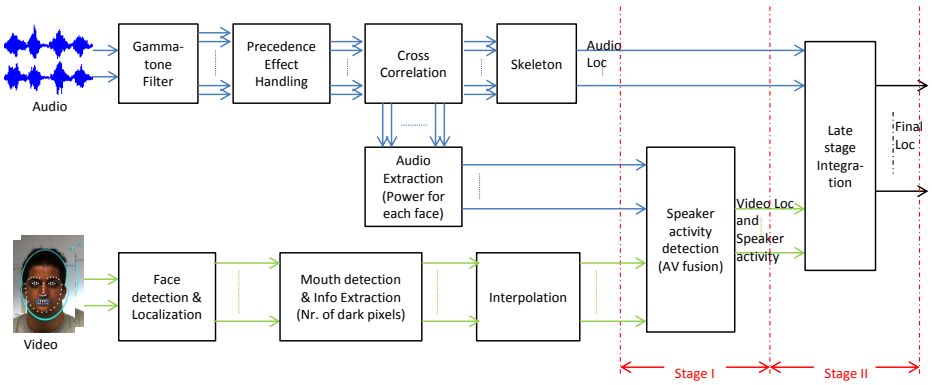

Fig. 2. *System architecture.*

to its corresponding *ITD* in each channel of the auditory model. The functions relating azimuth to *ITD* were trained using HRTF (Head-Related Transfer Function) simulation and typical mapping formulas [3]. For high frequencies, the cross-correlogram always exhibits multiple 'spines'. Here we choose the one 'spine' which is closest to the corresponding azimuth angle based on *ILD*. The *ILD* can be calculated by Eq. 3. The mapping from *ILD* to azimuth angles can be trained for each frequency [3].

$$ILD = 10\log_{10}\frac{\sum_t l^2(t)}{\sum_t r^2(t)}\,\text{dB}. \tag{3}$$

A further stage of processing is based on the *Skeleton* cross-correlation function [10]. For each channel of the cross-correlogram, a *Skeleton* function $S(i,j,\phi)$ is formed by superimposing Gaussian functions at azimuths corresponding to local maxima in the corresponding cross-correlation function $C(i,j,\phi)$. First, each function $C(i,j,\phi)$ is reduced to a form $Q(i,j,\phi)$, which contains non-zero values only at its local maxima, and the values are weighted by the energy of the current frame. Subsequently, $Q(i,j,\phi)$ is convolved with a Gaussian to give the *Skeleton* function

$$S(i,j,\phi) = Q(i,j,\phi)\exp\left(\frac{-\phi^2}{2\sigma_i^2}\right). \tag{4}$$

The standard deviations of the Gaussians σ_i, vary linearly with the frequency channel i, being 4.5 samples in the lowest frequency channel and 0.75 samples in the highest (these parameters were derived empirically using a small data set) [10]. This approach is similar in effect to applying lateral inhibition along the azimuth axis, and causes a sharpening of the cross-correlation response.

2) Precedence Effect Filtering and Weighting: In reverberant recordings, many time-frequency (T-F) units will contain cues that differ significantly from free-field cues. Including a weighting function or cue selection mechanism that indicates when an azimuth cue should be trusted can improve localization performance [11]. Motivated by the precedence effect [12], [13], we incorporate a simple cue weighting mechanism that identifies strong onsets in the mixture signal. We generate a real-valued weight $w_{i,j}$, that measures the energy ratio between unit $u_{i,j}$ and $u_{i,j-1}$.

Better performance can be achieved by keeping only those weights above a specified threshold ($Thres_{PE}$).The final audio sources localization results can be represented as $A(\phi)$, which is the sum of *Skeleton* functions $S(i,j,\phi)$ for all T-F units with precedence effect filtering and weighting:

$$A(\phi) = \sum_i \sum_j w_{i,j} S(i,j,\phi), \quad \text{if } w_{i,j} > Thres_{PE}. \tag{5}$$

We also found that the threshold 1.0 leads to the best performance in our recording environment for most candidates. The fixed threshold may cause too few frames above the threshold [13]. To avoid this problem an automatic threshold control is also applied that the remaining frames should have no less than $25\,\%$ of the overall signal energy. Moreover, precedence effect weighting and filtering can also reduce the disturbing peaks caused by reverberations.

3) Audio Data Extraction: In presence of multiple speakers, this step extracts the audio signal for each individual from the audio mixture. This information is then used in the speaker activity detection module. Thus, it makes sense to extract individual audio data for each detected speaker of the video modality. Compared with speech separation, audio signal extraction here can tolerate a little interference. So the audio signal extraction can be done based on the location cue of each T-F unit. Meanwhile, precedence effect filtering is also applied to handle reverberations.

After the precedence effect filtering, a T-F unit is chosen for a given speaker only when the location cue of this unit is within a threshold $Range_A$ which equals to the half range of the maximum errors in degree from the audio localization. All the chosen units in different frequency are summed together to build the audio signal for a time slot, shown as Eq. 6. Here $Loc_{i,j}$ denotes the location cue of the unit $u_{i,j}$ and Loc_{sp_k} is the location of speaker sp_k calculated by the video modality. All the locations mentioned here are represented in the azimuth domain.

$$Audio_{sp_k}(j) = \sum_i u_{i,j}, \quad \text{if } w_{i,j} > Thres_{PE} \text{ and } |Loc_{i,j} - Loc_{sp_k}| < Range_A. \tag{6}$$

Please also note that for the case of only a single speaker in a scene, the original audio signal can be used directly in the proposed speaker activity detection based on audio-visual fusion.

C. Video Modality

Besides the audio information, visual information can also be used to localize multiple subjects. To this end, we employ the two cameras that are available in our motorized robotic head. In the first step, we calibrate the cameras as will be described in the next subsection. Afterwards, we present our approach to detect faces in the images and to extract information for active speaker detection. Using two cameras allows us to triangulate detected speakers in both views and obtain approximate depth values. Furthermore, two cameras can capture a larger part of the scene.

1) Camera Calibration: The goal of camera calibration is to estimate camera parameters. There are extrinsic camera parameters, such as position and orientation of the camera, and intrinsic parameters, such as focal length, principal point offset, and radial distortion parameters. Popular and often-used approaches use a calibration pattern with known geometry for parameter estimation [14]. 2D-3D correspondences are extracted from images taken of such a pattern. They are used for estimating the camera parameters. Because of the large range of possible viewing directions, however, calibration of the rotating robotic head requires a special calibration procedure. Hence, in order to perform this calibration, we use the idea presented in [15]. In this approach multiple spatially distributed calibration patterns are used for camera parameter estimation. Tsai's approach [14] is used for parameter initialization. Afterwards, the spatially distributed patterns are related into a globally consistent coordinate system. Finally the parameters are optimized by bundle adjustment. In our case, we jointly estimate the intrinsic camera parameters (focal length, radial distortion) for all viewing directions.

2) Face Detection and Localization: For face detection we employ the approach [16] that is provided by the OpenCV library. The OpenCV face detector is a popular, easy-to-use, and robust method for detecting faces. It is based on Haar-like features for object detection, which are used in a classifier cascade. The cascade is trained on a large data set of positive images (those containing a face) and negative images (those not containing a face). This training makes it relatively robust to image degradations such as noise, blur, and illumination changes in the input images, and results in good detection rates for faces with different expressions and skin colour. For our detection we used the trained classifier for frontal faces (see Fig. 3).

Given pixel coordinates from the position of each detected face in the video we obtain the line of sight from camera calibration. By projection onto the reference plane, we get the azimuth for each detected face in each frame for both views. We then compute the azimuth with respect to the robotic head. There are two sources of inaccuracy in the computation of the azimuth: the estimation of camera parameters in camera calibration and the detection of faces in the frames. Because of the short focal length of the cameras ($f = 6\,\text{mm}$) a deviation of approximately 23 pixels translates into an angular error of one degree. In our experiments we found the error of the azimuth in visual localization to be below one degree.

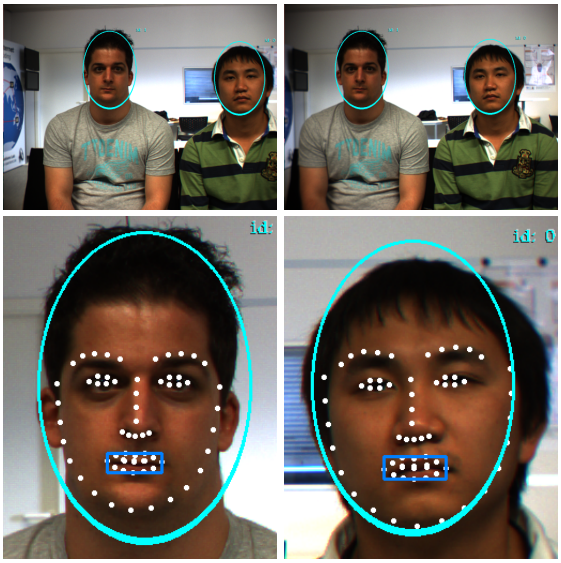

Fig. 3. *Top: Left and right frame of test sequence with detected faces marked; Bottom: Detected contours of face, eyes, nose, and lips, mouth region bounding box of faces detected in left frame.*

3) Mouth Detection and Mouth Information Extraction: We detect the mouth region of a detected face using the approach presented in [17]. This approach uses an Active Shape Model (ASM) for fitting and tracking facial features in image sequences. It is based on a parameterized shape model that is fitted to the locations of detected landmarks in the face. Furthermore, it is capable of identifying the silhouette of the face, the position of the eyes and eyebrows, the position of the nose, and the position and contour of the lips. We decided to use this approach because of its robust and reliable detection results for various poses of the head and facial expressions. Although the detection results for face, eyes, mouth, etc. were reliable, they were not precise enough for detecting visual lip activity. Consequently, we use the contours of the lips to determine a bounding box of the mouth region, which is used to extract information that will be used in active speaker detection (see Fig. 3).

Our approach for extracting visual information to detect active speakers is inspired by [4]. In this approach the active speaker is identified by computing and comparing the average fraction and the variance in the fraction of pixels with low intensities in the mouth region. In this context pixels with low intensities are those below a specified threshold in the greyscale image. The idea behind this approach is that while speaking, parts of the mouth cavity of the speaker are visible in image, which are not well illuminated and hence increase the fraction of dark pixels in the mouth region. In our approach we only use the number of pixels with low intensities in the mouth region.

The video modality alone can detect the movements of the lips. However, some lip movements make no voice, such as sigh, lip play, etc. (we call them false lip movements). As a consequence, both audio modality and video modality are needed to detect the speaker activity.

III. AUDIO-VISUAL FUSION IN TWO STAGES

As discussed previously, video localization for each face has a much higher accuracy than audio localization. However, this information is only of value if the person is actually speaking. Consequently, we propose a two stage audio-visual fusion approach. The first stage has the goal to detect which person is active (speaking). The second stage integrates the localization estimates from both modalities.

A. Audio-visual Fusion based Speaker Activity Detection

Apparently the video modality alone cannot distinguish between false lip movements and real lip movements of speaking because there is no visual difference. As a consequence, audio information is needed to detect the activity of speakers.

Fig. 4 (top) shows the audio signal for a given speaker and we can see that the amplitudes (corresponding to signal power) during the talking and silence period show large differences. Similarly, the number of pixels with low intensities of mouth region for open and close mouth also have a large difference, as shown in Fig. 4 (center). To detect speaker activity, cross-correlation can be used to explore the correlation between audio information and video information. We also found that cross-correlation has a better performance than mutual information here. The detection procedure is as follows:

Algorithm I:

1) Calculate the audio power for each time unit and divide them in segments A_{seg} (with a length of e.g. 0.5 second).
2) Interpolate the video data according to the audio data and divide them in segments V_{seg} having the same length as the audio segments.
3) Compute the cross-correlation $xcorr(A_{seg}, V_{seg})$ between audio and video data for each time segment.
4) Find the maximum value of the cross-correlation results and compare it with a threshold to determine the speaker activity for the current segment by

$$Active_{sp_k}(seg) = \begin{cases} 1, & \text{if } \max\big(xcorr\big(A_{seg}, V_{seg}\big)\big) \\ & \quad > ActiveThres_{sp_k} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

A fixed threshold $ActiveThres_{sp_k}$ is sufficient for a speaker at a particular position as experimental results show. A location-adaptive threshold can be automatically computed based on the audio power and video data.

B. Audio-visual Results Integration

Based on the audio-visual fusion, the speaker activity is plausible and the location of each active speaker in view can be used as the final localization result. However, when the speakers are sheltered in view, this method so far cannot localize the active speakers. To increase the robustness of the system, we also integrate the localization results from both modalities in the last stage.

First, we represent video results as a probabilistic function of azimuth angles. To compensate for potentially missing detections of the video modality, the probability of all the unclear azimuth angles is set to 0.5. So the video localization results can be represented as follows:

$$V(\phi) = \begin{cases} p_\phi, & \text{active speaker at } \phi \\ 0.5, & \text{otherwise,} \end{cases} \tag{8}$$

where p_ϕ denotes the probability of the speaker activity detected by the audio-visual fusion in the first stage, e.g. 0.99 in our work. As discussed in the above sections, the localization results of the audio modality have larger deviation than the video detections (where the error is below one degree), especially in reverberant environments. So the representation of the video results is expected to have the ability to improve the accuracy of the audio results. We replace the pulses in Eq. 8 with smooth peaks. Inspired by the *Skeleton* method, we propose a Gaussian representation of the video detections as follows:

$$V(\phi) = 0.5 + (p_\phi - 0.5)\, Gau(\phi, [\sigma, \phi_0]) \tag{9}$$

where σ equals to the half range of the maximum errors in degree from the audio modality and ϕ_0 is the location of the visually detected faces. In this way, the video representation has the ability to cut off the deviated audio peaks over a wider azimuth range. Note that the accuracy of the video detection is not reduced by the smoothing with a Gaussian kernel.

In order to build a new probability curve in the azimuth domain, we first multiply audio and video probabilities and then smooth the curve by median filtering. The position of the new peaks indicates the final localization results. The integration procedure is shown as follows:

Algorithm II:

1) Multiply both audio (Eq. 5) and video (Eq. 9) results in the azimuth domain:

$$F(\phi) = A(\phi) \cdot V(\phi). \tag{10}$$

2) Remove small and side peaks based on a specified threshold. This is to remove fake and disturbing sources from audio or video modality.
3) The indices of the residual peaks are the final localization results of active speakers.

Please note that beside the improvement of accuracy and robustness, the Gaussian fusion have also the ability to distinguish between close speakers. The performances of the fusions in both stages will be evaluated in the next section.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Our dummy head, Bob, resides in a normal office meeting room of size 10×6 m and a reverberation time $RT_{60} = 0.4$ s. The audio signals are recorded via two microphones in Bob's ears. For our experiments, we invited various subjects for our audio/video recording. The subjects are located approximately two metres away from Bob. Scenes with one or two sound sources are considered. For scenes with two

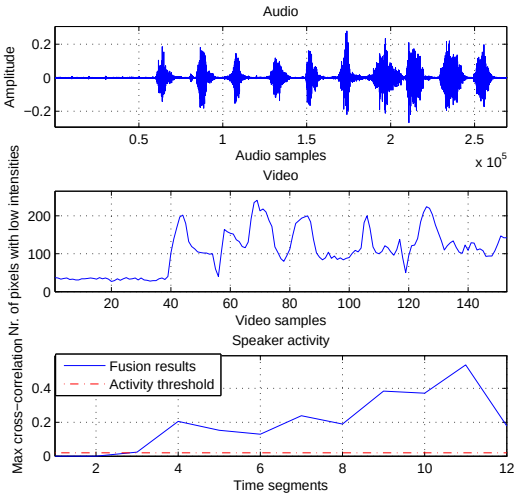

Fig. 4. *Speaker activity detection: a normal case.*

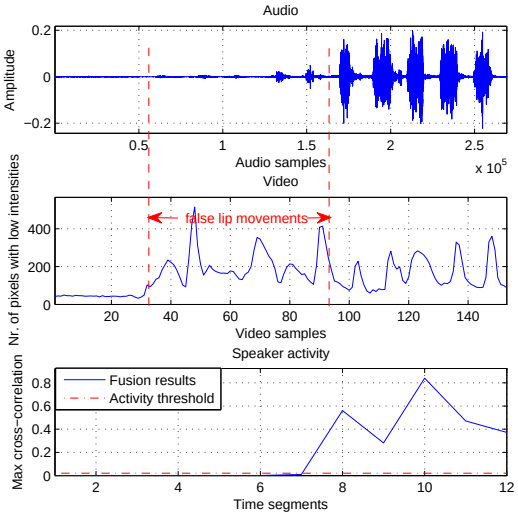

Fig. 5. *Speaker activity detection: with false lip movements.*

or more sound sources the deviation of audio localisation is a little larger than for scenes with only one sound source. Moreover, movements of the sound sources also degrade the accuracy of audio localisation. Static or slow moving subjects are considered. The localisation system is triggered when the power of audio signals exceeds a specified threshold.

A. Audio-visual Fusion based Speaker Activity Detection

Fig. 4 shows the case of a single speaker without false lip movements in a scene containing two speakers. Audio signal for this speaker is extracted from the mixture audio signal based on its location following Eq. 6. As expected, the maximum value of the cross-correlation between audio data and video data shows a very different distribution. And hence a simple filter based on a fixed threshold can determine speaker activity effectively.

Fig. 5 shows the case of one speaker with false lip movements in a scene containing two speakers. It can be observed that the proposed algorithm can handle this situation. The audio signal for the silent speaker (with false lip movements)

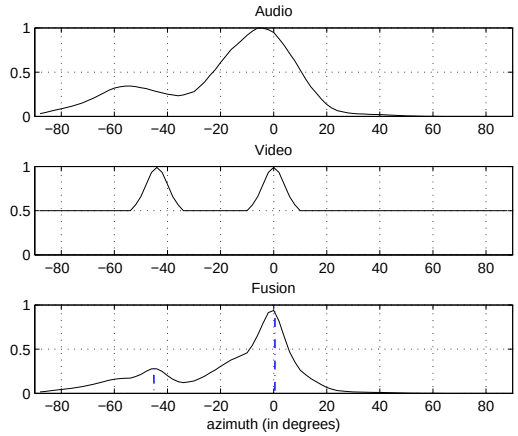

Fig. 6. *Audio-visual results integration: speakers at* 0 *and* -45 *degrees for a normal case where both modalities perform well; Top: probability of the audio localization from Eq. 5; Centre) probability of the video localization using Gaussian extension from Eq. 9; Bottom: fusion results by Algorithm II; The blue lines denote the final localization results. (The layout remains the same for the figures below.)*

is weak and the maximum value of the cross-correlation between audio data and video data with false lip movements still shows a very different distribution. This is due to the fact that both audio data and video data contribute to the final curve and the loss of either leads to an obvious decrease of the final value. The filtering with a threshold can correctly determine speaker activity even in the presence of false lip movements.

Overall, from our experiments with about 30 recording, corresponding to about $1\,000$ segments, the accuracy of speaker activity detection based on audio-visual fusion, is above $99\,\%$. The failure occurs only when a speaker with false lip movements is too close to other active speakers, which occurred seldom in our recordings. In this case, the audio signal extracted for this false active speaker has still high power due to reverberations. Further improvement focusing on the handling of reverberations is left as future work.

B. Audio-visual Results Integration

Fig. 6 shows a fusion result for a simple case where both modalities perform well. Two speakers are located at 0 and -45 degrees respectively. The audio modality alone does detect two peaks but the localization is not very accurate. The video modality can localize the speakers accurately (below one degree deviation in our work) but the speaker activity is not $100\,\%$ plausible. Using the proposed fusion method, the peaks of audio results are correctly adjusted and lead to a more precise localization result.

Fig. 7 shows the case of two active speakers but one of them is occluded in the camera view. We can see that the audio peaks still remain large enough after the fusion for a robust speaker detection. Fig. 8 shows the case of the audio modality failing to distinguish between different audio sources. This may be due to the fact that the voice of a speaker is too weak or the speakers are too close to each other. In this case, the fusion method can distinguish these

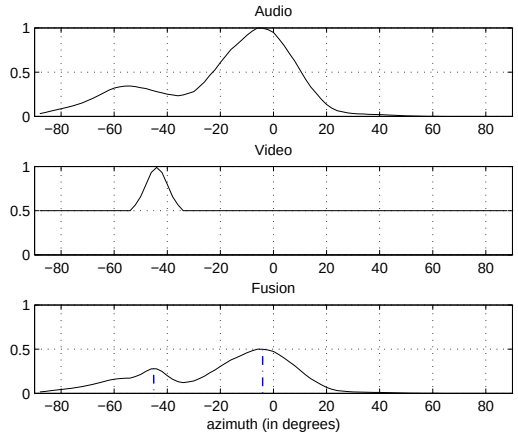

Fig. 7. *Audio-visual results integration: speakers at* 0 *and* -45 *degrees with a sheltered face.*

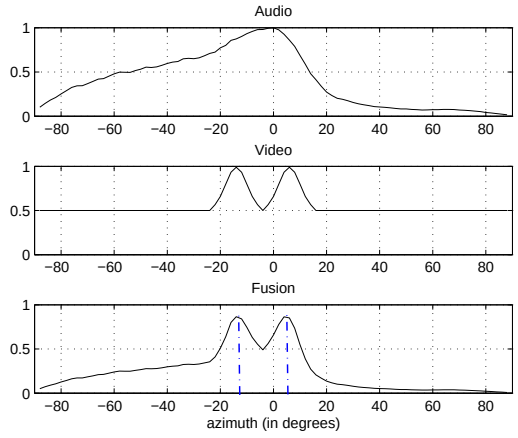

Fig. 8. *Audio-visual results integration: with close active speakers (at* 7 *and* -16 *degree).*

speakers and create corresponding peaks with the help of the video modality.

Overall, for most cases where subjects are in view of the cameras, the final accuracy is as good as visual localisation which is below one degree. For the rare case where subjects are not in view of the cameras, the final accuracy depends only on the audio modality which is about 1 to 10 degrees of deviation for locations from 0 to ± 90 degrees, respectively.

V. CONCLUSION

In this work we first proposed a robust system for speaker localization in reverberant environments that is based on audio-visual fusion in two stages. The audio modality performs the audio source localization based on advanced audio processing algorithms and extracts audio signal for each visual face to help the video modality with speaker activity detection. The video modality localizes the detected faces and computes the number of pixels with low intensities in the mouth region of speakers, where the latter one is used to detect speaker activity. In the first stage, speaker activity is detected based on the audio-visual fusion which can handle false lip movements. In the second stage, estimates from both modalities are integrated by the proposed Gaussian fusion method that audio localization deviations are compensated while the cases of occluded faces in view are well handled. As a consequence, the localization accuracy and robustness compared to the audio/video modality alone is significantly increased. Experimental results for different scenarios confirmed the improved performance of the proposed system.

Future work includes improving audio sources localization by monaural grouping and onset filtering and improving audio extraction against reverberation. Another future research direction is speech separation based on audio-visual fusion.

ACKNOWLEDGMENT

Many thanks to Daniel Gnad, Zheng Xu and other colleagues in the Telecommunications Lab for the help in the experiments related to this work.

REFERENCES

[1] P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Trans. on Multimedia*, vol. 10, no. 1, pp. 63–73, 2008.

[2] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," in *IEEE Trans. on Systems, Man, and Cybernetics*, 2009, pp. 799–807.

[3] S. Kümmel, E. Haschke, and T. Herfet, "Human inspired auditory source localization," in *Digital Audio Effects*, 2009, pp. 20–27.

[4] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," in *Circuits and Systems for Video Technology*, 2009, pp. 133–137.

[5] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Advances in Signal Processing*, 2007.

[6] T. Butz and J.-P. Thiran, "Feature space mutual information in speechvideo sequences," in *International Conference on Multimedia and Expo*, 2002, pp. 361–364.

[7] M. Beal, H. Attias, and N. Jojic, "Audio-visual sensor fusion with probabilistic graphical models," in *ECCV*, 2002.

[8] J. Blauert, *Spatial Hearing – The Psychophysics of Human Sound Localization*. Cambridge, UK: MIT Press, 1997.

[9] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Applied Psychology Unit (APU), Report 2341*, 1988.

[10] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, pp. 361–378, 2004.

[11] K. W. Wilson and T. Darrell, "Learning a precedence effectlike weighting function for the generalized cross-correlation framework," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, pp. 2156–2164, 2006.

[12] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *Journal of the Acoustical Society of America*, vol. 106, pp. 1633–1654, 1999.

[13] J. Woodruff and D. L. Wang, "Integrating monaural and binaural analysis for localizing multiple reverberant sound sources," in *Proc. of the ICASSP*, 2010, pp. 2706–2709.

[14] R. Tsai, "An efficient and accurate camera calibration technique for 3d machine vision," in *CVPR*, 1986, pp. 364–374.

[15] M. Grochulla, T. Thormählen, and H.-P. Seidel, "Using spatially distributed patterns for multiple view camera calibration," in *Computer Vision/Computer Graphics Collaboration Techniques and Applications (Mirage)*, 2011, pp. 110–121.

[16] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001, pp. 511–518.

[17] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *ICCV*, 2001, pp. 1034–1041.