AUDIO-VISUAL MULTIPLE ACTIVE SPEAKER LOCALISATION IN REVERBERANT ENVIRONMENTS

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

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

ABSTRACT

Localisation of multiple active speakers in natural environments with only two microphones is a challenging problem. Reverberation degrades the performance of speaker localisation based exclusively on directional cues. This paper presents an approach based on audio-visual fusion. The audio modality performs the multiple speaker localisation using the *Skeleton* method, energy weighting, and precedence effect filtering and weighting. The video modality performs the active speaker detection based on the analysis of the lip region of the detected speakers. The audio modality alone has problems with localisation accuracy, while the video modality alone has problems with false detections. The estimation results of both modalities are represented as probabilities in the azimuth domain. A Gaussian fusion method is proposed to combine the estimates in a late stage. As a consequence, the localisation accuracy and robustness compared to the audio/video modality alone is significantly increased. Experimental results in different scenarios confirmed the improved performance of the proposed method.

1. INTRODUCTION

The problem of localising the active speakers in reverberant and clustered 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. Because of the potentially large number of subjects moving and speaking in such cluttered environments the problem of robust speaker localisation is challenging.

In many systems that handle speaker localisation, audio and video data are treated separately. Such systems usually have subsystems that are specialised for the different modalities and are optimised 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. The reason is that each modality may compensate for weaknesses of the other one. Thus, whereas a system using only video data may mistake the background for the object or lose the object altogether due to occlusion, a system also using audio data could continue to focus on the object by following its sound pattern. Conversely, video data could help where an audio system alone may lose track of the object as it is masked by background noise and reverberation.

The problem of multimodal multiple speaker localisation 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. 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 localisation, 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 the active speaker among a few face candidates, where it is assumed that all the faces of the speakers can be successfully detected by the video modality. However, this assumption does not always hold in practise, especially in cluttered environments.

In this paper we present an approach that fuses the estimates of both modalities in a late stage. The audio modality performs the multiple speaker localisation using the *Skeleton* method [8, 9], energy weighting, and precedence effect filtering and weighting. The video modality performs the active speaker detection based on the increased average value and standard deviation of the number of pixels with low intensities in the mouth region of speakers. The results of both modalities are represented as probabilities in the azimuth domain. Inspired by the *Skeleton* method, the Gaussian distribution is used for the representation of the video results to compensate for the localisation deviation of the audio modality. Meanwhile, the audio modality has the ability to correct the false detection of the video modality.

In our human-machine interaction scenario, a motorised human dummy head with three degrees of freedom (called Bob) 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 by two microphones 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.

In summary, this paper presents the following contributions. Firstly, we propose a robust system for speaker localisation 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. Furthermore, it can handle the most difficult scenario where multiple speakers are talking at the same time. Finally, the late fusion approach allows the simultaneous improvement of estimation accuracy and robustness. If both modalities are available, the estimation accuracy is improved due to the accurate video localisation. Nevertheless, the approach is also robust if only a single modality contributes information.

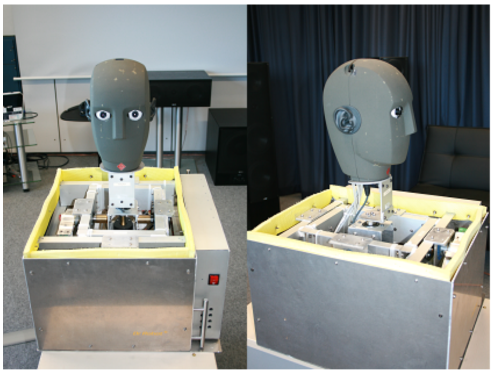

Figure 1: *Bob — the movable human dummy head.*

The rest of this paper is organised as follows. Section 2 and Section 3 present the audio modality and video modality, respectively. The proposed fusion method is described in Section 4. In Section 5, we will show experimental results of the audio modality and the fusion method. The last section provides the conclusion and future work.

2. AUDIO SOURCE LOCALISATION IN REVERBERANT ENVIRONMENTS

It is widely acknowledged that for human audition, Interaural Time Differences (ITD) are the main localisation cues used at low frequencies ($< 1.5\,\mathrm{kHz}$), whereas in the high frequency range both Interaural Level Differences (ILD) and ITD between the envelopes of the signals are used [10]. The resolution of the binaural cues has implications for both localisation and recognition tasks. Headphone experiments show that listeners can reliably detect 10–15 μs ITDs from the median plane, which correspond to a difference in azimuth of between 1 and 5 degrees. On the other hand, the smallest detectable change in ILD by the human auditory system is about 0.5 to 1 dB at all frequencies. Resolution deteriorates as the reference ITD gets larger, and the difference limen can be as much as 10 degrees when the ITD corresponds to a source located far to the side of the head [10].

2.1. Auditory Periphery

Human cochlear filtering can be modeled by a bank of bandpass filters. The filterbank employed here consists of 128 fourth-order gammatone filters [11]. The impulse response of the i^{th} filter has the following form:

$$g_i(t) = \begin{cases} t^3 \exp(-2\pi b_i t)\cos(2\pi f_i t + \phi_i), & \text{if } t \geqslant 0 \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where b_i is the decay rate of the impulse response related to the bandwidth of the filter, f_i is the centre frequency of the filter, and ϕ_i is the phase (here ϕ_i is set to zero). The Equivalent Rectangular Bandwidth (ERB) scale is a psychoacoustic measure of auditory filter bandwidth. The centre frequencies f_i are equally distributed on the ERB scale between 80 Hz and 5 kHz. We specifically set the bandwidth according to the following equations for each filter [12]:

$$\mathrm{ERB}(f_i) = \quad 24.7\left(4.37\,\frac{f_i}{1\,000} + 1\right), \tag{2}$$

$$b_i = \quad 1.019\,\mathrm{ERB}(f_i). \tag{3}$$

In order to simulate the middle-ear transfer function, the gains of the gammatone filters are adjusted according to the data provided by Moore et al. [13]. We include this middle-ear processing for the purpose of physiological plausibility. In the final step of the peripheral model, the output of each gammatone filter is half-wave rectified in order to simulate the firing rates of the auditory nerve [8, 9]. Saturation effects are modeled by taking the square root of the rectified signal.

2.2. Azimuth Localisation and the *Skeleton* Method

Current models of azimuth localisation almost invariably employ cross-correlation, which is functionally equivalent to the coincidence detection mechanism proposed by Jeffress [14]. Cross-correlation provides excellent time delay estimation for broadband stimuli and for narrow band stimuli in the low-frequency range. However, for high frequency narrow band signals it produces multiple ambiguous peaks. 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

$$C(i,j,\tau) = \sum_{n=0}^{M-1} l_i(jT-n)\,r_i(jT-n-\tau)\,\mathrm{win}(n), \tag{4}$$

where win is a window of 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. For efficiency, the Fast Fourier Transform is used to evaluate function 4 in the frequency domain. Computing $C(i,j,\tau)$ for each channel i ($1 \leqslant i \leqslant N$) gives a cross-correlogram, which is computed at 10 ms intervals of the time index j.

Ideally, the cross-correlogram should exhibit a 'spine' (sharp peak) at the delay τ corresponding to the ITD of a sound source. This feature can be emphasised by summing the channel cross-correlation functions, giving a pooled cross-correlogram, $P(j,\tau)$, which is shown as follows:

$$P(j,\tau) = \sum_{i=0}^{N} C(i,j,\tau). \tag{5}$$

In free-field listening conditions, diffraction effects introduce a weak frequency-dependence to the ITDs which is evident in the

HRIR (Head-Related Impulse Responses)-filtered stimuli used here. As a result, the 'spine' can be unclear and Eq. (5) does not exhibit a clear peak at the ITD. Here, we address this issue by warping each cross-correlation function to an azimuthal axis, resulting in a modified cross-correlogram of the form $C(i, j, \phi)$, where ϕ is the azimuth in degrees. The azimuth is quantised to a resolution of 1 degree, giving 181 points between -90 and $+90$ degrees. Warping is achieved by a table look-up, which relates the azimuth in degrees 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 'spine' which is closest to the corresponding azimuth angle based on ILD. The ILD can be calculated by Eq. (6). The mapping from ILD to azimuth angles can be trained for each frequency [3].

$$\mathrm{ILD} = 10 \log_{10} \frac{\sum_n l^2(n)}{\sum_n r^2(n)} \,\mathrm{dB}. \tag{6}$$

A further stage of processing is based on the Skeleton cross-correlation function [9]. 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 $S(i, j, \phi)$:

$$S(i, j, \phi) = Q(i, j, \phi) \exp\left(\frac{-\phi^2}{2\sigma_i^2}\right). \tag{7}$$

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) [9]. This approach is similar in effect to applying lateral inhibition along the azimuth axis, and causes a sharpening of the cross-correlation response.

2.3. Precedence Effect Filtering and Weighting

The term 'precedence effect' refers to a group of psychophysical phenomena which are believed to underlie the ability of listeners to localise sound sources in reverberant environments [10, 15]. In such environments, direct sound is closely followed by multiple reflections from different directions. However, listeners usually report that the sound has originated from a single direction only. The perceived location corresponds to the direction of the first wavefront. Hence it appears that the directional cues in the first-arriving sound are given 'precedence' over cues contained in the later reflections.

In reverberant recordings, many time-frequency units $u_{i,j}$ 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 localisation performance [16]. Motivated by the precedence effect [15, 17], 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 that lie above a specified threshold ($\mathrm{Thres}_{\mathrm{PE}}$). The final results of audio source localisation can be represented as $A(\phi)$, which is the sum of Skeleton functions $S(i, j, \phi)$ for all time-frequency units with precedence effect filtering and weighting:

$$A(\phi) = \sum_i \sum_j w_{i,j} S(i, j, \phi), \qquad \text{if } w_{i,j} > \mathrm{Thres}_{\mathrm{PE}}. \tag{8}$$

Fig. 2 shows the precedence effect filtering and weighting for two male speaking sources at 0 and -45 degrees. From the results, we first can see that it is difficult to determine the non-dominating audio source without precedence effect filtering or weighting. Multiple audio sources are easier to distinguish with precedence effect weighting and filtering. We also found that a threshold of 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 [17]. To avoid this problem, an automatic threshold control is applied. It ensures that the remaining frames 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.

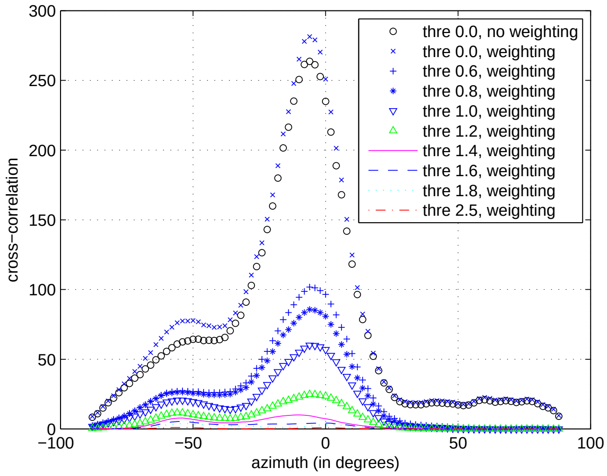

Figure 2: *Precedence effect filtering and weighting (allows amplification of the non-dominating audio source in this example).*

3. VISUAL ACTIVE SPEAKER DETECTION

Besides the audio information, visual information can also be used to localise multiple subjects. To this end, we employ the two cameras that are available in our motorised 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 determine the active speaker.

3.1. Camera Calibration

The goal of camera calibration is to estimate camera parameters. Typical camera parameters are position and orientation of the camera (extrinsic parameters) and focal length, principal point offset, and radial distortion parameters (intrinsic parameters). Popular and often used approaches use a calibration pattern with known geometry for parameter estimation [18]. From one or several images of such a calibration pattern, 2D-3D correspondences can be extracted. These correspondences are then used to estimate the camera parameters. However, calibration of the rotating robotic head requires a special calibration procedure because of the large range of possible viewing directions. Hence, in order to perform this calibration, we use the idea presented in [19]. In this approach multiple spatially distributed calibration patterns are used for camera parameter estimation. Initial camera parameters are estimated with Tsai's approach [18]. Afterwards, the spatially distributed patterns are related into a globally consistent coordinate system. Finally the parameters are optimised by bundle adjustment. In our case, we jointly estimate the intrinsic camera parameters (focal length, radial distortion) for all viewing directions.

3.2. Active Speaker Detection

Our approach for visually detecting active speaker consists of the following three steps:

- Face detection,
- Mouth region detection, and
- Active speaker detection.

3.2.1. Face Detection

We apply a face detection approach [20] that is provided by the OpenCV library. The OpenCV face detector is a popular, easy-to-use, and robust method for face detection. It is based on Haar-like features for object detection, which are used in a classifier cascade. The classifier cascade is trained on a large data set of positive images (those containing a face) and negative images (those not containing a face). Training the classifier on a large data set of images makes it relatively robust to image degradations such as noise, blur, and illumination changes in the input images, and gives a good detection rate for faces with different expressions and skin colour. For our detection we used the trained classifier for frontal faces (see Fig. 3), which worked well as long as the face of a stationary or moving speaker is facing the camera.

3.2.2. Mouth region detection

Within each face found in the image we locate the mouth region. We detect the mouth region using the approach presented in [21]. This approach uses an Active Shape Model (ASM) for fitting and tracking facial features in image sequences. It is based on a parameterised shape model that is fitted to the locations of detected landmarks in the face. The approach 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 the 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 the detection of visual lip activity. Consequently, we use the contours of the lips to compute a bounding box of the mouth region, which is used as input for the last step of active speaker detection (see Fig. 3).

3.2.3. Speech Detection

For detecting active speakers, we employ the main idea of [4]. In this approach the active speaker is identified by computing 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 average and variance are computed over a time window of several frames. Lip activity is detected in case where both values exceed specified thresholds. These thresholds can be determined by test sequences in which the persons are silent. The idea behind this approach is that while speaking, parts of the mouth cavity of the speaker are visible in the image, which are not well illuminated and hence increase the fraction of dark pixels in the mouth region (see Fig. 4).

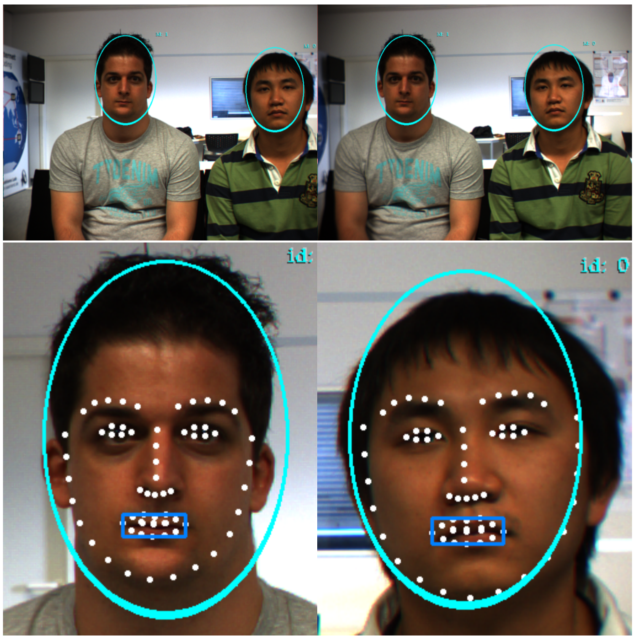

Figure 3: *Top: Left and right frame of test sequence with detected faces marked; Bottom: Contours of face, eyes, nose, lips, and mouth region bounding box of faces detected in left frame.*

3.3. Detection Results

To combine audio and visual localisation, we compute the azimuth angle for the detected speakers. 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. From the azimuth of a detected face in the left and right image we then compute the azimuth with respect to the robotic head. Two sources of inaccuracy occur in the computation of the azimuth: the estimation of camera parameters in camera calibration and the detection of faces in the frame. Because of the short focal length of the cameras ($f = 6$ 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 localisation to be below one degree.

Finally, we combine the localisation information with the visual active speaker detection to determine the azimuth of the active speaker.

We tested this approach on 10 sequences with speakers of different ethnicities. For each sequence, we captured two synchronised RGB video streams with 1024×768 pixels at 7.5 frames per second resulting in sequences between 40 and 60 seconds in

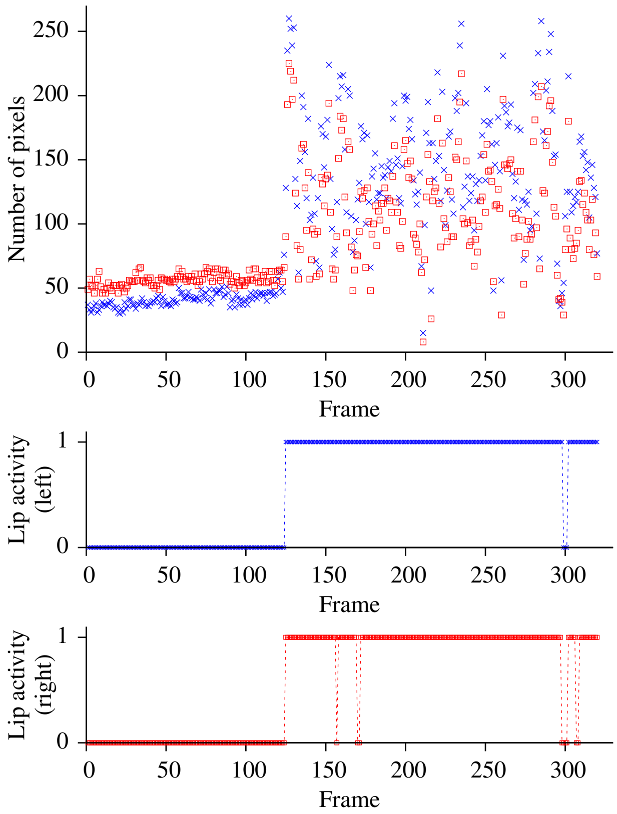

Figure 4: *Top: Number of low greyscale intensity pixels in the mouth region of one subject in left (blue, ×) and right (red, □) frames of a test sequence. Centre and Bottom: Corresponding visual lip activity detection result for left (blue, ×) and right (red, □) frames. The observed subjects started to talk at frame no. 125.*

length. In each sequence two persons are visible. In the first part of the sequence both persons are quiet. After a signal, both persons start to talk at the same time until the end of the sequence. The sequences consisted of approximately 4 000 frames in total. With two video streams for each test sequence and two speakers observed in each video stream, lip activity detection has been evaluated on approximately 16 000 instances. We evaluated the number if times visual lip activity was correctly detected in the test sequences and achieved a detection rate of 95 %.

There are three main reasons for a slightly lower detection rate in comparison to the performance reported in [4]. First, the camera's automatic shutter adjusts the brightness of the image to maintain an average brightness in the entire image, which sometimes leads to poor contrast in the faces. Second, imprecision in fitting the lip contour to the image sometimes leads to poorly detected mouth regions. Third, face and especially mouth region detection might fail in case of motion blur.

4. AUDIO VISUAL FUSION

In this work, we propose a new method to fuse audio and video results, the goal of which is to deal with disadvantages of both audio and video modalities. We fuse both results and build a new probability curve in the azimuth domain. The new peaks show the final localisation results. The motivation is to keep partial information from both modalities. Audio localisation deviations are adjusted by video results, while video detection failures are compensated for by audio results.

4.1. Probabilistic Representation of Video Detections

In order to fuse the results of both audio and video modalities, the video results have to be represented 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 localisation results can be represented as follows:

$$V(\phi) = \begin{cases} p_\phi, & \text{if speaker at } \phi, \\ 0.5, & \text{otherwise,} \end{cases} \tag{9}$$

where p_ϕ denotes the probability of the speaker activity from video results, e.g. $0.90, 0.95$. As discussed in the above sections, the localisation 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. (9) with smooth peaks. Inspired by the *Skeleton* method, we propose a Gaussian representation of the video detections as follows.

$$V(\phi) = \begin{cases} 0.5 + (p_\phi - 0.5)\,\mathrm{Gau}(\phi, [\sigma, \phi_0]), & \text{if speaker at } \phi \text{ and} \\ & |\phi - \phi_0| < \mathrm{Range}_A, \\ 0.5, & \text{otherwise,} \end{cases} \tag{10}$$

where Range_A is equal to the half range of the maximum errors in degrees from the audio modality. 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.

4.2. The Fusion Procedure

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 localisation results. The fusion procedure can be summarised as follows:

1. Multiply both audio and video results in the azimuth domain:
$$F(\phi) = A(\phi) \cdot V(\phi). \tag{11}$$
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 localisation results of active speakers.

We will show and discuss the fusion results for different scenarios in the next section.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In this section we present audio localisation results and the final fusion results.

5.1. Auditory Localisation Results

Our robotic head, Bob, resides in a normal office meeting room of size 10 × 6 m with a reverberation time of $RT_{60} = 0.4$ s. The audio signals are recorded by two microphones in Bob's ears. For our experiments, we invite various candidates from among our lab members to do the audio/video recording. These data are denoted as meeting room data. We also use HRTF to generate audio signals without reverberations ($RT_{60} = 0$ s), denoted as anechoic data. In our experiments, scenes with one or two sound sources are considered. For scenes with two or more sound sources the deviation of audio localisation is a little larger than for scenes with only one sound source. Moreover, movement of sound sources also degrade the accuracy of audio localisation. The overall audio modality results with and without precedence effect handling are shown in Tab. 1.

	0°	±10°	±20°	±30°
anechoic−PE	0.13	0.20	0.22	0.20
anechoic+PE	0.13	0.13	0.20	0.22
meeting−PE	1.7	2.2	2.3	3.5
meeting+PE	0.5	0.9	1.7	2.5
	±45°	±60°	±80°	
anechoic−PE	1.8	2.8	8.4	
anechoic+PE	1.7	3.3	7.8	
meeting−PE	7.3	10.5	14.2	
meeting+PE	3.2	4.2	9.3	

Table 1: *Average azimuth errors in degrees* (−PE *and* +PE *denotes our audio modality without and with precedence effect handling, respectively).*

From Tab. 1 we can see that for the anechoic room the precedence effect weighting and filtering makes no big difference. This is because there are no reverberations in audio signals from the anechoic room. As expected, for the meeting room scenario the performance of the audio modality without PE degrades significantly. Meanwhile, our audio modality with PE still works well, thanks to the precedence effect weighting and filtering.

It is also confirmed in our experiments that the *Skeleton* method and precedence effect weighting/filtering are helpful to distinguish weak peaks and to reduce disturbing peaks. To further improve the localisation performance, we need help from the video modality, where the localisation errors can be as low as one degree. The performance of the proposed fusion method will be shown in the following subsection.

5.2. The Fusion Results

Fig. 5 shows the fusion result for the case where both modalities perform well. Two speakers are located at 0 and −45 degrees, respectively. The audio modality alone detects two peaks, but the localisation is not very accurate. The video modality can localise the speakers accurately (below one degree of 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, which leads to a more precise localisation result.

Fig. 6 shows the case where the video modality has a false positive active speaker detection. We can see that the audio results have the ability to remove these false peaks of the visual results. The audio modality is sometimes more robust for determining speaker activity than the video modality, because there could be lip movements without sounds. Fig. 7 and 8 show the cases where the video modality misses an active speaker. We can see that the audio peaks still remain large enough after the fusion for a robust speaker detection.

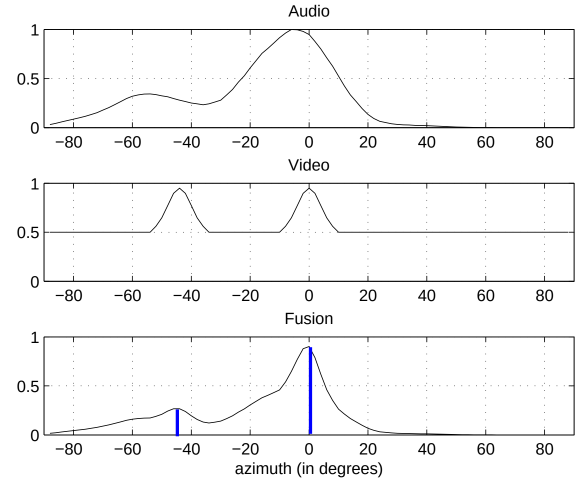

Figure 5: *Experiment 1: speakers at* 0 *and* −45 *degrees, for the case where both modalities perform well; Top: probability of the audio localisation from Eq. (8); Centre: probability of the video localisation using Gaussian extension from Eq. (10); Bottom: fusion result, azimuth of detected speakers indicated by blue lines. (The layout remains the same for the figures below).*

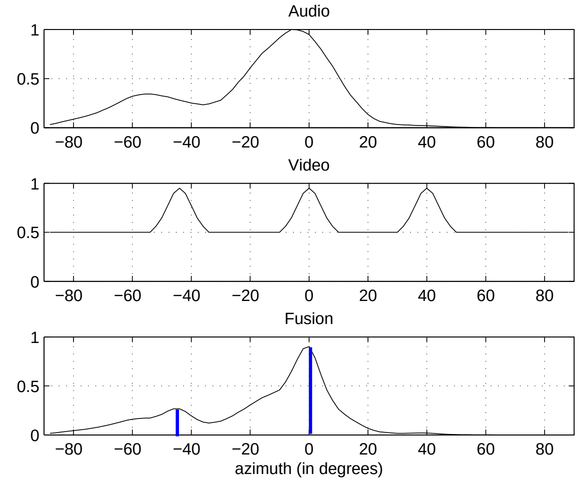

Figure 6: *Experiment 2: speakers at* 0 *and* −45 *degrees, for a case where the video modality has a false positive active speaker.*

Fig. 9 shows a case where the audio modality fails to detect an audio source. This may be due to the voice of this speaker being too weak. In this case, the fusion method can create a peak with the help of video modality. Fig. 10 shows a case where two audio sources are too close and the audio modality alone fails to distinguish them. In this case ,the proposed fusion method can also distinguish the audio sources with the help of the video modality.

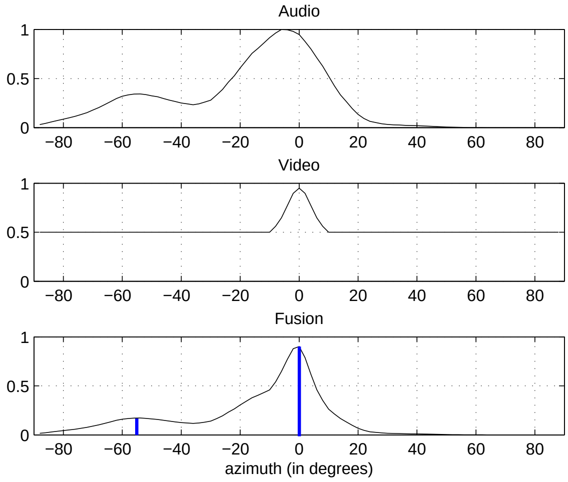

Figure 7: *Experiment 3: speakers at* 0 *and* -45 *degrees, for a case where the video modality misses one non-dominating speaker.*

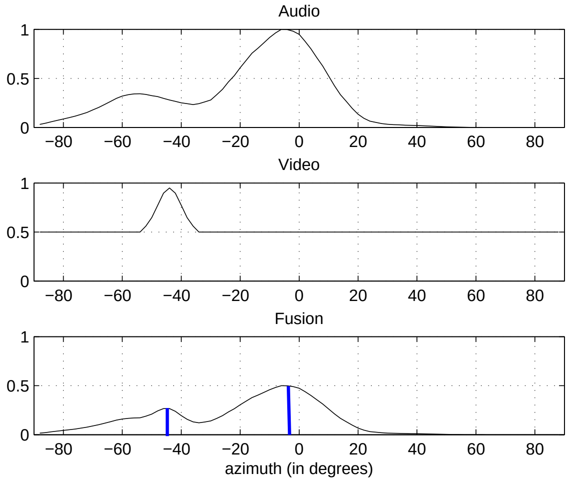

Figure 8: *Experiment 4: speakers at* 0 *and* -45 *degrees, for a case where the video modality misses one dominating speaker.*

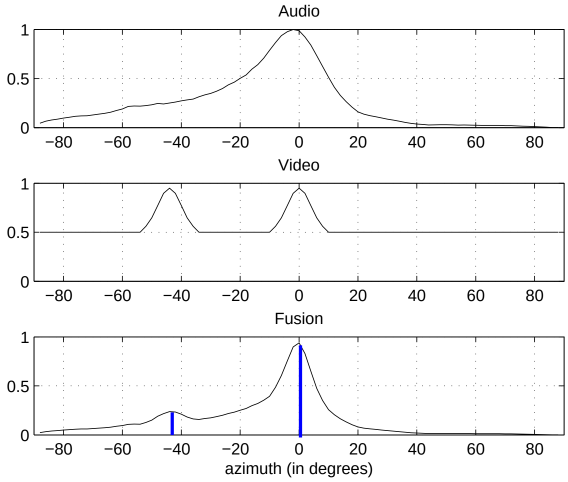

Figure 9: *Experiment 5: speakers at* 0 *and* -45 *degrees, for a case where the audio modality misses one dominating speaker.*

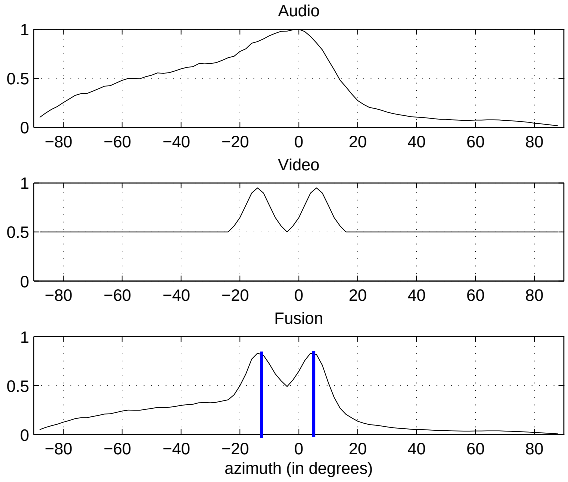

Figure 10: *Experiment 6: speakers at* 7 *and* -16 *degrees, for a case where the audio modality fails to separate speakers.*

6. CONCLUSION

In this work, we first proposed a robust system for speaker localisation in reverberant environments that is based on the combination of advanced audio and video processing algorithms. Multiple speaker localisation is performed by the audio modality using the *Skeleton* method, energy weighting, and precedence effect filtering and weighting. The video modality performs an active speaker detection and localisation as well. Detection of an active speaker is based on the increased average value and variance of the number of pixels with low intensities in the lip region. Camera calibration allows the localisation of the speaker. The localisation results of both modalities are represented as probabilities in the azimuth domain. A Gaussian fusion method is used to fuse the estimates in a late stage. As a consequence, the localisation accuracy and robustness compared to the audio/video modality alone can be significantly increased. Experimental results for different scenarios confirmed the improved performance of the proposed method.

Future work includes improving the audio source localisation by monaural grouping and onset filtering, and threshold optimisation for visual lip activity detection. Another future research direction is speech separation based on audio-visual fusion.

7. ACKNOWLEDGMENTS

This work has been carried out in the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) and is funded by the German National Science Foundation DFG. Many thanks to Daniel Gnad, Zheng Xu and other colleagues in the Telecommunications Lab for their help in the experiments related to this work.

8. 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 Transactions 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 Transactions 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 *European Conference on Computer Vision*, 2002.

[8] N. Roman, D. L.Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, pp. 2236–2252, 2007.

[9] 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.

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

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

[12] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 44, pp. 99–122, 1990.

[13] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, 1997.

[14] L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative & Physiological Psychology*, vol. 41, pp. 35–39, 1948.

[15] 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.

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

[17] J. Woodruff and D. L. Wang, "Integrating monaural and binaural analysis for localizing multiple reverberant sound sources," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 2706–2709.

[18] R. Tsai, "An efficient and accurate camera calibration technique for 3d machine vision," in *Computer Vision and Pattern Recognition*, 1986, pp. 364–374.

[19] 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.

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

[21] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *International Conference on Computer Vision*, 2001, pp. 1034–1041.