Algorithms for Audiovisual Speaker Localisation in Reverberant Acoustic Environments

Christoph Voges[a], Volker Märgner[a], Rainer Martin[b]

[a]Braunschweig Technical University, Germany, e-mail: {c.voges, v.maergner}@tu-bs.de
[b]Ruhr-Universität Bochum, Germany, e-mail: rainer.martin@rub.de

Abstract - Innovative and future human-machine interfaces or video conference systems require knowledge of the speaker's position for automatic beamformer- and camera-steering purposes. To determine this position, acoustical as well as visual localisation techniques can be applied, and the aim of this project was to develop suitable algorithms for such an audiovisual speaker localisation. Furthermore, an experimental setup had to be developed in order to test these algorithms under realistic conditions. Since both, acoustical and visual localisation techniques have specific advantages and disadvantages, depending on the situation, their location estimates were combined, thus forming a robust overall joint location estimate of the speaking person.

1 Introduction

To record the sound of a speaking person in interactive human-machine interfaces or video conference systems, microphone arrays are increasingly employed. In contrast to single microphones, these microphone arrays can be combined with beamformers with variable directional characteristics and are therefore able to effectively suppress reverberation and noise coming from undesired directions. Hence, the position of the speaker in relation to the microphone array has to be known in order to steer the beamformer towards the speaker's direction. Furthermore, the estimated position might be used to steer a pan tilt zoom (PTZ) camera. Unlike a fixed-view camera, such a camera is capable of obtaining detailed visual information about an object of interest. To determine the position of the speaker, acoustical as well as visual localisation techniques can be applied. Furthermore, by combining acoustical and visual information, improved location estimates can be obtained.

This paper will focus on two-dimensional audiovisual localisation methods for a system comprising two microphone arrays and two cameras and is organised as follows: The next section is devoted to acoustical speaker localisation techniques, particulary the generalised cross correlation method (GCC). Section 3 suggests an approach to visual speaker localisation based on background subtraction and section 4 describes the experimental setup and the experiments that were performed to evaluate the performance of these algorithms. Based on the experiments, an approach will be presented in section 5 that combines both acoustical and visual location estimates to form a robust audiovisual overall location estimate. The last section draws conclusions from the experiments and gives suggestions for further developments.

2 Acoustical Speaker Localisation

One possibility of estimating the speaker's position would be an analysis of the audio signals as received by a microphone array. Since the microphones within the array are separated by a certain distance d, the acoustical waves originating from a speaking person will reach one of the microphones first and the second one after a time interval Δt. This time interval is generally referred to as the TDOA (Time Difference Of Arrival) and defined with respect to a certain reference microphone. Accordingly, the angle φ between a plane acoustical wave and a microphone pair is given by

$$\varphi = \arccos\left(\frac{c \cdot \Delta t}{d}\right) \tag{1}$$

where c denotes the velocity of the sound waves ($c \approx 343$ m/s under normal conditions [4]). The TDOA can be estimated by means of the *generalised cross correlation* (GCC) [3], that is

$$R^{(\mathrm{g})}(\tau) = \int_{-\infty}^{\infty} \psi_{\mathrm{g}}(f)\, G_{x_1 x_2}(f)\, e^{j2\pi f \tau} \mathrm{d}f \tag{2}$$

with the *general frequency weighting* $\psi_{\mathrm{g}}(f)$ and the cross power spectral density $G_{x_1 x_2}(f)$ between the two microphone signals $x_1(t)$ and $x_2(t)$. The GCC performs a correlation in the frequency domain whereby the weighting term $\psi_{\mathrm{g}}(f)$ emphasizes or suppresses certain frequency components. One example of such a weighting function is the so-called Smoothed Coherence Transform (SCOT) given by

$$\psi_{\mathrm{scot}}(f) = \frac{1}{\sqrt{G_{x_1 x_1}(f)\, G_{x_2 x_2}(f)}} \tag{3}$$

with the power spectral density functions $G_{x_1 x_1}(f)$ and $G_{x_2 x_2}(f)$ of the microphone signals. Other common choices for $\psi_{\mathrm{g}}(f)$ are the Roth weighting, the Phase Transform (PHAT) and the Hannan and Thomson weighting (HT). In general, these weighting functions improve the performance of the GCC estimation procedure in the presence of noise and reverberation. All these weightings are functions of the power spectral densities or cross power spectral densities of the microphone signals. These densities can be obtained through a short-time spectral analysis comprising windowing and recursive periodogram averaging. A detailed description of all the above-mentioned weighting functions can be found in [3]. The SCOT, Roth, PHAT and HT weighting functions were implemented and evaluated in this study.

3 Visual Speaker Localisation

Alternatively, the speaker's position can be estimated using visual information, for example if the speaker is located within the view of a camera. Various means are discussed in the literature in order to detect a person within an image or a video sequence. One such method is the so-called background subtraction approach [2],[5] which compares the current image $\boldsymbol{P}[x, y]$ with a known background image $\boldsymbol{P}_{\mathrm{bg}}[x, y]$.[1] Therefore, the absolute value of the difference between both images is calculated:

$$\boldsymbol{P}_{\mathrm{diff}}[x, y] = \Big|\boldsymbol{P}[x, y] - \boldsymbol{P}_{\mathrm{bg}}[x, y]\Big|. \tag{4}$$

The background image may either be recorded separately or be obtained by background estimation techniques such as the temporal median filter [5].

[1] When using a black and white camera, the values of $\boldsymbol{P}[x, y]$ and $\boldsymbol{P}_{\mathrm{bg}}[x, y]$ directly represent the brightness values for each pixel. Alternatively, when using colour cameras, these matrices may represent a single colour channel. For this paper only the green channel of the cameras' colour information was evaluated.

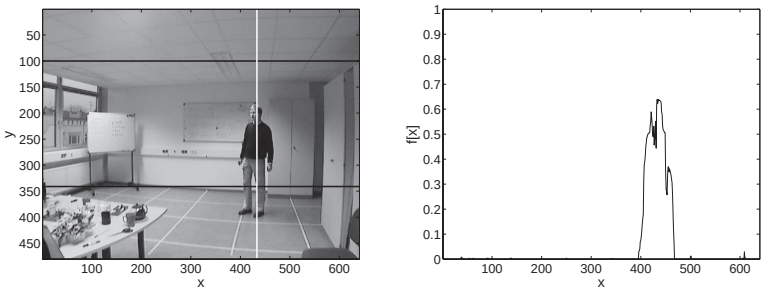

Figure 1: An example for a detected speaker and the corresponding visual localisation function $f[x]$. The black horizontal lines mark the evaluated range and the white vertical line shows the estimated speaker position.

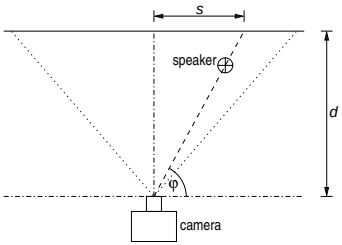

Figure 2: A camera observing a speaker (top view): The speaker direction (dashed line) is projected on an imaginary wall (solid line) running perpendicular to the camera's viewing direction at a distance d to the camera.

In order to transform the visual information into a lateral direction estimate, the visual localisation function

$$f[x] = \frac{1}{y_2 - y_1 + 1} \sum_{i=y_1}^{y_2} \boldsymbol{P}_{\text{th}}[x, y = i] \tag{5}$$

is introduced [8]. The matrix $\boldsymbol{P}_{\text{th}}[x, y]$ is calculated by thresholding $\boldsymbol{P}_{\text{diff}}[x, y]$. That is, for each pixel of $\boldsymbol{P}_{\text{diff}}[x, y]$ the value $\boldsymbol{P}_{\text{th}}[x, y]$ is set to zero if it is below a certain threshold and to one otherwise. It is reasonable to restrict the analysis to a certain range defined by y_1 and y_2 thus discarding regions where it is unlikely for a person to be situated. Normally, this is the case for the ceiling and the floor of a room. The peak of this localisation function gives the desired estimate of the speaker position. An example to illustrate this procedure is presented in Fig. 1.

For further processing, the location estimates within the image have to be transformed into geometrical values with respect to a certain coordinate system. Within the scope of this project, only the lateral position of the speaker is examined. Therefore, the position within the image will be transformed into an angle φ specifying the speaker's direction as depicted in Fig. 2. It is obvious from this figure, that

$$\varphi = \frac{\pi}{2} - \arctan\left(\frac{s}{d}\right). \tag{6}$$

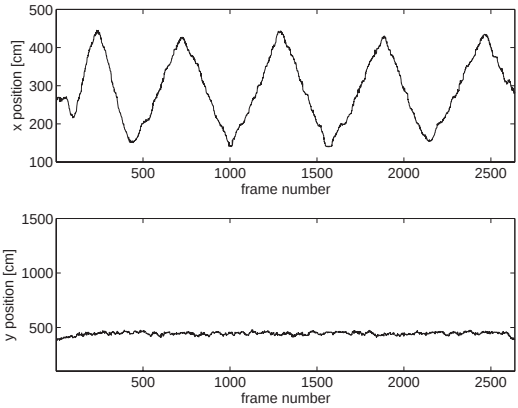

Figure 3: Estimated x- and y-positions using visual analysis.

This expression is used during a calibration step that generates a lookup table containing values φ corresponding to each pixel column of the camera's image sensor chip.

4 Experimental Setup and Measurements

In order to test and evaluate the presented techniques and algorithms under realistic conditions, a room was prepared with microphone arrays, cameras and the corresponding recording hardware. The acoustical reverberation time inside this room was approximately 1200 ms, thus it represents a quite strongly reverberant environment. A precise depth information was obtained through triangulation between two cameras and two microphone arrays. Now the results for an example measurement will be presented and interpreted. For this measurement, the speaker was reading a text and walking up and down in a room at a constant x-position and an approximately constant velocity. This is clearly recognisable when regarding the estimated speaker positions shown in Fig. 3 and 4. It is obvious, that in this case the visual detection process works more reliably with relatively small errors compared to the acoustical localisation. A number of other experiments were performed to evaluate the benefits of acoustical and visual localisation and to guide the development of a joint localisation algorithm. The experiments have also shown, that the SCOT and the PHAT weighting functions lead to good results under the experiments acoustical conditions [8].

5 Combined Approach

Through the experiments, the introduced acoustical and visual localisation techniques proved to be well suitable for speaker localisation purposes although each has its specific advantages and disadvantages. Acoustical localisation techniques may be sensitive to noise effects such as reflections or background noise and are naturally unable to locate a speaker during speech pauses. Visual localisation techniques require the observed person to be within the view of the camera and not to be concealed by obstacles. A suitable approach to combine both methods

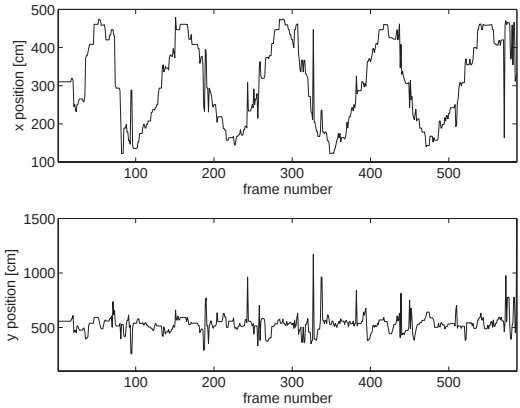

Figure 4: Estimated x- and y-positions using acoustical analysis (weighting function: SCOT).

is to use the visual localisation function's peak amplitude as a measure of the visual location estimate's reliability [8]. This is reasonable since - due to (5) - a speaker that covers the whole detection range between y_1 and y_2 will lead to large peak values of $f[x]$. On the other hand, if the speaker is out or almost out of the detection range and the location estimates become inaccurate, these peaks will tend to very small values. Hence, a system can rely on the visual estimates as long as they are considered as reliable according to this criterion and otherwise switch to acoustical localisation. This can be achieved by comparing the peaks of $f[x]$ to a threshold value. An example is presented in Fig. 5: Around frame indices between 600 to 700 the speaker moved out of the visual detection range resulting in increased localisation errors and smaller values of the localisation function's amplitudes. Then, the estimated positions can be smoothed using a Kalman or a Particle Filter as proposed in [6],[7]. The reference positions for the calculation of the visual localisation errors were determined by manually labelling the visual information.

6 Conclusions

It has been the aim of this paper to develop suitable algorithms in order to locate a speaking person by visual and acoustical localisation techniques. An experimental setup was described in order to test and evaluate these algorithms in practice and results for acoustical and visual localisation experiments were presented. Through the experiments, both methods proved to be able to locate a speaking person. However, the visual speaker localisation works definitely more accurately compared to acoustical localisation whenever the speaker is within the cameras' field of view. Thus, an approach was introduced to combine both methods to form an overall audiovisual joint location estimate.

The algorithms developed here together with the experimental setup form an excellent basis for further developments and extensions. The acoustical signal is important, for instance, to identify a speaker among several persons located by visual means. Furthermore, the obtained location estimates could further be processed by a Kalman Filter or a Particle Filter. Whatever

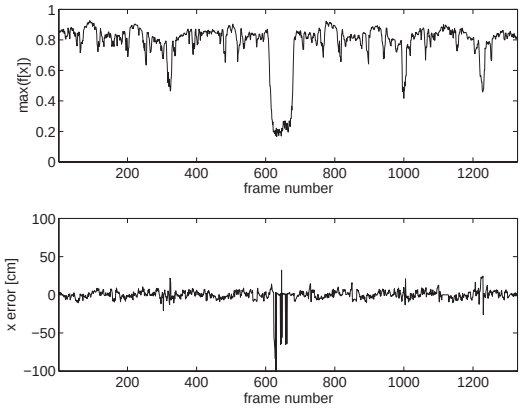

Figure 5: Amplitude of a visual localisation function (top) and the corresponding localisation error (bottom).

further developments will emphasise, both audio and video signals play an important role for the audiovisual speaker localisation process.

References

[1] M. Brandstein, D. Ward, *Microphone Arrays-Signal Processing Techniques and Applications*. Springer, Berlin, 2001.

[2] A. Hampapur et al., *"Smart Video Surveillance"*. IEEE Signal Processing Magazine, Vol. 22, Number 2, pp. 38–51, March 2005.

[3] C. H. Knapp, G. C. Carter, *"The Generalized Correlation Method for Estimation of Time Delay"*. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 4, pp. 320–327, August 1976.

[4] I. Malecki, *Physical Foundations of Technical Acoustics*. Pergamon Press, Oxford, 1969.

[5] M. S. Nixon, A. S. Aguado, *Feature Extraction and Image Processing*. Newnes, Oxford, 2002.

[6] N. Strobel, S. Spors, R. Rabenstein, *"Joint Audio-Video Object Localization and Tracking"*. IEEE Signal Processing Magazine, Vol. 18, Issue 1, pp. 22–31, January 2001.

[7] J. Vermaak, A. Blake, *"Nonlinear Filtering for Speaker Tracking in noisy and reverberant environments"*. Proceedings ICASSP'01, pp. 3021–3024, May 2001.

[8] C. Voges, *Development of Algorithms for the Audiovisual Speaker Localisation*, Diplomarbeit, Braunschweig Technical University, 2005.