The Video Conference Tool Robot VICTOR

Tom Goeckel[1], Stefan Schiffer[2], Hermann Wagner[1], and Gerhard Lakemeyer[2]

[1] Institute of Biology II
RWTH Aachen University, Aachen, Germany
{goeckel,wagner}@bio2.rwth-aachen.de

[2] Knowledge Based Systems Group (KBSG)
RWTH Aachen University, Aachen, Germany
{schiffer,gerhard}@cs.rwth-aachen.de

Abstract.
We present a robotic tool that autonomously follows a conversation to enable remote presence in video conferencing. When humans participate in a meeting with the help of video conferencing tools, it is crucial that they are able to follow the conversation both with acoustic and visual input. To this end, we design and implement a video conferencing tool robot that uses binaural sound source localization as its main source to autonomously orient towards the currently talking speaker. To increase robustness of the acoustic cue against noise we supplement the sound localization with a source detection stage. Also, we include a simple onset detector to retain fast response times. Since we only use two microphones, we are confronted with ambiguities on whether a source is in front or behind the device. We resolve these ambiguities with the help of face detection and additional moves. We tailor the system to our target scenarios in experiments with a four minute scripted conversation. In these experiments we evaluate the influence of different system settings on the responsiveness and accuracy of the device.

1 Introduction

Today, in many situations people who want to lead a discussion are unable to meet at the same place. Often some of the participants have to remotely connect to the discussion via a video link. Going beyond more traditional video conference tools, it would be convenient for a remote participant to have a physical representation at the distant site, for example some kind of robot that allows full acoustic and visual perception of the discussion. This can be achieved by a device that actively and autonomously orients itself towards the current speaker to transmit their facial expressions and gestures in addition to the audio signal.

In this paper, we are concerned with such an autonomous robotic remote presence device for video conferencing. In particular, we developed a robotic tool that uses binaural sound source localization as its main input cue. We supplement our sound localization with additional methods to arrive at a successful system for the given target scenario. First, to the improve responsiveness of the system, we implement a simple onset detector, and second, we add a time integration process that acts as source detector to reduce the number of noisy localizations. Lastly, we additionally use face detection to resolve ambiguities that occur in the sound localization. To optimize our system for the target scenario we conduct an experimental evaluation in a simulated discussion between three speakers. In these experiments we try to find a parameter set that allows for fast reaction times and a good accuracy.

2 Background and Related Work

In this section, we describe the methods underlying the components of our approach before we briefly review related work.

2.1 Sound Source Localization

We use a biologically inspired model for *binaural sound source localization*. Our system relies on interaural time differences (ITD) to determine the azimuthal angle of incidence of a sound source. The algorithm performs a normalized running cross-correlation on the recorded signal over the ITD range of our system. The ITD range is defined by the distance between the microphones, which is about $20\ \mathrm{cm}$ in our case, yielding an approximated range of $[-583\ \mathrm{\mu s}, +583\ \mathrm{\mu s}]$. The sound is recorded at a frame rate of $44100\ \mathrm{Hz}$, which yields an ITD resolution of about $22\ \mathrm{\mu s}$. This can be improved by interpolating between samples [13] or by increasing the sampling rate, but we wanted to keep the complexity of the algorithm low. As the relation between ITD and the azimuth of the incoming sound is sinusoidal, the system has a better resolution close to the midline (less than $2°$), while lateral sources (above $70°$) have a relatively low resolution (about $10°$ in the worst case). Thus, the best strategy for our system is to face the active speaker. When turning towards lateral sound sources, in most cases additional movements will be required to compensate for the localization error. A problem with ITD detection is the so-called front/back confusion because a single ITD can be mapped to two directions in the horizontal plane, one in the front and one in the back. We resolve this with a combination of movements of our robotic device and face detection (see Section 3.3). Localization in office rooms is problematic because of the high degree of reverberations. To reduce their effect on localization, we included a model of the precedence effect [3, 12], a modified version of a model by Faller and Merimaa [7]. The main idea is that the amount of correlation between the left and right signals indicate the reliability of the extracted ITD cues. In reverberant and noisy conditions correlation is low. Our model tries to weight ITD cues according to their amount of correlation. We showed that this model does indeed improve sound localization in noisy and reverberant conditions and, in addition, has a low computational complexity [8]. In our system, the left and right channels of the recording are subdivided into time frames of 2024 samples (about $46\ \mathrm{ms}$). Longer time frames would result in more reliable ITD information but slower reaction times. Processing a time frame of the signal takes about $11\ \mathrm{ms}$, resulting in a theoretical detection time of less than $57\ \mathrm{ms}$ for the sound localization part of the system. Recording and processing is done concurrently, thus every part of the signal gets evaluated and no information is lost. A simple energy-based *signal detector* precedes the sound localization stage to avoid that localization takes place during periods of silence. It has an adaptive energy threshold based on the model presented by [15]. Before considering an ITD estimate of the sound localization step as possible source direction, there is an additional *source detection* stage. It reduces the impact of short and potentially noisy estimates with a histogram filter. This filter, which also takes the current orientation of the device into account, establishes a distribution of ITD estimates that were detected during previous time frames. An exponential decay function makes sure that the accumulated estimates lose influence on the distribution over time. The distribution then acts as a weight to the output of the sound localization stage, and finally a threshold value determines whether a source is considered as reliable or if it should be ignored. During the evaluation, we tried to find a decay constant that allows fast response times without causing too much noise in the sound localization results.

2.2 Onset Detection

While the previously described source detection stage helps to further reduce the impact of noise and reverberations, it also extends the detection time of new sources because it mainly enforces that sources appear for several consecutive time frames before it considers them as reliable. To improve the detection time for new sound sources, we included a simple *onset detector* based on the rising slope of the energy envelope of the input signal and the ITD. The idea for our onset detector stems from psychoacoustical studies with humans that have shown that a rise in the energy envelope is a salient auditory cue that allows glimpsing to infer a source direction in noisy and reverberant conditions [5, 4]. We adapted this idea by determining the maximum slope of the signal envelope during a time frame for each ITD value. The ITD is taken into account by shifting left and right signals to each other and computing the average of the energy envelopes. For each ITD estimate of the sound localization stage, a threshold value for the maximum slope of the energy envelope determines whether the estimate is considered as an onset or an ongoing source. If an onset has been detected it is assigned a weight that is equal to our minimum source selection criterion (0.25, see previous section), and is thus selected as a reliable direction estimate. Our approach might not be as sophisticated as more elaborate onset detectors (e.g. [17]) and it does not replace complex auditory saliency models that also take spectral components into account [6, 10] but it was sufficient to emphasize new sources in our setting. During our evaluation, we apply different slope thresholds to find a value that improves our source detection times without reducing the reliability of the direction estimate.

2.3 Related Work

There exist commercially available tools for video conferencing like Microsoft RoundTable [18]. It features a 360° camera and a set of microphones to provide a panoramic view of the scene and a video with a higher resolution of the active speaker. The system follows the currently speaking person by means of active speaker detection. One of the main differences to the work presented in this paper is that RoundTable uses integrated fusion of the visual and the auditory cue while we perform sound source localization and use face detection independently to resolve ambiguities. Also, RoundTable does not move and uses multiple microphones to detect the active speaker while we deliberately limit ourselves to only using two. There exist also actuated systems like Kubi

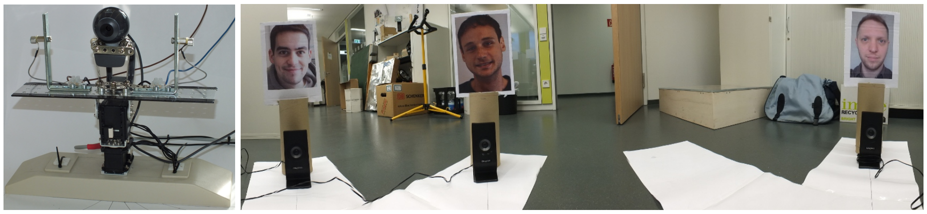

Fig. 1. Our video conferencing tool robot prototype VICTOR and a panorama view of the meeting scenario situation it is being tested in as seen by the PTU.

from revolve robotics.[3] It is a stand to put a smartphone or tablet on that can be remote controlled to pan and tilt. Another example is MeBot [1], which can even move around and whose robot arm can be controlled remotely as well. A survey of mobile telepresence robots can be found in [11]. In this paper we stick with an actuated but stationary platform. Our focus is to make the robot follow a conversation autonomously. That is, the robot should automatically orient towards the currently active speaker.

In [14] the authors describe their approach to multi-modal speaker detection on a Nao robot using two cameras and two microphones. They use multi-modal Gaussian mixture models to fuse the auditory and visual cues. Using two cameras allows the localization of a potential speaker in 3D, which they fuse with binaural cues to prime the probabilities of sound sources. We, however, only have monocular vision and use it to resolve ambiguities in our sound source localization and to provide an image of the speaker. The two perceptional cues work independently.

3 An Autonomous Video Conferencing Tool Robot

After describing the hardware of our device, we present a system overview, the target scenario, and the algorithm used for movement control.

3.1 VICTOR-Hardware

The VICTOR hardware is based on CAPTURE [16], a configurable audio PTU specifically designed to conduct experiments in sound source localization. We added a simple web cam to the CAPTURE system to enable an additional cue, namely visual input that is used for face detection. The camera is a Logitech QuickCam C200 with a diagonal field of view (FOV) of $60°$. With its $4:3$ aspect ratio this results in a horizontal FOV of $48°$. VICTOR is shown on the left side of Figure 1.

3.2 System Overview

Our aim is to tailor a system that mainly uses binaural sound source localization to work well for telepresence in video conferencing by improving on the auditory perception and by adding (existing methods for) face detection to resolve ambiguities.

[3] http://www.kubi.me/ and http://www.revolverobotics.com/

Target Scenario We aim at a robotic tool that acts as a remote presence of a participant in a meeting. VICTOR is placed on the table and it should autonomously follow the currently active speaker to provide an image to the person it represents. Figure 1 shows a 180° panorama view of what VICTOR is able to see of its communication peers.

Architecture In a first step, we filter time frames with a low amount of energy with a simple *signal detector* to avoid false localizations. We then estimate the *ITD* as described in Section 2.1. At the same time the maximum slope of the energy envelope is determined. A slope threshold determines whether an onset has been detected for the current ITD estimates (see Section 2.2). If no onset has been detected, the ITD estimates are integrated over time in the *source detection* stage to reduce the impact of short noisy localizations. The threshold value for source detection was fixed at 0.25 with 1 being the maximum weight a source can achieve. In the case of an onset, the weight of the ITD is directly set to 0.25, and thus qualifies as reliable ITD estimate. Reliable ITD estimates are then combined with face detection to select a source direction and resolve possible front/back confusions.

Face Detection For reasons of simplicity, we take off-the-shelf technology for the vision component. That is, we use the widely-known AdaBoost face detector [9], an implementation of which is freely available in OpenCV.[4] It uses a cascade of haar features to classify regions as face or non-face. In our system it is running with a frequency of 30 Hz on an image of size 320 by 240 pixels. If no face has been detected at the position of the currently selected sound source, we check if a face was detected in the previous time frame at the same position. This reduces errors caused by missed faces.

3.3 Movement Control

Algorithm 1 shows the main procedure for controlling the orientation of the PTU. First, if several sound sources have been detected, the one that is closest to the current position will be selected. If a detected source position lies within the FOV of the camera, the face detector checks if a corresponding face can be found. If so, the PTU directly moves to this position and the system has reached its target. If the source is outside of the FOV, the PTU has no additional cues to evaluate and also has to directly move to that position. However, in case the source is within the FOV and the detector is unable to find a face, we are dealing with an ambiguity. The system tries to resolve this by projecting the position to the back of the PTU and checks, whether it is within the allowed range of the setup. If this is the case, the PTU rotates in this direction, otherwise the current time frame is ignored. Algorithm 1 is used to let VICTOR follow a conversation autonomously.

4 Experimental Evaluation

We ran experiments in a reverberant office environment with a variety of different settings to evaluate the system. To yield reproducible results, instead of letting the system

[4] http://www.opencv.org/

Algorithm 1: The basic algorithm for movement control

```
proc main,
    fetch_sources();                 %% from binaural localizer
    pick_source();                   %% select closest source
    if(!check_fov())                 %% source in field of view?
        move_ptu();                  %% instruct PTU to turn
    else
        if(check_face())             %% if face in FOV
            move_ptu();              %% instruct PTU to turn
        else if(resolve_ambiguity()) %% is 2nd option valid
            move_ptu();              %% turn to 2nd poss. angle
        endif                        %% otherwise ignore frame
    endif
endproc
```

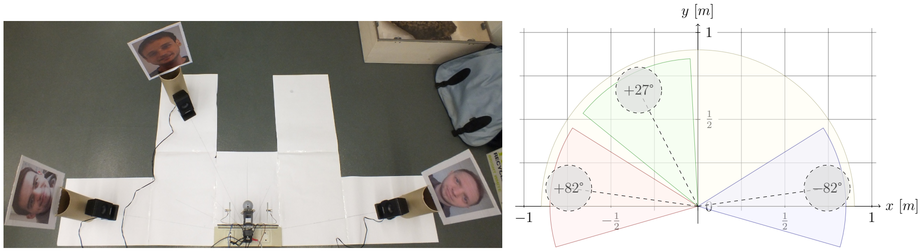

Fig. 2. Bird's eye view and schematic view of the experimental setup. The colored cones represent the FOV of the camera when looking at the $82°$, $27°$, and $-82°$ positions respectively.

perform in a real conversation, our setup uses a single audio file (played back by a Dolby 5.1 speaker system) and a fixed speaker setup positioned around the PTU. The audio file was constructed by arranging spoken sentences from the VoxForge[5] repository in a multi-channel file. The speakers take turns, uttering a random sentence. However, there are several instances where two speakers start speaking at the same time. One of the speakers then usually speaks longer than the other.

4.1 Setup

The setup consists of three speakers placed around the PTU at a distance of 75 cm each. With the PTU placed in the middle, zero degree pointing forward, positive angles to the left the three sources are placed at $-82°$, $+27°$, and $+82°$. Figure 2 shows a schematic view of the setup. We picked the setup described above to make sure we have both cases, one where ambiguities can be resolved, and one where they cannot be resolved with the help of face detection. We only consider the frontal $180°$ to contain valid sound sources. Although the system would be able to handle a full circle area and the PTU

[5] http://www.voxforge.org/

is able to turn in a range of 270°, we restrict ourselves to half a circle for reasons of simplicity and clarity.

4.2 The Evaluation Data Set

Instead of a real conversation, we constructed an audio file to mimic a conversation between three persons. Each speaker only talks for a period of 2 to 15 seconds, requiring a high reactivity of the system. We tried to have equally distributed turn taking between any two speakers. Additionally, we also included cases where two of the three speakers start talking simultaneously. The activity of speakers over time in this audio file is depicted in Figure 3.

There are two possible transition targets for each of the three speakers to another speaker, resulting in a total of six possible combinations. Overall, there are 46 events. We have 40 normal switches between speakers (six or seven for every of the possible six combinations) and six additional cases where two speakers are talking simultaneously and one of the two stops earlier than the other. In the latter case we want to PTU to orient towards the remaining speaker. If we take into account that the PTU should always orient towards the closest source given its previous position, it should react to 44 events.

4.3 Methodology

We measure the performance of the system by determining its responsiveness, accuracy, and consistency. Responsiveness is determined by measuring detection and reaction times. The *detection time* indicates the mean time it took the localization system to detect the direction of a source after its onset with a given tolerance level. The *reaction time* consists of the mean time the system required to orient itself towards the currently active speaker. Responsiveness was tested with three different tolerance levels, 5°, 10°, and 15°. Even the highest tolerance level was sufficient to capture the face of the active speaker with the web cam. For each tolerance level we performed a different trial. To determine the accuracy, we look at the angular precision in azimuth by measuring the absolute mean difference between the active source and the orientation of the PTU. With consistency we imply that the system should stay on its target and the number of

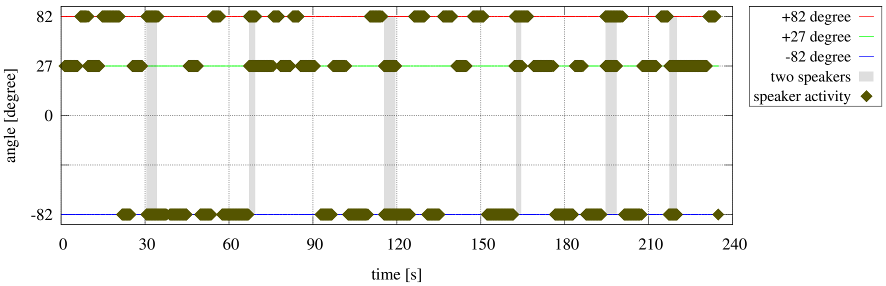

Fig. 3. Activity of speakers at different angles over time in our data set. The dark yellow dots indicate speaker activity at $+82°$, $+27°$, and $-82°$ respectively, times where two speakers are active simultaneously have a gray background.

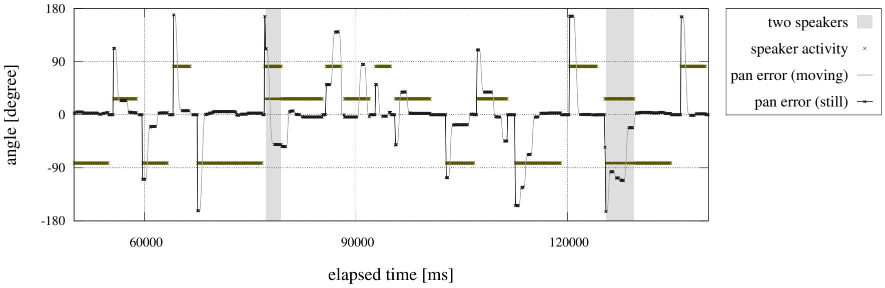

Fig. 4. Error in orientation towards the currently active speaker over time for a part of the conversation. Every time a new speaker starts talking the error is high until the PTU turns towards it. When face detection is used to resolve ambiguities, an additional delay can be noticed.

erratic moves of the PTU should be as low as possible. In part, this can be deduced from the mean error, but we also look at the total number of PTU moves during a trial.

With these criteria in mind, we thoroughly tested the individual components of the system and tried to assess a parameter set that leads to a good overall performance of the system. Some of the parameters used here already proved to be reliable in a series of preliminary tests that will not be shown here. Given this parameter set, we tried to optimize the interaction between the onset detector, the time constant of the source detector, and reduce the influence of movements as a source of noise on the results. For the onset detector, we varied the slope threshold to find a value that would prove to be neither too sensitive nor too insensitive. At the source detection stage we changed the integration constant and tried to find a setting that provides a good balance between responsiveness, accuracy, and consistency. Overall, we tried to showcase the performance of VICTOR while optimizing some of its critical parameters.

4.4 Results

Preliminary experiments A series of preliminary tests were done to find a parameter set for our system that yields the best performance according to the previously described criteria. These settings were used for the trials that are described in the next sections. In addition, we were able to verify that face detection is able to improve responsiveness of the system by reducing the number of moves that are required to solve ambiguities. We omit more detailed results for space reasons and concentrate on the calibration of our onset and source detector models that were developed for VICTOR.

Movements & Onsets We evaluated the responsiveness and accuracy of the system with the PTU facing an azimuth of $0°$ to assess the performance in the absence of movements (see Table 1). The decay constant of the time integration was set to 0.1. We performed these tests with and without onset detection to measure its influence on responsiveness. With onset detection (slope threshold was set to 0.0002) the detection times were about 300 ms for the $15°$ and $10°$ tolerance levels, while it slightly decreased for $5°$. Also note that the responsiveness could only be measured for sources that were actually detected. Mean error was at about $13°$ azimuth in each case. The error was

Table 1. Performance without movements, with enabled and disabled onset detector.

Condition	Tolerance	Detection time	Mean error	Missed sources
No onsets	15° / 10° / 5°	1.38 s / 1.59 s / 3.00 s	22.25° / 23.27° / 20.58°	6 / 8 / 33
Onsets	15° / 10° / 5°	0.28 s / 0.31 s / 0.70 s	13.12° / 13.53° / 13.24°	0 / 0 / 3

Table 2. Performance with a moving PTU with enabled and disabled onset detector. The *wait* condition told the system to stop localization during a movement.

Condition	Tolerance	Detection time	Reaction time	Mean error	# moves	Missed sources
No onsets, continue	15°	1.15 s	1.64 s	42.03°	171	9
	10°	1.05 s	1.58 s	38.46°	186	9
	5°	1.43 s	2.17 s	41.41°	175	12
No onsets, wait	15°	1.43 s	2.15 s	44.27°	79	10
	10°	1.49 s	2.13 s	47.1°	87	6
	5°	1.92 s	2.56 s	43.77°	76	14
Onsets, wait	15°	1.02 s	1.77 s	35.79°	132	5
	10°	1.24 s	1.87 s	44.76°	130	1
	5°	1.42 s	2 s	44.61°	149	9

relatively high because two of the three speakers were situated at lateral positions, where the ITD resolution is lowest (see Section 2.1). Also, situations with two active sources have a negative influence on both the mean error and response times, especially if one source dominates the other source in terms of the localization cues. These situations produced outliers that significantly reduced the measured responsiveness and accuracy. Without onset detection the system had low detection times, which also had a negative impact on the mean error. This shows that onset detection does indeed improve the overall performance.

Table 2 summarizes results with movements. Best performance was achieved with an activated onset detector and by interrupting localization during movements. Mean errors were relatively high because at the onset of a new source the error in azimuth stays high until the PTU has oriented itself towards the source (see peaks in Figure 4). The difference between detection and reaction times lies in the time it takes the PTU to move towards a source. On average, it took the PTU about $550\ \mathrm{ms}$ to execute a single movement. For some of the source transitions at least two moves were required to solve ambiguous cues in the absence of visual cues. Even after a correct orientation move, the PTU did small corrections to its position, which added up to the reaction time, especially for the lowest tolerance level. For tolerance levels of $10°$ and $15°$ responsiveness was very similar. Trials with continuous localization and enabled onset detector were omitted because the system tended to detect onsets during a movement, which significantly decreased the overall performance. During a rotation of the PTU, the microphones are exposed to correlated vibrations and noise caused by the servos of the unit. This provoked erratic movements of the PTU when the onset detector was

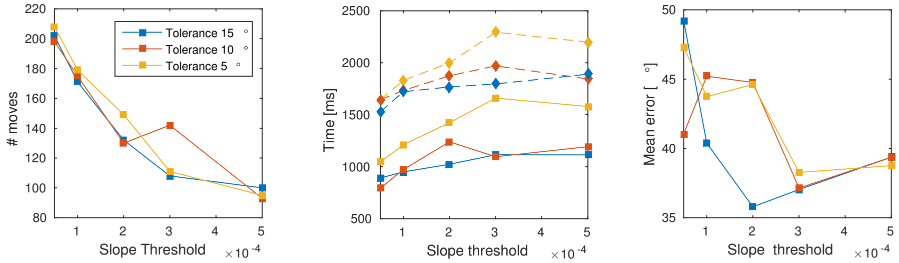

Fig. 5. Influence of onset threshold on #moves, detection (solid / squares) and reaction time (dotted / diamonds), and the mean error in orientation.

activated, mostly to $+90°$ or $-90°$. These also caused some of the observable trial to trial variations with enabled onset detector. This can for example be noticed by different mean errors for the same parameters (the tolerance level has no influence on mean error and moves).

Overall, onset detection improved the responsiveness of the system but also made it more sensitive to noise produced by the movements of the PTU.

Slope Threshold for Onset Detection The next set of trials was intended to find a slope threshold that yields a high reactivity without becoming too sensitive to noise. Figure 5 shows the influence of different slope thresholds on the responsiveness, number of moves, and the mean error. As expected, both detection and reaction times improved when the threshold was lowered. However, the number of moves, and thus the number of errors, did also increase because the device became much more sensitive to (correlated) noise. This can also be seen in the increased mean error. However, also note that the variability between trials increased, which caused the large differences in mean error for the three different tolerance levels. If we reduced the onset threshold, the behavior of the PTU became more stable, the variability between runs was lower, and the responsiveness decreased and slowly converged towards the situation without onset detection because only very strong onsets could still be detected during the discussion. We select 0.0002 as the default slope threshold as it improves responsiveness and still provides a reasonably stable turning behavior.

Decay Constant Based on the previously determined parameter set, we varied the decay constant that is used to integrate source estimates over time (decay was set to 0.01, 0.05, 0.1, 0.2, 0.3, and 0.5). A large decay constant reduces the amount of time integration and vice versa. The decay constant is only important for time frames that did not include an onset (which is the vast majority of time frames) as onsets are assigned a higher weight.

Results are presented in Figure 6. By increasing the decay constant, the responsiveness significantly increased, but at the same time the behavior of the PTU became less stable, which can be seen in an increase of moves and a larger mean error (especially for a decay of 0.5). Finally, we opted for a decay constant of 0.1, resulting in a time

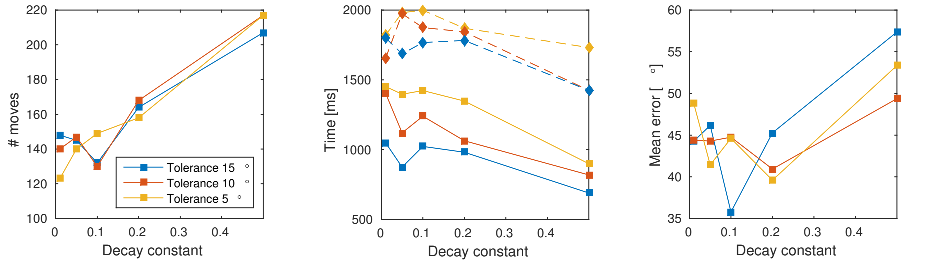

Fig. 6. Influence of the integration time on #moves, detection (solid / squares) and reaction time (dotted / diamonds), and the mean error in orientation.

constant of about 460 ms with the given time frame length. It was the highest decay that yielded a relatively low number of moves. Responsiveness was sufficient to react to most of the sentences of the speakers, even in the case of two simultaneous sources. By further reducing the decay source estimates would have an even longer impact on results, but it would also take more time for new sources to be noticed in the absence of a strong onset.

5 Discussion

We presented our work on a video conference tool robot called VICTOR. It is based on biologically inspired sound source localization as its primary input cue and additional face detection to solve ambiguities. A source detector is used to emphasize sources that are consistently detected over longer periods of time. Additionally, an onset detector was implemented to improve responsiveness for new sources that show a significant rise in the energy envelope. The overall aim was to find a parameter set for VICTOR that allows for a responsive, accurate, and stable tracking of speakers in a remote discussion. We tested different sets of parameters, aiming for configurations with minimal response time, minimal angular error, and as an additional criterion minimal number of moves. We could show that both onset and source detectors work as intended and improve the overall performance of the system. Despite variability caused by noise and vibrations, we were able to optimize the parameter set for the onset detection and time integration to yield consistently good results in our target scenario.

Problems with correlated noise and vibrations that the system currently suffers from could be overcome with a band pass filter in a future version. Robustness to noise could also be further increased by making sure that the correlation at the maximum slope is above a certain threshold level to improve onset detection. The current system has no attention guidance that could, for example, focus on a single source while there are two active speakers. Also, it has no memory regarding the positions of the previously detected faces. Such an attention model could further improve responsiveness of the system as it is able to directly relate a new source to a previously detected speaker position. Adding an artificial head to resolve the front/back confusion with the help of spectral cues would reduce the number of required movements. Instead of only face detection we could also use face recognition (e.g. [2]) to display who is currently talking. Additional speech recognition could be used to create a transcript of the conversation that is being followed. We will address these challenges in future iterations of VICTOR.

References

1. Adalgeirsson, S.O., Breazeal, C.: MeBot: A Robotic Platform for Socially Embodied Presence. In: Proc. 5th Int'l Conf on HRI (HRI'10). pp. 15–22. IEEE Press (2010)
2. Belle, V., Deselaers, T., Schiffer, S.: Randomized trees for real-time one-step face detection and recognition. In: Proc. Int'l Conf. on Pattern Recognition (ICPR'08). pp. 1–4. IEEE Computer Society (December 8-11 2008)
3. Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press (1997)
4. Dietz, M., Klein-Henning, M., Hohmann, V.: The influence of pause, attack, and decay duration of the ongoing envelope on sound lateralization. Journal of the Acoustical Society of America 137(2), EL137–EL143 (2015)
5. Dietz, M., Marquardt, T., Stange, A., Pecka, M., Grothe, B., McAlpine, D.: Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds ii: Single neuron recordings. Journal of Neurophysiology 111(10), 1973–1985 (2014)
6. Elhilali, M., Xiang, J., Shamma, S.A., Simon, J.Z.: Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biology 7(6), e1000129 (2009)
7. Faller, C., Merimaa, J.: Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. J Acoust Soc Am 116(5), 3075–3089 (nov 2004)
8. Goeckel, T., Lakemeyer, G., Wagner, H.: Echo suppression for sound localization with a model of the precendence effect. Tech. rep., Biology II, RWTH Aachen University (2014)
9. Jones, M., Viola, P.: Face recognition using boosted local features. In: Proc. ICCV (2003)
10. Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology 15, 1943–1947 (2005)
11. Kristoffersson, A., Coradeschi, S., Loutfi, A.: A review of mobile robotic telepresence. Adv. in Hum.-Comp. Int. 2013, 3:3–3:3 (Jan 2013)
12. Litovsky, R.Y., Colburn, H.S., Yost, W.A., Guzman, S.J.: The precedence effect. J Acoust Soc Am 106(4), 1633–1654 (Oct 1999)
13. May, T., van de Par, S., Kohlrausch, A.: A probabilistic model for robust localization based on a binaural auditory front-end. IEEE Trans. Audio, Speech, Language Process. (1) (2011)
14. Sanchez-Riera, J., Alameda-Pineda, X., Wienke, J., Deleforge, A., Arias, S., Cech, J., Wrede, S., Horaud, Radu, P.: Online Multimodal Speaker Detection for Humanoid Robots. In: Proc. Int'l Conf. on Humanoid Robotics (Humanoids 2012). pp. 126–133. IEEE (Dec 2012)
15. Sangwan, A., Chiranth, M., Jamadagni, H., Sah, R., Prasad, R., Gaurav, V.: Vad techniques for real-time speech transmission on the internet. pp. 46–50 (2002)
16. Schiffer, S.: cAPTUre: a configurable audio pan-tilt unit for repeatable experimentation. Tech. rep., Knowledge-based Systems Group, RWTH Aachen University (2012)
17. Supper, B., Brookes, T., Rumsey, F.: An auditory onset detection algorithm for improved automatic source localization. IEEE Trans. Audio, Speech Language Process. 14(3), 1008–1016 (2006)
18. Zhang, C., Yin, P., Rui, Y., Cutler, R., Viola, P., Sun, X., Pinto, N., Zhang, Z.: Boosting-based multimodal speaker detection for distributed meeting videos. IEEE Transactions on Multimedia 10(8), 1541–1552 (2008)