Improving Robustness against Environmental Sounds for Directing Attention of Social Robots

Nicolai B. Thomsen, Zheng-Hua Tan, Børge Lindberg, and Søren Holdt Jensen

Dept. Electronic Systems, Aalborg University, Fredrik Bajers vej 7, 9220 Aalborg Ø, Denmark

Abstract. This paper presents a multi-modal system for finding out where to direct the attention of a social robot in a dialog scenario, which is robust against environmental sounds (door slamming, phone ringing etc.) and short speech segments. The method is based on combining voice activity detection (VAD) and sound source localization (SSL) and furthermore apply post-processing to SSL to filter out short sounds. The system is tested against a baseline system in four different real-world experiments, where different sounds are used as interfering sounds. The results are promising and show a clear improvement.

1 Introduction

In the past decade much research has been conducted in the field of human-robot interaction (HRI) [1, 2, 3] and especially social robots [4], which are to operate and communicate with persons in different and changing environments, have gained much attention. Many different scenarios arise in this context, however in this work we consider the case where a robot takes part in a dialog with multiple speakers. The key task for a social robot is then to figure out when someone is speaking, where the person is located and whether or not to direct its attention toward the person by turning. In uncontrolled environments like living rooms, offices etc. many different spurious non-speech sounds can occur (door slamming, phone ringing, keyboard sounds etc), making it important for the robot to distuingish between sounds to ignore and sounds coming from persons demanding the attention of the robot. Unlike humans, robots are often not able to quickly classify an acoustic source as human or non-human using vision due to limited field-of-view and limited turning speed. If this ability is missing the behaviour of the robot may seem unnatural from a perceptional point of view, which is undesirable.

In [1], an anchoring system is proposed, which utilizes microphone array, pan-tilt camera and a laser range finder to locate persons. The system is able to direct attention to a speaker and maintain it, however non-speech interfering sounds are not considered and the system is only evaluated for persons talking for approximately 10s. The work in [5] introduces a term called *audio proto objects*, where sounds are segmented based on energy and grouped by various features to filter out non-speech sounds. Good results are reported for localization, however no reults are reported for an actual real-world dialog including interfering non-speech sounds.

In this work we focus on a the sound source localization (SSL) part of the system and use standard method for face detection. We specifically propose a system where a voice activity detector (VAD) and SSL are used to award points to angular intervals spanning $[-90°, 90°]$. These points are accumulated over time, enabling the robot only to react to persistent speech sources.

The outline of the paper is as follows: the baseline system will be described in Sect. 2 followed by a description of the proposed system in Sect. 3. Section 4 states results for both a test of the localization system and test of the complete system in different real-world scenarios. Section 5 concludes on the work and discuss how to proceed.

2 Baseline System

We developed a baseline system which is shown in Fig. 1. It is inherently sequential, where first SSL is used to determine the direction of an acoustic source (if any), and then after having turned face detection is used to verify the source and then possibly adjust further. Face detection is done according to [6] and is implemented using OpenCV.

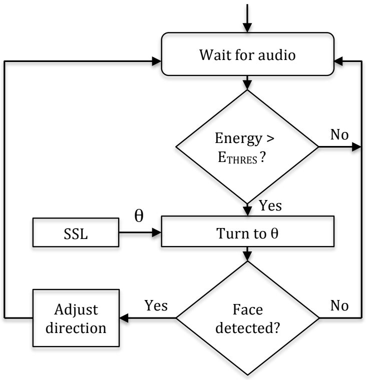


Fig. 1. Flowchart of baseline system.

2.1 Sound Source Localization

For sound source localization (SSL) we use the steered response power method with phase transformation (SRP-PHAT) [7]. It is a generalization of the well-known generalized cross-correlation method with phase transform (GCC-PHAT) [8], when more than one microphone pair is utilized. Furthermore it takes advantage of the whole cross-spectrum and not only the peak value. The basic idea is to form a grid of points (most commonly in spherical coordinates) relative to some point, which is typically the center of the microphone array, and then steer the microphone array toward each point in the grid using delay-sum beamforming and at last find the output power. After all points have been processed, the three-dimensional (azimuth, elevation and distance) power map can now be searched for the maximum value, indicating an acoustic source at that point. It is computationally heavy to consider all points assuming a fine grid of points, however in this work we are only interested in the direction, and not elevation, hence we can disregard this. Assuming that the source is located in the far-field, i.e. the microphone spacing is much smaller than the distance to the source, we can use only one value for distance.

3 Proposed System

Figure 2 shows the structure of the proposed system. It has the same overall sequential structure as the baseline where audio is first used to roughly estimate the direction of the person, and afterwards vision is used to verify the existence of a speaker and possibly adjust the direction further. The two differences between the baseline system and the proposed system are; first, the use of a better VAD to increase robustness against environmetal sounds, and second, post-processing of SSL to increase robustness against short speech segments and short sounds, which are misclassified by the VAD.

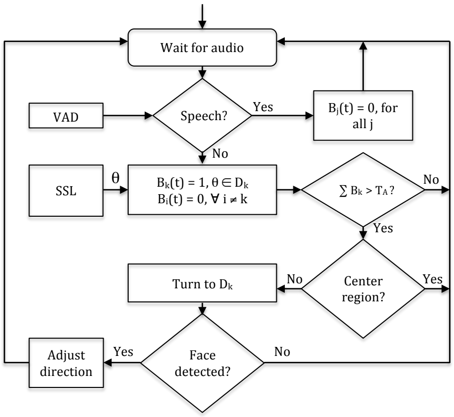


Fig. 2. Flowchart of proposed system. The post-processing using $B_i(t)$ is explained in Sect. 3.2.

3.1 Voice Activity Detection

In this work a variant of the voice activity detector (VAD) described in [9, 10] is utilized. Results show a good trade-off between accuracy and low complexiy, which is of high importance, because the robot has limited ressources and heavy processing tasks such as image processing and speech recognition (not included in this work) should run simultaneously. The algorithm is based on a posteriori SNR weighted energy difference and involves the follwoing step, which are performed on every audio frame.

1. Compute the a posteriori SNR weighted energy difference given by
$$D(t) = \sqrt{|E(t) - E(t-1)| \cdot \mathrm{SNR}_{\mathrm{post}}(t)} \ . \tag{1}$$
where $E(t)$ is the logarithmic energy of frame t and $\mathrm{SNR}_{\mathrm{post}}(t)$ is the a posteriori SNR of frame t.
2. Compute the threshold for selecting the frame given by
$$T = \overline{D(t)} \cdot f(\mathrm{SNR}_{\mathrm{post}}(t)) \cdot 0.1 \ . \tag{2}$$
where $\overline{D(t)}$ is an average of $D(t), D(t-1), ..., D(t-T)$, and $f(\mathrm{SNR}_{\mathrm{post}}(t))$ is piece-wise constant function, such that the threshold is higher for low SNR and lower for high SNR. If $D(t) > T$, then $S(t) = 1$ otherwise $S(t) = 0$.
3. Perform a prior moving average on $S(t)$ and compare to threshold, T_{VAD}. If above threshold, the frame is classified as speech and otherwise as non-speech.

It should be noted that the VAD is only performed on one of the four channels from the microphone array.

3.2 Post-Processing of SSL

The range of output angles, $[-90°, 90°]$, from SSL is divided into non-overlapping regions, e.g. the first region could be $D_1 = [-90°, -85°[$. This is motivated by the fact that even during short speech segments ($\sim$ 1s) the speaker is not standing completely still and likewise the head is also not completely fixed, thus SSL estimates which are very close should not be assigned to different sources, but are most likely to be caused by the same source. In this work we have split the range of angles into regions of 5° except for the center region which is defined as $[-5°, 5°[$, thus the total number of regions is 35. For each of the aforementioned regions we assign a vector $\boldsymbol{B}_i(t) = [B_i(t-T+1) \ \ B_i(t-T+2) \ ... \ B_i(t)]$, where t denotes the tth audio frame and T denotes the length of the vector in terms of audio frames. Whenever an audio frame is classified as speech by the VAD, SSL is used to estimate the angle of the supposed speaker relative to the robot. The current element of the vector corresponding to the region, in which the estimated angle belongs, is then set to 1 for the current frame, t, and all current elements of vectors for the other regions are set to 0. If the frame is classified as non-speech, then the current element of all vectors are set to 0. Attention is then given to region i if the sum of the corresponding vector is above some threshold, i.e. $\sum_{m=T-1}^{0} B_i(t-m) > T_{\mathrm{A}}$. If a vector exceeds the threshold thus making the robot turn, the vectors for all regions are set to zero. The motivation for this system is that it enables control over the duration of the sentences which should trigger the robot to turn toward a speaker.

4 Evaluation of the Systems

Two seperate test were performed. One test with the purpose of testing only the localization capabilities of both baseline and proposed system and that the robot was able to turn toward the sound source and adjust using vision, and a second test where the system was tested in four different types of scenarios with three speakers and interfering sounds.

4.1 Localization Performance

We test only the proposed system here, since for one speaker and no noise they are the same. The localization system was tested for five different angles by having a person speaking continuously at the angle at a distance of approximately 1.5m until the robot had turned toward the person. Here the angle between robot and person is defined as in Fig. 3, where positive angles are clockwise. The results are stated in Table 1. It is seen that the system is clearly able to turn toward the person with acceptable accuracy. It should be noted that this test is associated with some uncertainties, since it is very difficult to place the speaker at the exact angle, and it is difficult to measure the angle with high accuracy.

Table 1. Performance of localization system. Mean and standard deviation of angle between person and robot after localization and rotation. 10 repetitions were used for each angle.

	15°	30°	45°	60°	75°
μ	14.2	29.1	45.9	59.1	73.4
σ	0.9	1.3	1.3	0.9	2.7

4.2 Attention System Performance

The baseline and proposed system were tested through four different experiments, resulting in a total of eight trials. The four experiments are described below

1. The speakers take turn talking for approximately 10s.
2. The speakers take turn talking for approximately 10s and in between speakers interfering sounds are played (see Table 2).
3. The speakers take turn talking for either approximately 10s or 1s.
4. The speakers take turn talking for either approximately 10s or 1s and in between speakers interfering sounds are played (see Table 2).

In all four experiments a total of 20 time slots are used, where a slot can either be a speaker talking (10s or 1s) or an interfering sound, thus the slots are of varying length. We emphasize that there is no overlapping sounds. Information about the interfering sounds is listed in table 2. Each noise source is responsible for two different sounds, where sound 1 is always played as the first of the two. The test setup and the location of the robot, the noise sources and the speakers are shown in Fig. 3. All experiments were recorded using a seperate microphone and a seperate video camera and information about the direction of the robot was logged on the robot. This data was afterwards used to annotate precisely when different sounds occured, and the focus of attention of the robot was also annotated using this. The logged data from the robot was not used directly, as the absolute angle did not match reality due to small offsets in the base when turning, however it was used for determining the timeline precisely. We also emphasize that the annotation of a sound begins when the sound begins and is extended until the next sound begins, thus silence is not explicitly stated due to simplicity. Furthermore, the annotation of the robot starts when the robots has settled at a direction, thus turning is not stated explicitly. Figures 4-7 show the results for the four experiments for both baseline and proposed system, where "OOC" means out-of-category, "SP1" means speaker 1, "N1" means noise source 1 and so on. "Annotation" (light grey) shows who was active/speaking and "Robot" (black) shows where the attention of the robot was focused.

Table 2. Description of the six interfering sounds used in the experiments. Same ringtone used for both sound 1 and sound 2 from N3.

	Sound 1			Sound 2		
Source	Description	Duration	SPL (dB)	Description	Duration	SPL (dB)
N1	Coughing	≈ 0.7s	≈ 77dB	Door slamming	≈ 0.4s	≈ 90dB
N2	Scrambling chair	≈ 1.1s	≈ 80dB	Scrambling chair	≈ 1.1s	≈ 89dB
N3	Phone ringing	≈ 3.7s	≈ 75dB	Phone ringing	≈ 3.7s	≈ 75dB

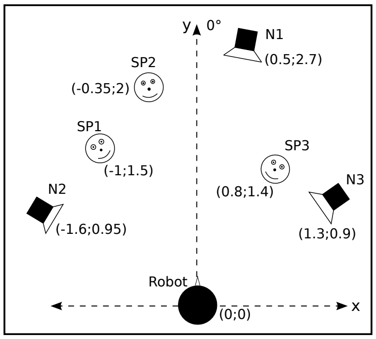
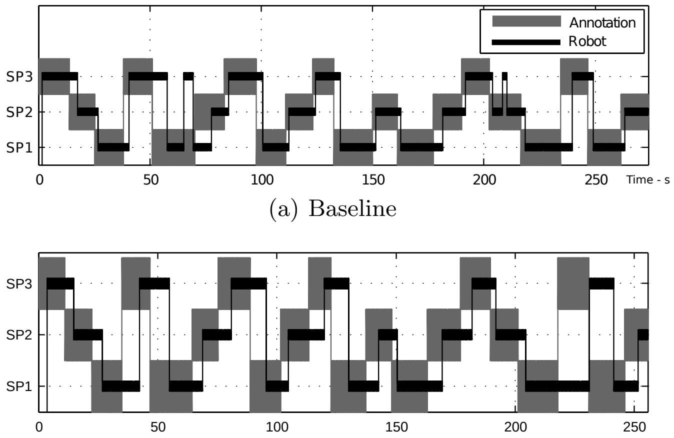


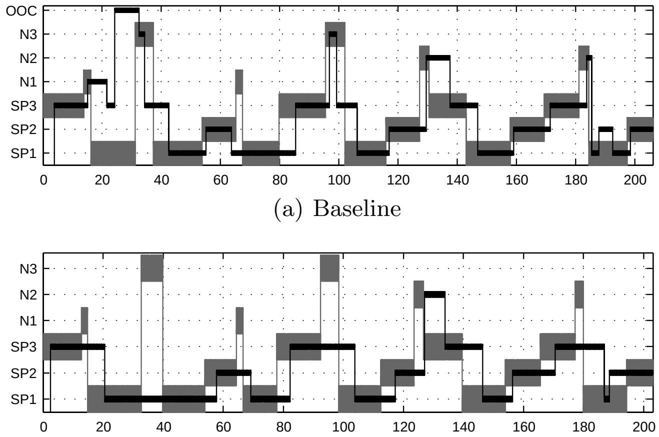
Fig. 3. Setup for attention experiment. XY-coordinates are given in metres.



(a) Baseline

(b) Proposed

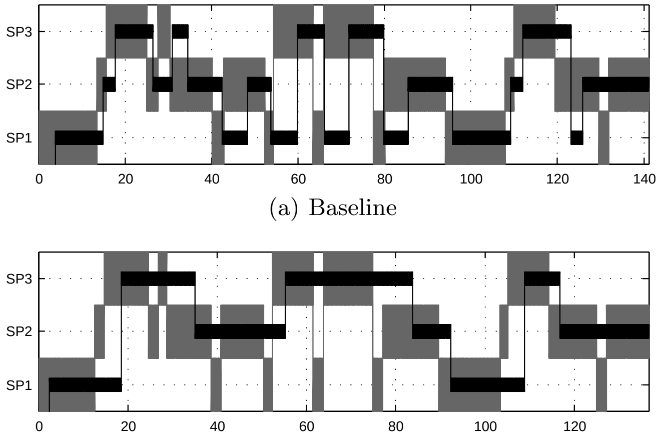
Fig. 4. Experiment 1. Figure 4(a) shows the baseline and Fig. 4(b) shows the performance of the proposed method. The two anomalous behaviours for the baseline are assumed to be caused by sounds, not related to the experiment, created from the direction of SP3. The much delayed transition in the proposed system in the end is caused by not triggering the VAD properly.



(a) Baseline

(b) Proposed

Fig. 5. Experiment 2. Legends and axis similar to Fig. 4. It is seen that for the baseline the robot turns toward SP3 after N3, which is due to detecting the face of SP3. A similar thing happens for both the baseline and proposed system in the second last time slot. We also note that the VAD used in the proposed method is triggered by the "sound 1" from N2 at $\sim$ 125s, which is unexpected, however this could most likely be avoided using pitch information too.



(a) Baseline

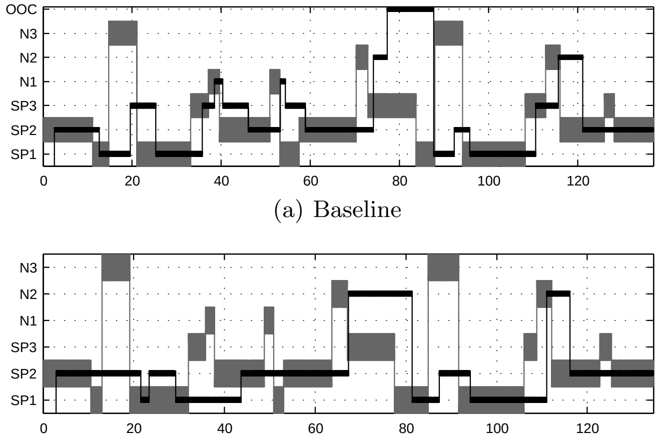
(b) Proposed

Fig. 6. Experiment 3. Legends and axis similar to Fig. 4. The anomalous behaviour for the baseline in the third last slot is caused by detecting the face of SP1.

Table 3 states the number of correct and incorrect transitions along with number of anomalous behaviours. A correct transition is when the robot turns attention to a person speaking for approximately 10s or ignores a short speech segment (approximately 1s) or an interfering sound. An example of the first case is seen in Fig. 5(b) at the start, where the robot turns toward SP3. An example of the second is seen in the same figure at slot 1 to 2, where the robot does not shift attention due to an interfering sound from noise source N1. An incorrect transition is when the robot turns toward a noise source, a person speaking for approximately 1s or out-of-category. The number of correct and incorrect transitions should add to 20. An anomalous behaviour is when the robot makes an unexpected turn during a slot. An example is seen in Fig. 5(b) in slot 19, where the robot turns toward SP2 while SP3 is speaking. We see in Table 3 that for the first experiment both systems perform equally well, which is too be expected. But as both short sentences and interfering sounds are added to the

Table 3. Number of correct and incorrect transitions and anomalous behaviours for the baseline and the proposed system for each experiment.

	Baseline			Proposed		
Experiment	#Correct	#Incorrect	#Anomalies	#Correct	#Incorrect	#Anomalies
1	20	0	2	19	1	0
2	13	7	3	18	2	1
3	12	8	1	20	0	0
4	8	12	0	14	6	3



(a) Baseline

(b) Proposed

Fig. 7. Experiment 4. Legends and axis similar to Fig. 4. In the beginning of the baseline, the robot turns toward SP3 instead of N3. This happens because N3 is located at an angle of $\sim +90^\circ$ relative to SP1, and since the SSL has lower resolution for large angles, the sound is perceived as coming from a smaller angle. It is seen that both systems behaves unexpectedly at $t \sim 75$s. This is caused by the fact, that SSL only covers $[-90^\circ, 90^\circ]$. Again, the VAD in the proposed system is triggered by the the sounds from N2, which is undesirable.

experiment, the proposed method generally performs better than the baseline. The relatively low number of correct transitions for both the baseline and the proposed method in experiment 4 is caused by being adressed by a speaker from a relative angle greater than $|90|^\circ$, which is a general limitation of the SSL algorithm used in both systems.

5 Conclusion

In this work we have presented a method for increasing robustness against environmental sounds and short speech segments for sound source localization in the context of a social robot. Different experiments have been conducted and they show an improvement over a baseline system. The method proposed is however based on a constant, T_A, set before deployment of the robot, which is not ideal. Future work should look into how this parameter can be learned during runtime. Furthermore, using a VAD designed for distant speech would improve the system.

References

[1] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, “Providing the basis for human-robot-interaction: A multi-modal at-

tention system for a mobile robot," in *in Proc. Int. Conf. on Multimodal Interfaces*. ACM, 2003, pp. 28–35.

[2] K.-T. Song, J.-S. Hu, C.-Y. Tsai, C.-M. Chou, C.-C. Cheng, W.-H. Liu, and C.-H. Yang, "Speaker attention system for mobile robots using microphone array and face tracking," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, May 2006, pp. 3624–3629.

[3] R. Stiefelhagen, H. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, "Enabling multimodal human robot interaction for the karlsruhe humanoid robot," *Robotics, IEEE Transactions on*, vol. 23, no. 5, pp. 840–851, Oct 2007.

[4] M. Malfaz, A. Castro-Gonzalez, R. Barber, and M. Salichs, "A biologically inspired architecture for an autonomous and social robot," *Autonomous Mental Development, IEEE Transactions on*, vol. 3, no. 3, pp. 232–246, Sept 2011.

[5] T. Rodemann, F. Joublin, and C. Goerick, "Audio proto objects for improved sound localization," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, Oct 2009, pp. 187–192.

[6] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[7] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2510–2526, Nov 2007.

[8] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[9] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 798–807, Oct 2010.

[10] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Ma, S. Cumani, O. Glembek, H. Hermansky, S. Mallidi, N. Mesgarani, R. Schwartz, M. Soufifar, Z. Tan, S. Thomas, B. Zhang, and X. Zhou, "Developing a speaker identification system for the darpa rats project," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 6768–6772.