ROBUST IMPULSIVE SOUND SOURCE LOCALIZATION BY MEANS OF AN ENERGY DETECTOR FOR TEMPORAL ALIGNMENT AND PRE-CLASSIFICATION

Machmer, T., Swerdlow, A., Kroschel, K.

University of Karlsruhe (TH)
Institute for Anthropomatics
76128 Karlsruhe, Germany
{machmer, swerdlow}@ies.uka.de
kristian.kroschel@iitb.fraunhofer.de

Moragues, J., Vergara, L., Gosálbez, J.

Polytechnic University of Valencia (UPV)
Communication Department
46022 Valencia, Spain
jormoes@upvnet.upv.es
{lvergara, jorgocas}@dcom.upv.es

ABSTRACT

In this paper, we present a novel approach for detection and localization of both impulsive and non-impulsive sound sources. At first, theoretical basics of the used algorithms are presented. Subsequently, we describe a standard SRP-PHAT based localization method and discuss occurring complications, especially for impulsive sound sources. Therefore, a modified approach is presented as a solution. It distinguishes between impulsive and non impulsive sound sources, and additionally aligns the detection window to the event. The pre-classification and alignment are done with the help of an energy detector.

1. INTRODUCTION

For a complete acoustic scene analysis, especially for surveillance applications or interaction with a humanoid robot [1], it is necessary to localize and detect all types of sound events which can happen in the proximity. Basically, two types of sound sources can be differentiated: impulsive and non-impulsive. In many cases only the non-impulsive ones, mainly speech, are taken into account. But especially for the detection of dangerous or unusual situations, it is often necessary to localize and detect also impulsive sound sources like slamming doors or breaking glass.

The detection problem is directly related to the knowledge of the signal we are interested in and the background noise characteristics. The easiest case would be to detect known events in a stationary white Gaussian background noise environment, but when the sound sources are not completely known, the design of the appropriate detector is more difficult [3, 5]. In this case, energy detection can be useful to collect more information about the actual event and improve the localization step.

While the localization of non-impulsive sound sources following the approach in [6] showed very stable results, impulsive sound sources were not localized reliably. In order to be able to handle both types of events, a modification of a standard SRP-PHAT localization method is required. Therefore, a novel approach using an energy detector for a temporal event alignment and a pre-classification is proposed.

This paper is organized as follows. Section 2 presents the principles of the Gaussian energy detector and in Section 3 the general idea of the localization algorithm is described. In Section 4 the experimental setup is presented. The modified localization method is proposed in Section 5. In Sections 6 and 7 achieved results are shown and a conclusion of our work is given.

2. ENERGY DETECTOR

The simplest detection problem is to decide whether a signal is embedded in noise or if only noise is present. One common method for detection of unknown signals is energy detection, which measures the energy in the received waveform over a specified observation time.

The energy detector is an optimum solution when the noise $\mathbf{w}$ and the signal $\mathbf{s}$ are considered zero-mean Gaussian random vectors with uncorrelated components, $\mathbf{w} : \mathcal{N}(0, \sigma_{\mathbf{w}}^2 \mathbf{I})$ and $\mathbf{s} : \mathcal{N}(0, \sigma_{\mathbf{s}}^2 \mathbf{I})$, where $\sigma_{\mathbf{w}}^2$ and $\sigma_{\mathbf{s}}^2$ are the noise and the signal variances. The optimum test under this conditions is [3, 5]:

$$\frac{\mathbf{y}^T \mathbf{y}}{\sigma_{\mathbf{w}}^2} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \tag{1}$$

where $\mathbf{y}$ is the observation vector, hypothesis H_1 corresponds to $\mathbf{y} = \mathbf{s} + \mathbf{w}$, and H_0 corresponds to $\mathbf{y} = \mathbf{w}$. The statistic $\frac{\mathbf{y}^T \mathbf{y}}{\sigma_{\mathbf{w}}^2}$ is chi-squared distributed with N degrees of freedom (χ_N^2) and λ can be set for a specific probability of false alarm (PFA).

The test (1) assumes that the components of $\mathbf{w}$ are independent and identically-distributed. However, real audio signals do not have white noise properties, as adjacent audio samples are highly correlated. In this case, some additional preprocessing is required to increase the detection performance significantly.

In this work, the background noise is assumed to be Gaussian and additive with zero-mean. In so doing, statistical independence and uncorrelation are equivalent, hence simple prewhitening is sufficient and the original observation vector $\mathbf{y}$ is transformed into a prewhitened observation vector $\mathbf{y}_p$ by means of

$$\mathbf{y}_p = \mathbf{R}_{\mathbf{w}}^{-1/2} \mathbf{y}, \tag{2}$$

where $\mathbf{R}_{\mathbf{w}} = E\big[\mathbf{w}\mathbf{w}^T\big]$ is the noise covariance matrix, which can be estimated from a training set of noise vectors $\mathbf{w}_k$, with $k = 1, \ldots, K$, using the sample estimate

$$\hat{\mathbf{R}}_{\mathbf{w}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{w}_k \mathbf{w}_k^T. \quad (3)$$

The test (1) can be rewritten as

$$\frac{\mathbf{y}_p^T \mathbf{y}_p}{\sigma^2_{\mathbf{w}_p}} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda. \quad (4)$$

Note that $\mathbf{R}_{\mathbf{w}} = E\big[\mathbf{w}_p \mathbf{w}_p^T\big] = \mathbf{I}$ and hence $\sigma^2_{\mathbf{w}_p} = 1$. The prewhitening transformation whitens and mean-power calculation normalizes the original observation noise.

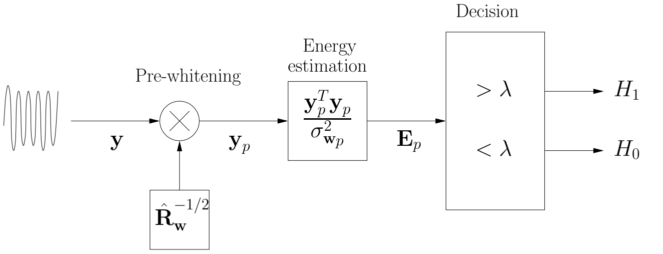

Figure 1: Block diagram of an energy detector.

In Figure 1 the complete energy detector procedure is depicted. The acoustic signal is divided into frames $\mathbf{y}$ of size N and then these observed vectors are linearly transformed ($\hat{\mathbf{R}}_{\mathbf{w}}$), so that a new white vector $\mathbf{y}_p$ is obtained. After that, the energy of the prewhitened data is calculated ($\mathbf{E}_p$) and compared with a threshold fixed by the PFA. The output of the energy detector will be 1 for H_1 and 0 for H_0.

3. LOCALIZATION

3.1 Time delay estimation

Today, the most commonly used methods for the acoustic localization are based on the estimation of the time difference of arrival (TDOA) of sound signals in a pair of spatially separated microphones. The time delay estimation is achieved by correlating the sound signals of two microphones in a microphone pair. The correlation function $R_{x_i x_j}(\tau)$ in frequency domain can be defined as

$$R_{x_i x_j}(\tau) = \int_{-\infty}^{+\infty} X_i(\omega) X_j(\omega)^* e^{j\omega\tau}\, d\omega, \quad (5)$$

where X_i is the Fourier transform of the given microphone signal x_i. In theory, the signals x_i and x_j in a given pair should be an exact copy of each other, with a time delay τ. However, in real environments we have to deal with noise and reverberation effects. This leads to the following system model:

$$x_i(t) = h_i(t) * s_i(t) + n_i(t) \quad (6)$$
$$x_j(t) = h_j(t) * s_j(t) + n_j(t), \quad (7)$$

where $h_i(t)$ is the acoustic impulse response of the room from the source to the i^{th} microphone, the additive term $n_i(t)$ summarizes the channel noise in the microphone system as well as the environmental noise for the i^{th} sensor, and s_j represents the delayed signal s_i, which is delayed by τ_{ij}. Having a closer look to the noise terms, they are represented by two different types: on the one hand, those which are correlated to each other, like the background noise of a running fan, on the other hand, those which are not correlated. If we assume that the noise is fully correlated and we have an ideal room with a Dirac impulse response, we can easily achieve a noise free estimation of the correlation in a given microphone pair by subtracting the correlation of the noise from the correlation of the received signal, analogously to the background noise suppression, presented in [6]:

$$R_{s_i s_j}(\tau) = R_{x_i x_j}(\tau) - R_{n_i n_j}(\tau) \quad (8)$$

In order to make the correlation more stable, we additionally use the so called Phase Transform (PHAT) $\psi_{x_i x_j}$ to weight the correlation function. This way of proceeding leads to the well known Generalized Cross Correlation (GCC) function [4]:

$$R^{(g)}_{x_i x_j}(\tau) = \int_{-\infty}^{+\infty} \psi^{PHAT}_{x_i x_j}(\omega) X_i(\omega) X_j(\omega)^* e^{j\omega\tau}\, d\omega, \quad (9)$$

with PHAT weighting function defined by

$$\psi^{PHAT}_{x_i x_j}(\omega) = \frac{1}{|X_i(\omega) X_j(\omega)^*|}, \quad (10)$$

which can also be regarded as a whitening filter. The combination of the GCC and (8) leads to:

$$R^{(g)}_{s_i s_j}(\tau) = R^{(g)}_{x_i x_j}(\tau) - R^{(g)}_{n_i n_j}(\tau). \quad (11)$$

The estimation of $R^{(g)}_{n_i n_j}(\tau)$ for each microphone pair is achieved during phases of no activity of the sound source, which are detected by using the energy detector, both described in detail in [6].

3.2 SRP-PHAT

Based on the time delay estimation, the spatial position of a sound source can be calculated. Therefore, the so called Power Field (PF) technique, also known as SRP (Steered Response Power), can be used. In this approach, beamforming is used to focus a microphone array to a specific spatial area. In order to find the exact position of a sound source, the entire environment is scanned searching for the spatial position with the highest acoustic power.

The combination of SRP and the TDOA based methods mentioned before leads to a called SRP-PHAT [2], which fuses the stability of the SRP against reverberation and the efficiency of the GCC method giving us the possibility to build a real-time system.

SRP-PHAT is computed as

$$P(\mathbf{s}) = \frac{1}{|M_p|} \sum_{(i,j) \in M_p} R^{(g)}_{s_i s_j}(\tau_{ij}(\mathbf{s})), \quad (12)$$

where $\tau_{ij}(\mathbf{s})$ denotes the theoretical delay between the microphones in pair (i, j) for the assumed spatial source

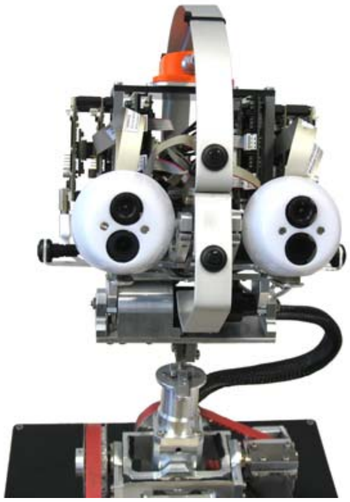

Figure 2: Head of the humanoid robot ARMAR III.

position $\mathbf{s} = (s_x, s_y, s_z)$. M_p represents a given set of microphone pairs. In order to estimate the source position $\hat{\mathbf{s}}$, the position of the maximal value in $P(\mathbf{s})$ has to be found in a given search space $\mathbf{S}$:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}\in\mathbf{S}} P(\mathbf{s}). \tag{13}$$

4. EXPERIMENTAL SETUP

In order to evaluate the localization of impulsive and non-impulsive sound sources, recordings were done with and without background noise. Different signal-to-noise ratios are of particular interest because of various noise sources, which can exist for example in the proximity of a humanoid robot. In our application, such a typical case is represented by the cooling fans of the robot.

For the evaluation, impulsive sound sources like putting a cup on the table, opening and closing a door, dropping a spoon on a table, and a toaster were analyzed. A mixer and human speech were used as non-impulsive sound sources.

The microphone array used for our experiments was built according to the head geometry of a humanoid robot (Figure 2) and consisted of four omni-directional electret condenser microphones. It is roughly an inverse t-shape geometry with a total width of 20 cm and a height of 5.5 cm. Sound data were acquired by using a multichannel audio data acquisition unit with the sampling frequency of 48 kHz. The window size used for the Gaussian energy detector was about 5 ms (256 samples); the amount of noise vectors required to estimate the whitening matrix was 1024 with a re-estimation period of 2 seconds. The source position was estimated by means of a 3D-grid search with grids of 5 cm and a total grid dimension of 3.60 m x 1.80 m x 1.20 m.

Furthermore, it was necessary to define a correct localization of the sound source position. Due to the small concentrated array used, it was not possible to determine the distance to the sound source, and only the azimuth and elevation angles were taken into account. The localization was deemed correct, if the Euclidean distance between the estimated and the real angle was below 10 degrees.

5. MODIFIED LOCALIZATION METHOD

For the impulsive sound sources, SRP-PHAT does not reach the high accuracy that is obtained with non-impulsive events. This is the reason why even in scenarios without any background noise, the mislocalization rate is very high [6]. In order to be able to localize both impulsive and non-impulsive sound sources, we modified the standard SRP-PHAT technique, described in Section 3. The basic idea thereby is to distinguish between the different types of sound sources and to adapt the localization algorithm accordingly.

5.1 Pre-classification

In the first step, the pre-classification phase, the sound source which should be localized is classified as an impulsive or non-impulsive event. This is done by measuring the length of the event counting detections of the energy detector in a specific time interval. In our case, this interval has a length of 256 detector windows and corresponds to approximately 1.37 seconds. An event is handled as an impulsive event if the totalized time duration of all detections in the time interval amounts less than one second.

5.2 Temporal event alignment

Mislocalizations of impulsive events can be mainly reasoned by reverberation. Because of the fact that all reflected possible paths of the sound are longer than the direct path, the first wave which arrives to the microphone pair, is not influenced by reverberation. In order to benefit from this knowledge, it is necessary to align the correlation window exactly to the event. This is done by positioning the beginning of an event in the middle of the correlation window. We do this by using the detections of the energy detector which uses smaller windows of only 256 samples. This alignment is done for both sound source types. In the case of an impulsive event, we additionally decrease the window size to a quarter (43 ms, 2048 samples) to gain more influence of the first wave.

After the first localization, the localization algorithm has to handle two different event types: for an impulsive one, it terminates and is waiting for the next event, for a non-impulsive event type, it waits for a half correlation window (85 ms, 4096 samples) and the pre-classification is repeated. This whole procedure is iterated until the classification result is impulsive again. That means that the on-going event is finished and the algorithm is waiting for the next event.

6. RESULTS

As a baseline localization method we used the standard SRP-PHAT approach. Thereby, sound data are divided in windows of a specific size with an overlap factor of 0.5. Each time an event inside of such a window is detected by the energy detector, the data is first multiplied by a Hamming window and then passed to the localization algorithm described in Section 3. The window size, used

source	baseline method		modified method	
	correct [%]	RMS [°]	correct [%]	RMS [°]
cup	29.23	42.08	93.41	10.91
door	60.81	39.77	79.77	17.46
spoon	48.18	37.22	100.00	3.26
toaster	73.39	23.78	97.87	6.35
mixer	96.89	7.12	98.75	4.58
speech	94.98	8.90	97.78	6.28

(a) without background noise

source	baseline method		modified method	
	correct [%]	RMS [°]	correct [%]	RMS [°]
cup	30.38	42.08	93.26	6.19
door	65.43	31.29	74.97	26.06
spoon	27.28	34.28	97.78	4.02
toaster	55.64	30.53	80.22	17.94
mixer	79.44	16.04	80.22	12.65
speech	96.09	9.38	95.65	9.04

(b) with background noise

Table 1: Percentage of correct localizations and the corresponding RMS in degrees without (a) and with background noise (b), in comparison between the baseline and the modified localization method.

in this case, was 8192 samples and corresponds to 170 ms, according to 11.7 localizations per second.

For non-impulsive sound sources like speech or a mixer, on which we concentrated in the past [6], this setup delivers high correct localization rates of over 95% with a relatively small root mean square error (RMS), under both conditions, i.e. with and without background noise. But the disadvantages of this setup can be clearly seen in the results for impulsive sound sources. In this case, the localization rate drops partially under 50% and also the RMS increases significantly (Table 1).

For non-impulsive sources, the modified algorithm, proposed in Section 5, results in a slight improvement of 1-2%. However, the localization rate and the RMS can be improved significantly for impulsive sound sources (Table 1). In this case, an absolute improvement of up to 71% is reached, with a partially significant decrease of the RMS. For example, using the modified localization method, the RMS decreases from 37° to 3° for the spoon case. Fig. 3 highlights this fact, comparing the baseline and the modified localization method for the azimuth estimation.

7. CONCLUSION

In this paper, we proposed and evaluated a modified localization algorithm which is able to localize reliably both impulsive and non-impulsive sound sources. Based on the standard SRP-PHAT localization approach we showed that a much higher correct localization rate with and without background noise can be reached using an energy detector for the temporal alignment and the pre-classification of an event. In so doing, an absolute improvement of the localization accuracy up to 71% could be achieved.

However, as pointed out before, the pre-classification results in an additional delay of 1.37 seconds. For real applications it is still acceptable, but it does not represent an optimal solution. Further investigations will try to minimize the time needed for the pre-classification.

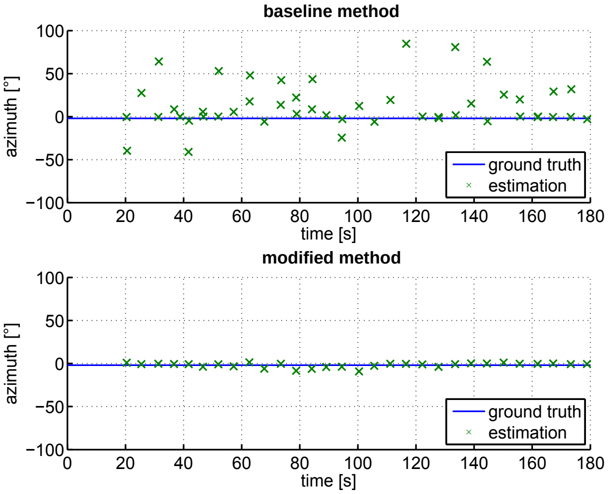

Figure 3: Localization azimuth results for the case of a dropping spoon in a noisy environment; comparison between the baseline and the modified localization method.

8. ACKNOWLEDGMENT

This work has been supported by the German Science Foundation within the Sonderforschungsbereich 588 “Humanoid Robots” and the Spanish administration under the project TEC 2008-02975. This paper was financially supported by the European Cooperation Exchange Program HD2008-0062. Additionally, we want to thank B. Kühn for his technical assistance.

REFERENCES

[1] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann. ARMAR-III: An integrated humanoid platform for sensory-motor control. *In Proceedings of the 2006 IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS), Genoa, Italy*, December 2006.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. *"Robust localization in reverberant rooms"*, chapter 8, pages 157–180. Springer, Berlin, 2001.

[3] S. M. Kay. *"Fundamentals of Statistical Signal Processing: Detection Theory"*. NJ: Prentice-Hall, 1st edition, 1998.

[4] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics*, 24(4):320–327, August 1976.

[5] K. Kroschel. *"Statistische Informationstechnik: Signal- und Mustererkennung, Parameter- und Signalschätzung"*. Springer, Berlin, 4th edition, 2004.

[6] J. Moragues, T. Machmer, A. Swerdlow, L. Vergara, J. Gosálbez, and K. Kroschel. Background noise suppression for acoustic localization by means of an adaptive energy detection approach. *In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, USA*, March-July 2008.