BACKGROUND NOISE SUPPRESSION FOR ACOUSTIC LOCALIZATION BY MEANS OF AN ADAPTIVE ENERGY DETECTION APPROACH

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

Machmer, T., Swerdlow, A., Kroschel, K.[*]

Universität Karlsruhe (TH)
Institut für Nachrichtentechnik (INT)
76128 Karlsruhe, Germany
{machmer, swerdlow, kroschel}@int.uka.de

ABSTRACT

A microphone array can be employed to localize dominant acoustic sources in a given noisy environment. This capability is successfully used in good signal to noise ratio (SNR) conditions but its accuracy decreases considerably in the presence of other background noise sources. In order to counteract this effect, a novel approach that combines the information provided by a Gaussian energy detector (GED) [1] with the approved localization method SRP-PHAT [2] is presented in this paper. To evaluate the presented technique, several acoustic sources (speech and impulsive sounds) were considered in a variety of different scenarios to demonstrate the robustness and the accuracy of the system proposed.

Index Terms— energy detector, background noise suppression, SRP-PHAT, acoustic localization.

1. INTRODUCTION

There are a lot of areas, in which acoustic scene analysis is required. One of the most important is the interaction between man and machine. Such situations occur e.g. in scenarios where a human cooperates with a so called *humanoid robot*, or is assisted by one [3]. In this case several active sound sources can exist in the robots proximity, for example in a kitchen, which contains many different appliances that can be acoustically observed. However, the presence of background noise normally decreases the performance of the detection and the localization of desired sound sources. The goal of the detection theory is to be able to decide when an event of interest takes place and then to collect more information about it. 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. The detector under these considerations is known as a matched filter. When the sound sources are not completely known, the design of the appropriate detector is more difficult. An example is speech, for which the waveform of a speaker is completely unpredictable and depends on many factors. In this case energy detection is of interest in detecting departures from a known background due to imprecisely defined changes (event or novelty detection) [1, 4]. However, in real situations the acoustic events are corrupted by non-stationary and non-white background noise caused, for example, by the fans placed in the robot or by kitchen appliances.

This paper is organized as follows. Section 2 presents the principles of the Gaussian energy detector and in Section 3 the localization algorithm used is described. In Section 4 the combination of both techniques is presented, in Sections 5 and 6 the experimental setup and achieved results are shown. Finally, the conclusion of our work is summarized in the last section.

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.

More formally, energy detectors are optimum solutions for both, Bayes and Neyman-Pearson criteria, for the following detection problem [1]:

$$\begin{array}{ll} H_0 : \mathbf{y} = \mathbf{w} & \mathbf{w} : N(0, \sigma_{\mathbf{w}}^2 \mathbf{I}) \\ H_1 : \mathbf{y} = \mathbf{s} + \mathbf{w} & \mathbf{s} : N(0, \sigma_{\mathbf{s}}^2 \mathbf{I}), \end{array} \tag{1}$$

where $\mathbf{y}$ is the observation vector (dimension N), $\mathbf{s}$ is the signal vector and $\mathbf{w}$ is the noise background vector. In model (1), both the noise and the signal are considered zero-mean multivariate Gaussian random vectors with uncorrelated components. $\sigma_{\mathbf{w}}^2$ and $\sigma_{\mathbf{s}}^2$ are the noise and the signal variances. The

[*]This work has been supported by the German Science Foundation DFG within the Sonderforschungsbereich 588 "Humanoid Robots".

optimum test for (1) is:

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

where the statistic $\frac{\mathbf{y}^T\mathbf{y}}{\sigma^2_{\mathbf{w}}}$ is chi-squared distributed with N degrees of freedom (χ^2_N) and λ can be set for a specific probability of false alarm (PFA). Resuming, the energy detector is optimum to detect zero-mean uncorrelated Gaussian signals, and is a generalized likelihood ratio test (GLRT) detector to detect any unknown deterministic signal. In both cases, $\mathbf{w}$ must be zero-mean uncorrelated Gaussian noise.

2.1. Non-independent background noise

The test (2) assumes that the components of $\mathbf{w}$ are i.i.d. (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 significantly increase the detection performance.

In this work, the background noise is assumed to be Gaussian and additive with zero mean. In this case, 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

$$\mathbf{y}_p = \mathbf{R}_{\mathbf{w}}^{-1/2}\mathbf{y}, \tag{3}$$

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

$$\hat{\mathbf{R}}_{\mathbf{w}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k\mathbf{w}_k^T. \tag{4}$$

The test (2) can be rewritten as

$$\frac{\mathbf{y}_p^T\mathbf{y}_p}{\sigma^2_{\mathbf{w}_p}} \underset{\mathbf{H_0}}{\overset{\mathbf{H_1}}{\gtrless}} \lambda \qquad \Leftrightarrow \qquad \mathbf{y}^T\mathbf{R}_{\mathbf{w}}^{-1}\mathbf{y} \underset{\mathbf{H_0}}{\overset{\mathbf{H_1}}{\gtrless}} \lambda. \tag{5}$$

Note that $\mathbf{R}_{\mathbf{w}} = E\left[\mathbf{w}_p\mathbf{w}_p^T\right] = \mathbf{I}$ and hence $\sigma^2_{\mathbf{w}_p} = 1$. The prewhitening transformation whitens and mean-power calculation normalizes the original observation noise. However, one consideration is pertinent for the design of energy detectors: the components of $\mathbf{w}_p$ must be not only independent but also identically distributed. The set of training observation noise vectors $\{\mathbf{w}_k\}, k = 1 \dots K$ is grouped in the matrix $\mathbf{W} = \left[\mathbf{w}_1 \dots \mathbf{w}_K\right]$. In the following it is assumed that $\{\mathbf{w}_k\}$ denotes independent vector observations. In practice this implies that the vectors $\{\mathbf{w}_k\}$ must correspond to non-overlapped (and rather well separated) segments of the noise record, or, preferably, that different noise records for every $\mathbf{w}_k$ are used.

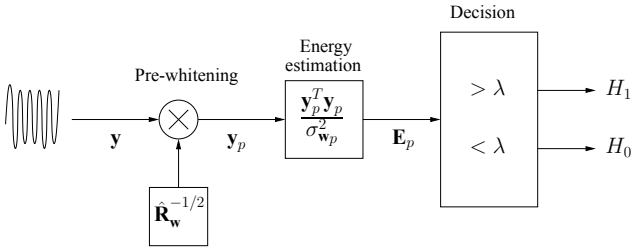

Fig. 1. Block diagram of an energy detector.

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 and compared with a threshold fixed by the PFA.

2.2. Non-stationary background noise

Energy detection principles have been described in the previous section, however the whitening process described in (4) and (5) assumes stationarity in the background noise environment. This assumption is not always correct and could considerably reduce the performance of the energy detector when used in real scenarios where the characteristics of the noise considerably change over the time. For that reason, knowledge about the noise altering the desired sound sources at any time is important in order to adapt the estimation of $\hat{\mathbf{R}}_{\mathbf{w}}$ dynamically, making our decision even more robust in the presence of background noise.

For that purpose, the initial estimation of instationary noise covariance matrix must be computed and then the last K noise vectors (according to our energy detector decision), saved in matrix $\mathbf{W}$, are used to reestimate $\hat{\mathbf{R}}_{\mathbf{w}}$ every T seconds. This period of time (T) will depend on the characteristics of the noise and specially on its stationarity.

3. LOCALIZATION

Today two approaches for acoustic localization are mainly used. The first is based on the estimation of the time difference of arrival (TDOA) of sound signals in a pair of spatially separated microphones. The most common technique for the determination of TDOAs is the generalized cross correlation (GCC). The GCC function $R_{ij}^{(g)}(\tau)$ is defined as

$$R_{ij}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_{ij}(\omega)X_i(\omega)X_j(\omega)^* e^{j\omega\tau}\, d\omega, \tag{6}$$

where $X_i(\omega)$ is the Fourier-Transforms of given microphone signals. ψ_{ij} is a weighting function which intends to decrease the noise and reverberation influences and tries to emphasize the GCC peak at the true TDOA. For real environments, the

Phase Transform (PHAT) technique has shown the best performance [5]. The PHAT weighting function is defined as

$$\psi_{ij}^{PHAT}(\omega) = \frac{1}{|X_i(\omega)X_j(\omega)^*|} \tag{7}$$

and can be regarded as a whitening filter.

The other well known technique for the acoustic localization is the so called Power Field (PF), also known as SRP (Steered Response Power) [2]. In this approach, beamforming is used to focus a microphone array to a specific spatial area. Using SRP, it is possible to "scan" the environment to search for the spatial position with the highest acoustic power in order to find the exact position of a sound source.

A combination of both techniques mentioned before leads to a method called SRP-PHAT, which fuses the stability of the SRP against reverberations and the efficiency of the GCC method which gives us the possibility to build a real time capable system. At a given time t, SRP-PHAT is computed as

$$P(t, \mathbf{s}) = \frac{1}{|M_p|} \sum_{(i,j)\in M_p} R_{ij}^{(g)}(t, \tau_{ij}(\mathbf{s})), \tag{8}$$

where $\tau_{ij}(\mathbf{s})$ denotes the theoretical delay between the microphones in pair (i, j) for the assumed spatial source position $\mathbf{s} = (s_x, s_y, s_z)$. M_p represents a given set of microphone pairs. To estimate the source position $\hat{\mathbf{s}}(t)$ at time t, the position of the maximal value in $P(t, \mathbf{s})$ has to be found in a given search space $\mathbf{S}$:

$$\hat{\mathbf{s}}(t) = \arg\max_{\mathbf{s}\in\mathbf{S}} P(t, \mathbf{s}). \tag{9}$$

4. BACKGROUND NOISE SUPPRESSION

In a kitchen scenario, background noise like the fan of an air conditioner or long lasting sounds of kitchen appliances is ubiquitous. Depending on the type of sound, the correlation of the background noise could be higher than the correlation of the desired sound source to be localized. Therefore, this problem can commonly lead to a high amount of mislocalization. To avoid this, a method to suppress these background noise sources was developed by using an adaptive Gaussian energy detector (GED) described in section 2.

To suppress the background noise, it is necessary to collect information about it. By using the energy detector it is possible to distinguish between a stationary background noise source and another active source. Therefore, when no desired sound source is detected our system estimates SRP-PHAT of the current background noise (P(t, **s**)) and stores it in a buffer $B(i, \mathbf{s})$ of size H where $i = 1, \ldots, H$. When a desired sound source is detected the mean of the last H SRP-PHAT computations of the noise, stored in the buffer, is computed and subtracted from the current SRP-PHAT estimation in the following way:

$$P^{(BNS)}(t, \mathbf{s}) = P(t, \mathbf{s}) - \frac{1}{H} \sum_{i=1}^{H} B(i, \mathbf{s}), \tag{10}$$

where the resulting Power-Field $P^{(BNS)}(t, \mathbf{s})$ at time t is estimated using background noise suppression (BNS). To estimate the sound source position $\hat{\mathbf{s}}^{(n)}(t)$, the same maximum search as described in (9) is used.

5. EXPERIMENTAL SETUP

To evaluate the performance of the combined technique explained in the previous sections and the improvement introduced in the localization phase in comparison to the case without background noise suppression, two recording sets were tested. In the first one, the performance of the presented localization system was assessed when a speaker was active in several room positions. For better evaluation of the system proposed, a second signal source was conceived using a toaster as an impulsive sound source. Three kinds of typical kitchen background noise sources were studied in each setup in order to simulate different signal to noise ratios and to evaluate the system ability to ignore the noise source, once it was localized for the first time. In the first scenario (S1) only the desired sound source without any additional background noise was recorded. In the second one (S2) a fan was additionally used as a background noise, and in the third scenario (S3) a kitchen grinder was activated in addition to the fan in S2.

The microphone array used for our experiments was built according to the head geometry of a humanoid robot 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. The data was 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 1000 with a reestimation period of 2 seconds. For the SRP-PHAT method a frame size of about 170 ms (8192 samples) was used. The source position was estimated by means of a 3D grid search with grids of 5 cm and a total grid dimension of 3 m x 3 m x 2 m.

To measure the accuracy of the proposed approach, 60 seconds of audio data were acquired for each of the three defined scenarios. About 45 seconds of natural human speech, and of an impulsive sound source, respectively, were recorded in each scenario. Five recordings were done within the same setup but placing the desired sound source and the active background noise generators at different positions. The mean signal to noise ratio was 31.35 dB for S1, 23.70 dB for S2, and 9.65 dB for S3.

Furthermore it was necessary to define the assignment of each localization to the real active sound source position. Because of the small concentrated array used, according to the robot's geometry, 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 obtained and the real angle was below 10 degrees.

6. RESULTS

In this section the results of our experiments are presented and discussed. Table 1 shows the averaged results over all measurements for both sound source types, speech (a) and the impulsive sound source (b), using the SRP-PHAT method and the presented background noise suppression approach SP-BNS (SRP-PHAT plus BNS). The absolute percentage of the correct localization rate is given for all combinations of active sound sources (noise and desired) in the corresponding scenario (S). In order to clarify the performance of the SP-BNS method, normalized values considering only the correct localization of active sound sources are presented in brackets.

S	Method	Fan	Grinder	Speaker	Wrong
S1	SRP-PHAT	off	off	**95.0 (100.0)**	5.0 (–)
	SP-BNS	off	off	**93.2 (100.0)**	6.8 (–)
S2	SRP-PHAT	19.1 (20.7)	off	**73.0 (79.3)**	7.9 (–)
	SP-BNS	0.0 (0.0)	off	**95.4 (100.0)**	4.6 (–)
S3	SRP-PHAT	0.0 (0.0)	83.7 (84.0)	**15.9 (16.0)**	0.4 (–)
	SP-BNS	0.0 (0.0)	1.2 (1.4)	**88.3 (98.6)**	10.5 (–)

(a) Speech

S	Method	Fan	Grinder	Toaster	Wrong
S1	SRP-PHAT	off	off	**39.5 (100.0)**	60.5 (–)
	SP-BNS	off	off	**38.7 (100.0)**	61.3 (–)
S2	SRP-PHAT	23.3 (48.1)	off	**25.1 (51.9)**	51.6 (–)
	SP-BNS	0.0 (0.0)	off	**46.6 (100.0)**	53.4 (–)
S3	SRP-PHAT	0.2 (0.2)	40.0 (56.8)	**30.2 (43.0)**	29.6 (–)
	SP-BNS	0.0 (0.0)	0.0 (0.0)	**35.8 (100.0)**	64.2 (–)

(b) Impulsive sound source

Table 1. Percentage of averaged measurement results for speech (a) and an impulsive sound source (b).

In case of speech, the correct localization rate was about 95% for both methods in the scenario without any background noise (S1). As expected, the localization rate for the desired source decreases dramatically for the scenarios with active background noise (S2 and S3) when using the SRP-PHAT method without BNS. However, nearly 100% suppression of the background noise was achieved when SP-BNS method was used. This leads to an improvement of the correct localization rate from 16% to 88% in case of S3. For the impulsive sound sources like a toaster, SRP-PHAT does not reach the high accuracy that is obtained with speech. This is the reason why the mislocalization rate amounts about 60% already in the scenario without any background noise (S1). Since the purpose of this paper was not the improvement of the acoustic localization using SRP-PHAT, the normalized values, given in brackets, should be considered in order to evaluate our approach (SP-BNS) and to avoid the influence of the mislocalization on the results. Then it can be seen that the combination of SRP-PHAT and BNS obtains an improvement of the correct localization rate from 43% to 100% for S3, e.g.. In scenarios S2 and S3, background noise is totally suppressed.

7. CONCLUSION

An adaptive energy detector which is able to adapt to a given stationary background noise and distinguishes between the background noise and a real acoustic event was presented. A combination of this detector with the common SRP-PHAT technique for the acoustic source localization was investigated and an improvement up to 72% has been achieved with speech data.

Further investigations will consider other microphone constellations and various parameter settings in order to improve the reliability of the localization of impulsive sound sources using SRP-PHAT. To reduce the mislocalization rate, a tracking algorithm could be used, for example a Kalman filter [6].

8. REFERENCES

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

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

[3] A. Swerdlow, K. Kroschel, T. Machmer, and D. Bechler, "Localization and identification of persons and ambient noise sources via acoustic scene analysis," *In Proceedings of the 14th International Congress on Sound and Vibration (ICSV), Cairns, Australia*, vol. 1, July 2007.

[4] M. Markou and S. Sameer, "Novelty detection: a review-part 1: statistical approaches," *Signal Processing*, vol. 83, pp. 2481–2497, November 2003.

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

[6] D. Bechler, M. Schlosser, and K. Kroschel, "Acoustic 3d speaker tracker for humanoid robots with a microphone array," *Proc. of the Int. Conf on Humanoid Robots (Humanoids)*, 2003.