ACOUSTIC DETECTION AND CLASSIFICATION OF SOUND SOURCES USING TEMPORAL MULTIPLE ENERGY DETECTOR FEATURES

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

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

Moragues, J., Serrano, A., Vergara, L.

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

ABSTRACT

In this work, a classification method using a novel approach for acoustic feature extraction is proposed. Therefore, a multiple energy detector structure (MED) is utilized and features called *temporal MED* (TMED) features are introduced. The usage of an energy detector enables a pre-classification for the differentiation between impulsive and non-impulsive acoustic events. The actual classification task can be performed by using a MED. Furthermore, investigations regarding the classification accuracy using more than one microphone are presented.

1. INTRODUCTION

There are a lot of areas, in which the acoustic scene analysis is required. One of the most important is the interaction between man and machine. Appropriate situations occur e.g. in scenarios, where a human cooperates with a *humanoid robot*, or is assisted by one [1]. In this case, several active sound sources can exist in the robot's proximity, for example in a kitchen, which contains many different acoustically observable appliances. Thereby, the presence of background noise normally decreases the performance of the detection and the classification of desired sound sources.

In real acoustic applications, the trade-off between the recognition rate of real events and the proportion of data rejected (background noise) is of particular importance. Because of uncertainty and noise inherent in any pattern recognition task (classification), errors are generally unavoidable. The option to reject is introduced to safeguard against excessive misclassification. Novelty or event detection is a solution to this problem, because of its ability to determine the presence of the event of interest inside a background noise which is already categorized [5].

As the sound sources are not completely known, the design of an appropriate detector is more difficult and in this case energy detection is of interest. However, as we do not know the duration of the novel event, a multiple energy detector (MED) structure with different time durations can be used in order to fit the window size of the detector to the length of the novelty [3].

The MED can also provide information about the detected event. By using the shape that one event produces when it is processed by the MED, some appropriate novel features can be extracted in order to train a Gaussian Mixture Model (GMM) [7] classifier.

This paper is organized as follows. Section 2 presents the principles of the novelty detection . In Section 3, the general idea of a multiple energy detector is described. Section 4 introduces the classification approach. In Section 5, the experimental setup is presented. Finally, achieved results and a conclusion of our work are given in Sections 6 and 7.

2. NOVELTY DETECTION

In this part of the paper, we deal with the particular case of having only one known class, the background noise. The aim thereby is to detect a possible novelty due to the presence of an event that is to be classified. The most appropriate criterion for this detection problem is the Neyman-Pearson (NP), which tries to maximize the probability of detection (PD) for a given probability of false alarm (PFA). This hypothesis testing approach can be defined as follows:

$$\begin{aligned} H_0 &: \mathbf{y} = \mathbf{w} \\ H_1 &: \mathbf{y} = \mathbf{s} + \mathbf{w}, \end{aligned} \tag{1}$$

where $\mathbf{y}$ is the observation vector (dimension N), $\mathbf{s}$ is the signal vector and $\mathbf{w}$ is the background noise vector.

It is well known that the optimum NP test can be expressed as [3, 4]:

$$\Lambda(\mathbf{y}) = \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \tag{2}$$

where $p(\mathbf{y}|H_i)$ is the probability density function (PDF) of $\mathbf{y}$ conditioned on hypothesis H_i, $\Lambda(\mathbf{y})$ is the so called likelihood ratio, and λ is a threshold which depends on the required PFA. This optimum test enhances the fact that the detection problem depends on $p(\mathbf{y}|H_1)$ which is unknown and cannot be estimated from training data. Hence (2) cannot be implemented in a practical case, but even if $p(\mathbf{y}|H_1)$ is known, the threshold λ must be selected to fit a required PFA, and this requires knowledge of the PDF of the likelihood-ratio $\Lambda(\mathbf{y})$ in (2), which is in general not available.

Taking into account the above limitations, we have to think about other detection alternatives noting that the PFA must be under control and that some kind of optimality must be achieved without having knowledge of $p(\mathbf{y}|H_1)$. One common method for detection of unknown signals is energy detection, which measures the

energy in the received waveform over a specific observation time.

In this work, a simple energy detector (ED) test given in (3) is used to distinguish between the background noise and the novel events that are to be classified [6]:

$$\frac{\mathbf{y}^T\mathbf{y}}{\sigma^2_{\mathbf{w}}} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda, \tag{3}$$

where $\sigma^2_{\mathbf{w}}$ is the noise variance and $\mathbf{y}$ is the observation vector supposing uncorrelated samples. The components of the background noise are assumed Gaussian distributed. But as real audio signals have highly correlated samples, some additional preprocessing is required to increase the detection performance significantly by means of a whitening matrix.

In order to summarize the detection performance of the ED, the PD will be plot versus the PFA for the detection problem described in (1). For a given threshold λ we have:

$$PFA = Q_{\chi^2_N}\left(\frac{\lambda}{\sigma^2_{\mathbf{w}}}\right) \tag{4}$$

$$PD = Q_{\chi^2_N}\left(\frac{\lambda}{\sigma^2_{\mathbf{s}} + \sigma^2_{\mathbf{w}}}\right), \tag{5}$$

where χ^2_N represents a chi-squared distribution with N degrees of freedom. In Fig. 1 the receiver operating charasteristic (ROC) is plotted for different signal-to-noise ratios (SNR). Each point on the curve corresponds to a value of (PFA, PD) and by adjusting λ, any point of the curve may be obtained. As expected, as λ increases, PFA decreases but so does PD and vice-versa.

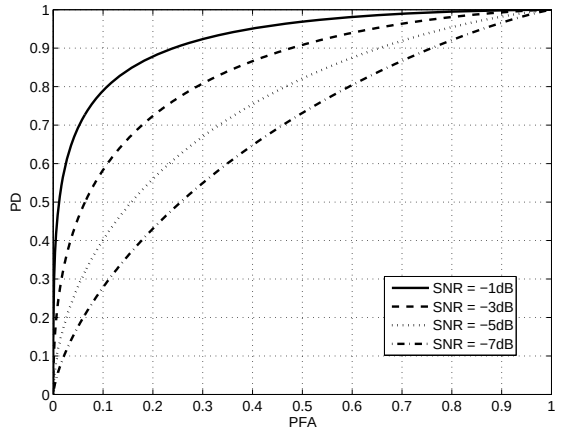

Figure 1: ROC curves for different signal-to-noise ratios (SNR).

3. MULTIPLE ENERGY DETECTORS

However, there is an issue which must be considered for the practical application of EDs. As we ignore a-priori the novelty duration, we do not know the most appropriate size N of the observation vector $\mathbf{y}$ for implementing the detector. This question is addressed in this section leading to a method based on using multiple ED matched to multiple novelty durations.

Let us assume that the observation vector $\mathbf{y}$ corresponds to N time samples. In principle, the only constraint that we could have for fitting N, is the maximum delay allowed by the particular application to make a decision about the possible presence of a novelty. However, if N is much larger than the duration of the novelty, the SNR under H_1 will be much lower than in the case that N is near to the novelty duration. In consequence, the PD will be much lower compared to an appropriate N. Similarly, if we select a value for N being too small in comparison to the event duration, we also have a loss in PD. This can be verified, for example, for large N, by using the approximation given in [3] for the performance of an ED:

$$\mathrm{PD} \approx Q\left(\frac{Q^{-1}(\mathrm{PFA}) - \sqrt{\frac{N}{2}\mathrm{SNR}}}{\mathrm{SNR}+1}\right), \tag{6}$$

where Q is defined as:

$$Q(\gamma) = \frac{1}{\sqrt{2\pi}}\int_{\gamma}^{\infty} exp(-\frac{1}{2}x^2)dx. \tag{7}$$

Studying the dependence of the PD with N for a SNR $=$ 1 and a PFA $= 10^{-4}$, notice that if the actual duration of the novelty is 60 samples, and we select $N = 60$, the PD is 0.978, meanwhile if we select $N = 30$, the PD is reduced to 0.810.

Taking into account this consideration, a multiple energy detector (MED) structure is presented in this paper formed by different EDs with varying the sample size N of $\mathbf{y}$.

Many different strategies for building the initial observation vector could be used, but in the absence of any a-priori information we will consider L layers of partitions. Each partition $u(l,k)$ corresponds to the output of the ED in level l and to the k-th partition in this level. At level 1, we will have the original interval of N samples. In level 2, we have 2 non-overlapped intervals of $N/2$ samples each and so on until L levels of successive divisions by 2. This produces a partition like the one represented in Fig. 2.

We assume that $\mathrm{PFA}_l = \mathrm{PFA}, \forall l$, where l represents the number of different levels of the MED structure. This implies that a different threshold will be required for every different interval size N_l. In every selected interval, an ED of the form given in (3) is implemented considering the observation vector $\mathbf{y}_l$, which is a portion of the initial observation vector $\mathbf{y}$. Then the corresponding statistic is a $\chi^2_{N_l}$ random variable, with $N_l \leq N$ the dimension of $\mathbf{y}_l$. For large N_l, the $\chi^2_{N_l}$ PDF can be approximated by a Gaussian PDF having mean N_l and variance $2N_l$. Hence the threshold λ_l in (3) corresponding to the l-th ED can be obtained from:

$$\mathrm{PFA}_l = \mathrm{PFA} = Q\left(\frac{\lambda_l - N_l}{\sqrt{2N_l}}\right). \tag{8}$$

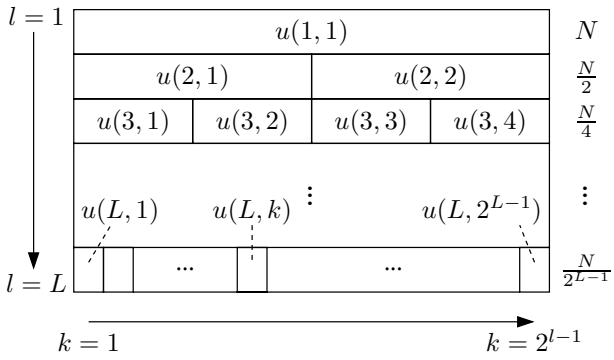

Figure 2: Multiple energy detector (MED) structure with L levels.

4. CLASSIFICATION

4.1 Pre-classification

After a detection of an event by using the MED described in Section 3, the detected sound source is pre-classified as an impulsive or non-impulsive event. This is done by measuring the length of the event, counting the detections of the energy detector in a specific time interval. In our case, this interval has a length of 512 detector windows and corresponds to approximately 2.73 seconds. An event is handled as an impulsive one if the totalized time duration of all detections in the time interval amounts less than one second.

4.2 TMED features

The information provided by the MED can be used to classify different events that take place in the robot's proximity. Therefore, we propose appropriate novel features, which can be extracted from the MED structure. They are calculated in the following way:

$$H(l) = \sum_{k=1}^{2^{l-1}} u(l,k) \quad \forall l = 1, \ldots, L \tag{9}$$

$$V(k) = \sum_{l=1}^{L} u\left(l, \left\lceil \tfrac{k}{2^{L-l}} \right\rceil\right) \quad \forall k = 1, \ldots, 2^{l-1}, \tag{10}$$

where $H(l), \forall l = 1, \ldots, L$, is the representation of the energy distribution of the event in each level of the pyramid, and $V(k), \forall k = 1, \ldots, 2^{l-1}$, provides information of the temporal distribution of the event detected.

Subsequently, the actual training vector in each time step is formed by applying the Discrete Cosine Transform (DCT) on (9) and (10) in order to reduce the dimensionality of the feature vector. Afterwards, the first 10 coefficients from both are concatenated. This way of proceeding results in a 20-dimensional vector $\mathbf{x}$, consisting of features which we call *temporal MED* (TMED) features.

4.3 Statistic modeling

The individual sound sources can be distinguished on the basis of their specific feature vectors. Therefore, an individual statistical model is required for each sound source. Over the past decades, the Gaussian Mixture Model (GMM) approach [7] has become the method par excellence for the classification task with context-independent sound data.

Thereby, for a statistical sound source model with M mixtures, a GMM probability density function can be defined as

$$f(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}), \tag{11}$$

with p_i the probability for the mixture i, a Gaussian density function $b_i(\mathbf{x})$

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu_i)^T(\Sigma_i)^{-1}(x-\mu_i)}, \tag{12}$$

with the mixture-dependent mean vector μ_i and the covariance matrix Σ_i, and

$$\lambda = (p_i, \mu_i, \Sigma_i), i = 1, \ldots, M \tag{13}$$

representing the parameters of the GMM. By $f(x|\lambda)$, the probability is given that an unknown feature vector $\mathbf{x}$ is generated by a specific GMM.

In order to determine the model parameters of the GMM for each sound source, a training phase is required. For this purpose, we drew on the Expectation-Maximization (EM) algorithm [2]. The parameters of GMMs are determined on the basis of TMED feature training vectors by the iterative application of the EM algorithm. The general GM modeling supports full covariance matrices. Contrary to that, we used diagonal covariance matrices only. On one side, this way of proceeding resulted in a higher computational efficiency; on the other side, empirical investigations showed that diagonal-matrix GMMs normally outperform full-matrix GMMs.

If S sound source models $\{\lambda_1, \ldots, \lambda_S\}$ are available after the training, the identification of the observed source source can be executed based on a new feature vector $\mathbf{x}$. The sound source model $\hat{s}$ is determined, which maximizes the a posteriori probability $P(\lambda_s|\mathbf{x})$. The mixed form of the Bayes rule yields the following result:

$$\hat{s} = \max_{1 \le s \le S} P(\lambda_s|\mathbf{x}) = \max_{1 \le s \le S} \frac{f(\mathbf{x}|\lambda_s)}{f(\mathbf{x})} P(\lambda_s)\,. \tag{14}$$

Assuming the equal probability of all sound sources and the statistical independence of the observations, the decision rule for the most probable sound source can be redefined:

$$\hat{s} = \max_{1 \le s \le S} f(\mathbf{x}|\lambda_s), \tag{15}$$

with $f(x_t|\lambda_s)$ given by Equation (11).

4.4 Channel combination

Due to the fact that the robot is equipped not only with one microphone, but with a microphone array, our investigations concentrated on the redundancy of acquired audio signals, with a view to improving the classification

accuracy. For this purpose, we evaluated a channel combination approach. Thereby, the channel combination was applied on the result level, to wit: after training of a separate GMM per channel, one global classification decision is calculated by combining the classification results from all channel-based classifiers. In so doing, Equation 15 is expanded to:

$$\hat{s} = \max_{1 \leq s \leq S} \sum_{t=1}^{C} \log f(\mathbf{x}|\lambda_s), \tag{16}$$

with C representing the number of channels.

The energy detector handles all channels independently. This way of proceeding leads to the fact that the prewhitening matrix is estimated for each channel separately. This behavior leads to dissimilar detector outputs on different channels for the same event. In order to deal with this, a synchronization step was required. Thereby, only detections within a maximum delay of 53 ms (according to 10 detector steps) are handled as the same event and taken into account while forming a TMED feature vector in one processing step.

5. EXPERIMENTAL SETUP

In order to evaluate the classification accuracy for 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 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, real experiments were carried out in a typical office room and a sound source database was collected using all six microphones of the robot (Figure 3, 4). Two of the microphones are placed on the positions of the human's ears, one on the forehead, one on the chin, and finally two further microphones are located on the back of the robot's head. The distance between the two ear microphones is 19 cm, between the front and back microphones 23 cm, between both front microphones 6 cm, and 4.5 cm between both microphones on the back of the head, respectively. To simulate the background noise emitted by the cooling fans of the robot, a 12 cm fan was placed near the microphone array.

Impulsive sound sources like putting a cup on the table, opening and closing a door, and dropping a spoon on a table were used. Additionally, a mixer and human speech were analyzed as non-impulsive sound sources. Thereby, three different data sets were generated: without background noise for the training process, without background noise for the evaluation, and with background noise for the evaluation. Each data set consisted of recordings of all sound sources in three different room positions. For each position and each sound source, 30 events were recorded with the sampling frequency of 48 kHz. Hence 1620 events (3 data sets, 6 sound sources, 3 positions, 30 events) were recorded in total.

For the sample frequency selected, a MED structure of 10 levels was used. This leads to 5.46 seconds of time samples for the highest level and to an energy detector of a window size of 256 samples (5 ms) for the lowest one. The PFA was set to 10^{-8}.

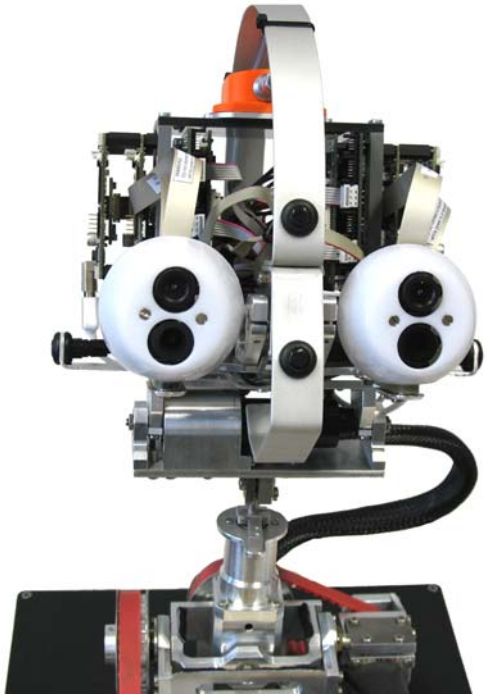

Figure 3: Front view of the head of the humanoid robot ARMAR III.

Figure 4: Lateral view of the head of the humanoid robot ARMAR III.

Depending on the pre-classification result, GMMs were trained with different numbers of mixtures for impulsive and non-impulsive sources. Thereby, 6 mixtures were used for impulsive events and 10 mixtures for non-impulsive ones, respectively.

6. RESULTS

In this section, experimental results are presented. Table 1 summarizes the correct classification rates using two different techniques (single channel, channel combination) for the evaluation without (a) and with background noise (b), as described in Section 4. As mentioned before, the training phase was performed using the training sound data set without background noise.

The given results are averaged over three room positions. At first, the results for the single channel case are given. Thereby, the worst and the best individual classification rates from all channels are presented. It can be seen, that the choice of a "bad" channel can result in a really poor classification performance. This situation reflects the challenge of selecting the right microphone for the classification task, especially in cases when the robot moves its head.

However, using the channel combination approach presented in 4.4, the classification reliability can be increased significantly. It can be seen, that the classification rate is always higher for the channel combination in comparison to the worst classification rate in the single channel case. This circumstance ensures reliable classification results for all sound sources and each situation without taking the knowledge of the robot's head position into account.

source	**single channel**		**channel combination**
	worst	best	
cup	0.97	0.99	1.00
door opening	0.98	1.00	1.00
door closing	0.54	0.66	0.65
spoon	0.72	0.82	0.89
mixer	0.68	0.90	0.93
speech	0.77	0.98	0.88

(a) evaluation without background noise

source	**single channel**		**channel combination**
	worst	best	
cup	0.98	0.99	0.98
door opening	0.82	0.92	0.86
door closing	0.48	0.82	0.80
spoon	0.29	0.73	0.79
mixer	0.57	0.80	0.86
speech	0.82	0.95	0.98

(b) evaluation with background noise

Table 1: Correct classification rates for the single channel approach, in comparison to the channel combination, for the case without (a) and with background noise (b). The results are averaged over three different room positions.

7. CONCLUSION

In this paper, a classification method for impulsive and non-impulsive sound sources by means of a multiple energy detector (MED) structure for forming novel temporal acoustic features (TMED features) was proposed. The classification using only one microphone could not achieve reliable results for all sound sources. A partially significant improvement, especially in comparison to the single channel case with a non-optimal channel choice, was achieved using the channel combination approach.

Future work will investigate the possibility of using not only the temporal features generated by the MED, but also in combination with other ones.

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] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

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

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

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

[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.

[7] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):73–83, January 1995.