Published in IEEE European Signal Processing Conference EUSIPCO X, 2000
which should be used for any reference to this work

AUTOMATIC SOUND DETECTION AND RECOGNITION FOR NOISY ENVIRONMENT

Alain Dufaux, Laurent Besacier, Michael Ansorge, and Fausto Pellandini*

Institute of Microtechnology, University of Neuchâtel
Rue A.-L. Breguet 2, 2000 Neuchâtel, Switzerland
Phone: +41 32 7183420; Fax: +41 32 7183402
alain.dufaux@imt.unine.ch, laurent.besacier@imag.fr

* L. Besacier is now with CLIPS Laboratory, University Joseph Fourier, BP 53 -38041 GRENOBLE Cedex 9.

ABSTRACT

Keywords: *Impulsive sound detection – Sound recognition – Gaussian Mixtures – Hidden Markov Models – Multimodels – Robustness – Background noise – Telesurveillance – Tele-assistive technologies.*

This paper addresses the problem of automatic detection and recognition of impulsive sounds, such as glass breaks, human screams, gunshots, explosions or door slams. A complete detection and recognition system is described and evaluated on a sound database containing more than 800 signals distributed among six different classes. Emphasis is set on robust techniques, allowing the use of this system in a noisy environment. The detection algorithm, based on a median filter, features a highly robust performance even under important background noise conditions. In the recognition stage, two statistical classifiers are compared, using Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), respectively. It can be shown that a rather good recognition rate (98% at 70dB and above 80% for 0dB signal-to-noise ratios) can be reached, even under severe gaussian white noise degradations.

1. INTRODUCTION

The use of a sound detection and recognition system can offer concrete potentialities for surveillance and security applications, by contributing to alarm triggering or validation. Furthermore, these functionalities can also be used in portable tele-assistive devices, to inform disabled and elderly persons affected in their hearing capabilities about relevant environmental sounds (warning signals, etc.) [1].

After a system overview (Section 2), and a description of the sound database (Section 3), the paper describes the efficient method used as a pre-processing, for detecting impulsive sounds (Section 4). Then, the pattern recognition stage (Section 5) is considered, starting with the signal analysis scheme (Subsection 5.1), and followed by the two considered classifiers (Subsection 5.2), respectively based on Gaussian Mixtures Models (GMM) and Hidden Markov Models (HMM). The main contribution of the paper is exposed in Subsection 5.3, dealing with robustness improvement based on multimodels. Section 6 discusses the global performance of the detection and recognition system operating under gaussian white noise, whereas Section 7 reports results achieved with real background noise. Finally, concluding remarks and future works considerations appear in Section 8.

2. SYSTEM OVERVIEW

The on-line surveillance system, depicted in *Figure 1*, is made up of a microphone recording the sound activity. Whenever the detection module is finding discontinuities or anomalies in the input signal, the recognition process is activated. A time-frequency analysis of the signal is then performed, and the class of the detected sound is determined after comparison with different sound models, trained from a database. Adequate human intervention (eg. intervention patrols, fire brigade, etc.) can then be undertaken according to the automatic system verdict.

3. SOUND DATABASE

The database used in the experiments reported in this paper contains 822 sounds of 6 different classes associated to intrusion or aggression situations : 314 door slams, 88 glass breaks, 73 human screams, 62 explosions, 225 gun shots and 60 other stationary noises. The sounds were taken from different sound libraries (BBC, Warner, Noisex-92 [2]) and there is a lot of variability within a same class (differences of quality, differences in signal length, different signal energy levels, etc). Some of the signals were also manually recorded. All signals were digitized and sampled at 44.1 kHz.

4. IMPULSIVE SOUND DETECTION

The detection module involves a non-linear median filter analyzing the energy variations in the 44.1 kHz-sampled input signal, with the effect of selectively amplifying the

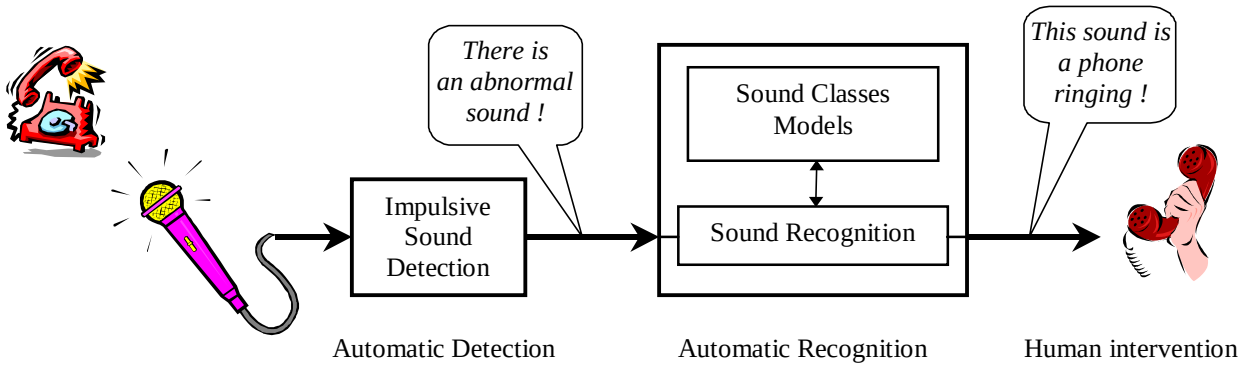

***Figure 1**: Overview of the surveillance system*

pulses occurring in the temporal energy sequence (see *Figure 2*). In detail, the detection process proceeds as follows. In a first step, the signal energy is estimated for every successive 100 ms block. Next, the obtained energy sequence is median-filtered (with ten taps), and the output of the filter is subtracted from the energy. This results in a new sequence that is normalized, emphasizing the relevant energy pulses. An adaptive thresholding – depending on the standard deviation of a past long-term windowed energy sequence – is then applied.

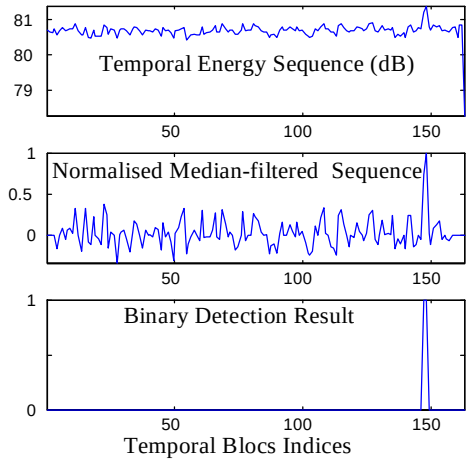

***Figure 2**: Detection of an impulse, placed in a –8 dB SNR random white noise background environment*

This method provides a tunable and very sensitive detection scheme for impulsive signals, where the pulses can be detected under quite adverse background noise conditions, with a signal-to-noise ratio (SNR) becoming as low as –10 dB (*Figure 2*). It must be noted that SNR values are measured over a window that includes the decreasing part of the signal.

Figure 3 shows the achieved performance evaluated on signals of the impulsive sound database for a variable level of gaussian white noise. A 100% correct detection rate above 0 dB SNRs and a very low false detection risk are guarantied (0% over 5 dB, 0.7% at max below).

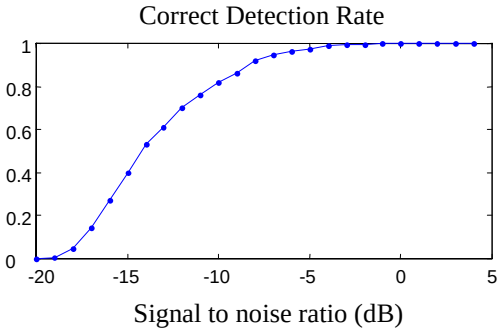

***Figure 3**: Impulsive sound detection algorithm: Performance evaluation on the impulsive sound database.*

5. IMPULSIVE SOUND RECOGNITION

5.1. Features Extraction

The first step of the recognition algorithm consists in an analysis of the signal to be classified, in view of extracting some typical features. In this work, the spectrum of the signal is calculated for every successive time frame of 512 samples. For each frame, the energy of N spectral bands is then derived, covering the frequency range from 0 to 20 kHz in a uniform manner. In this way, every feature vector is composed of N parameters, representing the spectral energy distribution of one time frame.

5.2. Statistical Classifiers

Two statistical pattern recognition techniques have been compared in this work. Their implementation was done in a mixed C / Matlab language, using the *h2m* toolbox [3].

5.2.1. Gaussian Mixtures Models (GMM)

For each sound class, the statistical behavior of the features (Probability Density Functions, pdf) can be modeled with a mixture of Gaussians. This model is characterized by the number of Gaussians, their relative weights, and their mean / covariance parameters [4].

During a training process, the system learns the GMM parameters, by analyzing a subset of the sound database. To find the best model for each class of sounds, the likelihood is maximized using 20 iterations of the Expectation Maximization (EM) algorithm [5]. In the recognition process, the signal to be classified is compared to the models of each class, so as to find the most probable one.

5.2.2. Hidden Markov Models (HMM)

Using on the other hand left-right HMMs [6] for the pattern recognition stage, offers the advantage that the time evolution of the signal features is taken into account [7]. In this paper, M=3 successive states are considered for the signal features, approximately corresponding to the pulse attack, steady state, and fading phases.

During the training process, the system learns the HMM characteristics of each considered signal class, by estimating mono-gaussian pdf of the features, and the transition probabilities between states. This training is done with 20 iterations of the Baum-Welsh recursion [8].

During the pattern recognition process, the most probable class of signal is determined by a log-likelihood estimation. Instead of the Forward-Backward Algorithm, the likelihood is evaluated using the Viterbi approximation [8], reducing the computation complexity.

5.3. Robustness Improvement

During the study, it was noticed that the trained sound models were highly dependent of the background noise level. The recognition results rapidly decrease when the background noise level of the unknown sound does not correspond to the noise level present during training. Then, one strategy to overcome the noisy environment problem was experimented, considering different steps of gaussian background white noise levels, and building one model for each of them. Practically, for every sound class, one independent model is built for SNR values belonging to the range –10 to 70 dB, with a step of 10 dB.

This solution produces a good recognition rate, even for important noise levels. The drawback is that the recognizer has to test 9 models for each class of sounds, increasing the computation load and the processing time. In this work, a coarse estimation of the SNR is done at detection stage, and only the models corresponding to the nearest SNR values are tested for recognition. The class that maximizes the likelihood over all considered models is selected as the winner. This method seems to be a good compromise between complexity and high recognition rates in corrupted environment. The recognition performance is shown in *Table 1*, for the GMM (with 8 gaussians) and the HMM (with 3 states) techniques. The recognition rates are presented for different white noise background levels. Those results were obtained by testing the half part of the database, the other part being used for training. The features number N, leading to the best recognition results, was found to be 10 and 40 frequency bands for the GMM and HMM cases, respectively. The reason for this difference, is the reduced performance of the GMM, due to calculation precision limits, when N exceeds 10.

Signal-to-Noise Ratio (dB)	Rec. Rate (%) GMM Classifier	Rec. Rate (%) HMM Classifier
70	98.29	98.54
60	96.83	97.07
50	94.39	95.85
40	91.71	95.37
30	89.76	94.63
20	86.83	90.73
10	80.73	88.54
0	65.85	81.95
-10	52.44	61.95

***Table 1**: Performance of the sound recognition algorithms: Identification rates, according to different gaussian white noise levels – 412 tests at each SNR level.*

This table shows that the HMM classifier with 40 frequency bands features is more robust to background noise degradation than the GMM using 10 frequency bands features, especially for low SNR values around 0 dB. However, considering that its computational load is three times larger than for the GMM, a compromise could be interesting, using the GMM models for higher SNR and the HMM for lower noise levels.

6. DETECTION AND RECOGNITION SYSTEM

When the detection stage is cascaded with the subsequent recognition stage, the classification performance listed in *Table 1* is observed to decrease, due to possible mismatches between the true pulse time-location of the input signal, and the detected one. The beginning and duration of the signal window used for classification are both important, especially for low SNR signals. For that reason, the time location of the pulse start must be refined (time resolution of 200 samples) with an adaptive amplitude-based thresholding process. Global system recognition results are shown on *Table 2*, using a fixed signal duration of 0.75 seconds. In *Table 2*, bad-detection situations are ruled out, when a detection position error higher than 1 second appears on the attack of the signal. SNR estimations of the detected pulse are calculated based on past noise level.

SNR (dB)	Bad-detected Signals (%)	GMM Rec. Rate (%)	HMM Rec. Rate (%)
70	0	97.32	98.54
60	0	94.88	96.10
50	0	91.71	95.37
40	0	90.93	96.32
30	0	90.89	94.53
20	0	86.54	91.54
10	0	77.02	85.09
0	0.26	63.57	68.30
-10	18.24	44.06	49.15

***Table 2**: Performance of the whole sound detection and recognition system after removal of bad-detected signals: Detection / identification rates, according to different gaussian white noise levels – 412 tests at each SNR level.*

7. REAL WORLD BACKGROUND NOISE

Real world background noises are generally more structured than gaussian white noise, possibly with limited bandwidth. Therefore, real world background noises are often less critical than white noises, for recognition systems based on spectral features. However, the robust method proposed in the above sections can be used all the same, after a background noise whitening operation. In this work, the proposed whitening technique replaces portions of signals whose absolute amplitudes are lower than a given threshold, by a white noise. In this way, the significant part of the signal (pulse) to be recognized remains unchanged. The threshold and the replacing gaussian white noise level both depend on the original background noise variance.

SNR (dB)	Bad-detected Signals (%)	HMM Rec. Rate (%) No whitening	HMM Rec. Rate (%) Whitening
70	0	97.80	96.10
60	0	96.59	94.88
50	0	93.41	93.41
40	0	84.39	93.17
30	0	68.05	82.44
20	0	52.20	81.71
10	0	42.44	66.34
0	18.29	33.13	42.69
-10	94.39	52.17	39.13

***Table 3**: Performance of the sound recognition system, for musical background noise, with and without whitening process – 412 tests at each SNR level.*

In *Table 3*, HMM recognition rates are compared for musical background disturbances, with and without noise whitening process. For SNRs between 0 and 40 dB, an important performance improvement is observed. At –10 dB, the detection module performance is very bad and recognition results are not significant.

8. CONCLUSION AND FUTURE WORK

This work has shown that a good recognition rate (98% at 70dB and above 80% for 0dB SNR) can be reached, even under important noise degradation conditions. The study is presently going on, taking more complex and real noise environment types into account. Other robust recognition techniques, like Perceptron Neural Networks, will be considered. Hybrid solutions seem to be interesting in order to increase robustness and reduce the overall system complexity load. At system level, detection precision improvement and rejection of signals not belonging to the ensemble of considered classes, will be examined.

ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation under Grant FN 20-53'843.98.

REFERENCES

[1] A. Dufaux, L. Besacier, M. Ansorge, F. Pellandini, "Automatic Classification of Wideband Acoustic Signals", *Joint 137th meeting of the Acoustical Society of America and Forum Acusticum 99*, Berlin, Germany, pp. 14-19, March 1999.

[2] Leonardo Software, Santa Monica, CA 90401, http://www.leonardosoft.com

[3] O. Cappe, "h2m : A set of MATLAB functions for the EM estimation of hidden Markov models with Gaussian state-conditional distributions". ENST/Paris, http://tsi.enst.fr/~cappe/h2m/index.html

[4] L. Besacier, A. Dufaux, M. Ansorge, F. Pellandini, "Automatic Sound Recognition Relying on Statistical Methods, with Application to Telesurveillance", *Proc. of COST 254, Int'l. Workshop on Intelligent Communication Technologies and Applications, with Emphasis on Mobile Communication*, Neuchâtel, CH, May 5-7, 1999, pp. 116-120.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. of the Royal Statistical Society B*, vol. 39, pp.1-38, 1977.

[6] A. B. Poritz, "Hidden Markov models: A guided tour", in Proc. of the IEEE Int'l. Conf. on Acoustics, Speech and Signal Processing (ICASSP '88), May 1988, pp. 7-13.

[7] C. Couvreur, *Environmental Sound Recognition : a Statistical Approach"*, PhD thesis, Faculté Polytechnique de Mons, Belgium, June 1997.

[8] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in *Proc. of the IEEE*, vol. 77, n°2, pp 257-286, February 1989.