Joint detection and identification of emergency vehicle alarm sound

Mamadou Mboup†,* — Monia Turki** — Fatimetou Mint†

† CReSTIC - Université de Reims Champagne Ardenne - BP 1039 Moulin de la Housse, 51687 Reims cedex 2 - France
* Non-A, Inria Lille-Nord Europe
Mamadou.Mboup@univ-reims.fr
** U2S - École Nationale d’Ingénieurs de Tunis - Tunis Bevédère - Tunisie
m.turki@enit.rnu.tn

ABSTRACT. Two algorithms for joint detection-identification of emergency vehicle sound alarm signals are presented. The algorithms exploit the time-frequency characteristics of the signals and both are based on the computation of the Short-time Fourier Transform. Two different approaches are then considered template matching and parameter estimation.

RÉSUMÉ. Deux algorithms de détection-identification conjointe de signaux d’alarme sonore sont présentés. Les deux algorithmes exploitent les caractéristiques temps-fréquence des signaux et ils reposent tous les deux sur le calcul de la transformée de Fourier à court-terme. Par la suite, on distingue une approche de type template matching et une approche de type estimation paramétrique.

KEYWORDS: Detection, Estimation, Identification, Audio alarm signals, Short-term Fourier Transform

MOTS-CLÉS: Détection, Estimation, Identification, Signaux d’alarme audio, Transformée de Fourier à court-terme
1. Introduction

1.1. Problem description

This study participates to the enhancement of safe outdoor mobility for people with presbycusis and in particular for elderly people. More precisely, the objective is to help people for a better positioning and moving in the sounding city in a driving context. In such situation, a correct perception of the audio ambiance surrounding the car is necessary for a safe driving: it allows the driver to react in due time to a given event and even more, to anticipate on different possible scenarios. Emergency sound alarms are of particular interest in such audio ambiance. When an alarm sounds in the street, there is always a period of time before we realize it, and especially, before we know where the signal comes from. This period of time lengthens with age. Now, a fast detection and spatial location [4] of the alarm source is critical not only for a safe and smooth mobility, but also for an appropriate response with respect to the emergency.

In its most basic form, the detection step produces a binary result: alarm/No alarm. Note however that such binary information is incomplete. Indeed, all alarm sirens do not have equal priority: two-tone alarms, like ambulance, car fire and police sirens, have absolute priority while three-tone alarms (e.g. organ transplant emergency siren) do not. Therefore, the detection step needs to be completed by identifying the type of alarm. Traditionally, the two tasks are considered successively. The detection problem is solved first, using e.g. measures of the sound energy variations [6]. Then the detected alarm is identified using some classification technique such as K-means, Neural networks or Support Vector Machine [9]. Note however that these two complementary tasks should ideally be jointly achieved. In this direction, the problem is most commonly approached in terms of pattern recognition, taking advantage of the well developed machinery of automatic speech recognition [2], [5], [8] (see also [3] and references therein). Many other recognition methods are also developed such as those borrowed from the genetic motif analysis [1] or the well elaborated auditory scene analysis.

The purpose of the paper is to present a simple solution to the joint alarm detection and identification problem. The proposed method does not rely on the extraction of global and sophisticated features as the Mel frequency cepstral coefficients or the use of elaborated tools as Non-negative matrix factorization. The method is rather straightforward and it consists in retrieving the time-frequency parameters of the alarm from a simple Short-term Fourier Transform of the recorded signal. These parameters define the alarm characteristics, which are described in the next subsection. The characteristics are very typical and they constitute adequate signatures of the alarms. Two algorithms are then presented in sections 2 and 3 respectively. The first algorithm uses the alarm signature as a template while the second is based on parameter estimation.

1.2. Alarm signal characteristics

The alarm signals of interest are essentially constituted of a successive transmission of two amplitude-modulated fundamental tones. This structure is repeated at the different harmonics of the fundamental tones. Figure 1 gives a sketch of typical representation of such a signal in the time-frequency plane (spectrogram representation).
The fundamental frequencies of the tones and their transmission frequency depend on the alarm type.

To account for both the time and the frequency features of the alarm signals, we consider here the Short-time Fourier Transform. This reveals the structure of the alarms as schematized in Figure 1. Two joint detection and identification algorithms are proposed. Both are based on the computation of the Short-time Fourier transform of the recorded signal. The choice of the window type is not critical here, and so, we use the Hanning window.

2. Template matching

As already mentioned, each emergency alarm signal is characterized in the time-frequency plane by a typical form as represented in Figure 1.

The distribution of the maxima of the spectrogram, as previously illustrated, represents an adequate signature of the alarm. A simple joint alarm detection and identification method then follows by searching the alarm signature in the spectrogram of the recorded signal. The corresponding algorithm is:

Algorithm 1 - Searching alarm signature

**Inputs:** \( x(t) \): signal recorded at one microphone and alarm parameters: \( f_1, f_2, T_1, T_2 \).

**Outputs:** Detection/Non detection of the alarm signal

**Step 1:** Compute \( X(f, t) \): the Short time Fourier transform of the signal \( x(t) \).

Window: length = 50ms, type = 'Hanning', \( N \) = number of frequency bins

**Step 2:** Compute the distribution of the local maxima in the frequency band of interest:

\[ [f_1 - \Delta f, f_2 + \Delta f] \], where \( \Delta f = 5Hz \) is a tolerance parameter. To make this part more robust, instead of the local maxima, we compute for each \( t \), the mean values of the spectrogram around each of the frequencies of interest:

\[
X_i(t) = \frac{1}{\Delta f} \sum_{f = f_i - \Delta f}^{f_i + \Delta f} |X(f, t)|; \quad i = 1, 2.
\]  

(1)
Step 3: Compute the characteristic function defined by

$$\chi_{[f_1, f_2]}(t) = \begin{cases} f_1 & X_1(t) > X_2(t) \\ f_2 & \text{else} \end{cases} \quad (2)$$

Step 4: Decision - Template matching

The signature of the alarm corresponding to the parameters $f_1$ and $f_2$ is compared with the above characteristic function. This alarm is detected and then automatically identified when there is a matching in between the template (as illustrated in figure 1.a) and the characteristic function. A matching is meant when the time lags separating a sequence of rising-falling-rising edges corresponds respectively to $T_1$ and $T_2$ up to some user selected tolerance error $\mu$.

The following figure represents the sonogram (sound spectrogram) of a complex audio scene, recorded in some street in Paris. The record includes several segments of the SAMU emergency signal and also fire brigades emergency alarm. Other various noises are also present, e.g. the traffic noise.

![Sonogram of an audio scene recorded in some street in Paris.](image)

The characteristic functions corresponding to four different alarm signals are computed from the above audio scene and plotted in figures 3. Note however that the characteristic functions considered here are slightly different from the definition above: we replace here the values $f_1$ and $f_2$ by 0 and 1 respectively. This is without any loss of generality.

First, we note that an alarm signal is present during the beginning of the scene. This alarm is identified as the SAMU ambulance emergency warning (see the top plot of figure 3) and this is the only alarm that has been detected during the time segment [0, 8s]. The same alarm signal is detected again in several other places in the scene. The corresponding time slots are indicated in figure 4. The top plot of this figure reproduces the top plot of figure 3, along with the time slots with positive detection. The second plot of the figure displays the time lags as described in step 4 above. They correspond to the widths of the “1” and the “0” of the associated characteristic function. Recall that a detection is meant when the time lag, say $\tau$, satisfy $|\tau - (T_1 + T_2)/2| \leq \mu(T_1 + T_2)/2$. Here we have set $\mu = 0.05$. Note that a second alarm is also detected during the time segment [53.5s, 58s], as shown in the second plot of figure 3.
Figure 3. Characteristic functions of 4 different alarm signals.

Figure 4. Detection criterion: Template matching

Figure 5. Detection criterion: Template matching

This second alarm, which is identified as the fire vehicle emergency siren, is actually present in the scene although it is almost hidden by the SAMU emergency alarm. This
conclusion is confirmed in the top plot of figure 5 which shows a good matching in the quoted time interval. The remaining plots of the figure show that the Police and Ambulance alarms are not present in the scene. This experiment on real world measurement shows the sensibility and the robustness of the detection/identification. It also shows that the proposed algorithm 1 is able to detect and identify separately simultaneous alarms.

3. Parameter estimation

The second algorithm is based on a direct estimation of the four parameters \( f_1, f_2, T_1, \) and \( T_2 \) of the alarm signals. The algorithm is devised from the following observations:

1) The frequency switches of an alarm signal carry part of the signal energy and, this is reflected by the vertical lines on the corresponding sonogram (see the zoom of the sonogram in figure 6). Now, the values of \( T_1 \) and \( T_2 \) can be deduced from the positions of these vertical lines.

2) Assume that an emergency alarm, with frequency parameters \( f_1 \) and \( f_2 \), occurs during some time period \([a,b] \). Consider the projection on the frequency axis, of the related part of the sonogram (restricted to \([a,b]\)). As this projection obviously reveals the local frequency distribution of the energy, it has the typical form depicted in figure 1. The local maxima of the pulses correspond to the values of the parameters \( kf_1 \) and \( kf_2 \), \( k = 1, 2, \ldots \).

![Figure 6. Zoom in the sonogram of figure 2](image)

These observations lead to the following algorithm:

Algorithm 2 - Parameter estimation

**Inputs:**
- \( x(t) \): signal recorded at one microphone
- \( \varepsilon \): Tolerance parameter
- \( \Delta \): Width of the time interval \([a, b]\)
- Alarm signal parameters: \(f_1, f_2, T_1\) and \(T_2\) for all alarms of interest.

**Output**: Detection/Non-detection of the alarm signal

**Step 1**: Compute the Short time Fourier transform of the signal \(x(t)\).
- Window: length = 50ms, type = 'Hanning', \(N\) = number of frequency bins
- Result: \(X(f, t)\)

**Step 2**: Compute a projection of the Sonogram on the time axis by:

\[
p(t) = \sum_f |X(f, t)|
\]  
(3)

**Step 3**: Detect the positive pulses of \(p(t)\), and compute their locations. Let \(\{t_k\}_{k \geq 0}\) be the sequence of the locations. Then, alarm detection will be meant if this sequence is of the form:

\[
\begin{align*}
t_{2k} &= t_0 + k(\hat{T}_1 + \hat{T}_2) \\
t_{2k+1} &= t_0 + (k + 1)\hat{T}_1 + k\hat{T}_2 \\
& \quad k = 0, 1, 2, \ldots
\end{align*}
\]  
(4)

for some \(\hat{T}_1\) and \(\hat{T}_2\), which incidentally, provide us with some estimates of the corresponding alarm parameters \(T_1\) and \(T_2\). Of course, the equalities above have to be considered up to some uncertainty \(\varepsilon\) in order to obtain a robust detection.

**Step 4**: Identification

The objective of this step is twofold: 1) confirm the detection step and 2) identify the detected alarm by estimating the corresponding two tones. To proceed, we compute the following projection on the frequency axis:

\[
\Phi(f, t) = \sum_{\tau=t-\Delta}^t |X(f, \tau)|.
\]  
(5)

If, for a given time index \(t\), \(\Phi(f, t)\) is a sequence of pulses as in figure 1, then the detection is confirmed and we conclude that an alarm occurs during the time interval \([t - \Delta, t]\).

The so-detected alarm is finally identified from \(\hat{T}_1\) and \(\hat{T}_2\) above and from the two tones \(f_1\) and \(f_2\) estimated from the location of the local maxima of \(\Phi(f, t)\).

The algorithm is now applied to the preceding audio scene, recorded in Paris. Figure 7(a) below displays the pulse \(p(t)\) computed in Step 2 within the same time interval as that of the zoom in figure 6. (Actually, the figure plots a crude numerical differentiation of \(p(t)\), using the difference operator. A more elaborated numerical differentiation method as e.g. [7], would be more adequate to enhance the pulses while reducing the noise.) The locations \(t_k\) satisfy \(|t_{k+1} - t_k| = 0.54 \pm \varepsilon\) where \(0 < \varepsilon < 0.015\). This corresponds to the characterization (4) in Step 2 above, with \(\hat{T}_1 = \hat{T}_2 = 0.54s\). Since these values are compatible with the parameters of some alarm signal, we conclude that an alarm occurs during the current period of time. In figure 7(b), the projection \(\Phi(f, t)\), defined in (5) in Step 4 above, is plotted, for \(t = 35s\) and \(\Delta = 2s\). The maxima of the pulses correspond respectively to: \(\hat{f}_1 = 435.1\); \(\hat{f}_2 = 652.2\); \(\hat{2f}_1 = 878.6\); \(\hat{2f}_2 = 1320.3\)

Comparing with the parameters of the alarm signals, we identify the SAMU emergency
Figure 7. Alarm signal parameters estimation

alarm for which $f_1 = 435$ and $f_2 = 651$.
As opposed to Algorithm 1, Algorithm 2 does not need to run several parallel detectors,
one for each type of alarm for the detection and identification. The estimation of
the parameters directly allows for the joint detection and identification of the alarm present
in the scene. However, only one alarm can be handled at a time. When more than one
alarm occur the less dominant one may behave as a perturbation, and this can affect the
estimation accuracy. Now, it is often the case that both police and ambulance alarm signals
are associated in a same audio scene. In such case, Algorithm 1 is preferable. Comparison
with other methods in the literature will be done in an extended version of the paper.

Acknowledgements

This work is supported by the Ambient Assisted Living Project AAL 2011-4-056,
'CityForAll-Age Sentitive ICT Systems for Intelligence City For All.'

4. References

using Multiple Audio Words and Contextual Bayesian Networks. In CBMI - 11th Int. Workshop
on Content Based Multimedia Indexing, Veszprém, Hongrie, June 2013.
[9] VLADIMIR VAPNIK. The Nature of Statistical Learning Theory. Information Science and