Detecting Alarm Sounds

Dan Ellis • Columbia University • dpwe@ee.columbia.edu

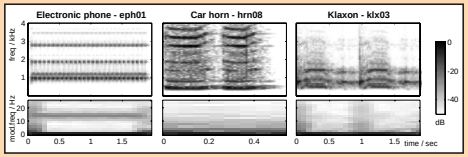
Summary: Alarms sounds (telephone rings, sirens etc.) carry important information. Automatic systems to detect them in high-noise conditions would be useful for hearing prostheses and intelligent machines. We characterize alarm sounds in general, and compare two approaches to recognition.

Introduction

- **Alarm sounds** (bells, phones, buzzers etc.) are important to listeners
  → Automatic recognizers would have many applications
- Listeners can recognize 'new' sounds as alarms
  → Are there some **general characteristics** common to all alarms?
- There is **no existing standard task** for alarm sounds
  → **Collect a corpus** consisting of different alarm sounds
  + **Inspect examples** to find common characteristics
  + Try a **baseline recognizer** using standard pattern recognition
  + Compare to a **source-separation approach** that tries to isolate alarms from background noise

Alarm sound corpus

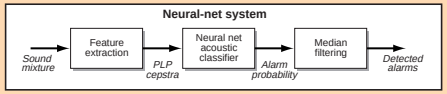
- A set of 50 short **alarm sound examples** was collected ... from the **web**, and by **making recordings** at home and in the office
- Examples include: Car horns, emergency sirens, fire alarms, doorbells, mechanical and electronic telephones, smoke alarms etc.



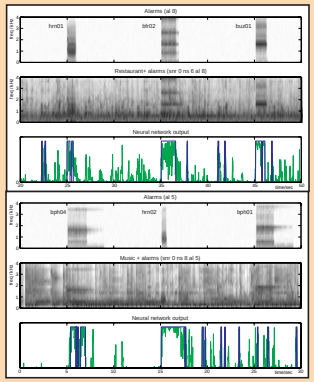
- Browsing the data suggests several candidate **alarm characteristics**:
  - Alarm sounds often have a strong and stable **pitch**. This appears as **pronounced horizontal harmonic structure** in spectrograms.
  - Alarms start abruptly and **sustain at a constant level** for hundreds of milliseconds. (Natural sounds more often decay away.)
  - There is often a significant **energy component around 3-4 kHz**, the peak of human sensitivity.
  - Some classes of alarm – e.g. phone rings – have characteristic **amplitude modulation** in the 8-30 Hz range.
    This is visible in the summary modulation spectrograms above.
  - Alarm sounds often **repeat** at a 0.5-4 Hz period.
- Can these characteristics be used to build a general alarm detector?

Baseline pattern-recognition neural-net system

- Recognizing acoustic patterns is addressed in **speech recognition**. A simple starting point for alarm detection is to **adapt** those techniques.
- We trained a multi-layer perceptron **neural net acoustic model** (as used in our connectionist speech recognition approach) to estimate the **posterior probability of an alarm** being present.
- The output probability was smoothed with a **median filter** over a 100 ms window. A filtered probability above 0.5 was taken as a detected alarm event.



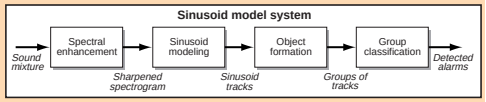
- The net was **trained** by back-propagation on alarm examples artificially mixed with a range of background noises.
- An alarm detector must work in high levels of ambient noise. Thus the alarm amplitudes were adjusted for an overall **SNR of 0 dB**.



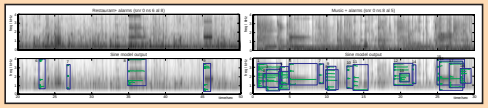
- In this example, 3 randomly-chosen alarm sounds (top panel) are mixed with **restaurant noise** giving the middle panel.
- The network correctly locates all 3 alarms, but inserts many extra **false alarms**.
- Three different alarms are added to a **pop music** background.
- The third alarm is not detected, and there are once again numerous falsely detected insertions.

Sinusoid model source-separation system

- A major weakness in speech recognition is that pattern recognition is applied to **global signal characteristics** – there is normally no way to separate target from other sounds (background noise).
- For alarm detection, we are particularly interested in spotting the alarm sounds **regardless of background**, and at poor signal-to-noise ratios.
- We therefore experimented with an alternative approach based on sinusoid modeling and object grouping, borrowed from **Computational Auditory Scene Analysis** (CASA).



- **Spectral enhancement** filters the spectrogram to emphasize the horizontal structures that are characteristic of alarms.
- **Sinusoid modeling** represents prominent, spectral energy in the signal as a set of individual sinusoids with slowly-varying amplitudes and frequencies. This stage ignores unstructured background noise.
- **Object formation** groups together sine tracks that appear to come from a single source on the basis of synchronous onset and similar durations.
- **Group classification** calculates various statistics for each sine-track group and decides whether to label it as an alarm, on the basis of training examples. This stage should remove pitched non-alarm background elements.
- Useful statistics include **spectral moment** (a measure of how widely spaced and discrete the spectrum is) and **duration-normalized frequency variation** (which detects the long, stable harmonics of alarms).



- Against the restaurant noise, the sine model does well, although the horn is rejected as a non-alarm object.
- The pop music background results in many false alarms as the instrument harmonics are mistaken for alarms.

Evaluation & Results

- The fifty example alarm sounds were divided into two sets of 25 for **training** and **test** respectively. Each set was mixed with **four background noises**, with different noises used for training and test.

Index	*Training set noise*	*Test set noise*
1	Aurora station ambience	Aurora airport ambience
2	Aurora babble	Aurora restaurant
3	Speech fragments	Different speech
4	Pop music excerpt	Different pop music

- The **results** of both systems tested on all 100 examples (in 20 groups of 5) are shown below:

Noise	*Neural net system*			*Sinusoid model system*		
	Del	*Ins*	*Tot*	*Del*	*Ins*	*Tot*
1 (amb)	7 / 25	2	36%	14 / 25	1	60%
2 (bab)	5 / 25	63	272%	15 / 25	2	68%
3 (spe)	2 / 25	68	280%	12 / 25	9	84%
4 (mus)	8 / 25	37	180%	9 / 25	135	576%
Overall	**22 / 100**	170	**192%**	**50 / 100**	147	**197%**

- Due to the **large number of false-alarms** (insertion errors) committed by both systems, the overall error rate is close to 200% in both cases!
- The **neural net** system makes many insertions on noises 2 and 3, presumably because they **did not resemble the noises used in training**. This is a major weakness of the global-features pattern-recognition approach.
- The **sine model** system makes the vast majority of its errors on the **music example**, where instrument notes are mistaken for alarms. It should be possible to improve the "group classification" stage using the training data to discriminate between true alarms and music notes.
- Looking only at **deletions** (false rejections), the neural net system makes fewer than half the errors made by the sine model system. However, many alarms (such as sirens) were actually rejected by the "group classification". We will pursue correcting these errors.
- Other future work will include investigating the variation of error rate with SNR, and differentiating between specific alarms.

for CRAC-2001/Eurospeech, Aalborg • 2001-08-28 dpwe@ee.columbia.edu