DETECTING ALARM SOUNDS

Daniel P.W. Ellis

Department of Electrical Engineering, Columbia University, New York NY USA

dpwe@ee.columbia.edu

ABSTRACT

Alarms such as phone rings, smoke alarms and sirens are an important component of the acoustic world, designed to convey urgent information in an efficient and unambiguous manner. We are investigating automatic recognition of this class of sounds both because of the practical applications for the hearing impaired, and also because alarm sounds are deliberately constructed to be easily heard, making them a promising target for detection in adverse circumstances. We compare two different approaches to alarm detection, one based on techniques and representations borrowed from speech recognition, and the other more specifically designed to exploit the structure of alarm sounds and minimize the influence of background interference. In this preliminary work, both approaches achieve similarly poor error rates but with different patterns in response to alarm type and background noise.

1. INTRODUCTION

Informal observation suggests that we are able to identify a particular sound as an alarm even when we have never heard it before, and in spite of significant background noise; indeed, when providing some device with an alarm, audibility and identifiability are key criteria. However, hearing loss can disproportionately affect the perception of alarm sounds, and a device that could reliably recognize such sounds would have many applications, both as a kind of special-purpose hearing aid, and for intelligent systems that need to respond to their acoustic environments.

Since by their nature alarm sounds are intended to be easily detected, we could expect that alarm sound detection is simpler than, say, speech recognition. However, the distinctive characteristics of alarm sounds are not formally defined, and it is not obvious that such sounds do indeed share common characteristics, rather than being learned by listeners as the conjunction of a set of more special-purpose sound types.

The goal of this work is to produce a device to detect alarm sounds ‘in general’ i.e. with an ability to generalize away from the specific examples used in development to be able to cover a reasonable class of real-world alarm sounds. This work is also motivated by an interest in object-based sound analysis, as exemplified by computational auditory scene analysis (CASA [1]): Rather than looking for *global* characteristic of the input sound that can be correlated with the intended classification (the approach of speech recognition), the alternative approach is to first decompose the input sound into distinct acoustic events, in imitation of human auditory perception (in so far as we understand it). Classification is then performed on the properties of these separated events, which should be far less influenced by background sounds than any global feature.

Section 2 reports our observations on the general nature of the class of alarm sounds, and section 3 describes our two different algorithms for alarm sound detection, one based on a speech recognition neural network, and the other based on sinusoid modeling, which should be better able to separate the properties of different components in the sound. Section 4 presents the results of our evaluation and discusses the different properties of the two algorithms. We draw conclusions and suggest future directions in section 5.

2. THE ACOUSTIC PROPERTIES OF ALARM SOUNDS

To investigate whether a set of general characteristics can be enumerated that define the class of ‘alarm sounds’, we collected a small database of examples and performed some analyses to look for common properties. Sounds were collected by searching the web for sound clips of alarms (vetted by ear), and by making new recordings of various alarm sources around the home and office. (The most significant challenge in alarm detection is reliable detection in high background noise levels, so a low level of background noise in the examples was not a significant problem, since the examples were mixed with much higher levels of noise before any testing). The breakdown of this informal but broadly representative collection of 50 examples is shown in table 1.

Alarm class description	*# examples*
Car & truck horns	7
Emergency vehicle sirens	6
Fire alarms/klaxons	5
Door bells & buzzers	4
Mechanical bell telephones	5
Electronic phone rings	10
Alarm clocks (electronic)	5
PDA organizer alarms	6
Smoke alarms	2

Table 1: Alarm sound corpus breakdown.

Figure 1 shows the spectrogram of three examples: a car horn, a klaxon and an electronic phone ring.

Our analysis revealed the following characteristics, visible in the examples:

Energy in 3 kHz region: Alarm sounds typically include frequency components around this frequency, which is close to the region of greatest hearing sensitivity.

Narrow-band, fixed-frequency signals: Almost all alarms are perceived as strongly pitched, and the pitch is very stable. This is manifested by well-defined horizontal energy concentrations in the spectrograms, corresponding to single, unchanging frequencies.

Amplitude modulation: Both mechanical and electronic telephones, as well as the smoke alarms and some other examples, show strong amplitude modulation in the 4-30 Hz range. Modula-

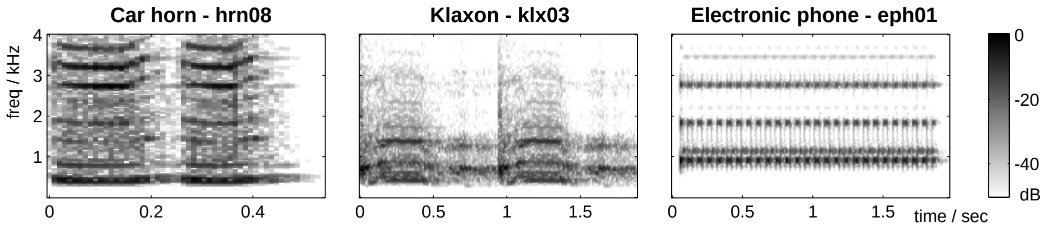

Figure 1: Spectrograms of three examples from the alarm corpus.

tion in this range is associated with perceived 'roughness' in a sound [2].

Abrupt onset, sustained level: In common with many acoustic events, alarm sounds have an abrupt onset. More distinctively, they often have sustained energy at a near-constant level for hundreds of milliseconds, rather than decaying away immediately after the initial transient.

The correspondence between these attributes and some of the different alarms we examined is shown in table 2.

Alarm	*3 kHz region*	*Fixed spectra*	*Amplitude modulation*	*Abrupt, sustained*
Car horns	yes	yes	sometimes	yes
Door bells	yes	yes	no	~ 3 s decay
Bell phone	yes	yes	yes ≈ 20 Hz	yes + slow decay
Electric phone	mostly	yes 1 or 2 tones	yes 8-20 Hz	yes ~ 2 s bursts
Smoke alarm	yes	yes	yes ≈ 4 Hz	yes

Table 2: Common characteristics of alarm sounds**.**

While based on a small set of examples, this investigation revealed some distinctive attributes for the general class of alarm sounds that could support the development of a special-purpose detector able to recognize and perhaps categorize such sounds without being specifically familiar with each example.

3. EXPERIMENTS IN ALARM SOUND DETECTION

To evaluate the feasibility of automatic alarm sound detection, we performed some experiments to measure the accuracy of detecting alarm sounds in high-noise conditions. We are not aware of any existing published work in this area, so a new evaluation task was developed. In order to provide some kind of reference point for our results, we implemented two quite different alarm detection schemes, one based on speech recognition techniques, and one attempting to separate the alarm sounds from the background.

3.1 Alarms-in-noise sound examples

The task was to be simple detection of an alarm sound, rather than any kind of classification or discrimination between alarm sounds. To make this challenging, the alarms were artificially mixed with a variety of background noises; in the current experiments, all the mixes were constructed to have a signal-to-noise ratio of 0 dB.

The 50 alarm sounds described in section 2 were divided into two random subsets of 25 examples each, for detector training and testing respectively. Each of the training examples was combined with 4 different, essentially continuous, background noises, intended to present increasingly difficult 'camouflage' for the alarms. A different but similar set of four different background noises was used for the test examples, as shown in table 3:

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

Table 3: Background noises used in constructing the evaluation examples. "Aurora" noises are drawn from the ETSI Aurora task [3]; speech was recorded at random from the radio; music is taken from a pop CD.

Each sound example consisted of 5 alarm sounds uniformly spaced within 55 seconds of background noise (i.e. 10 seconds of noise around each alarm, with a 5 second lead-in). There were thus five sound examples based on each background noise type, or a total of 20 sound examples containing altogether 100 alarms, for both training and test. Note that there is no overlap in alarms or background noises between the training and test sets. Ground truth transcripts were created that recorded the exact times during which the added alarm examples were within 10 dB of their peak energy; the test system outputs were scored against these transcripts.

3.2 Baseline detector (Neural net)

Our baseline detection results came from a straightforward adaptation of a standard connectionist speech recognizer [4], as illustrated on the left of figure 2. The mixture of alarm and noise is analyzed into 8th order PLP cepstral features, calculated every 10ms over a 25ms frame. These features along with their time-derivatives (deltas) are presented to the input of a feed-forward multi-layer perceptron neural network with a single hidden layer. The network classifies based on the features of five adjacent time frames (about 50 ms of sound), for a total of 90 input units. The network has 100 hidden units and two output units, corresponding to "alarm" and "not alarm".

The network was trained via backpropagation using a minimum-cross-entropy criterion on the 20 training sound examples; the outputs of the trained network may be regarded as estimates of the posterior probability that an alarm sound is present (or absent). For recognition, this probability was median filtered over 11 steps (about 110 ms) and thresholded at 0.5; a transition to above this threshold was taken as a detected alarm sound.

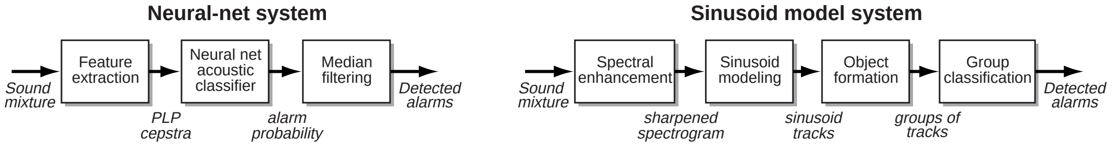

Figure 2: Block diagrams of the two alarm detection systems. Left: baseline system, adapted from a connectionist speech recognizer. Right: Sinusoid model based system.

3.3 Sinusoid model system

A major weakness of the speech-recognizer-derived detector (and of speech recognizers too) is that cepstral features describe the *global* properties of the spectrum at each time slice, and therefore confound the contributions of the alarm 'target' and whatever background noise may be present. As a contrast, we developed a system intended to *separate* the alarm sounds from other sound sources.

The system is based on sinusoid modeling [5], where the sound is represented with a relatively small number of pure tones i.e.

$$\hat{x}(t) = \sum_i a_i(t) \cdot \cos(\omega_i(t)t + \phi_i)$$

where $\hat{x}$ is an approximation consisting of the sinusoids defined by the slowly-varying amplitudes $\{a_i(t)\}$ and frequencies $\{\omega_i(t)\}$ (with initial phases ϕ_i). The rationale behind this approach is that, as observed in section 2, alarm sounds frequently exhibit a sparse and stable spectrum, concentrating their energy into a few spectral locations. If this energy can be picked up by sinusoid modeling, that representation should be largely invariant to the background noise, since it is only the spectral peaks that are being described. As long as these peaks have locally more energy than the background, they are left relatively untouched.

The block diagram of the sinusoid modeling system is shown on the right of figure 2. Since we are concerned only with extracting the sustained and prominent harmonic components arising from alarms, we can improve the detectability of these components by filtering the initial time-frequency energy surface of the spectrogram to enhance horizontal structures (the 'spectral enhancement' stage). Thus, the tracks generated by the sinusoid modeling stage correspond only to components in the original sound with well-defined spectral prominences and with static or slowly-varying frequencies.

Of course, the background noise may contribute pure-tone components that are also extracted by sinusoid modeling. To discriminate between these and true alarms, two further stages are applied. First, in the "object formation" stage, the sinusoid tracks are assembled into groups that are judged to relate to a single source. Tracks are grouped together based on simple heuristics that look for tracks that start and end at about the same time. Specifically, a similarity score is calculated between each pair of tracks as the weighted sum of the squared onset time difference and the squared difference between the ratio of the track durations and unity. If this distance is below a threshold, the tracks are placed in a single group.

The final "Group classification" stage calculates attributes for each group designed to discriminate between alarms and other sounds. We have experimented with a range of different statistics that aim to capture the characteristics described in section 2. The results reported in the next section are based on a combination of two statistics: the **spectral moment** is large for groups consisting of a few sinusoids widely spaced in frequency, which is commonly the case with bells and electronic alarms, although not some others. The **duration-normalized frequency variation** measures the steadiness the frequencies of the sinusoid components relative to their duration. Groups of sinusoids that vary very little in frequency over a long duration are common in many of the alarm sounds. Other statistics we investigated include the **spectral centroid** (known to correlate well with perceived timbre), the **onset time variation** (measuring how closely in time all the tracks in a group started) and several versions of **magnitude variation**, which sought to capture the amplitude modulation seen e.g. in telephone ringing.

The threshold used to discriminate between alarms and nonalarm groups was set by inspection on some of the training examples then used for the test examples. The sinusoid model system did not otherwise take advantage of the training data.

Figure 3 illustrates stages in the sinusoid model alarm detection.

4. RESULTS AND DISCUSSION

Each system was run over the 20 test examples to generate a detection 'transcript' for each, indicating the times and durations of each detected alarm target. For the neural net system, these were the times that the median-filtered alarm probability exceeded 0.5. For the sinusoid model system, these were the enclosing times of the track groups labelled as alarms. These transcripts were scored against the ground-truth by treating any detected target that overlapped with a reference event for more than half of its total dura-

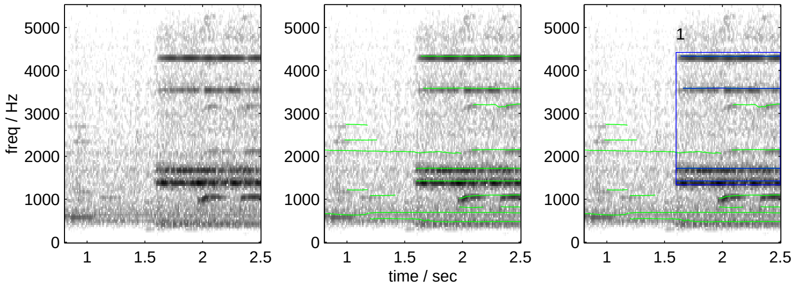

Figure 3: Stages in the sinusoid modeling system. The spectrogram on the left shows the strong horizontal energy components indicating an alarm sound. In the middle panel, the extracted sinusoid tracks are overlaid, finding both alarm and background components. Track group properties allow correct identification of the alarm energy in the third panel.

tion as correct, otherwise it counted as a false alarm (insertion). Any reference event for which no target was detected counted as a false rejection (deletion). Multiple detections were allowed to match against the same reference event since some alarms (ringing phones, tooted car horns) could result in several detected targets. The overall results are shown in table 4:

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%

Table 4: Alarm detection results for both systems, broken down by the background noise conditions (all at SNR = 0 dB). "Del" indicates a missed target (false reject); "Ins" refers to erroneously reported targets (false alarm); "Tot" is the sum, as a percentage of the total true targets.

We see that both systems performed rather poorly in terms of the bottom line result, achieving overall error rates of 192% for the neural net and 197% for the sinusoid modelling approach. Error rates greater than 100% reflect the very large number of insertion errors (false alarms) committed by both systems. Considering only the deletion errors, we see that neural net system is performing significantly better than the sinusoid model system, missing only 22 of 100 alarms, compared to 50 of 100 for the sinusoid system. In part, this is because the object classification criteria used in the sinusoid system covered only a subset of the alarm types, and were unable to detect alarm sounds with significant continuous frequency variation such as sirens and some klaxons.

The two systems also show very different behavior across the different noise types. Noise 1 (airport ambience) proves the easiest background for both systems. Noises 2 and 3 (restaurant babble and speech excerpts) are of broadly similar difficulty for the sinusoid model system, but cause large numbers of insertion errors for the neural net system. We attribute this to the way that the network has used the training set to learn the properties both of alarms and of non-alarms i.e. the backgrounds used in training. When confronted with a different set of backgrounds in the test set, the network is confused; by contrast, the deliberate discarding of background in the sinusoid system makes it relatively insensitive to changes in background. Figure 4 shows some detail of the two systems' outputs for a part of one of these examples.

Noise 4, the music background, proves to be the sinusoid model's nemesis. Because musical notes are also constructed of sustained spectra with stable frequency characteristics, they are frequently mistaken for alarms, leading to a huge number of false alarms. The neural net, however, finds the music a little easier to cope with than noises 2 and 3.

5. CONCLUSIONS AND FUTURE WORK

Based on our small corpus of sound examples, we conclude that both the global-features pattern-recognition approach borrowed from speech recognition and the signal-separation approach based on sinusoid modeling show promise as techniques for automatic alarm detection in high-noise conditions.

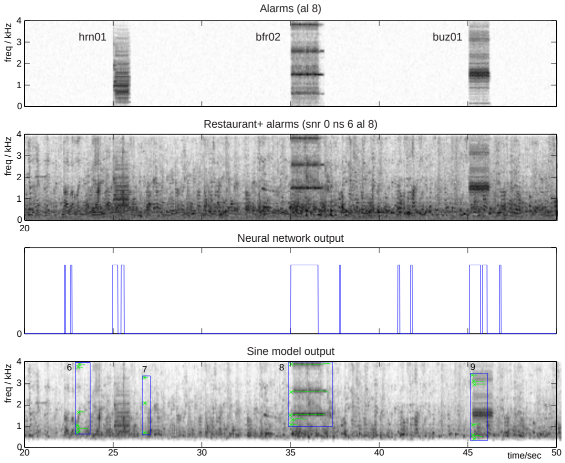

Figure 4: Comparison of systems' performances. Three alarm sounds (top panel) are mixed with the restaurant babble giving the second panel. The filtered alarm probability from the neural network is shown in panel 3 (showing multiple false alarms due to the mismatch between training and test background noise). The bottom panel shows the alarm groups located by the sinusoid modeling system (which fails to detect the horn because it is inconsistent with the hand-defined 'alarm-like' criteria).

False alarms are the largest component in the high error rates we have reported. Training with a wider range of noises might allow the neural net to generalize over more test conditions, although this approach seems inherently limited. Further development of the sinusoid group classification, particularly employing machine learning to allow the exploitation of training data, should dramatically improve the sinusoid modeling approach.

Future work will investigate the trade-off between insertions and deletions, and further characterize the variation of errors with signal-to-noise ratio and background noise type. Recognition of different types within the class of alarms will also be pursued.

ACKNOWLEDGMENTS

Thanks are due to Victor Stein for his support and contributions, including the core of the alarm sound database.Supported in part by the European Commission under the Project Respite (28149).

REFERENCES

[1] M. Cooke & D. Ellis (2001). "The auditory organisation of speech and other sound sources in listeners and computational models", *Speech Communicatio*n, accepted for publication.

[2] B.C.J. Moore (1997). *An introduction to the psychology of hearing*, 4th ed., Academic Press.

[3] D. Pearce (1998). "Aurora Project: Experimental framework for the performance Evaluation of distributed speech recognition front-ends," *ETSI working paper.*

[4] N. Morgan & H. Bourlard (1995). "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach," *Signal Processing Magazine*, 25-42, May 1995.

[5] D. Ellis (2001). "A tutorial on sinusoid modeling in Matlab," http://www.ee.columbia.edu/~dpwe/resources/matlab/