# Unifying Background Models over Complex Audio using Entropy

Simon Moncrieff, Svetha Venkatesh, Geoff West
Department of Computing, Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{simonm,svetha,geoff}@cs.curtin.edu.au

## Abstract

*In this paper we extend an existing audio background modelling technique, leading to a more robust application to complex audio environments. The determination of background audio is used as an initial stage in the analysis of audio for surveillance and monitoring applications. Knowledge of the background serves to highlight unusual or infrequent sounds. An existing modelling approach uses an online, adaptive Gaussian Mixture model technique that uses multiple distributions to model variations in the background. The method used to determine the background distributions of the GMM leads to a failure mode of the existing technique when applied to complex audio. We propose a method incorporating further information, the proximity of distributions determined using entropy, to determine a more complete background model. The method was successful in more robustly modelling the background for complex audio scenes.*

## 1. Introduction

In audio surveillance and monitoring applications, we would like to segment infrequent or unusual sound from the audio signal. Such sounds, considered to be foreground, are relevant to higher level analysis, such as sound event classification. Consequently, a useful first stage of the audio analysis is the detection of the background of the signal, i.e. sounds that dominate the signal. To do this we model the background, which we define as recurring and persistent audio characteristics that dominate a portion of the signal. We classify background audio into two types, *simple*, audio emanating from a single source, and *complex*, audio from multiple superimposed sound sources. For example, in the case of an industrial processing plant, various processing units produce different sounds that combine to form a complex audio background.

Existing learning techniques for determining the background explicitly model background or foreground sounds [5]. Such methods require prior knowledge of the audio. A simple technique involves determining high energy segments of the audio [9], e.g. sound level sensing. Such methods can be unsupervised but lack the sophistication required to model complex background audio. Recently, an online Gaussian Mixture model (GMM) video background modelling technique [8], was adapted to accommodate the processing of audio data [7]. This technique has the advantage of being adaptive and unsupervised, and the use of multiple statistical models to characterise the states of the data enables the method to be applied to more complex data. However, the complex and dynamic nature of audio results in background modelling for audio differing significantly from video. Consequently, complex audio backgrounds form a source of error for the adapted algorithm. Thus, an extension to the underlying theory of background modelling is necessary to more thoroughly account for the complex audio backgrounds.

One failure mode is the fragmentation of the background representation across multiple distributions within the GMM. Typically, the fragmented background representation consists of a distribution of large weight, the dominant distribution, and a number of lower weighted distributions that are similar to the dominant distribution (see fig. 1). Algorithms that solely consider the dominant model to contribute to the background classify the lower weighted similar distributions as foreground. Thus, the fragmented background representation results in a portion of the background not being included in the background model. We propose a method to integrate the fragmented background representation to form a unified background model. To achieve this

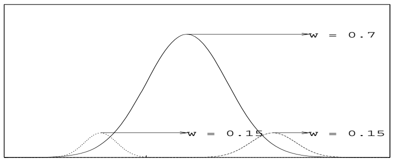

**Figure 1. Mixture of 3 Gaussians**

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

we consider both the dominance and proximity when determining the background model. The proximity is used to cluster similar distributions, and is determined using the probabilistic entropy.

In this paper we extend a background modelling algorithm for application to complex audio environments. The modelling of the background enables the detection of foreground events, which serves to highlight sections of interest within the audio signal and focus higher level analysis. The audio context related to foreground events can itself be used as a tool for analysis, such as for content based browsing. The extended algorithm enables the determination of the background model from a fragmented background representation. This results in a more robust method for modelling complex audio background, accounting more comprehensively for the variability present in the audio data. We test the resulting algorithm on a number of complex data sets that represent instances of monitoring applications.

The layout of this paper is as follows. Section 2 details related work in audio surveillance and monitoring, and the audio background modelling algorithm. Section 3 describes our approach to unifying the background model. Section 4 then details the experimentation exhibiting the improvement in the background modelling over a number of audio data sets.

## 2. Background

### 2.1. Audio Surveillance and Monitoring

Audio analysis methods for surveillance and monitoring include a tele-monitoring system for the detection of sound events such as cries for help [9], and the detection of alarm sounds [5]. Cowling [2] proposed a taxonomy for the classification of environmental sounds for audio surveillance. These methods focus on the detection of specific sound events. In contrast our approach is an initial stage in a framework with which to analyse the global and contextual information within the audio.

### 2.2. Audio Background Modelling

We adapted the video method proposed by Stauffer *et al.* [8] for application to the audio domain [7]. To model the background, the incoming audio signal is segmented into fixed duration audio clips. A number of features are calculated for each clip and combined to form the observed feature vector for the current clip, $X_t$. A single multidimensional GMM, that accounts for dependencies between features, is then used to model the background. The recent history is modelled by a mixture of $K$ Gaussian distributions. The probability of observing $X_t$ is

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}). \qquad (1)$$

The weight of each model, $w_{i,t}$, is related to the proportion of recently observed feature vectors at time $t$ that are accounted for by model $i$. Each $X_t$ is associated with a model within the GMM using on-line K-means approximation, comparing $X_t$ to each model in the GMM. A model represents $X_t$ if $X_t$ is within $P$ standard deviations of its mean. The highest ranking model that represents $X_t$ is selected as the matching model, with models ranked in descending order according to $\omega_i/\sigma_i$. If no match is determined for $X_t$, the model of lowest weight is replaced by a new model with $\mu = X_t$, a high initial variance, and a low initial $\omega$. The GMM is then updated, the weights for the $K$ distributions at time $t$, are

$$\omega_{k,t} = (1 - \alpha_\omega * M_{k,t})\omega_{k,t-1} + \alpha(M_{k,t}), \qquad (2)$$

where $\omega_{k,t}$ is the weight of the $k^{th}$ model at time $t$, and $M_{k,t}$ is 1 for the matched model, and 0 otherwise. The weights are subsequently normalised. The Gaussian distribution parameters for the matched model are updated to reflect $X_t$:

$$\begin{aligned} \mu_t &= (1-\rho)\mu_{t-1} + \rho X_t, \text{ and} && (3)\\ \Sigma_t^{i,j} &= (1-\rho)\Sigma_{t-1}^{i,j} + \rho(X_t^i X_t^j), && (4)\\ \text{where} \quad \rho &= \alpha_g e^{-\frac{1}{2d}(X_t-\mu_{t-1})^T \Sigma_{t-1}^{-1}(X_t-\mu_{t-1})} && (5) \end{aligned}$$

and $\Sigma_t^{i,j}$ is the $(i,j)$ element of the covariance matrix, $X_t^n$ is the $n^{th}$ element of $X_t$ and $d$ is the dimension of $X_t$. The value $\alpha_\omega$ determines the rate of adaptation of a model to the background, and $\alpha_g$ determines the update rate of the Gaussian distribution parameters. Foreground classification is determined using $FG = \sum_{k=K}^{k_{hit}} \omega_k < T$ where $k_{hit}$ is the matched model and $K$ corresponds to the model of lowest rank, with ranking determined in descending order by $w_i$. The remaining models are considered background. The threshold $T$ represents the background classification tolerance, with a lower $T$ resulting in more distributions being regarded as background.

## 3. Unification of Background

### 3.1. Background Fragmentation

As described above, the models of the GMM are ranked according to $w_i/\sigma_i$ when determining the matching model for $X_t$. The $\sigma_i$ acts as a constraint on the model, restricting the growth of the variance by decreasing rank as variance increases. The absence of this constraint leads to the formation of models with high variance, which are less sensitive to foreground events and changes in the background.

This is due to the high degree of variation present in complex audio backgrounds. However, this constraint can result in the fragmentation of the background representation across a number of distributions. An increase in the variance of a model will decrease the model's rank, irrespective of the support, $\omega$, for the model. As a result, a model of lower weight can have a higher rank compared to the dominant model within the background representation. If such a model has a high degree of similarity to the dominant model in the feature vector space, both models will be deemed to represent $X_t$. The model of lower $\omega$ will be selected as the matched model for $X_t$ in preference to the dominant model if the dominant model has a lower rank due to a larger variance. We term the lower weighted matched model a *satellite distribution* (see fig. 1). This results in a background representation that is fragmented, with dominant and satellite distributions present. As the satellite models are of lower weight, considering only the dominance by $\omega_i$ in determining foreground classification results in the satellite models being erroneously classified as foreground. This error is implicit due to the proximity of the satellite and dominant distributions of the background representation.

### 3.2. Background Unification using Entropy

We incorporate the satellite distributions into the background model using a combination of proximity and dominance to determine a unified background representation. We use entropy to determine the similarity between the matched model for $X_t$ and the remaining models of the GMM. The *Information Radius* (*IR*) [6], a symmetric form of the Kullback-Leibler divergence ([1], pg. 18) is used to calculate entropy. The combined weight of the dominant and satellite distributions, $W(E)$, is then used to determine the foreground classification:

$$FG = \left( W(E) + \sum_{k=1}^{K} W_{\notin IR}(k) \right) < T \qquad (6)$$

$$\begin{aligned} \text{where } W(E) &= \sum_{k=1}^{K} W_{IR}(k) \\ W_{IR}(i) &= \begin{cases} \omega_i, & \mathit{IR}(k_{hit}, k_i) < T_{IR} \\ 0, & \text{otherwise.} \end{cases} \\ W_{\notin IR}(i) &= \begin{cases} \omega_i, & W_{IR}(i) = 0 \cap \omega_i < W(E) \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

$IR(i,j)$ denotes the Information Radius metric between two Gaussian distributions $i$ and $j$, calculated as follows [6]

$$IR(i,j) = \frac{1}{2}\left[ D_{kl}\left(i \| avg(i,j)\right) + D_{kl}\left(j \| avg(i,j)\right) \right] \quad (7)$$

where $avg(i,j)$ is the average of the two distributions, formed by averaging the mean and covariance matrices for $i$ and $j$. $D_{kl}$ is the Kullback-Leibler (*KL*) divergence. The threshold $T_{IR}$ is used to determine if two distributions, $i$ and $j$, are similar given $IR(i,j)$. $T_{IR}$ is a property of the Information Radius, and is defined as $T_{IR} = \frac{d\bar{n}^2}{2}$ where $\bar{n}$ is the average number of standard deviations, $\sigma_{avg(i,j)}$, of $\mu_i$ and $\mu_j$ from the mean of the average distribution, $\mu_{avg(i,j)}$. The threshold represents an upper bound for the information radius for a given $\bar{n}$. That is, at least one distribution will be within $2\bar{n}\sigma$ of the other distribution.

## 4. Experimentation

### 4.1. Data

Continuous, unedited audio streams were used as test data, with foreground events present at the time of capture. Three data sets ($44.1kHz$, $16bit$, mono, wave format) of differing levels of audio complexity were used for analysis. The lab data (10.6 minutes) consisted of a simple background recorded in a computer lab. The traffic data (12.1 minutes) consisted of a complex background of traffic noises from a busy road recorded outside. In processing, the majority of the audio was considered to be background. The kitchen data (19.9 minutes) consisted of multiple backgrounds, both simple and complex, recorded in a kitchen environment, with multiple foreground events ($55s$ in total). A fourth data set ($16kHz$), consisting of 195 minutes of industrial noise from a processing plant, was used. The data consisted of a complex background of machinery and wind noise, with foreground events including a jet of steam.

The ground truth for the data sets was defined in terms of the foreground events, short duration events that are meaningful in the context of the surrounding audio. The remaining audio was classed as background. For example, in the context of the kitchen data, sound associated with the kitchen was considered to be foreground (predominantly the result of a user interacting with the environment). Traffic noises such as sirens and car crashes are considered background. However, in the context of traffic, such sound would be foreground. This restricts the concept of foreground to events or activities of semantic interest. This results in the labelling of the ground truth background differing from the algorithmic definition of background, the latter forming a subset of the former. This analysis provides a more accurate indication of the real world performance and usefulness of the algorithm for a given application as the context of the application is accounted for.

### 4.2. Procedure

Each data set was divided into audio clips of duration $t$ ($1.0s$ and $0.25s$). For each clip, $d$ audio features were determined. A $d-D$ GMM was used to model the background of the audio signal, classifying each audio clip in sequence. Two parameters determine the clustering of $X_t$; the number of standard deviations used to determine if a model represents $X_t$, $P$, and the model update parameter $\alpha_g$. The values used for these parameters were optimised for each feature set across all data sets, accounting for the accuracy in modelling the background and sensitivity to changes in the background. A value of $T = 0.5$ was used to enable multiple models to be classed as background. A value of $\bar{n} = 1$ was used to determine similarity, with an $\alpha_\omega$ of $0.01$. Ten distributions were used in the GMM.

### 4.3. Evaluation of Results

The BG/FG classification result for each clip was then compared with the ground truth. The accuracy of the detection of the background clips was calculated according to $BG_{acc} = \frac{TP_{BG}}{N_c - FG_D}$ where $TP_{BG}$ is the number of background clips classified as background, $N_c$ is the total number of clips, and $FG_D$ is the total number of foreground clips correctly detected. A foreground event was considered to have been detected if one or more clips were classified as foreground within the duration of a ground truth foreground event.

### 4.4. Audio Feature Set

Two feature sets were used to encapsulate the characteristics of the audio signal content. The *WE* feature set (7 features) consisted of the mean wavelet energy for seven frequency sub-bands. Wavelet coefficients were constructed, using six levels of decomposition, using the Daubechies wavelet transform [3]. The sum of the absolute values of the wavelet coefficients within each sub-band was averaged by the number of coefficients within the band to calculate the sub-band energies. The *RF* feature set (10 features) consisted of predominantly frequency based features. The features were selected using an attribute selection technique [10] over a number of temporal and frequency domain features, calculated for background audio extracted from the traffic and lab data sets. The feature set consists of selected mel-cepstrum coefficents (MFCC), the mean and standard deviation of the zero crossing rate (ZCR [11]), and the ZCR and mean energy of selected wavelet sub-bands. The MFCCs for each clip were generated using a 25 order MFCC set [4] and averaging corresponding coefficients over the clip.

### 4.5. Results

Figure 2 shows an example of the background modelling process for the first $80$ minutes of the industrial data set for the *RF* set. The figure shows the foreground audio (top) as determined by the algorithm, and the original waveform (bottom). The foreground at the beginning of the sequence corresponds to the algorithm adapting to the background. The plant had an anomaly in sound when a jet of steam was released, resulting in a foreground event surrounded by complex industrial noise at 31 minutes (point A). The detection of the steam event demonstrates the advantage of using audio background modelling over sound level sensing. The $16$ minutes of audio preceding the steam event, background, has a maximum energy of $70.9dB$, and a mean of $58.0dB$. The steam event, $4$ minutes long, has a maximum energy of $66.7dB$, and a mean of $57.9dB$.

Table 1 shows selected results for background classification accuracy and foreground recall, both with the use of entropy information (unified background results), and without (fragemented background results).

**Table 1. Accuracy with and without entropy.**

| Data (s) | Foreground ( % ) | | Background ( % ) | |
|---|---|---|---|---|
| | without | with | without | with |
| Lab WE 0.25 | 100 | 100 | 90.8 | 92.3 |
| Kitchen WE 1 | 100 | 100 | 62.3 | 62.3 |
| Kitchen RF 1 | 94.1 | 94.1 | 78.1 | 83.8 |
| Kitchen WE 0.25 | 100 | 100 | 74.6 | 77.9 |
| Kitchen RF 0.25 | 100 | 100 | 92.3 | 94.1 |
| Traffic WE 1 | - | - | 59.6 | 65.9 |
| Traffic RF 1 | - | - | 79.0 | 81.7 |
| Traffic WE 0.25 | - | - | 69.8 | 83.3 |
| Traffic RF 0.25 | - | - | 93.1 | 93.2 |
| Industrial WE 1 | 99.4 | 99.4 | 86.3 | 86.3 |
| Industrial RF 1 | 95.2 | 95.2 | 99.9 | 99.9 |
| Industrial WE 0.25 | **100** | **100** | **77.3** | **96.7** |
| Industrial RF 0.25 | 95 | 94.8 | 99.4 | 99.7 |

### 4.6. Analysis

Due to the stability of the simple background of the lab set, the use of the entropy offers little advantage. The low variance of the data set results in a single distribution being sufficient to model the background, with few satellite models present in most cases.The presence of satellite models is linked to the variability of both the data set, and the feature set with respect to the data set, which results in the background fragmentation. This is observed in the marginal increase in performance for the lab data.

For the background accuracy, the use of entropy leads to an overall improvement in the performance, with the most

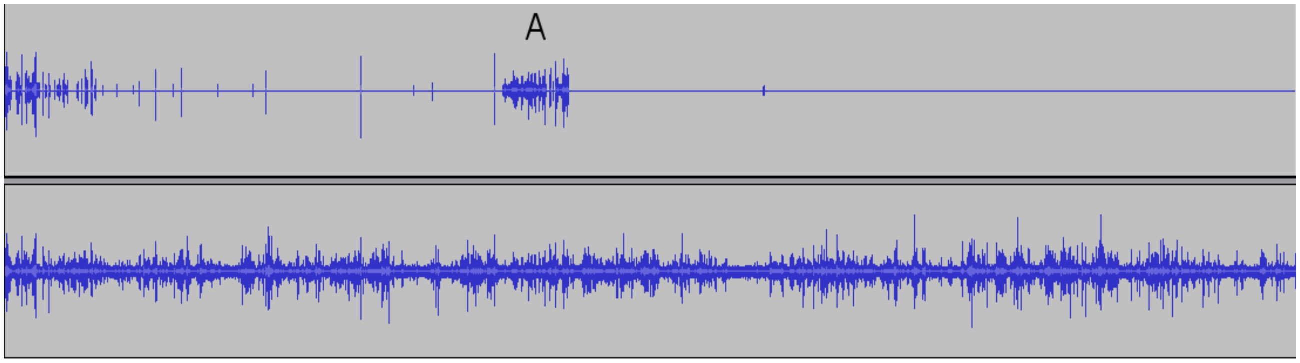

**Figure 2. Industrial data foreground audio.**

significant gain in performance occurring for the more complex data, e.g. the industrial data set. A second result evident from table 1 is the greater increase in performance for the *WE* set in comparison with the *RF* set, and the $0.25s$ clip size compared to the $1s$ clip size. These results justify our motivation for using entropy as the improved performance is due to the incorporation of satellite models into the background model. The foreground detection accuracy was unaffected for the data sets examined. The exception is the *RF* set for the industrial data at $0.25s$, which is attributed to the reduced sensitivity to the start and end times for foreground events due to the averaging of features over clips.

The addition of entropy results in the *WE* becoming a viable feature set, particularly at the higher clip resolution ($0.25s$). The improved detection of foreground events for the *WE* set is necessary for certain applications. While the *RF* set has a high background accuracy for the industrial data set, the high foreground detection of the *WE* is required for applications such as hazard detection, so that no hazards are missed. Thus, the combination of entropy and *WE* calculated at $0.25s$ is an appropriate solution for this problem. Furthermore, a decrease in clip size increases sensitivity to short duration foreground events.

## 5. Conclusion

This paper proposes improvements to a previous approach to audio background modelling. We solve the misclassification of non-dominant background distributions by considering the dominance of a cluster of distributions, grouped by similarity, to determine the unified background model. Similarity was determined as a property of the clustering phase of the algorithm, matching distributions in a similar manner to the matching of an observation to a distribution in the GMM. This method was successful in combining fragmented background representations where present. This increased the robustness of the background modelling, particularly with respect to more complex audio data, and variability encapsulated in the audio features and the analysis resolution.

## References

[1] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.

[2] M. Cowling and R. Sitte. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895–2907, 2003.

[3] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.

[4] J. R. Deller, J. G. Proakis, and J. H. Hansen. *Discrete-Time Processing of Speech Signals*. Maxwell Macmillan International, 1993.

[5] D. P. W. Ellis. Detecting alarm sounds. In *Consistent and Reliable Acoustic Cues for sound analysis*, Aalborg, Denmark, 2001.

[6] L. Lee. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, 1999.

[7] S. Moncrieff, S. Venkatesh and G. West. Persistent audio modelling for background determination. In *IEEE International Conference on Multimedia and Expo (ICME 2005)*, Amsterdam, Netherlands, 2005.

[8] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, pages 246–252, Fort Collins, CO USA, 1999.

[9] M. Vacher, D. Istrate, L. Besacier, J. F. Serignat, and E. Castelli. Life sounds extraction and classification in noisy environment. In *5th IASTED-SIP*, Hawaii, 2003. ACTA Press, Calgary.

[10] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, Morgan Kaufmann, 2000.

[11] T. Zhang and C.-C. Jay Kuo. Hierarchical classification of audio data for archiving and retrieving. In *IEEE International Conference On Acoustics, Speech, and Signal Processing*, volume 6, pages 3001–3004, Mar. 1999. Phoenix.