Privacy-aware environmental sound classification for indoor human activity recognition

Wei Wang*
w.wang-1@utwente.nl
University of Twente
Enschede, The Netherlands

Nirvana Meratnia
n.meratnia@utwente.nl
University of Twente
Enschede, The Netherlands

Fatjon Seraj
f.seraj@utwente.nl
University of Twente
Enschede, The Netherlands

Paul J.M. Havinga
p.j.m.havinga@utwente.nl
University of Twente
Enschede, The Netherlands

ABSTRACT

This paper presents a comparative study on different feature extraction and machine learning techniques for indoor environmental sound classification. Compared to outdoor environmental sound classification systems, indoor systems need to pay special attention to power consumption and privacy. We consider feature calculation complexity, classification accuracy and privacy as evaluation metrics. To ensure privacy, we strip voice bands from sound input to make human conversations unrecognizable. With 5 classes of 2500 indoor audio events as input, our experimental results show that using SVM model with LPCC feature, 78% classification accuracy can be reached. Furthermore, the performance is improved to more than 85% by combining several simple features and dropping unreliable predictions, which only slightly increase the complexity.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; • **Applied computing** → **Sound and music computing**; • **Human-centered computing** → *Ubiquitous and mobile computing theory, concepts and paradigms*.

KEYWORDS

Smart Buildings, Privacy-aware Environmental Sound Recognition, Voice Bands Stripping, Internet Of Things, Computational Efficiency, Web Crawling, Mel Frequency Cepstral Coefficients, Linear Predictive Cepstral Coefficients, Support Vector Machine

ACM Reference Format:

Wei Wang, Fatjon Seraj, Nirvana Meratnia, and Paul J.M. Havinga. 2019. Privacy-aware environmental sound classification for indoor human activity recognition. In *Proceedings of PETRA '19, (Conference'19)*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3316782.3321521>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'19, June 5–7, 2019, Rhodes, Greece,
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6232-0/19/06...\$15.00
<https://doi.org/10.1145/3316782.3321521>

1 INTRODUCTION

Recent studies show that nowadays people spend in average 80%-90% time indoor and the indoor environments have large effect on the comfort, health and productivity of the occupants [4]. A new concept of "smart buildings" therefore has emerged, referring to building automation systems (BAS) with automatic centralized control for heating, ventilation and air conditioning, lighting, etc. Smart buildings are designed to improve occupant comfort, efficient operation of building systems, reducing energy consumption and operating costs as well as to improve the life cycle of utilities [10]. Energy efficiency impacts positively the portfolio of both buildings owners and utility companies, with benefits ranging from reduction in utility bills to grid stability and sustainability [10]. In order to achieve "the smartness", the building first needs sensors to collect wide and relevant ambient information. These data then need a fully connected network to be gathered. Finally an intelligent control system will learn from these data and react to environmental change automatically. Among all the needed information, data about occupants is the key component, as human comfort is the reason these systems exist in the first place. Thus knowing the human activity, location and mobility patterns inside the building can substantially increase the effectiveness of such solutions. This is particularly important because knowing of the activity of a person, more information can be provided to the building management systems on how to operate more efficiently.

Numerous technologies are used to automatically detect human presences and activities. Such technologies include camera systems that recognize human presence and actions from images, wearable devices that track users movement, PIR sensors that give binary information about people presence [21][18][25]. As many papers have proposed, these sensors have pros and cons in different aspects. For instance, PIR sensors are cheap, non-intrusive but only detect the presence of stationary people [16]. A wearable device such like a badge needs carefully maintenance and often let users feel it troublesome [2]. Camera or infra-red camera systems can recognize human activities in fine grain but they are expensive and need line-of-sight, more importantly many people worry that their privacy are exposed [14][26].

Another type of sensors worth mentioning is the acoustic sensors. Compared to the aforementioned sensors, they offer many advantages such as being pervasive and rich information medium

while cheap to deploy. Beside the rich information that sound contains, there are also obvious challenges such as noise interference and high attenuation. To apply sound in indoor applications, sound sensors also introduce some critical privacy concerns. People could feel stalked and spied on when microphones are around them. Although many research initiatives have focused on environmental sound recognition for general purposes, very few are specific for human activity detection. There is still a gap between general sound classification techniques and human activity specific techniques inside buildings, e.g. how could privacy concern being preserved and how to make the model lightweight to fit in IoT devices. In this paper, we will fill in this gap, by exploring the possibilities of accurately using sound sensors to recognize human activity in privacy.

Our idea to preserve privacy is to strip voice bands from input audio stream, as human voice largely falls into the range of 80HZ - 3KHZ while environmental sound events have much wider range. We achieve this by using a hardware band-pass acoustic sensor that omits human voice frequencies in the device layer. Being a challenging task, we conduct comparative experiments to find the best suitable model and features that provide presence information in absence of human voice bands. We make use of machine learning techniques to classify different human activity classes based on sound. Both the classification accuracy and calculation density results are given in the evaluation. We also compute the highest rating model that can be used by the IoT devices with sound capable functions to automatically detect ambient human behaviors in real time.

The remainder of the paper is organized as follows: Section 2 gives an overview of the related works on environmental sound recognition. Section 3 describes the methodology used to define the characteristics, extract features and model the candidate sound event classifiers. Section 4 describes the experimental and evaluation phase of this work. We conclude this paper with our open discussions in Section 5.

2 RELATED WORK

Classification of human activities based on sound falls under category of '*Environmental Sound Recognition*' (ESR). ESR aims to automatically detect and identify audio events from captured audio signal. Compared to other areas in audio presence such as speech or music recognition, general environmental sound recognition has received less attention and is less effective because the sound input is unstructured and arbitrary in size and pattern.

Eronen et al. [11] developed an environmental context classification system based on audio. With low-order hidden Markov models and MFCC (Mel Frequency Cepstral Coefficients) features, they achieved an accuracy of 82% for 6 classes, which dropped to 58% when number of classes was increased to 24. Cai et al. [6] proposed a key audio effect detection framework for context inference, where movies are used as the dataset. The advantage of their model over others is to first infer the context by carefully picked and distinguishable key sounds events (like explosion and gun shot means an action movie). However these 'key events' need manually and carefully chosen and is hardly scaleable. Chu et al. [7] proposed an environmental sound recognition model to classify context, with

Wei Wang, Fatjon Seraj, Nirvana Meratnia, and Paul J.M. Havinga

continuous audio recordings from 14 different scenarios. Their innovation was a matching pursuit algorithm to extract features from time-frequency domain to complement MFCC. Compared to previous works, the rich features used helped improved the classification accuracy. Heittola et al. [17] developed a context-dependent sound event detection system. With context label which provided prior probabilities to HMM model to improve event detection accuracy. The system also adopted an iterative algorithm based on original HMM to model the overlapping sound events.

The aforementioned papers mainly focus on classifying the sound context but not the events. Later people use the features and models from these works to solve another problem, where the focus is the shorter duration of an audio or so called events. From data processing perspective, algorithms presented in these papers normally cut audio streaming into fixed length frames (millisecond level) and build models with feature arrays extracted from them. This is because the time-varying characteristic of audio signal, short frame based features can approximate time invariant functions and represent the details well.

Cowling et al. [8] compared several feature extraction and classification techniques typically used in speech recognition for general environmental sound recognition. Particularly, tempo-frequency features such as STFT (Short Time Fourier Transformation) were found to perform better than stationary MFCC features. Dalibor et al. [23] compiled a survey on the audio feature extraction techniques for audio retrieval systems. They proposed a taxonomy of audio features together with the relevant application domain. Temporal domain, frequency domain, Mel-frequency and harmonic features were the most popular features in environmental sound recognition. Tran et al. [30] proposed a probabilistic distance SVM (Support Vector Machines) for online sound event classification. This method has comparably good performance with very short latency and low computation cost compared to other mentioned approaches. STE (Short Time Energy) was the only feature used in their model since in real online applications, the sound power is highly related to the distance of the sound source, using this feature alone could be highly biased in sound recognition task. Piczak et al. [24] applied CNN (Convolutional Neural Network), a method usually applied to image processing, to classify environmental sound and achieved results comparable with traditional methods (64.5% accuracy for 10 classes). Adavanne et al. [3] used multi-channel microphones and RNN (Recurrent Neural Networks) for sound event detection, which improved the accuracy on detecting overlapping sound events. Lopatka et al. [20] proposed an acoustic surveillance system to recognize the acoustic events for possible threats like gunshot and explosion. The accuracy was not so encouraging compared to state of the art, but the advantage is the low-cost model and realtime processing capacity. Dong et al. [9] used sensor fusion on PIR, CO₂, temperature data combined with acoustic sensors to detect and predict user occupancy patterns. They used only sound volume and no further exploration of other sound-based feature. Yaniv et al. [33] used a combination of sound and floor vibrations to automatically detect people fall events to monitor elderly people living alone. Sound event length, energy, and MFCC features were extracted and used for event classification. Table 1 shows the summary of the related works.

Table 1: Features and Methods in Related Works

References	Scenario /Data Source	Event /Context	Platform	Online /oOffline	Features	Models
[11]	General	C	None	F	MFCC	HMM
[6]	Movie	C	None	F	Spectral Flux,Harmonicity,MFCC	SVM, BN
[7]	General	C	None	F	MFCC,temporal signatures,time-frequency	SVM, BN
[17]	Transport,Shop,Open Space	C&E	None	F	MFCC	GMM,HMM
[3]	General	E	None	F	MFCC,Pitch,TDOA	RNN
[8]	General	E	None	F	FFT subband,STFT subband,MFCC	ANN, GMM
[24]	General Urban	E	None	F	MFCC	CNN
[9]	Conference Room	E	SN	O	Volume	SMM
[33]	People Falling	E	None	O	Length,STE,MFCC	HMM
[29]	General	E	None	O	STE	SVM

C = Context, E = Event, SN = Sensor Network, BN = Bayes Network, SMM = Semi-Markov Model, GMM = Gaussian Mixture Model

Figure 1. audio recognition system

In General, many models have provided good results in sound events recognition problem, especially deep learning based models have shown great potentials in this field [3][24]. However to use sound in smart building applications there are still several drawbacks. One biggest problem is that sound may expose more privacy information of occupants than cameras. Another drawback of deep learning based model is the lack of huge training dataset. Compared to the popular speech recognition problem, environmental sound recognition problem is less attractive to researchers while the data source is more diverse than speech. What is more, the models also need to be light weighted in order to operate in IoT devices. Our work will fill in these gaps and provide solutions to these problems. In our work, we first strip the voice bands of audio stream as human conversations are the most concerned privacy issue. Our training data is crawled from multiple free audio websites so that the data source is diverse enough for model training. Regarding to the models, we mainly investigate the low cost classifiers and efficient feature combinations to improve the performance at low computation complexity.

3 METHODOLOGY

Our overall aim is to build an algorithm capable of classifying human activities based on environmental sound in real-time. The methodology is shown in Figure 1. The raw audio stream is first segmented into short events, from which important features are extracted, and finally each event will be classified into an activity class using a classification model.

3.1 Preprocessing

3.1.1 Segmentation. In this step, the duration of events are extracted from continuous audio streaming from background noise (or silence). The algorithm to extract these events is called segmentation. Although segmentation is as important as other steps in our algorithm, it is not the focus of this paper since numerous general approaches exist to tackle the problem already exist. Here we provide one simple segmentation approach, which works as follows: (1) The audio stream is smoothed in time domain, and cut into fixed short frames (20ms) while the power is calculated for each frame. (2) The frames with power higher than a threshold are labeled as 'active'. The threshold can be preset static value or dynamic adjusting. (3) Adjacent 'active' frames are combined to form an event. (4) Events shorter than a given duration are dropped, while the long frames are truncated such that the events are between 1 to 3s. The reason to choose this duration range is from practical experience, as human can identify sound segments with such length quite well. Figure 2 shows an example of the segmentation results.

3.1.2 Voice bands stripping. The purpose of stripping human voice is to preserve privacy, as human conversations are the one of the most critical privacy concerns in indoor environments. Here we implement a software band-stop filter to eliminate the human voice while in a real world implementation this can be implemented by acoustic sensor physically, so that the privacy is protected by the device layer. A typical band-stop filter can achieve this function, which let pass only the bands from zero up to its lower cut-off frequency F_{low} and the bands above its upper cut-off frequency F_{high} . This results in a rejection of frequencies in the band $F_{low}:F_{high}$.

Figure 2. An Example of Audio Segmentation

In our case $F_{low} = 300\text{Hz}$ and $F_{high} = 3\text{KHZ}$, this range is often referred to as the voice bands [31]. By filtering out this band the human speech content become unrecognizable.

Figure 3 shows the signal before and after voice-bands truncation in time and frequency domain. Obviously many information is removed to preserve privacy, resulting in a more challenging classification problem.

3.2 Feature Extraction

Audio data is a continuous stream of high sampling rate information. This stream of continuous data can be transformed into a reduced set of features, which contain the most important and the most relevant information for the classification task. To provide an intuitive impression, we choose a representative sample from each class and plot the signal from different domain, as shown in Figure 6 [23].

A single feature from a single domain only represents limited information, thus it will be hard to classify. However, by combining multiple features, the class characterization becomes conspicuous. In this paper, we first cut audio stream to smaller frames of fixed length and partly overlapped (e.g. 20ms frame length with 10ms overlap). We then calculate the statistics (mean and variance) of all frame features as the representation of the whole event. There are several benefits for the aggregation of short frames: firstly the features extracted from any event have the same length, i.e. irrelevant to event duration, secondly the statistics stand for global representation which treat the event as a whole. Figure 7 shows the feature extraction flow.

Figure 4 and 5 show the box-plot of per-frame value and the scatter-plot of statistic values of features: 'spectral-spread'. At the first glance, basically no conclusion can be drawn from a single frame feature value, but the statistic feature of an event shows much clearer pattern and would be easier to classify.

3.2.1 Features Lists. While there are numerous audio features, we mainly select features that prove to be highly efficient in audio

recognition tasks. We also choose features from different domains and with different characters since machine learning works better by combining features with low correlation. The features used in our paper are listed in Table 2, including temporal-domain, frequency-domain, mel-frequency and other audio features.

3.2.2 Temporal domain features. The temporal domain is the native domain of audio signals. The features are extracted directly from raw signal without any pre-processing. Consequently, the computational complexity and delay tends to be low. We select the following features from temporal-domain to describe different aspects of signal:

- (i) **Short-time energy (STE)** is a widely used feature in audio analysis which describes the energy of signal, calculated by mean-square of signal per frame [23].
- (ii) **Zero crossing rate (ZCR)** is the rate at which the signal changes from positive to negative or vice versa. This feature is one of the simplest and widely used in speech recognition, which characterizes the dominant frequency of signal [19].
- (iii) **Temporal entropy (TE)** is the entropy of temporal domain per frame, which characterizes the dispersal of acoustic energy [28].

3.2.3 Frequency domain features.

- (i) **Spectral centroid** is calculated as the weighted mean of the frequencies while magnitudes are the weights. It indicates where the "center of mass" of the spectrum is located. Sometimes the median is used for "center" rather than mean [23].
- (ii) **Spectral spread** is the magnitude-weighted average of the differences between the spectral components and the spectral centroid, together they describe how disperse and wide the frequency bands are [23].
- (iii) **Spectral entropy** is calculated as the entropy of spectrum it reflects the flatness but spectrum [28].
- (iv) **Spectral flux** also describes the flatness in spectral domain, but across frames. It calculates as the 2-norm Euclidean distance between the power spectrum of adjacent frames [28].
- (v) **Spectral rolloff** represents the point where N% power are concentrated below that frequency. Spectral rolloff is extensively used in music information retrieval [23].

3.2.4 Other features.

- (i) **Mel-frequency cepstral coefficients (MFCC)** is the cepstral representation of Mel-Frequency. Compared to original linear frequency bands, Mel-Frequency are equally spaced on the Mel scale, which approximates the human auditory system's response. MFCC describes the spectral envelope and is commonly used in speech recognition [7].
- (ii) **Linear Predictive Coding (LPC)** is an auto-regression model in which a predictor estimates a sample by linear combination of previous sample values (1), s is the signal sequence and a_1 to a_p are the coefficients. LPC can be used in audio compression and speech recognition to represent the spectral envelope. However, LPC is prone to disturbance where small changes in LPC value would cause large deviation in

Figure 3. Voice bands truncation

Figure 4. Feature 'frequency-spread' on frame basis

Figure 5. Feature 'frequency-spread' statistics on event basis

Figure 6. An example from each event class

Figure 7. Feature extraction flow

spectrum [23].

$$s[n] \approx a_1 * s[n-1] + a_2 * s[n-2] + \dots + a_p * s[n-p] \quad (1)$$

- (iii) **Linear predictive cepstral coefficients (LPCC)** is the cepstral representation of LPC [23].
- (iv) **Line Spectrum Frequencies (LSF)** is the roots of 2 polynomials decomposed from the LPC polynomial [23].
- (v) **Chromagram** is a well-established tool for processing and analyzing music data which capture harmonic and pitch characters from sound. Apart from music application, Chroma features are also powerful mid-level feature representations in content-based audio retrieval or audio matching [23].

Because both LPCC and LSF are derived from LPC, they are alternative representation of LPC, thus containing the same amount of information. However for LPC, a small disturbance from input

Table 2: Used feature list

Domain transformation			
Nr.	Temporal	Frequency	Mel & Linear predictive
1	ZCE	centroid	MFCC
2	STE	medium	LPCC
3	TE	spread	LSF
4		entropy	Chromogram
5		flux	
6		rollof	

may incur a big difference to output, while LPCC and LSF are more robust in this aspect, which make them better for the classification task.

3.3 Classification

After features have been extracted, a classifier should be learned so that new coming data can be classified. This classification function which is also called representation, has different forms for each model, e.g. a search tree in decision tree or a hyperplane in SVM.

Learning is therefore a process performed by searching through the representation space to find the best hypothesis that better fits the solution.

Classification algorithms can be categorized into two types [5]: (i) stateless algorithms, in which events to be classified are treated as unrelated to each other and (ii) stateful algorithms, in which the algorithms treat the events as related to each other and put them into a context while updating the memory. A stateful model works best in the scenario where the output is not only decided by current input, but also by previous input in timeline (state). Stateless models such as GMM, SVM, Random Forest and Neural Network are widely used in sound events classification, while in context classification, stateful models such as Bayes Networks, HMM and RNN are more preferred. In our task, we think there is no need to remember the past information, since it's enough for human being to tell what's happening when a sound is event heard, even without much context.

While there might be a plethora of machine learning algorithm variations with different type of representation space, the effectiveness for different scenarios can only empirically be confirmed. Hence, we select several commonly used algorithms to conduct an empirical comparison and find the best candidate for the problem. The chosen models in this paper are listed below, all of which are stateless models:

- (1) Decision tree
- (2) Random Forest
- (3) Mixed Gaussian
- (4) Naive Bayes
- (5) SVM(Linear & RBF kernel)
- (6) Artificial Neural Network

4 EXPERIMENTAL EVALUATION

In this section, we present the experimental results over our labeled audio dataset. Discussions are given to gain instructive insights about how to build real-time indoor environmental sound recognition applications. The entire dataset is split into 2 parts: training and test set. For the training sets, 5 folds cross-validation is used to build the best fitting model, which is subsequently applied on the test set for validation.

4.1 Dataset Preparation

In order to build a general human activity classification model using sound, we first prepare data sets from different indoor environments. The audio data must be about indoor sound events and highly relevant for human activities and location. Unfortunately we could not find databases that perfectly match these objectives. Most famous environmental sound dataset are for very general purposes, for instance the urban sound dataset from NYU [27]. Part of our data comes from TUT [22], which contains labeled sound events recorded from streets and residential houses. Other part of our data come from online (www.freesound.org, www.freesfx.co.uk)[12][1], as shown in 3. As these audio stream have different formats and sampling rates, we first need to convert them into a unified form, i.e mono, 44.1KHZ and '.wav' format with the same average volume. Through the segmentation algorithm, these long audio stream were further segmented to short sound events. Initially we selected 9

Wei Wang, Fatjon Seraj, Nirvana Meratnia, and Paul J.M. Havinga

classes of sound events including: Speech, Crowd Chatting, Cheering, Walking, Applause, Door slamming, Chair moving, Elevator, Trolley. However, after looking into the dataset, we merged some classes resulting in 5 classes mainly because of 2 reasons: (I) Some objects make very similar sounds, which is hard to differentiate even by human. For example some door and chair sometimes can make similar crisp and sharp sound, (II) some sounds always appear simultaneously or very closely in certain scenarios, for example in a party when there is cheering, there is always applause accompanied with. Since these sounds are hard to separate, we merge them together into one class. After merging into 5 classes, the data samples distribution is shown in Figure 8, 150 samples from each class are chosen for training for balance, the rest are used for testing.

Table 3: Data Source

Ref	Scenario/ Data source used in this paper	Format	SR (khz)	Length (hour)
[22]	residential house	wav	44.1	24
[12]	random sound rec.	wav, mp3, aiff	22,44.1	20
[1]	random sound eff.	wav, mp3	22,44.1	5

Figure 8. Data samples number of each class

4.2 Feature Comparison

In this subsection, we compare different feature extraction techniques from performance and complexity perspectives. The performance is represented by the overall classification accuracy and F1-score, and the complexity is calculated as the ratio of: feature extraction time divided by audio duration, which should be as short as possible for realtime processing. Through all the comparative experiments, we choose SVM (with RBF kernel) model together with feature MFCC as the baseline. This baseline function has been used in both environmental sound classification and sound context classification and has achieved good results [23][30].

Figure 9 shows the performance of each single feature with and without voice bands. The MFCC of the baseline is the best before

Figure 9. Performance and complexity of single features

voice bands truncation, while LPCC is the best after truncation. These results show a significant performance drop of MFCC after voice bands truncation (from 82% to 77%), which is reasonable because MFCC is designed for human speech recognition as such it provides better resolution in lower frequency bands. While LPCC portraits the smoothed spectral envelope for entire frequency bands without discrimination, it works nearly good with or without voice bands.

In machine learning, multiple features could be combined together to reach better performances, as more features provide more information. However, this does not mean the more features the better, with a fixed number of training samples. With the predictive power would normally first increase as feature numbers goes up then decreases[13], mainly because duplicated and irrelevant features only introduce noise to the model. In offline audio classification, this is normally achieved by techniques like PCA (Principle Component Analysis) to first trim redundant information. However, in real time applications, that need to be efficient, selection of features should be proactively decided.

While brute force all the combinations is too time costing, we use heuristic greedy search to find the local optimum instead. It starts with selecting the best single feature, then at each iteration one more feature is added depending on the classification accuracy. Table 10 shows results of each iteration of the greedy algorithm for both with and without voice truncation, added with the all-features result for comparison. In both cases, combination of features can improve the performance significantly. Looking into the voice-bands truncated case, the best combination contains 4 features: LPCC, spectral-flux, STE and time-entropy. This result matches our expectations, as these features portrait different perspectives of the audio signals. Not surprisingly, using all-features for model input is not preferred, as high dimension data will cause overfitting problem and is much less efficient. The calculation complexity of features combination is smaller than the simple accumulation of each since they share some steps such as FFT for all frequency domain features.

Figure 10. Greedy Search For Feature Combination

In speech recognition, common techniques such as STFT are used to extract spectrum from overlapping frames, normally with half or 3/4 frame length overlapped. This is because audio signals especially speech are highly time varying, and need more sophisticated analysis to reflect the details. However using overlapping frames also duplicate or quadruple the complexity. We compared the results from different overlapping frame length, with the best features combination found above. Figure 11 shows the results with different overlapping frame lengths. Results show that it's best to use half frame length overlapping, which leads to approximately the same result as 3/4 overlapping, but is much more efficient.

4.3 Comparison of classification models

We also compare multiple classification models with the same input features. In total we adopted 6 stateless models, using LPCC feature in input.

The performance of classifiers is shown in figure 12. Results show that SVM(RBF kernel) is the best, and the second best is ANN. We used one hidden layer with 50 neurons for the ANN model based on the criterion in [15]. It is likely that our training data was not large enough, which has led to slightly worse results compared to SVM.

4.4 Confidence of classification results

In our application, knowing the classification results is not enough, we also need to score how reliable the results are. The high score classification results are kept, while the low score ones will be discarded. This is because there are a lot of noise and irrelevant events in the environment, which are hard to be filtered beforehand. However, they are likely to get low score in the classification model, since these sound carry very different features. On the other hand, it's acceptable if some interested events are discarded by mistake as knowing part of the events is enough for detecting human behaviours. Classification algorithms such as Naive-Bayes give both classification result and the probability, i.e. a degree of certainty

Figure 11. Frame overlapping test

Figure 12. Performance of Classifiers

Figure 13. Confidence of prediction

about the result, which just corresponds to our needed score. However, SVM algorithm does not provide the prediction probability directly, it only gives the support vectors, in which larger margin means more confident. Hence, we need to calibrate the support vectors into prediction probability, i.e. the score.

We use the algorithm from Wu [32] to calibrate the SVM support vectors into prediction probability. This algorithm is the multiclass version of Platt Scaling: applying logistic regression on the SVM's scores, fit by an additional cross-validation on disjoint training data. However, normally this method needs a large dataset (1000 samples for each class) to work well, otherwise the probability estimated may be inconsistent with the classification result. Figure 13 shows the real accuracy versus the calibrated probability of our classification results, where the green bars show the samples distribution. The real accuracy and prediction probability are roughly in positive proportion, which means the higher the score the more accurate the results. The violation of monotonic increasing of prediction accuracy could be from multiple reasons: our dataset is not big enough and the calibration method is not perfect. This result shows that nearly half the results are given 90% probability, whose real accuracy is at the same level.

Figure 14 shows the confusion matrix of our best model for training and test set. The 'applause' and 'crowd' are the best two classes in terms of accuracy. 'Door' and 'footsteps' are most likely to be mistaken for each other, which makes sense, since one single footstep and door slapping may sound similar. This mistake is mostly like to happen when people walk very slowly. Another finding is that 'speech' doesn't perform in the test as well as in the training, partly because speech sound characters are more variable and complicated compared to others. So 150 samples are not enough for training, partly because the voice bands is truncated which harms most this class.

5 OPEN DISCUSSION

In order recognize indoor human activities using acoustic sensors, performance is not the only concern, efficiency and privacy are equally important considerations. Our results show that with voice bands stripped off for the privacy concern, the system could still detect human activities with quite high accuracy (86%). In order to find the best features for this task, we first conducted a comparative test for all single features. Results show that LPCC performs best

Figure 14. Confusion matrix of training and testing dataset

for non-voice, and MFCC performs best for full signal bands. These results match our expectation that MFCC is designed for speech recognition and voice bands details. We also tested combinations of features in order to further improve the accuracy. In a greedy searching experiment, LPCC together with one frequency domain feature(flux) and two time-domain features(STE, entropy) perform best, since they portrait different aspects of signal: LPCC and 'flux' each portraits the static and dynamic characters of frequency domain, while STE and entropy do the similar job in time-domain. The third feature test is about the frame overlapping, the trade-off is between complexity and performance. In our experiment, half the frame length overlapping (0.02s frame length, 0.01s overlapping) is the best.

The classification process show that SVM performs the best, and the prediction probability can be calibrated through Platt Scaling algorithm to filter out unreliable results. If we only keep half of the best predictions, the accuracy increases to more than 90%. However, the real accuracy is not monotonically increasing when the prediction probability increases. We think it is because our dataset is not sufficiently large for the Platt Scaling algorithm to work well.

The most misclassified classes were 'door' and 'footsteps', this happens when people walk very slowly, our segmentation algorithm sometimes falsely split a series of footsteps into smaller events other than considering them as a whole. While in real applications,

we could differentiate the two classes with the help of other methods, for example with sound localization algorithm since door could not move. Even though our model is not perfect, we think this accuracy makes it competent for a real indoor events recognition system. Also our model is proved to be general since the dataset is from different context and different sound sources of high diversities. In real applications the accuracy could be higher, since the microphone embedded devices are normally fixed installed, where the sound source should be more homogeneous and easier to classify.

6 ACKNOWLEDGEMENT

This work is a part of the COPAS (Cooperating Objects for Privacy Aware Smart public buildings) project.

REFERENCES

- [1] [n. d.]. freeSFX.co.uk - FREESFX.CO.UK CONTENT PUBLISHER LICENCE AGREEMENT. ([n. d.]). <https://www.freesfx.co.uk>
- [2] G. Acampora, D. J. Cook, P. Rashidi, and A. V. Vasilakos. [n. d.]. A Survey on Ambient Intelligence in Healthcare. 101, 12 ([n. d.]), 2470–2494. <https://doi.org/10.1109/JPROC.2013.2262913> 3.
- [3] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, and Tuomas Virtanen. 2017. Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features. *arXiv:1706.02293 [cs]* (2017). <http://arxiv.org/abs/1706.02293>
- [4] Mohammed Arif, Martha Katafygiotou, Ahmed Mazroei, Amit Kaushik, Esam Elarrag, et al. 2016. Impact of indoor environmental quality on occupant well-being and comfort: A review of the literature. *International Journal of Sustainable Built Environment* 5, 1 (2016), 1–11.
- [5] Christopher M. Bishop. 2006. *Pattern recognition and machine learning*. Springer, New York.
- [6] R. Cai, Lie Lu, A. Hanjalic, Hong-Jiang Zhang, and Lian-Hong Cai. 2006. A flexible framework for key audio effects detection and auditory context inference. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 3 (2006), 1026–1039. <https://doi.org/10.1109/TSA.2005.857575> 7–5.
- [7] S. Chu, S. Narayanan, and C. C. J. Kuo. 2009. Environmental Sound Recognition With Time #x2013;Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 6 (2009), 1142–1158. <https://doi.org/10.1109/TASL.2009.2017438> 2-3.
- [8] Michael Cowling and Renate Sitte. 2003. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters* 24, 15 (2003), 2895 – 2907. [https://doi.org/10.1016/S0167-8655\(03\)00147-8](https://doi.org/10.1016/S0167-8655(03)00147-8)
- [9] Bing Dong and Burton Andrews. 2009. Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings. In *Proceedings of building simulation*. 1444–1451. <https://pdfs.semanticscholar.org/de95/b672e9e30b04749623c2d92c89f256eedda4.pdf> 1.
- [10] Monica Drăgoicea, Laurențiu Bucur, and Monica Pătrașcu. 2013. A service oriented simulation architecture for intelligent building management. In *International Conference on Exploring Services Science*. Springer, 14–28.
- [11] Antti J Eronen, Vesa T Peltonen, Juha T Tuomi, Anssi P Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, and Jyri Huopaniemi. 2006. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1 (2006), 321–329.
- [12] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound Technical Demo. In *ACM International Conference on Multimedia (MM'13)*. ACM, ACM, Barcelona, Spain, 411–412. <https://doi.org/10.1145/2502081.2502245>
- [13] Jerome H. Friedman. 1997. On Bias, Variance, 0/1–Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery* 1, 1 (01 Mar 1997), 55–77. <https://doi.org/10.1023/A:1009778005914>
- [14] Timothy D Griffiths, Adrian Rees, Caroline Witton, A Shakir Ra'ad, G Bruce Henning, and Gary GR Green. 1996. Evidence for a sound movement area in the human cerebral cortex. *Nature* 383, 6599 (1996), 425.
- [15] M.T. Hagan, H.B. Demuth, and M.H. Beale. 2014. *Neural network design*. Martin Hagan. <https://books.google.nl/books?id=4EW9oQEACAAJ>
- [16] Ebenezer Hailemariam, Rhys Goldstein, Ramtin Attar, and Azam Khan. [n. d.]. Real-time Occupancy Detection Using Decision Trees with Multiple Sensor Types. In *Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design (2011) (SimAUD '11)*. Society for Computer Simulation International, 141–148. <http://dl.acm.org/citation.cfm?id=2048536.2048555>
- [17] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. 2013. Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing* 2013, 1 (2013), 1–9–8 Add context-label to detect overlapping sound event by HMM.
- [18] Eun Jeon, Jong-Suk Choi, Ji Lee, Kwang Shin, Yeong Kim, Toan Le, and Kang Park. 2015. Human detection based on the generation of a background image by using a far-infrared light camera. *Sensors* 15, 3 (2015), 6763–6788.
- [19] Benjamin Kedem. 1986. Spectral analysis and discrimination by zero-crossings. *Proc. IEEE* 74, 11 (1986), 1477–1493.
- [20] K. Lopatka, J. Kotus, and A. Czyzewski. 2016. Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. *Multimedia Tools and Applications* 75, 17 (01 Sep 2016), 10407–10439. <https://doi.org/10.1007/s11042-015-3105-4>
- [21] Konstantinos Makantasis, Antonios Nikitakis, Anastasios D Doulamis, Nikolaos D Doulamis, and Ioannis Papaefstathiou. 2018. Data-driven background subtraction algorithm for in-camera acceleration in thermal imagery. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (2018), 2090–2104.
- [22] A. Mesaros, T. Heittola, and T. Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*. 1128–1132. <https://doi.org/10.1109/EUSIPCO.2016.7760424>
- [23] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. 2010. Features for Content-Based Audio Retrieval. In *Advances in Computers*. Vol. 78. Elsevier, 71–150. <http://linkinghub.elsevier.com/retrieve/pii/S0065245810780037> 8–6 features.
- [24] K. J. Piczak. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. 1–6. <https://doi.org/10.1109/MLSP.2015.7324337>
- [25] Rashmi Priyadarshini and RM Mehra. 2015. Quantitative review of occupancy detection technologies. *Int. J. Radio Freq* 1 (2015), 1–19.
- [26] Fariba Sadri. [n. d.]. Ambient Intelligence: A Survey. 43, 4 ([n. d.]), 36:1–36:66. <https://doi.org/10.1145/1978802.1978815> 3.
- [27] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1041–1044.
- [28] Jérôme Sueur, Sandrine Pavoine, Olivier Hamerlynck, and Stéphanie Duval. 2008. Rapid acoustic survey for biodiversity appraisal. *PLoS one* 3, 12 (2008), e4065.
- [29] Huy Dat Tran and Haizhou Li. 2011. Sound event recognition with probabilistic distance SVMs. *IEEE transactions on audio, speech, and language processing* 19, 6 (2011), 1556–1568.
- [30] Huy Dat Tran and Haizhou Li. 2011. Sound Event Recognition With Probabilistic Distance SVMs. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 6 (2011), 1556–1568. <https://doi.org/10.1109/TASL.2010.2093519>
- [31] Wikipedia contributors. 2018. Voice frequency — Wikipedia, The Free Encyclopedia. (2018). https://en.wikipedia.org/w/index.php?title=Voice_frequency&oldid=834458520 [Online; accessed 30-May-2018].
- [32] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5, Aug (2004), 975–1005.
- [33] Y. Zigel, D. Litvak, and I. Gannot. 2009. A Method for Automatic Fall Detection of Elderly People Using Floor Vibrations and Sound—Proof of Concept on Human Mimicking Doll Falls. *IEEE Transactions on Biomedical Engineering* 56, 12 (Dec. 2009), 2858–2867. <https://doi.org/10.1109/TBME.2009.2030171>