Localising Speech, Footsteps and Other Sounds using Resource-Constrained Devices

Yukang Guo
School of Computing and Communications
Lancaster University, UK

Mike Hazas
School of Computing and Communications
Lancaster University, UK

ABSTRACT

While a number of acoustic localisation systems have been proposed over the last few decades, these have typically either relied on expensive dedicated microphone arrays and workstation-class processing, or have been developed to detect a very specific type of sound in a particular scenario. However, as people live and work indoors, they generate a wide variety of sounds as they interact and move about. These human-generated sounds can be used to infer the positions of people, without requiring them to wear trackable tags. In this paper, we take a practical yet general approach to localising a number of human-generated sounds. Drawing from signal processing literature, we identify methods for resource-constrained devices in a sensor network to detect, classify and locate acoustic events such as speech, footsteps and objects being placed onto tables. We evaluate the classification and time-of-arrival estimation algorithms using a data set of human-generated sounds we captured with sensor nodes in a controlled setting. We show that despite the variety and complexity of the sounds, their localisation is feasible for sensor networks, with typical accuracies of a half metre or better. We specifically discuss the processing and networking considerations, and explore the performance trade-offs which can be made to further conserve resources.

Categories and Subject Descriptors

F.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity

General Terms

Algorithms, Design, Experimentation, Measurement, Performance

1. INTRODUCTION

Sensors and signal processing can be used to capture and interpret audio events for a number of applications, including audio surveillance [20], intelligent auditory interfaces [34], speaker positioning [6], and habitat monitoring [22, 35]. In many of these scenarios, the audio events being monitored also need to be localised.

IPSN'11, April 12–14, 2011, Chicago, Illinois.

Broadly speaking, there have been two approaches to such *acoustic source localisation*. The first is to use microphone arrays. These have been shown to have high positioning accuracy, and have the potential to identify and localise a wide variety of sounds. However, they require relatively static installations of precision microphones, wired to multi-channel sound cards and processed by high-throughput computing systems. The systems can also be difficult to install, due to the sensor hardware and cabling, and the tight calibration in microphone position and orientation. The second approach is to use comparatively low resource, distributed nodes with sensing and algorithms optimised to detect very specific types of sound, such as gunshots or animal calls. The advantage of these systems is that the wireless, battery-powered nodes are less expensive, easier to install, suited to deployment in a wider range of environments, and have less strict calibration requirements (centimetre rather than millimetre).

In this paper, we describe a sensor network–based acoustic source localisation strategy which can cope with a wider variety of sounds, closer to that encountered in everyday, indoor living and working environments. Consider for example people talking, walking, using computer keyboards, closing doors, placing items on table or shelf surfaces, or handling kitchen utensils. With such a diversity and complexity of indoor acoustic events, it is not feasible to build an embedded, wireless sensor system tailored to all types of sound. We propose using a more generalisable method, wherein events are put into very broad classification categories, and then an appropriate acoustic source localisation method is applied.

To show proof-of-concept, we select four different types of sound based on our own and others' observation of indoor office environments, and we investigate the classification, localisation and node resource trade-offs. We have generated these four types of sound repeatedly within a measurement volume in an office, which has been accurately surveyed to provide reliable ground truth. This has resulted in a small corpus of 2800 unique audio passages for evaluation of our chosen methods.

We rely upon modifications of existing classification, time difference of arrival (TDOA) and position estimation methods. As such, this paper represents little innovation from a core algorithmic standpoint. Our contributions are rather in three other areas: (1) we have surveyed a wide range of classification and acoustic TDOA estimation techniques, and identified ones which we expect to be both effective for localisation of indoor audio, and also appropriate for resource-constrained nodes; (2) we have shown that different types of sound are best detected by different TDOA algorithms, and that performing broad classification first allows low-computation localisation methods to then be applied to simple types of sound (such as claps, thuds or clunks), leaving more processor and network resources for localisation of complex sounds that oc-

cur (such as speech); (3) our analysis using our corpus of audio passages shows the achievable accuracies and overhead implications for resource-constrained sensor nodes.

2. RELATED WORK

Both audio classification and localisation are well-investigated research areas. However most of the existing systems and algorithms have been designed for a single or limited number of acoustic event classes, to suit specific application scenarios.

Classification. Many audio detection and classification systems have been developed, supporting applications such as content-based audio classification [27, 30], audio surveillance [20] and audio event modelling [26]. Most of these classification systems adopt two-stage approach: feature extraction followed by classification.

A number of speech and music discriminators have been proposed [30, 31, 13], using features selected from time, frequency and cepstral domains, and classifiers such as the Gaussian mixture model classifier, spatial partitioning classifier and k-nearest-neighbor classifier. Goldhor described a system which could recognise environmental sounds generated by acoustic sources common in domestic, business, and outdoor environments [17]. In recent years, audio-based surveillance systems have become increasingly important and popular both in private and public environments [20]. These systems usually have been developed for identifying particular types of acoustic events in indoor scenarios, such as the sound of a pen dropping to the floor, or a person casually whistling [20]. Algorithms have also been developed to detect people shouting, knocking, talking, walking and running in a corridor [3], or to detect impulse-type sounds such as screams or gunshots [33, 10].

Localisation. Most existing acoustic localisation work either requires the use of customised hardware, such as microphone arrays, or focuses on a specific type of acoustic event. Ali et al. [2] deployed a system to locate marmot calls using a dedicated audio sensing platform [16], equipped with a four-channel microphone array. The marmot call can be detected using a constant false alarm rate algorithm and its location can be estimated using approximate maximum likelihood–based direction-of-arrival. Brandstein et al. [6] mounted a ten-element bilinear microphone array in a laboratory environment to locate people talking using relative time delay estimation followed by location estimation. However, the location estimation performance was not quantitatively analysed. Bian et al. [4] mounted a number of four-element microphone arrays to monitor a 38 m^2 area in home lab. While location accuracy was reported as 33 cm (95th percentile), this evaluation was performed using playback of pre-recorded impulsive sounds on a loudspeaker. Furthermore, the algorithm discard rate of inaccurate measurements was exceptionally high (60–80%).

Scott et al. [32] implemented an audio location system which uses low-cost PC microphones. With six microphones covering a 1.8×1.8 m area, the median error of 3D location estimation of a "click" is better than 27 cm (90th percentile). However because the time-of-arrival was estimated using a simple energy-based thresholding method, this system design only works for impulsive sounds. Ajdler et al. [1] investigated sound source localisation in a distributed microphone sensor network. The reported standard deviation of location estimation is at centimetre level. However, it was not clear what geometry their microphone sensor placement followed, nor which type of sound was used to evaluate the system. Moreover, the ground truth location of the sound event was used as an aid for the outlier rejection criteria.

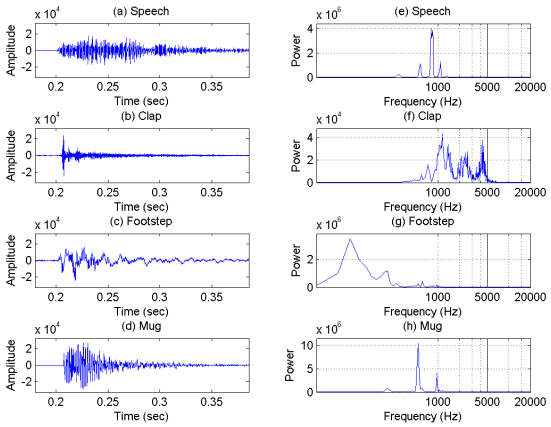


Figure 1: Time and spectral characteristics of some indoor sounds

3. DESIGN AND ALGORITHMS

Rather than tailor our design to a narrow range of acoustic events, we propose a more generalised hierarchical classification and localisation approach. As we illustrate below, appropriate acoustic event classification helps us choose the most accurate localisation algorithm. For experimental purposes, it was necessary to nominate a small number of candidate types of sound which represent the broad challenges found in real indoor auditory scenes. According to a 48-day recording [20] conducted in an office and our own observations from a 24-hour data set recorded in a large office, a number of sounds are commonplace in offices. These include human speech, low frequency sounds (footfalls, or a door closing), high frequency sounds (clacking of keys on a keyboard) and more generally sounds due to human movement, such as placing objects on tabletop or wheeled office chairs rolling. From our observations, we have chosen four broad types of sound (figure 1), which we used as a basis for selecting four specific sounds to generate in our experiments (section 4.1). One commonly-occurring sound is speech, which is composed of both pseudo-periodic voiced signals and non-periodic unvoiced segments. The energy of voiced segments is concentrated into certain regions of the frequency domain (dependant on the speaker), while the frequency of the unvoiced part is spread more widely across the spectrum. A second type of sound is footsteps, which have frequency components spread over a relatively low frequency range (see figure 1(g)); the exact frequencies will depend on the person, their footwear and the floor. Our third nominated type of sound is those arising when objects are placed onto or dropped onto surfaces; while the composition of these tend to vary more, they typically are concentrated in specific frequency bands, depending on the object and the table material. For example, sound may be concentrated in a higher frequency range for a metal table, compared to a wooden one. A fourth type of sound is impulsive (such as hand clapping or finger snapping), which is a short burst normally lasting 10–100 ms, concentrated in the higher frequencies. While these do not occur as commonly as the first three types of sound, we have chosen impulsive sounds to provide a comparable reference with certain existing acoustic localisation systems, which were specifically designed to localise impulsive sounds only.

Our proposed system, illustrated in figure 2, consists of three stages: acoustic event detection, classification and localisation. The first task is to distinguish the event-of-interest from background noise. To achieve this purpose, audio input is segmented into frames,

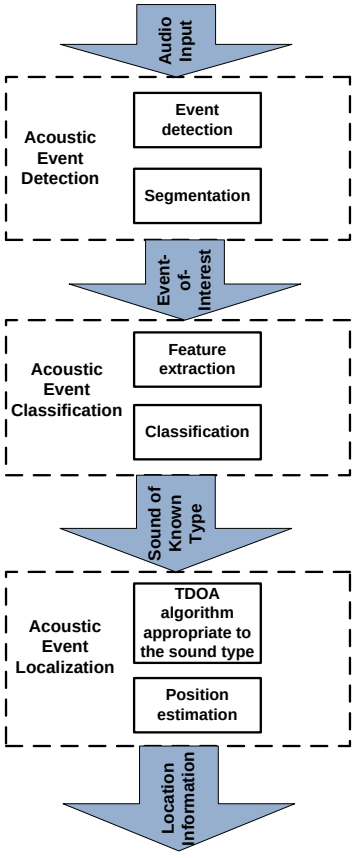


Figure 2: System architecture

and an event is detected when the frame energy exceeds a certain threshold. The rest of the frames are considered as silence and discarded. Thereafter, events-of-interest are fed into the classification module. At the classification stage, acoustic events are divided into a number of broad categories. After classification, a specific localisation algorithm may then be applied to the acoustic event, suited to its category's distinctive characteristics. The three stages are now discussed in detail.

3.1 Acoustic event detection

To distinguish an event-of-interest from background noise, an event detection algorithm must be run continuously to monitor the audio stream. Although frequency analysis has proven to be an effective method to detect acoustic events [20], the computation and analysis of the frequency spectrum are expensive [19]. Therefore we use a very low-complexity amplitude-based scheme for acoustic event detection, which might alternatively be implemented using very low power analogue hardware.

With the sampling rate at F_s, all recorded signals are first segmented into N sample frames, which are taken as the basic data processing unit. The nth segmented audio frame is defined as

$$x_n(m) = x(m)w(n-m), -\infty < m < \infty \quad (1)$$

where $x(m)$ is the recorded signal at the time index m, and $w(n)$ is the window function. We utilised a rectangular window, for simplicity:

$$w(n) = \begin{cases} 1 & 1 \le n \le N \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Index	Feature
1	root-mean-square
2	low short-time energy ratio
3	zero-crossing rate
4-6	linear prediction coefficients
7-9	linear spectral frequencies
10-16	mel-frequency cepstral coefficients
17	spectral centroid
18	spectral flux
19-23	frequency band energy

Table 1: Features used for classification

The root-mean-square (RMS) amplitude of the nth frame is

$$\text{RMS}_{n(x)} = \sqrt{\frac{\sum_{m=1}^{N} x_n(m)^2}{N}}. \quad (3)$$

It is compared with a pre-defined threshold RMS_0, according to the local noise level at an individual sensor node. The threshold can be based on several seconds of noise recorded by each sensor node. If the RMS value for a frame is less or equal than the threshold, the corresponding frame is considered as silence and discarded, otherwise, the frame is classified as an event-of-interest and passed to the following classification stage.

3.2 Acoustic event classification

At this stage, the event-of-interest window is classified into a category so that an appropriate TDOA algorithm can be subsequently applied. Past work [31, 18] on audio classification suggests that because the topology of the feature space is rather simple, there tends to be little improvement in classification performance arising from either (1) the use of different classifiers, or (2) particular settings for parameterised classifiers. Instead, it is the feature selection which is crucial to building an efficient and reliable classification system. Thus, our focus here is on choosing suitable features for classifying audio events indoors.

In typical audio classification, the first step is to represent the raw audio signal by using low-dimensional yet distinguishable feature vectors [30, 31, 13, 27]. This procedure also aims to remove information irrelevant to the audio classification task in order to avoid high computational complexity.

We have reviewed existing audio classification work, and surveyed a wide variety of features useful in distinguishing between different sounds [23, 13, 27, 31, 25]. From our survey, we have implemented and evaluated a total of fourteen temporal, spectral and cepstral features. From those, we chose nine features (table 1) based on two criteria: (1) their suitability for implementation on existing sensor nodes; and (2) their potential effectiveness for the diversity of sounds encountered indoors.[1]

The organisation of features is shown in figure 3. Five sets of features, marked as light grey and white, are extracted from each audio frame. These include short-term energy, zero-crossing rate, linear prediction coefficients (LPC), mel-frequency cepstral coefficients (MFCC) and short time Fourier (STFT). Based upon these features, six more feature sets, marked as dark grey, can be further extracted. Two of these (low short-term energy ratio and spectral flux) are obtained by analysing the difference between adja-

[1]Our five implemented features which did not make the cut were short-time energy, high zero-crossing rate ratio [27], pitch, spectral bandwidth, and spectral roll-off [25].

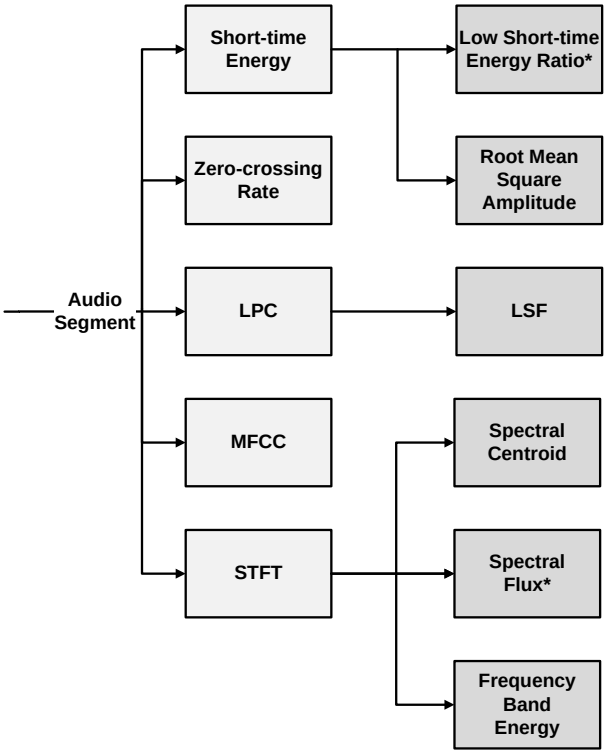


Figure 3: Fundamental and derived features for classification

cent subframes. (In extracting audio features for our classification scheme, the N sample audio frame was further divided into a number of non-overlapping subframes.)

As mentioned above, different classifiers tend not to exhibit much difference in performance [31]. After evaluating several classifiers, such as a multidimensional Gaussian maximum a posteriori estimator, a Gaussian mixture model classifier, a k-nearest-neighbour classifier and two discriminant analysis classifiers, we have favoured a quadratic discriminant analysis classifier (QDA). It yields satisfying classification results, and is feasible to implement on resource-constrained sensor nodes (see sec. 5). In the discriminant analysis, the conditional density function of the measurement is assumed to follow a multivariate Gaussian distribution. Statistical parameters (such as mean and covariance) of different classes are estimated using the training data set.

3.3 Time difference of arrival estimation

Existing source location estimation methods may be divided into three categories: steered beamforming, high-resolution spectral estimation, and time difference of arrival [6]. In the latter, time differences are computed between all microphone pairs that detected the arrival of the event, and then the location of the acoustic source is estimated using these TDOAs. Although suffering from moderate precision decline due to this two-step process, TDOA-based methods present a significant reduction in terms of computational complexity.

There are various TDOA estimation algorithms, ranging from simple thresholding methods [32], the average magnitude difference function (AMDF) method [11, 8] and generalised cross correlation (GCC) [24]. Thresholding methods have proven to be the most computationally efficient, but they have difficulties dealing with non-stationary signals, such as speech [32]. AMDF has relatively low computational cost, but is more sensitive to noise [14]. Among these three types of TDOA estimation method, GCC is the most computationally expensive. However, it is effective one for non-stationary signals in the presence of noise and room reverberance [5]. In this paper we outline and evaluate two TDOA estimation methods: thresholding and GCC.

Signal model and time difference of arrival. For an acoustic signal $s(t)$, in the presence of noise and room reverberance, the received signal $x_i(t)$ at each microphone i can be modelled as

$$x_i(t) = h_i(t) * s(t - \tau_i) + n_i(t), \tag{4}$$

where τ_i represents the time-of-flight from the sound source to the microphone, $*$ denotes convolution, $h_i(t)$ is the acoustic transfer function between the sound source and the microphone, and $n_i(t)$ models the sum of microphone channel noise and environmental noise. The noise $n_i(t)$ is assumed to be uncorrelated to the sound source $s(t)$.

The relative time difference of arrival between a pair of microphones i and j can be formulated as follows

$$\tau_{ij} = \tau_i - \tau_j. \tag{5}$$

In practicality, the TDOA should lie in the range of possible sound source positions (defined by the room's physical constraints), and corresponding to the spatial separation of the microphone pair.

$$\tau_{ij} \in [-\tau_{max}\tau_{max}] \tag{6}$$

$$\tau_{max} = \frac{\|\overrightarrow{m_j} - \overrightarrow{m_i}\|}{c} \tag{7}$$

where $\overrightarrow{m_i}$ and $\overrightarrow{m_j}$ are the positions of microphones i and j, c is the velocity of sound, and $\|\cdot\|$ is the Euclidean norm.

TDOA estimation using dynamic amplitude thresholding. The time-of-arrival at a microphone can be captured using a dynamic amplitude thresholding (DAT) scheme. The current noise profile $n_i(t)$ at microphone i is estimated using a weighted moving average of acoustic sample values:

$$\begin{aligned} n_i(1) &= s_i(1) \\ n_i(t) &= \alpha * s_i(t-1) + (1-\alpha) * n_i(t-1), \end{aligned} \tag{8}$$

where $s_i(t)$ is the current acoustic sample value, and $s_i(t-1)$ and $n_i(t-1)$ are the previous values of noise and microphone readings, respectively. The acoustic event is considered to arrive at the microphone when the current sound sample is some factor greater than the current noise profile ($s(t) > \beta * n(t)$) and the corresponding timestamp t_i is taken as the time-of-arrival. In our experiment set up described below, α and β are empirically determined; these values may vary if using different hardware. Using the DAT method, the TDOA between microphone i and j can be estimated as $\hat{\tau}_{ij} = t_j - t_i$. While this amplitude-based TDOA estimation is good in the presence of impulsive sounds such as hand clapping, it has been shown to be inadequate for low-amplitude or non-impulsive sounds.

TDOA estimation using generalised cross correlation. One of the most popular approaches for TDOA estimation is generalised cross correlation [24]. The GCC function is defined as

$$R_{x_i x_j}(\tau) = \int_{-\infty}^{\infty} W(\omega) X_i(\omega) X_j(\omega)^* e^{j\omega\tau} d\omega, \tag{9}$$

where $W(\omega)$ represents the frequency weighting function. Several GCC variants have been proposed, each of which has a different weighting function. The weighting functions of regular cross correlation, maximum likelihood (GCC-ML) and phase transform

(GCC-PHAT) are 1, $\frac{1}{|X_i(\omega)X_j^*(\omega)|}$ and $\frac{|X_i(\omega)X_j(\omega)|}{|N_i(\omega)|^2|X_j(\omega)|^2+N_j(\omega)|^2|X_i(\omega)|^2}$, respectively. While GCC-ML is reported to be robust to low signal-to-noise-ratio, GCC-PHAT performs better in the presence of room reverberation. The TDOA estimation τ_{ij} corresponds to the time lag at the global maximum peak of the GCC function R_{ij}^{τ}

$$\hat{\tau}_{ij} = \arg\max_{\tau} R_{ij}(\tau) \tag{10}$$

3.4 Location estimation

To compute the location of the sound source given the time differences of arrival, we specify modeling equations (below) and apply a non-linear regression algorithm which performs Levenberg-Marquardt gradient descent on the solution space to minimise the sum of squared residuals. However, factors such as multi-path due to room reverberation and non-stationary noise can result in inaccurate TDOA estimates, and microphones further from the source may capture events at low SNR and identify false TDOA peaks. To make our location algorithm robust, we apply an iterative outlier rejection algorithm (described below) which removes the TDOA with the highest residual and re-computes the regression solution.

Other options for computing the location solution are RANSAC and maximum likelihood estimators. However, these tend to function best when the number of data points (in our case, TDOAs) to be fitted is high. For I microphones which successfully measure a time-of-arrival, the number of resulting TDOAs is $\frac{I(I-1)}{2}$. Thus, even assuming six microphones detect the event, this results in only fifteen TDOAs, which is typically not enough for RANSAC or maximum likelihood to function well. We have tried RANSAC in similar acoustic time delay model-based positioning algorithms, but it was outperformed by simpler residual-based elimination techniques such as the one we describe here.

Non-linear regression. The TDOA τ_{ij} between microphone pair m_i and m_j generated from sound source is defined as

$$\tau_{ij} = \frac{\|\overrightarrow{m_i} - \overrightarrow{s}\| - \|\overrightarrow{m_j} - \overrightarrow{s}\|}{c}, \tag{11}$$

where $\overrightarrow{s}$ is the position of sound source, $\overrightarrow{m_i}$ and $\overrightarrow{m_j}$ are the positions of microphone m_i and m_j respectively. Therefore the TDOA estimation error $\epsilon_{\tau_{ij}}$ can be formulated as

$$\epsilon_{\tau_{ij}} = \tau_{ij} - \hat{\tau}_{ij}. \tag{12}$$

Given speed of sound c and positions of microphone pairs m_i and m_j, position of sound source can be estimated using a non-linear regression algorithm [29] to minimise the *standard error estimate*:

$$\text{errLoc} = \sqrt{\frac{\sum \epsilon_{\tau_{ij}}^2}{O-3}}, \tag{13}$$

where O is the number of observations (TDOAs) reported, and where the sum term occurs across all residuals of the O TDOAs.

Residual-based outlier rejection. The outlier rejection algorithm is shown in figure 4. If the observations given to the regression do not agree with each other, the standard error will be high. In this case, the observation with the largest residual error is discarded. Following this, the non-linear regression algorithm is applied on the remaining observations. This process is repeated until the standard error of the location estimation is small enough. If the standard error estimate never decreases below the threshold, then the estimation process is abandoned, and no location is returned for that event.

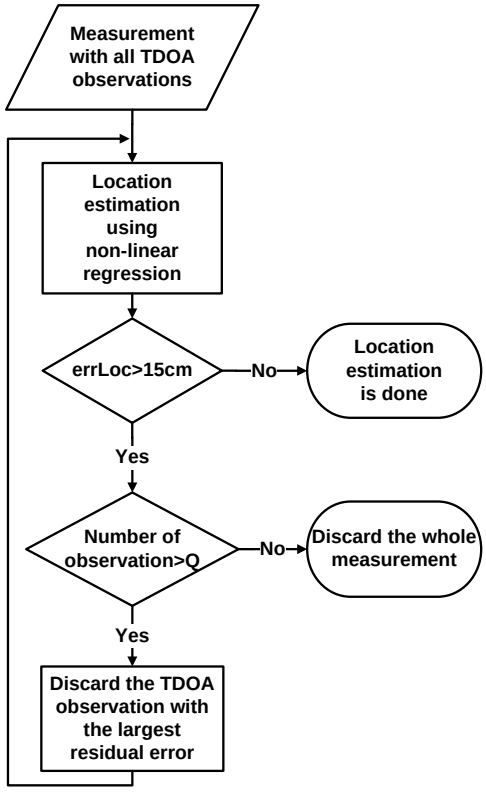


Figure 4: Outlier rejection. In our evaluation, we set $Q = 7$ and used a standard error threshold equivalent to 15 cm.

4. SYSTEM PERFORMANCE WITH PLENTIFUL RESOURCES

In this section, we illustrate the potential of our architecture and algorithms with the assumption that the sensor nodes have relatively high resources (computation, storage, and networking) to devote to the acoustic localisation process. We illustrate and discuss the impact of reducing available resources in the following section.

4.1 Experimental setup

To test the feasibility of the proposed system, we distributed six sensor nodes in a $5.98 \times 2.89 \times 2.69$ m office in a modern functional building. At one end of the room, the six nodes were widely spaced at approximately two different heights to cover a 5×7 grid of measurement points, approximately $3 \times 2 \times 1.8$ m, with about 0.5 m between each grid point (figure 5). The positions of both the nodes and grid points were accurately surveyed using an electronic theodolite (a Leica TotalStation TC500).

The sensor node used is built upon the Intel Mote2 (Imote2) platform. It is based around an XScale processor and a 15.4 radio module. Each Imote2 was fitted with a sensor board having an inexpensive Panasonic WM-62A omnidirectional microphone, and a 16-bit A/D converter (LTC1865L).

For the purposes of this formative work in generalised acoustic event classification and localisation, a traditional wired approach was used for synchronisation. Each node sampled its microphone signal at 40 kHz with 16-bit precision. As soon as the sample amplitude exceeded the pre-defined threshold on one of the six microphone sensor nodes, this node proclaimed itself the timing reference node and sent notification of such to the other nodes via a

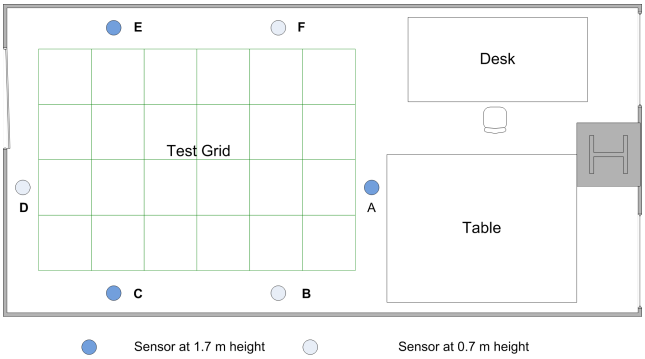


Figure 5: Experiment layout

general-purpose IO pin. Together with the last 204.8 ms of data kept in a circular buffer on each sensor node, 1638 ms of additional data was captured and stored for further analysis. Thus, for each audio passage a total of 1843 ms of samples was logged by each node.

Commensurate with the four types of sound discussed above, we created four specific sound events in our experiments: certain passages of speech, hand clapping ("clap"), a footstep and a mug being placed on a wooden table ("mug"). A female subject read a set of phonetically-compact sentences selected from the core test set of the TIMIT database [15], which provide a good coverage of phonemes. The footsteps were generated by the female subject by walking on the carpeted floor while wearing casual leather shoes. Each of the four specific types of sound was created twenty times at each of the thirty-five locations on the test grid. Thus, a total of 2800 of these *audio passages* comprise our data set.

4.2 Classification performance

To evaluate our system, all of the recorded audio streams were labelled automatically using the following procedure. Most of the acoustic energy of a typical event can be captured using a frame duration of about 200 ms. At this frame length, mismatching of the acoustic events among the different sensors can be avoided for microphones positioned less than about six metres apart, which is sufficient for many room-based sensor deployments. So, each 1843 ms audio clip was segmented into 204.8 ms frames (8192 samples, given the sampling rate of 40 kHz). If the frame RMS value exceeded a predefined threshold, it was passed to the classifier. Otherwise, it was considered as silence and discarded.

Because the size of the data set is relatively small (700 audio passages captured by each microphone across 35 locations, for each of the four sounds), we evaluated the classification performance using a K-fold cross-validation framework [21]. The whole data set is randomly partitioned into K subsets. Of the K subsets, a single subset k is retained as the validation data for testing the model, i.e. the test set, and the remaining $K-1$ subsets are used as training data to build the fitted model. For each k, the prediction error of the fitted model is calculated. The prediction errors are aggregated across all $k = 1, 2, \ldots, K$. The confusion matrix of classification results obtained is shown in table 2. The classification accuracy (i.e. the ratio of test events that are correctly classified) is better than 93%.

4.3 TDOA accuracy

For each passage, the first 204.8 ms segments which exceeded the pre-defined threshold at all of the microphones were grouped together to estimate the corresponding TDOAs. The cumulative

	Speech	Clap	Footstep	Mug
Speech	93.46 %	0.02 %	4.58 %	1.94 %
Clap	0.60 %	99.01 %	0.00 %	0.40 %
Footstep	1.00 %	0.27 %	98.06 %	0.68 %
Mug	1.30 %	0.01 %	0.26 %	98.44 %

Table 2: Classification results using QDA, based on nine features and a ten-fold cross-validation framework with 90% of data used as training.

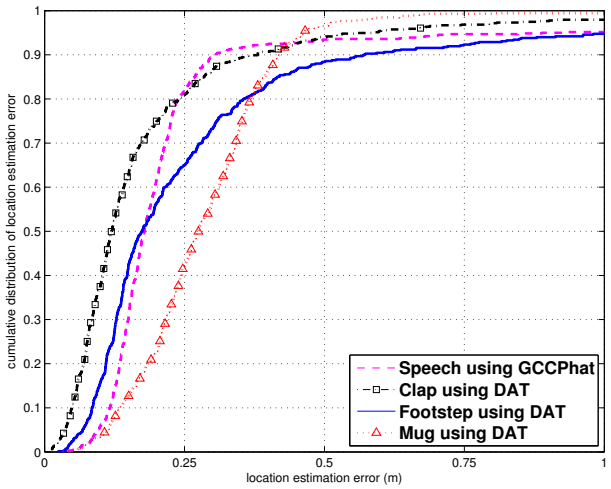


Figure 7: Location estimation error distributions

error distribution functions of TDOA estimation errors using the four discussed algorithms are given in figure 6. Figure 6(a) shows that speech signals achieve best performance using the GCC-PHAT method with an absolute TDOA accuracy of 60 cm 70% of the time. For clap and mug, as shown in figures 6(b) and 6(d), DAT performs the best, with 40 cm 90% of the time. Figure 6(c) shows that the TDOA estimation of footsteps yields the worst results compared to the other three sound categories. Still, the DAT method achieves 40 cm accuracy with 60th percentile confidence.

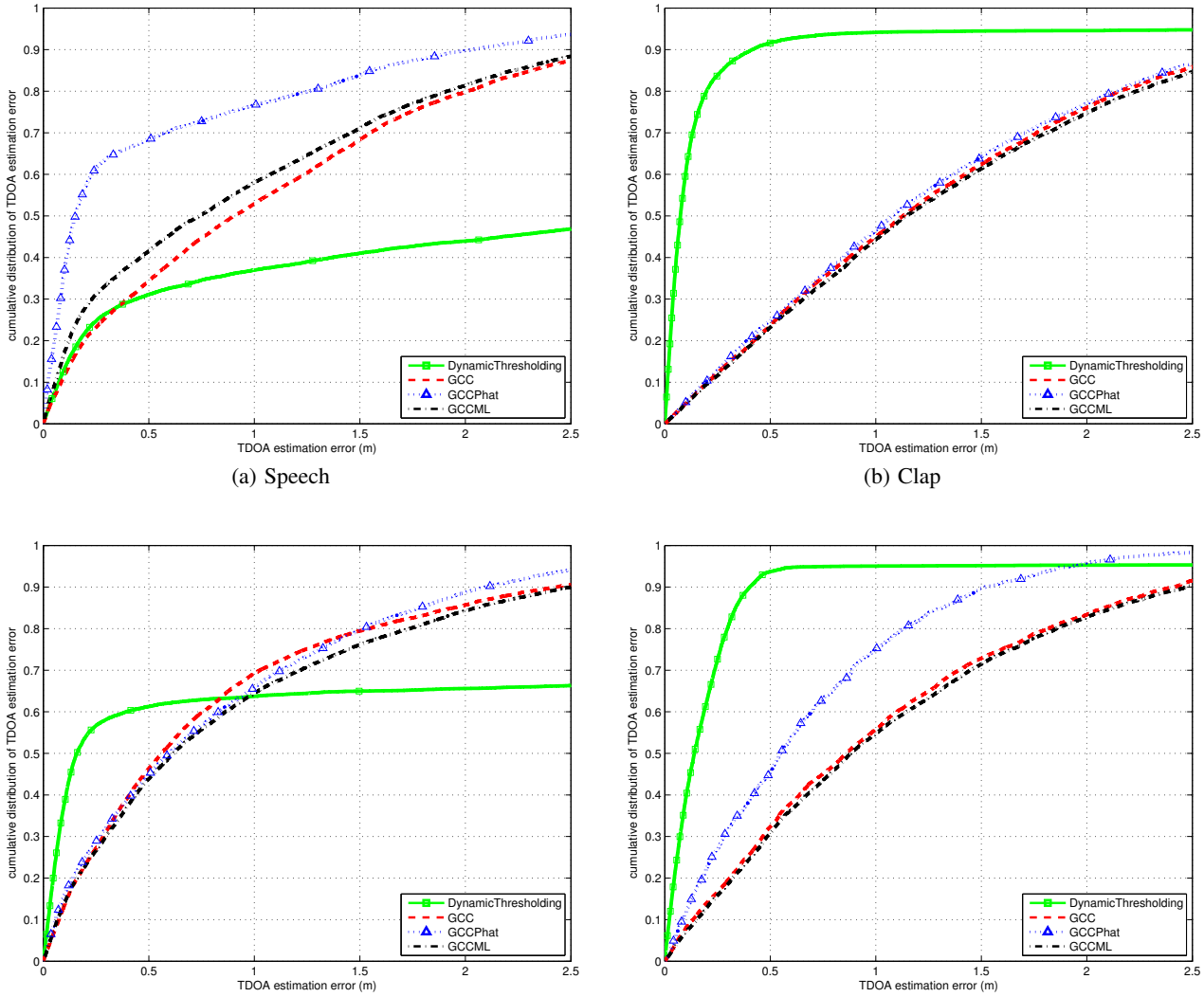
4.4 Location accuracy

Figure 7 shows the cumulative distributions of location errors (computed by model-based non-linear regression) for each type of sound using the most suitable TDOA estimation method for each type of sound. The 3D location estimation accuracy (table 3) for all types of sound is between 20 and 35 cm (75th percentile) and between 30 and 57 cm (90th percentile). However, the rejection rate of location estimates (due to high standard error) varies from just a few percent to as high as 40% for footstep.

4.5 Calibration considerations

As the size of training data affects the calibration effort required at system deployment, we investigate the effect of varying the amount of data used to train the classifier. Two approaches to reducing the amount of training data are explored here. One is to take fewer audio passages across all of the test points, and the other is to take audio passages from fewer test points.

Fewer audio passages from all of the test positions. The relationship between the percentages of data used as training set and



(a) Speech (b) Clap

(c) Footstep (d) Mug

Figure 6: TDOA estimation accuracy for the four types of sound

	Speech	Clap	Footstep	Mug
TDOA method	GCC-PHAT	DAT	DAT	DAT
75th percentile	23 cm	20 cm	31 cm	35 cm
90th percentile	30 cm	38 cm	57 cm	42 cm
rejection rate	20%	4.5%	40%	2%

Table 3: Localisation performance summary

training ratio (%)	1.25	5	10	25	50	75	90
training passages	9	35	70	175	350	525	630
cross-val. folds	80	20	10	4	2	4	10

Table 4: Relationship between the percentages of data used for training, the number of training passages per type of sound, and the number of cross-validation folds. Note that for classification purposes, the audio passages for speech and footstep typically contain multiple events. About 90% of the speech passages contained 4–7 events, and about 75% of the footstep passages contained 3–5 events.

the K folds of cross-validation is shown in table 4. As illustrated in figure 8, by using 25% to 90% of data as a training set, the classification results do not vary significantly. Even using less than 25% data as training set, the classification rates decline less than 5% for clap, footstep and mug data. However, the classification performance has slightly improved for speech when using smaller sets of data for training the fitted model. As noted in the caption of table 4, this improvement is because the number of speech events per audio passage is several times greater than that for other types of sound. Therefore, as the number of training passages of all types of sound decreased, the model built for speech remained more stable than the others.

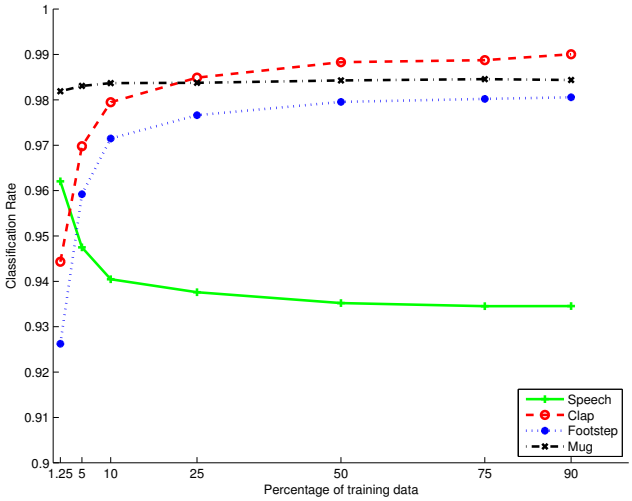


Figure 8: Classification and the ratio of training data

training points	Speech	Clap	Footstep	Mug
35	93.32%	98.68%	97.62%	98.17%
25	94.80%	98.83%	97.38%	98.52%
15	94.03%	98.60%	97.83%	98.33%
5	97.85%	87.04%	89.01%	96.67%

Table 5: Classification using audio passages from decreasing numbers of locations for the training data set. Regardless of the number of locations used for training, the total number of training passages was approximately one hundred for each type of sound. All remaining passages (across all locations) were used as test data to compute the above results. This corresponds to a training ratio of about 14%.

Training data from fewer locations. The other approach to reducing calibration effort is to take training passages from only a few test locations. The audio passages captured from 35, 25, 15, and 5 test locations (always starting from the middle of the test grid) were selected as the training data set, and the remaining locations served as test data. As shown in table 5, even using data from fewer positions to build the model, classification results remain above 87%.

Different speaking subjects. Another calibration concern is the number of different subjects required to provide training passages for speech. As mentioned in section 4.1, the four types of sound in our main experiment were generated by the same female subject. To evaluate whether our system will work for speech generated by different subjects, two male subjects read the same set of TIMIT sentences as the female subject had. This set of sentences was recorded ten times for both male subjects at each of 5 locations selected from the test grid. Here, the training data set was composed of one-third of the passages spoken by each subject (one female subject and two male subjects) at those five locations, and the remaining passages generated by the three subjects at those locations were retained as the test data set. For the other three types of sound, one-third of the passages at the five locations were selected as the training set and the remaining passages were the test data. The classification performance is shown in table 6. While the classification rates of footstep do decrease to 83.55%, the proposed system shows strong potential to work with speech generated by different subjects, even confining calibration to just a small number of convenient test locations.

	Speech	Clap	Footstep	Mug
Speech	91.30 %	0.00 %	2.42 %	6.28 %
Clap	1.00 %	99.00 %	0.00 %	0.00 %
Footstep	13.61 %	2.25 %	83.55 %	0.59 %
Mug	2.23 %	0.00 %	0.14 %	97.63 %

Table 6: Classification with different speakers

Subset	Features
time domain	root-mean-square (RMS) low short-time energy ratio (LSTER) zero-crossing rate (ZCR)
polynomial	linear prediction coefficients (LPC) linear spectral frequencies (LSF)
human auditory perception model	mel-frequency cepstral coefficients (MFCC)
spectral	spectral centroid spectral flux frequency band energy

Table 8: Subsets of features

5. THE PERFORMANCE/RESOURCE DESIGN SPACE

One purpose of our experiment is to validate the design of our algorithm architecture, particularly the choice of the classification features and the effectiveness of the TDOA algorithms. Thus, our experimental deployment discussed in section 4.1 focused on capturing the maximum fidelity of data. In this section, we discuss the trade-offs between performance achieved and resources required.

Table 7 shows a rough estimation of the operations required to accomplish the different algorithms we utilise. At higher sampling rates, the most computationally intensive tasks in the proposed system are feature extraction and TDOA estimation.

5.1 Constraining the feature set

Since extraction of all our chosen features can be an intensive task, here we explore the feasibility of reducing the number of features used for classification. With the features introduced in section 3.2, we form several subsets of the features. The subsets are shown in table 8. The classification accuracy when using different subsets is illustrated in figure 9; as before, classification was accomplished using QDA. By only using the subset of time domain features, speech and mug cannot be distinguished. The classification rate of mug can be largely improved by including either the polynomial or spectral subsets. The spectral features (which require an FFT) yield great improvement above time domain features for both mug and speech. While the cepstral features clearly benefit speech classification, all sets of features need to be computed for speech accuracy to exceed 90%.

Depending on the types of sound being classified, and the frame size N, it may be possible to choose reduced feature sets whilst maintaining acceptable performance. However, it is important to remember that accurate classification is needed to ensure that TDOA resources will be spent wisely (e.g. GCC-PHAT is only performed for speech).

Component	Operations	Notes	Requires
FFT	$N \log N$	N: the number of samples per frame, which depends	
ZCR	$3N$	on the frame duration and the sampling rate (see table 9)	
LPC [28]	$3Np - \frac{1}{2}(3D-1)p^2$	autoregressive modelling	
		p: order of autoregressive model	
		D: dimension of autoregressive model	
MFCC [36]	$2LM + MN - 1$	M: number of mel window filter banks ($M = 7$)	FFT
		L: coefficient order ($L = 13$)	
RMS	$2N$		
LSTER	$2N$		
Spectral Centroid	N		FFT
Spectral Flux	N		FFT
Frequency Band Energy	N		FFT
LDA/QDA	CF	C: number of categories ($C = 4$)	
		F: no. of feature parameters ($F = 23$ for full feature set; see tab. 1)	
DAT	$< 2N$		
GCC-PHAT [9]	$3N \log N - \frac{11}{2}N + 20$		
NLR	$135I(I-1) + 1016$	I: number of microphones which detect the event	

Table 7: Computational complexity

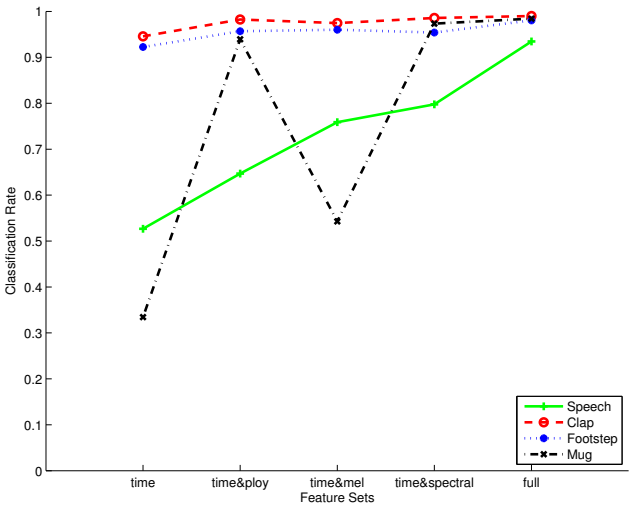


Figure 9: Classification with reduced features

Rate	Speech	Clap	Footstep	Mug	N
40 kHz	93.46%	99.06%	98.04%	98.44%	8192
20 kHz	92.53%	99.00%	97.41%	98.45%	4096
10 kHz	97.46%	98.64%	97.54%	97.57%	2048
5 kHz	88.28%	97.06%	94.19%	97.32%	1024

Table 9: Classification performance as sampling rate decreases. Also shown are the corresponding sample counts of our chosen frame duration of 204.8 ms.

		Speech	Clap	Footstep	Mug
40 kHz	75th %ile	23 cm	20 cm	31 cm	35 cm
	90th %ile	30 cm	38 cm	57 cm	42 cm
	rejection	20%	4.5%	40%	2%
20 kHz	75th %ile	24 cm	24 cm	32 cm	35 cm
	90th %ile	54 cm	47 cm	66 cm	41 cm
	rejection	26%	3%	25%	0.7%
10 kHz	75th %ile	38 cm	35 cm	33 cm	34 cm
	90th %ile	148 cm	66 cm	74 cm	40 cm
	rejection	55%	4%	20%	0.5%
5 kHz	75th %ile	139 cm	53 cm	36 cm	35 cm
	90th %ile	232 cm	108 cm	94 cm	42 cm
	rejection	80%	3%	10%	0.5%

Table 10: Localisation performance as sampling rate decreases

5.2 Lowering the sampling rate

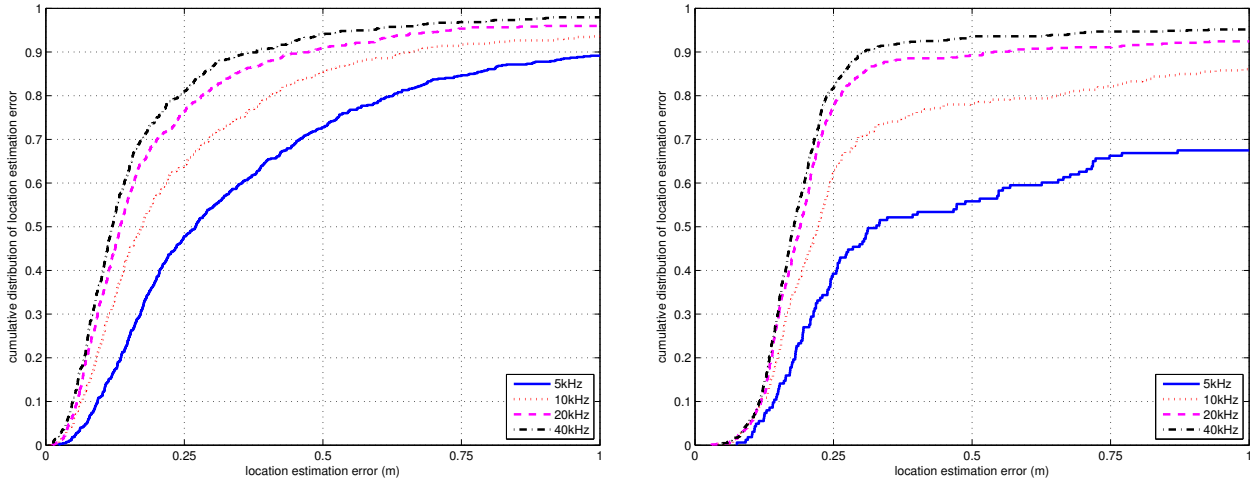
In our experiment, the acoustic events were sampled at 40 kHz, which was the maximum sample rate we could achieve on the Imote2. However, in practical wireless sensor systems, it is beneficial to have a lower sampling rate, as less data needs to be processed. Specifically, a lower sampling rate leads to a smaller frame size N (table 9). Here, we explore the impact of reducing the sampling rate on the performance of the proposed system.

We downsampled our 40 kHz data to 20 kHz, 10 kHz and 5 kHz, and compared the classification and location results. The classification accuracy is shown in table 9. The location accuracy CDFs for clap and speech are shown in figure 10. We do not show the CDFs for mug and footstep here, as the shape of their accuracy distributions did not change appreciably. However, we summarise the accuracy confidences and rejection rates for all four types of sound at different sampling rates in table 10.

In general, the classification and localisation performance declines with the sampling rate. For mug and footstep, the localisation performance does not vary significantly when the sampling rate is decreased. However, the localisation accuracies of speech and clap get dramatically worse. When sampling at 5 kHz, only 10% of speech events can be located to 35 cm; while sampling at 40 kHz, over 70% of speech events are within a similar error range.

5.3 Summary

Figure 11 summarises the above node resource considerations by illustrating how location fidelity and complexity can be traded off. Note that accurate localisation of very simple sounds (such



(a) Clap (b) Speech

Figure 10: Location accuracy as sampling rate decreases

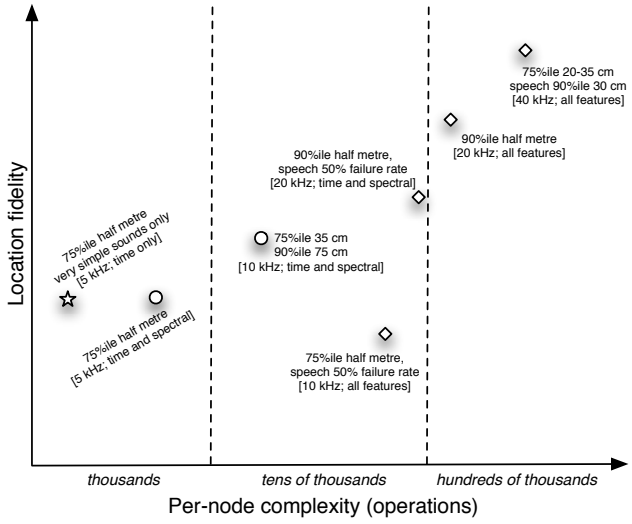


Figure 11: Summary of accuracy and complexity space. Diamonds mark computational levels which allow speech to be localised; these incur much larger communication overheads as nodes must share detailed information about audio passages to perform GCC-PHAT.

as the mug) is possible using only several thousand operations per node. If tens of thousands of operations are possible, then different sounds (footsteps, the mug, and clapping) can be accurately located. If desired, speech can also be located using tens of thousands of operations, albeit at reduced accuracy levels.

Hardware architecture. In our complexity analysis, we have not differentiated the operations which require floating point. Some of the more complex algorithm components (such as non-linear regression's matrix inversion) assume floating point capability. Architectures which natively support floating point typically have higher power requirements, and integer-based architectures often require ten or more times the instructions to implement floating point operations. Similarly, throughout the paper we have assumed that the node architecture in question supports 16-bit sampling and computation; reducing the sample depth to eight bits would allow lower-resource 8-bit hardware to execute the algorithms at speed. Regardless, we expect that even the most resource-constrained nodes will be able to localise simple sounds. Moreover, some architectures (like the XScale on the Imote2) are equipped with a co-processor which can be activated to efficiently perform operations such as the FFT and basic compression.

Communication requirements. In our analysis above, we have assumed that $\frac{I(I-1)}{2}$ TDOAs can be computed between all pairs of I microphones which detected the event. For sounds which can be localised using dynamic amplitude thresholding, nodes need only transmit their detected time-of-arrival (a few bytes per node) in order for all TDOAs to be computed. However to support this for sounds (such as speech) which require GCC operations, $I-1$ nodes must transmit the captured event so that correlations can be computed between all pairs of signals. Using our frame duration, 40 kHz sampling rate and 16-bit precision, this is potentially sixteen kilobytes of data for each of the $I-1$ transmissions.

There are a number of possibilities for reducing this, in practicality. First, we have shown acceptable performance results (even for speech) at 20 kHz. Second, the FFTs which are needed for the cepstral and spectral-based features are also an important component of correlation operations. Thus, it makes more sense to transmit the FFT of the event frame, rather than the raw time-domain representation. FFTs are also a very well-investigated technique for compressing audio signals. Very favourable communication savings can be made by low-overhead compression of these FFT signals, with little impact on TDOA accuracy once correlation is performed. Finally, we suspect that the events from certain microphones (for example with the most favourable SNRs) will tend to yield more accurate TDOAs when correlated. If for example only the three nodes with the strongest signals from the event were to transmit, this would again reduce the communication required, while still facilitating the computation of $3I-6$ TDOAs.

6. FUTURE CONSIDERATIONS

We have proposed an algorithm architecture where sound categories are used to help to select effective and efficient localisation methods in order to suit the application scenario. It was not feasible for us to cover all types of sounds in our experiments described above, due to the complexity of real world environments and the variability between environments. However, following our design, additional classes may be specified as a greater variety of sounds is encountered. In this case, further calibration (offline training) is required and different sets of features might be adopted, which could add to the computational load. Since QDA does not require too much processing power or storage, adding more categories would not introduce much overhead for the classifier, relative to the costs of operations such as the FFT or non-linear regression.

Issues we considered above include the classification features needed, suitable classifiers, and effective/efficient TDOA and location estimation methods. Furthermore our experiment was set up in a well-controlled environment. Therefore, a static thresholding method was applied to initiate the recording and avoid introducing processing overhead for sampling and logging. In the future, we will explore more robust and flexible dynamic thresholding methods, such as exponentially weighted moving average (EWMA) [19]. It is worth mentioning that in real settings, no matter which type of detection method is adopted, some events may be missing (false negatives) and at times background noise can be detected as events (false positives). Therefore, empirically evaluating thresholding methods according to the specific environment is critical.

There are two major drawbacks of the GCC-based TDOA estimation adapted in our system. First, it assumes the noise at different sensor nodes is independent. Second, it does not take into account room reverberance. In real world settings, correlated directional interference and room reverberance is very common. In the context of microphone arrays, some methods have been proposed to eliminate the localisation errors introduced by these issues [12, 7, 8]. Yet most of these methods have high complexity, and they need to be revisited with specific consideration for low-resource devices.

Our analysis in this paper is based on experiments which were performed in a relatively controlled environment in a small office. While common indoor background noise was present (due to things such as computer fans and building HVAC systems), there were no significant sounds which overlapped in time with the test events we generated. Many indoor living and working environments (with multiple people and devices in them) produce a more complex auditory scene, increasing the difficulty of proper *segmentation* and *matching* of the events detected at different microphones. To understand these challenges, we have since collected a 24-hour "real world" data set, in a large laboratory in which eight people work. We are conducting the analysis of this real world data, and preparing it for publication; the accuracy results are of similar magnitude (approximately half metre) with those we present above. The data collection for this paper was more controlled, by design, with the objective of understanding (1) what the algorithm options are for both classification and accurate TDOA estimation of different types of sound; and (2) the computational complexity required to support different levels of location fidelity.

7. CONCLUSION

In this paper we present a study of locating acoustic events appropriate to their diverse acoustic characteristics, using systems such as sensor nodes. The analysis of experimental data gathered in a controlled environment demonstrates the feasibility of implementing such methods on resource-constrained devices. Using the outlined methods at high sampling rates, the classification accuracy overall is better than 93% and the localisation accuracy is within 60 cm (90th percentile confidence) for sounds such as speech, footsteps, objects placed on surfaces, and claps. We have argued that classification should be used as an aid to identify an accurate yet efficient TDOA algorithm to apply to a given event. Further computation can be saved by choosing a sampling rate and feature sets which yield appropriate performance for the types of sound being localised.

In the future, we intend to address challenges in more realistic environments such as identifying the same acoustic event captured at different microphone sensors and the identification of multiple simultaneous sources. Additionally, to avoid the need for a labour-intensive survey of microphone positions, microphone nodes could self-localise via simultaneous localisation and mapping, or a minimal-effort offline method.

Acknowledgements. This PhD work was financially supported by the Faculty of Science and Technology at Lancaster University, the European Commission FP6 IST Programme (grant no. 013790), and Intel Research. The authors are also grateful to Intel Research for providing the Imote2 nodes and microphone sensor boards used in the experiments; to James Scott for his advice, enthusiasm and support in the early stages of this research; and to David Molyneaux for his insight and helpful comments on the experimental setup and analysis. Finally, we would like to thank the reviewers and our shepherd for their constructive feedback.

8. REFERENCES

[1] T. Ajdler, I. Kozintsev, R. Lienhart, and M. Vetterli. Acoustic source localization in distributed sensor networks. In *Proc. of Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1328–1332, 2004.

[2] A. M. Ali, K. Yao, T. C. Collier, C. E. Taylor, D. T. Blumstein, and L. Girod. An empirical study of collaborative acoustic source localization. In *Proc. of IPSN*, pages 41–50, 2007.

[3] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli. Audio based event detection for multimedia surveillance. In *Proc. of ICASSP*, May 2006.

[4] X. Bian, G. D. Abowd, and J. M. Rehg. Using sound source localization in a home environment. In *Proc. of Pervasive*, May 2005.

[5] M. S. Brandstein. Time-delay estimation of reverberated speech exploiting harmonic structure. *The Journal of the Acoustical Society of America*, 105(5):2914–2919, 1999.

[6] M. S. Brandstein, H. Silverman, and L. Engineering. A practical methodology for speech source localization with microphone arrays. *Computer Speech and Language*, 11(2):91–126, April 1997.

[7] B. Champagne, S. Bedard, and A. Stephenne. Performance of time-delay estimation in the presence of room reverberation. *IEEE Trans. on Speech and Audio Processing*, 4(2):148–152, March 1996.

[8] J. Chen, J. Benesty, and Y. Huang. Performance of GCC– and AMDF-based time-delay estimation in practical reverberant environments. *EURASIP J. Appl. Signal Process.*, pages 25–36, January 2005.

[9] J. Chen, J. Benesty, and Y. Huang. Time delay estimation in room acoustic environments: an overview. *EURASIP J. Appl. Signal Process.*, January 2006.

[10] C. Clavel, T. Ehrette, and G. Richard. Events detection for an

audio-based surveillance system. In *Proc. of ICME*, July 2005.

[11] R. Cusani. Performance of fast time delay estimators. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37(5):757–759, May 1989.

[12] T. G. Dvorkind and S. Gannot. Time difference of arrival estimation of speech source in a noisy and reverberant environment. *Signal Processing*, 85(1):177–204, 2005.

[13] El-Maleh, K. and Klein M. and Petrucci G. and Kabal P. Speech/music discrimination for multimedia applications. In *Proc. ICASSP*, pages 2445–2448, 2000.

[14] S. Furui and M. M. Sondhi, editors. *Advances in Speech Signal Processing*. Marcel Dekker, New York, 1991.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993.

[16] L. Girod, M. Lukac, V. Trifa, and D. Estrin. The design and implementation of a self-calibrating distributed acoustic sensing platform. In *Proc. of SenSys*, pages 71–84, 2006.

[17] R. Goldhor. Recognition of environmental sounds. *Proc. of ICASSP*, 1:149–152, 1993.

[18] S. Golub. Classifying recorded music. Master's thesis, University of Edinburgh, 2000.

[19] L. Gu, D. Jia, P. Vicaire, T. Yan, L. Luo, A. Tirumala, Q. Cao, T. He, J. A. Stankovic, T. Abdelzaher, and B. H. Krogh. Lightweight detection and classification for wireless sensor networks in realistic environments. In *Proc. of SenSys*, pages 205–217, 2005.

[20] A. Härmä, M. McKinney, and J. Skowronek. Automatic surveillance of the acoustic activity in our living environment. In *Proc. ICME 2005*, pages 1–8, 2005.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, first edition, July 2003.

[22] W. Hu, V. N. Tran, N. Bulusu, C. T. Chou, S. Jha, and A. Taylor. The design and evaluation of a hybrid sensor network for cane-toad monitoring. In *Proc. of IPSN*, page 71, 2005.

[23] Jeroen Breebaart and Martin Mckinney. Features for audio classification. In *Proc. Philips Symposium of Intelligent Algorithms*, 2002.

[24] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *Proc. of ICASSP*, 24(4):320 – 327, August 1976.

[25] S. Li. Content-based audio classification and retrieval using the nearest feature line method. *IEEE Trans. on Speech and Audio Processing*, 8(5):619–625, 2000.

[26] H. Lu, W. Pan, N. Lane, T. Choudhury, and A. Campbell. SoundSense: Scalable sound sensing for people-centric applications on mobile phones. In *Proc. MobiSys*, pages 165–178, 2009.

[27] L. Lu and H. Zhang. Content analysis for audio classification and segmentation. *IEEE Trans. on Speech and Audio Processing*, 10(7):504–516, October 2002.

[28] R. J. Martin and C. A. Openshaw. Autoregressive modelling in vector spaces: An application to narrow-bandwidth spectral estimation. *Signal Processing*, 50(3):189 – 194, 1996.

[29] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. CUP, second edition, 1992.

[30] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proc. of ICASSP*, pages 993–996, 1996.

[31] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. of ICASSP*, pages 1331–1334, 1997.

[32] J. Scott and B. Dragovic. Audio location: accurate low-cost location sensing. In *Proc. of Pervasive*, May 2005.

[33] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance. In *Proc. of AVSS*, September 2007.

[34] S. Vesa and T. Lokki. An eyes-free user interface controlled by finger snaps. In *Proc. of DAFx*, September 2005.

[35] H. Wang, C. Chen, A. Ali, S. Asgari, R. Hudson, K. Yao, D. Estrin, and C. Taylor. Acoustic sensor networks for woodpecker localization. In *Proc. of SPIE*, 2005.

[36] J.-C. Wang, J.-F. Wang, and Y.-S. Weng. Chip design of MFCC extraction for speech recognition. *Integration, the VLSI Journal*, 32(1-2):111 – 131, 2002.