1-2016

Smart ambient sound analysis via structured statistical modeling

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

Liqiang NIE

Tat Seng CHUA

DOI: https://doi.org/10.1007/978-3-319-27674-8_21

Published in MultiMedia Modeling: International Conference on Multimedia Modeling 2016: Miami, FL, January 4-6. pp. 231-243.
http://doi.org/10.1007/978-3-319-27674-8_21

Smart Ambient Sound Analysis via Structured Statistical Modeling

Jialie Shen[1(✉)], Liqiang Nie[2], and Tat-Seng Chua[2]

[1] School of Information Systems,
Singapore Management University, Singapore, Singapore
jlshen@smu.edu.sg
[2] School of Computing, National University of Singapore, Singapore, Singapore
nieliqiang@gmail.com, chuats@comp.nus.edu.sg

Abstract. In this paper, we introduce a novel framework called SASA (Smart Ambient Sound Analyser) to support different ambient audio mining tasks (e.g., audio classification and location estimation). To gain comprehensive ambient sound modelling, SASA extracts a variety of acoustic features from different sound components (e.g., music, voice and background), and translates them into structured information. This significantly enhances quality of audio content representation. Further, distinguished from existing approaches, SASA's multilayered architecture seamlessly integrates mixture models and aPEGASOS (adaptive PEGASOS) SVM algorithm into a unified classification framework. The approach can leverage complimentary strengths of both models. Experimental results based on three large test collections demonstrate the SASA's advantages over existing methods on various analysis tasks.

Keywords: Ambient intelligence · Environmental sound analysis

1 Introduction

Rapid advances in mobile computing and multimedia technologies have led to an explosive growth of various kinds of audio information related with our daily life. In particular, ambient audio (environmental sound) contains rich information (semantic concepts) about activity, event, emotion and venue. As a consequence, smart techniques for ambient sound understanding have become more and more important due to potential applications such as home care, health monitoring, intelligent personal assistant and security protection. Essentially, ambient audio understanding can be modelled as an S-class categorization. The performance of technical solutions is largely dependent on their capabilities to model and capture discriminative features to identify one category of sound from others. Although traditional audio analysis schemes or algorithms designed for speech or music recognition could be applied to solve the problem, it is difficult for them to achieve promising performance in terms of accuracy and robustness. This is because most of ambient sounds contain rich sets of basic audio components

coming from different sources (e.g., human voice, animal sound, music, background events or activities). The acoustic structure and interplay between the elements could be highly complex and dynamic. For example, from the sound track recorded in open market or restaurant, we can easily find that voice or music is often intertwined with the non-stationary background signals from different events or activities (e.g., car engine start or music from shop or party). To develop robust and effective modelling of ambient sounds, it is essential to identify those basic audio elements and design advanced approach to model the highly unstructured information.

In recent years, several approaches have been proposed to apply statistical models or machine learning techniques for ambient sound analysis [4,6,13]. These methods commonly consist of two main steps: audio modelling and label identification via machine learning algorithms. In audio modelling, low level feature is extracted and used as content representation of raw ambient sound. Based on the features extracted, specific statistical models or machine learning algorithms (e.g., SVM, KNN or artificial neural network) can be constructed to identify label of the ambient sound. However, the schemes based on this paradigm suffer from low accuracy and poor robustness. The main reasons are:

- many of them only use single type of acoustic feature, which is not able to characterize complex ambient audio comprehensively.
- as mentioned before, ambient sound's structure and content could be very complex and this requires combination multiple types of acoustic features as effective content signature. However, existing studies simply ignore the effects of multiple acoustic features.
- they are mainly based on simple machine learning algorithms instead of advanced scheme, which could lead to more accurate and robust performance.

Motivated by the above discussion, a novel system called SASA (Smart Ambient Sound Analyser) is proposed to facilitate ambient sound characterization and analysis. Distinguished from the previous approaches only considering very limited amount of features and directly applying classic machine learning algorithms, our main research contributions include:

- In order to achieve effective audio modelling, we propose a novel structural analysis framework using the multiple features extracted from various kind of components to improve the system's performance. To the best of our knowledge, this is the first attempt to characterize unstructured ambient sound using a structured way.
- A probabilistic sound characteristic modelling method is designed based on mixture models and aPEGASOS (adaptive PEGASOS) SVM classifier to bridge the "semantic gap" between low level features and high level audio concept.

2 Multilayer Based Ambient Sound Understanding

SASA applies multilayered architecture consisting of three major functionality modules: sound preprocessing, structured sound modelling and effective understanding with advanced SVMs. Figure 1 illustrates detail architecture of SASA.

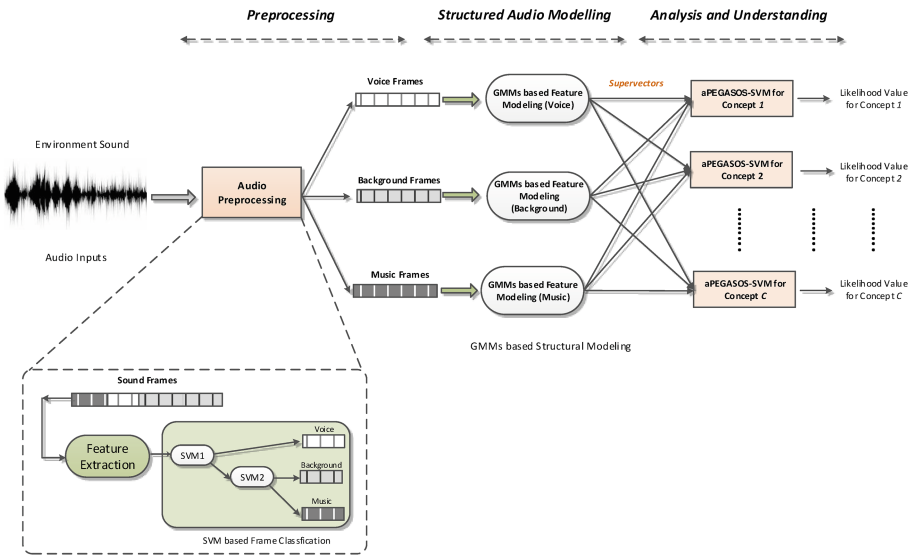

Fig. 1. Multilayer based ambient sound understanding framework

Audio preprocessing module aims to separate an incoming environmental sound into voice, music and background segments, and to extract related audio features from those segments. In the second layer of SASA, there are a collection of statistical models based on GMMs, one for each frame category. To analyse input environmental sound, different feature vectors are firstly extracted from the various segments. The feature vectors are then fed into the statistical models, generating a set of GMM supervectors, which serves as the input to the third module - effective understanding with advanced SVMs. The main functionality of the third module is to estimate which is the most relevant concept for input audio. The core of the third module is a set of SVM classifiers trained using aPEGASOS algorithm. Based on GMM supervectors, overall likelihood score for each concept is computed and the query audio is assigned to the k concepts (labels) with the top k likelihood scores. In the following subsections, we give a comprehensive introduction of the various modules and core algorithms used in the system.

2.1 Audio Preprocessing

In the first stage of the understanding process, our system classifies and labels the music, voice and background segments via the preprocessing module. We use the approach similar to the one presented in [9]. This process can be modelled as a problem of audio frame classification. A learning framework based on SVM can be applied and inputs are acoustic feature vectors combining MFCC and Wavelet. The algorithm demonstrates a promising performance because audio

segments containing human voices, music and other background events have very significant differences inside the spectral features.

The preprocessing phase consists of two sub-processes: acoustic feature extraction and frame classification with SVMs. After raw environment sound is received, it will be divided into multiple fixed length time-frames without overlapping. In our implementation, length of frame is set to be 0.75 sec because based on empirical study, it leads to the best performance in our experiments in terms of identification accuracy. The acoustic features extracted from each frame include: MFCC features and Wavelet. The acoustic features serve as input to a set of SVMs, which classify each frame into three possible categories: voices, music or background. We select SVM as classifier because it has demonstrated excellent performance on a range of categorization problems. In our framework, the LIBSVM [3] library is used for implementation and the kernel type is linear kernel. After the process, environment sound input au is segmented into three major components: voice frames au_v, music frames au_u and background frames au_b.

2.2 Structured Environmental Sound Modelling

Given that environmental sound can include multiple concepts and come from multiple sources, The second layer of SASA system consists of multiple modality models, one for each basic audio element (e.g., voice, music or background). Each model is made up of two parts: feature extractors and modelling via Gaussianization. In below, we will provide details about each component.

Acoustic Feature Extraction. To effectively represent and model the complex contents of ambient audio, our system extracts various features from different types of segments. In total, four different features are extracted to model ambient sound from various perspectives: timbre feature (TF), pitch feature (PF), instrument-based feature (IF) and wavelet-based feature (WF) [1]. In particular, the TF and PF capture information from the voice segments generated by human in ambient audio. The wavelet-based feature (WF) characterizes the music style and background acoustic events. The instrument feature (IF) is used to model the characteristics of typical instrument(s) in music frames. The details of the features used by the our framework:

- **Timbre Feature (TF):** Voice is a special instrument for human being and each person's timbre texture is unique due to the physical structure of voice fold. In SASA, LPCCs (Linear Prediction-based Cepstral Coefficients) from vocal segments are extracted to characterize this information (LPCCs are Linear Prediction Coefficients (LPCs) represented in the cepstrum domain).
- **Pitch Feature (PF):** To gain comprehensive modelling on human voice, it is crucial to take the harmonic and structural information about each person's voice into account. Since the algorithm proposed by Tolonen and Karjalainen

[1] Note that our method can be easily extended to consider more acoustic features.

is computational efficient and has superior capability in capturing human auditory perception, we apply it to extract pitch feature from human voice segments.

- **Instrument Based Feature (IF):** Instrument appearing in music segments can provide a lot of details about ambience. The main goal of IF is to characterize instrument configuration information of music frames. Our framework applies MFCC features as signature of instrument configuration. This is because MFCCs have been widely used to model timbre for purpose of instrument identification.
- **Wavelet based Feature (WF):** In SASA, WF is used as content representation to capture local and global dynamics of ambient sounds [1]. Wavelet analysis has been widely applied to model and characterize a wide range of audio information (e.g., background events [5] and music genre [7]). In our system, WF feature is based on the Daubechies Wavelet Coefficient Histograms (DWCHs) and mainly characterizes music genre and background events.

Multi-modality Based Modeling. The second layer of SASA system is a multi-modality based environmental sound characterization model. We apply the Gaussian Mixture Models (GMMs) for statistically modelling from different audio component perspectives since it has strong flexibility and effectiveness to represent complex data distribution. In SASA, one frame category corresponds to one GMMs and thus there are three GMMs. However, construction of GMMs based scheme on highly diverse distribution associated with environmental sound is not easy task. Since the number of sound frames for robust training is limited, parameter estimation of a GMMs robustly and accurately becomes very time-consuming. To solve this problem, a two-step adaptation approach is applied to develop GMMs including generative adaptation and sound segment based adaptation [2]. The key advantage of the approach include: (1) efficiency - a complex model can be developed using a small set of data and (2) simplicity - the model's output space is the Euclidean space, which can support fast search.

In generative adaptation, the GMMs is constructed with all training audio and then generate Universal Background Model (UBM). The UBM can be denoted as,

$$G = P(X|\theta) = \sum_{k=1}^{K} w_k N(X; \mu_k, \Sigma_k), \tag{1}$$

where w_k, μ_k and Σ_k are the weight, mean and covariance matrix of the kth Gaussian component, respectively. $X = \{x_1, x_2, ..., x_T\}$ is a set of input feature vectors extracted from audio segments. K is the total number of Gaussian components and the probabilistic density can be calculated using a weighted combination of K Gaussian densities,

$$N(X; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(z-\mu_k)^T \Sigma_k^{-1}(z-\mu_k)}. \tag{2}$$

The covariance matrix Σ_k is set to a diagonal matrix for reducing computational cost. To estimate the optimal values of key parameters $\{w_k, \mu_k, \Sigma_k\}$ of

UBM, we apply expectationmaximization (EM) algorithm, which is an iterative scheme to identify maximum a posteriori (MAP) estimates of model parameters.

2.3 Segment Based Adaptation

The goal of segment based adaptation is to modify the parameters of UBM to fit into the data distribution of audio segments. In SASA, each audio segment is represented as an collection of feature vectors, extracted from 30 ms window with step size of 5 ms. For different types of audio segments, we calculate different acoustics features and their details can be found in Sect. 2.2. The adaptation is carried out using maximum a posteriori (MAP). For Gaussian component k in the mixture model, we firstly compute,

$$pr(k, x_i) = \frac{w_k P_k(x_i|\theta_k)}{\sum_{j=1}^{K} w_j P_j(x_i|\theta_j)}, \tag{3}$$

$$\eta_k = \sum_{t=1}^{T} pr(k|x_i) \tag{4}$$

$$E_k(X) = \frac{1}{\eta_k} \sum_{t=1}^{T} pr(k|x_i)x_i, \tag{5}$$

The statistical values shown above can be then applied to adapt mean vector μ_k of each Gaussian component via the iteration process, in which $\hat{\mu_k}$ value at iteration l - $\hat{\mu_k}^l$ can be estimated by using:

$$\hat{\mu_k}^l = \alpha_k E_k(X) + (1 - \alpha_k)\hat{\mu_k}^{l-1} \tag{6}$$

where $\alpha_k = \eta_k/(\eta_k + r)$, $l = 1, ..., L$ is iteration number. r is smoothing factor, which can be fine-tuned empirically based on the total number of feature vectors extracted from each audio segment. It ranges from 5 to 20. We follow the approach introduced in [2] to gain an approximation of KL divergence of two models by:

$$d(\mu^\alpha, \mu^\beta) = \frac{1}{2} \sum_{k=1}^{K} w_k(\mu_k^\alpha - \mu_k^\beta) \sum_k^{-1} (\mu_k^\alpha - \mu_k^\beta) \tag{7}$$

where μ^α and μ^β denote supervector for model α and β. After audio segment based adaptation, the audio segment can be represented by super-vector,

$$SV = [v_1, v_2, \ldots, v_K] \tag{8}$$

where $v_k = \sqrt{\frac{w_k}{2}} \Sigma_k^{-\frac{1}{2}} \mu_k$. The supuervector serves as input to audio concept estimation module in the third layer of SASA.

2.4 Audio Concept Estimation Using SVM

The third layer of SASA system consists of a set of Support Vector Machines (SVMs) for the purpose of probabilistic estimation over different audio concepts. In order to support fast and effective SVM training, we develop an advanced

extension of PEGASOS [12] algorithm called aPEGASOS (adaptive PEGASOS), which enhances SVM based classification from two perspectives: (1) probabilistic based audio concept estimation and (2) adaptive sampling to effectively select discriminative audio segments as training examples instead of random projection.

With a given training set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where $\mathbf{x}_i \in R^n$ and $y_i \in \{+1, -1\}$, the PEGASOS algorithm is an efficient scheme aiming to effectively solve primal form of SVM $\mathbf{w}$ in an iterative fashion. At each iteration of the training algorithm, there are two key substeps: a stochastic gradient descent step and a projection step. The optimization goal of PEGASOS is to minimize the training error defined as:

$$f(\mathbf{w}; \mathcal{SA}_t) = \frac{\lambda}{2} \parallel \mathbf{w} \parallel^2 + \frac{1}{k} \sum_{(\mathbf{x},y)\in\mathcal{SA}_t} \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\} \tag{9}$$

The process requires T iterations and k the training samples for computing sub-gradients at each iteration. $\mathcal{SA}_t \subset \mathcal{S}$ consists of k samples selected using adaptive sampling scheme introduced late in the section from $\mathcal{S}$ at each iteration t. In the initial, we set the values of $\mathbf{w}$ to be zero. With learning rate $\eta_t = 1/(\lambda t)$ and a set of training samples $\mathcal{SA}_t^+$, parameter updating process at each iteration t has two steps,

$$\mathbf{w}_{t+\frac{1}{2}} = (1 - \eta_t\lambda)\mathbf{w}_t + \frac{\eta_t}{k} \sum_{(\mathbf{x},y)\in\mathcal{A}_t^+} y\mathbf{x} \tag{10}$$

$$\mathbf{w}_{t+1} = \min\{1, \frac{1/\sqrt{\lambda}}{\parallel \mathbf{w}_{t+\frac{1}{2}} \parallel}\}\mathbf{w}_{t+\frac{1}{2}} \tag{11}$$

$\mathbf{w}$ has non-zero training error when using $\mathcal{SA}_t^+$. Similar to the approach introduced in [15], once SVM training is completed, the learning method proposed in [10] is used to infer posterior probability $p^+(c = 1|x)$ of a given input belonging to certain class $c = 1$ as:

$$p^+(c = 1|x) = \frac{1}{1 + \exp(A\langle \mathbf{w}, \mathbf{x}\rangle + B)} \tag{12}$$

where scalar A and B can be estimated via seeking minimization of the error function by using the training data.

The quality of learning examples plays an important role in SVM training. Our system is trained using the acoustic features extracted from audio segments and thus notion of discriminative frame is very important because some of the segments enjoy more informative or distinctive cues about basic audio events or concepts (e.g., gun shot, party, dance and happy). However, how to select high quality training example is very challenging task because they generally need to satisfy two main requirements: (1) excellent representativeness and (2) high distinctiveness. To achieve this goal, we proposed a simple but effective algorithm based on semi-supervised principle. It involves two main steps:

– Seed selection: In our approach, human subjects are invited to provide a few good quality samples as seeds - a set of representative examples for three frame categories. In our current implementation, totally 5 human subjects are invited to select examples.
– Propagation: Based on seeds, we apply a common and robust neighborhood learning method proposed in [16] to construct n nearest neighborhood graph G_{ij} and identify training examples from unselected samples.

3 Experimental Configuration

In this section, we present the experimental settings for the performance evaluation, including competitive systems, testing datasets, evaluation task and performance metrics. All methods evaluated in this study have been fully implemented and tested on a server with 2.2 GHz Intel Xeon processor and 8 GB RAM.

3.1 Data Collections

Test collections play very important role in empirical study. The size of data sets used in existing study is quite small. To ensure accuracy, robustness and fairness of our empirical results, three large benchmark datasets are selected as testbeds in our evaluation. For three datasets, the sound files were converted to 16 kHz, 16 bit, mono audio files. More details about the three datasets can be found as below,

– Dataset I (DSI) is UrbanSound8K [11], which consists of 8,732 audio clips up to 4 sec in duration. The sound files are extracted from field recording crawled from the Freesound online archive [2]. Each clip contains one of 10 possible sound sources including: air conditioning, car horn, children playing, dog bark, drilling, idling engines, gun shot, jackhammer, siren, street music. Those sources are carefully selected from the Urban Sound Taxonomy [11] based on the frequency with which they appear in noise complaints provided by New York City's 311 service. All the sound clips have been manually annotated with human subject and a subjective judgment about whether the sound is in the foreground or background has been given.
– Dataset II (DSII): It consists of 1,873 audio clips and covers 25 concepts (e.g., dancing, singing, beach, playground, graduation,group of 3+,.....) belonging to 6 main categories including: activities, locations, occasions, objects, scenes and sounds [6]. The 25 concepts was defined by starting from a full ontology of over 100 concepts generated via user study done by the Eastman Kodak company [8]. To develop this dataset, totally 4,539 video was downloaded from YouTube by using most related keywords associated with the definition of these 25 concepts. To remove irrelevant commercial contents, raw dataset was manually checked before used for extracting accompanying sound tracks.

[2] http://www.freesound.org.

– Dataset III (DSIII) consists of 10,000 sound clip and covers 10 different recording locations including: Library, Office, Bathroom, Cafe, Restaurant, Kitchen, Living-room, Classroom, Subway, and Open mark. All of sound clips in the collection were recorded using Sony PCM-D100 high resolution audio-recorder. The duration of the audio files ranges from 5 sec to 30 sec. It covers 35 different concepts including high music, walking, chatting, cheering, typing, phone tone, door opening, door closing, crying, baby, male's voice, female's voice, TV sound, engine starting, crowd and others. Total duration of the whole collection is 20 h. All the concepts have been manually verified by human subjects and each sound in test collection could be associated with 1 - 5 concepts.

3.2 Methodology and Evaluation Metrics

Environmental sound analysis is one of the most fundamental components in various kinds of ambient intelligence applications. In order to conduct a comprehensive performance comparison of different schemes, our proposed system and the competitors are tested on the following two application driven tasks. They are,

– Task I - Sound understanding: The goal of the test is to evaluate and compare what are the accuracies achieved by different approaches in classifying the input environmental sounds. The datasets used for this task include DSI and DSII.
– Task II - Location estimation: Based on the input environmental sounds, we would like to test and compare how accurate different approaches can infer the venue. The dataset used for this test is DSIII.

As discussed above, the main goal of the system is to identify the suitable concepts related to input sound. Thus, our evaluation method focuses on how accurate the identification process is with different approaches for a particular database. We use the *accuracy* as the metric for evaluation: $Accuracy = \frac{NA}{NT}$, where NA is the number of sound correctly identified and NT is the total number of sounds used in the evaluation.

3.3 Competitors for Performance Comparison

We introduce several state-of-the-art methods on environmental sound recognition and location estimation based environmental sound analysis for comparison. For Task I (sound classification), we compare the performance of our system SASA against three state-of-the-art approaches including LEE [6], MP [4] and ESCLH [13]. For our, we consider three modality configurations (voice modality denoted by VM, music modality denoted by MM, background modality denoted by BM). SASA(VM), SASA(MM), SASA(BM), SASA(ALL) denote our proposed model built based on voice modality, music modality, background modality and the combination of all three modalities. To demonstrate advantages of our approach in location detection (Task II), we examine a wide range of possible

methods, including ABS [14], LEE [6], and MP [4]. For both Task I and Task II, mixture component number k for GMM in LEE and SASA is set to be 4, which is optimal value.

4 Experiment Results

This section presents a set of empirical studies to test and compare the performance of different systems on two tasks including environmental sound classification and location estimation.

On Environmental Sound Classification Table 1 shows the results of our experiments to test the accuracy of environmental sound classification using different schemes. The test was carried out on two different data sets - DSI and DSII. For each of the classifiers, fivefold cross validation is applied to ensure robustness of classification results.

Table 1. Environmental Sound Classification Accuracy Comparison.

Model	DSI	DSII
SASA(ALL)	84.3 %	89.3 %
SASA(BM)	73.5 %	77.5 %
SASA(MM)	50.2 %	56.9 %
SASA(VM)	50.6 %	55.1 %
ESCLH	73.2 %	80.2 %
LEE	74.2 %	81.2 %
MP	69.6 %	77.3 %

The first four rows of Table 1 indicate how the proposed SASA system performed using DSI and DSII. We find that the accuracies achieved by SASA(VM) and SASA(MM) (between 50 % and 60 % for all two datasets) are quite low. By taking background modality into consideration, SASA(BM) improves classification accuracy around 20 %. This result verifies the claim that background contains rich information about sound category and plays very crucial role in effective ambient audio classification and modelling. Further, SASA(ALL), which considers all three different sound elements, achieve a significant performance gain in classification accuracy, 84.3 % for DSI and 89.3 % for DSII. The results demonstrate that combining various acoustic cues from different sources can enhance classification effectiveness greatly. Meanwhile, the results provide strong empirical evidence that accurate classification cannot be achieved by considering only a single sound components.

In comparison with ESCLH, LEE and MP, sound classification with SASA(ALL) results in a great improvement in accuracy for all of the different datasets. For example, in case of DSI, SASA(ALL) improves accuracy by 16.1 % against MP. For DSI, the improvement is 15.5 % . Among all classification methods, SVMs give the best results, whatever kind of music descriptor is used. On

the other hand, good scalability is particularly important for large audio information systems, because the size of modern sound collection can be huge and changed frequently. Thus, it is important for analysis schemes to maintain stable accuracy against dataset size change. Based on Table 1, all the methods suffer from accuracy loss at some level when size of testbed becomes larger. However, SASA yields the lowest accuracy drop rates than do all other approaches. For example, when tested on DSI, SASA(ALL)'s accuracy is decreased by only 5 % comparing to SASA(ALL) on DSII. However, under same testing configuration, performance drops of other methods range from 8.6 % to 10 % , which is significantly higher.

On Location Estimation, Table 2 summaries location estimation effectiveness of the SASA, ESCLH, LEE and MP techniques. It is shown in the first four columns that comparing to other SASA variants, SASA(MM) built based on music components is the worst in terms of estimation accuracy rates. Furthermore, although the SASA(VM) demonstrates better performance than SASA(VM), gain is very marginal. This is because music and voice segments only capture very limited amount information about one location. In fact, the results clearly demonstrate that SASA(BM) outperforms the SASA(VM) and SASA(BM) greatly. Once again, this results provide strong support on the claim that background elements contains more information about one location than other two basic elements. More importantly, SASA(ALL) achieves the best estimation accuracy comparing to six other methods. In addition, it is worth noticing that integrating effects of additional sound elements bring SASA nice lift in accuracy improvement. For example, by considering two additional sound elements, accuracy improvement over SASA(VM), SASA(MM), and SASA(BM) is 54.3 %, 60.1 %, and 12.0 %, respectively.

Table 2. Location Estimation Accuracy Comparison.

Schemes	DSIII
SASA(ALL)	81.2 %
SASA(BM)	72.5 %
SASA(MM)	50.7 %
SASA(VM)	52.6 %
ABS	70.2 %
LEE	71.2 %
MP	56.6 %

5 Conclusions

In this paper, we present an intelligent framework, called SASA, to facilitate effective environmental sound analysis. The system has been fully implemented

and tested using different datasets. As shown in our experimental evaluation, the SASA system not only demonstrates significantly better effectiveness on audio classification and location estimation over the state-of-the-art systems, but also achieves good robustness against acoustic distortion. The research opens up several promising directions for future study. Firstly, we plan to test the framework over larger dataset with higher complexity. Further, it is interesting to investigate how to develop advanced acoustic modelling scheme to support accuracy and robustness improvement.

Acknowledgments. This work was partly supported by Singapore Ministry of Education Academic Research Fund Tier 2 (MOE2013-T2-2-156), Singapore.

References

1. Bailey, T., Sapatinas, T., Powell, K.J., Krzanowski, W.J.: Signal detection in underwater sound using wavelets. J. Am. Statist. Ass **93**, 73–83 (1998)
2. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
3. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM TIST **2**(3), 27 (2011)
4. Chu, S., Narayanan, S.S., Kuo, C.C.J.: Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1142–1158 (2009)
5. Feng, Z.R., Zhou, Q., Zhang, J., Jiang, P., Yang, X.W.: A target guided subband filter for acoustic event detection in noisy environments using wavelet packets. IEEE/ACM Trans. Audio Speech Lang. Proc. **23**(2), 361–372 (2015)
6. Lee, K., Ellis, D.P.W.: Audio-based semantic concept classification for consumer video. IEEE Trans. Audio Speech Lang. Proc. **18**(6), 1406–1416 (2010)
7. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: Proceedings of ACM SIGIR Conference, pp. 282–289 (2003)
8. Loui, A.C., Luo, J., Chang, S., Ellis, D., Jiang, W., Kennedy, L.S., Lee, K., Yanagawa, A.: Kodak's consumer video benchmark data set: concept definition and annotation. In: Proceedings of ACM MIR 2007, pp. 245–254 (2007)
9. Lu, L., Zhang, H., Li, S.Z.: Content-based audio classification and segmentation by using support vector machines. Multimedia Syst. **8**(6), 482–492 (2003)
10. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances In Large Margin Classifiers, pp. 61–74. MIT Press, Cambridge (1999)
11. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: 22st ACM International Conference on Multimedia (ACM-MM 2014) (2014)
12. Shalev-Shwartz, S., Srebro, N.: SVM optimization: inverse dependence on training set size. In: Proceedings of ICML 2008, pp. 928–935 (2008)
13. Su, F., Yang, L., Lu, T., Wang, G.: Environmental sound classification for scene recognition using local discriminant bases and HMM. In: Proceedings of ACM MM 2011, pp. 1389–1392 (2011)

14. Tarzia, S.P., Dinda, P.A., Dick, R.P., Memik, G.: Indoor localization without infrastructure using the acoustic background spectrum. In: Proceedings of ACM MobiSys 2011, pp. 155–168. ACM (2011)
15. Zhang, B., Shen, J., Xiang, Q., Wang, Y.: Compositemap: a novel framework for music similarity measure. In: Proceedings of ACM SIGIR Conference, pp. 403–410 (2009)
16. Zhou, D., Schölkopf, B., Hofmann, T.: Semi-supervised learning on directed graphs. In: NIPS 2004, pp. 1633–1640 (2004)