SOUND SOURCE RECOGNITION TECHNIQUE FOR CONSTRUCTION NOISE CONTROL

M Yang School of Architecture, University of Sheffield, UK
J Kang School of Architecture, University of Sheffield, UK

1 INTRODUCTION

It is known that the subjective evaluation/preference of soundscapes, i.e., holistic sound environment[1], operate on the basis of identification of physical sound sources[2]. Even with the same sound level, people's degree of tolerance varies among different types of environmental sound, such as sounds from nature or transport[3].

In the large construction project of London Bridge Station (LBS) redevelopment, although conventional noise monitoring and predicting on sound level have been taken, complaints have still been received concerning the construction/transportation noise from the local residents. Moreover, the noise might not only result from the construction site, but also the nearby transportations and other construction sites, which brings difficulties in the noise management and control. Thus, there is a recognised need to develop innovative techniques for monitoring the holistic sound environment and recognising automatically the noise sources, in order to help noise management and control, not only for LBS project, but also many other projects alike.

This paper explores the possibility of automatically identifying construction noise sources from general urban background noise, based on the techniques of single sound source recognition developed in the previous research of the authors[4, 5]. This paper uses the construction project of LBS redevelopment, carried out by Costain, a leading British construction and civil engineering company, as a case study site.

2 PREVIOUS RESEARCH OF SOUNDSCAPE RECOGNITION

A number of studies have aimed to build a system that can become the basis for an automatic analysis tool by identifying sound events in soundscapes. Basically, sound recognition (both for speech/music and environmental sounds) is achieved by two phases: first feature extraction, followed by classification. The feature extraction (or say parameterisation) stage produces a set of characteristic features for sound to reduce the complexity of the data before it reaches the classifier. The classification stage then recognises the sound based on the extracted features[6].

For single environmental sound recognition, Cowling and Sitte[6] comprehensively compared the different techniques that were typically used in speech and musical instrument recognition in their suitability for environmental sound identification. From the combinations of feature extraction techniques (such as frequency extraction, homomorphic/ Mel frequency/ Bark frequency cepstral coefficients, linear prediction cepstral (LPC) coefficients, perceptual linear prediction (PLP) features, short-time Fourier transform (STFT), fast (discrete) wavelet transform (FWT), and continuous wavelet transform (CWT)) and classification techniques (such as dynamic time warping (DTW), hidden Markov models (HMM), learning vector quantization (LVQ), self-organising maps (SOM), artificial neural networks (ANN), long-term statistics, maximum likelihood estimation (MLE), Gaussian mixture models (GMM), and support vector machines (SVM)), Cowling and Sitte found that the combination of CWT or Mel frequency cepstral coefficients (MFCCs) with DTW produced the best results, with a classification rate of about 70%.

For real-world environmental sounds, i.e., multiple sound sources that are mixed together, Bunting et al.[7], in the project of instrument for soundscape recognition, identification and evaluation (ISRIE), used time-domain signal coding (TDSC) combined with LVQ network, and employed a source separation algorithm prior to the feature extraction and classification stages. The accuracy varied among sound categories, and was not high for some categories. Krijnders et al.[8] also improved the signal-driven classification, performed by segment and feature extraction from a time-frequency cochleogram and machine learning techniques, by creating expectancies of sound events based on context information through a dynamic network. In contrast to these methods, Aucouturier et al.[9, 10] proposed to directly recognise soundscapes holistically, using the “bag-of-frames” approach, without the prior identification of constituent sound sources. It represented signals as the long-term statistical distribution of frame-based MFCC vectors, using GMMs, and proved a precision of 0.9 in the first five nearest neighbours.

In a previous research of Yang and Kang[4, 5] of single environmental sound identification, without using the MFCCs for feature extraction as in majority of the above studies, a range of psychoacoustic, music, and 1/f noise indicators were considered, such as loudness, pitch, and fluctuation strength[11]. Combined with machine learning or mathematical models for classification, such as discriminant function analyses and ANNs, the prediction accuracies were above about 98% for the three natural sound categories and one urban sound category (when fountain were labelled as water sound in one natural sound category). The method achieved high prediction accuracies, although the accuracies were not directly comparable across the different studies, since the sound samples and statistic methods of accuracy calculation differed. However, while MFCCs represent the spectrum of spectrum of signal on Mel frequency scale[12], they may be more suitable for feature extraction of speech and music rather than environmental sounds, since speech/music often consists of harmonic tonal components, whereas environmental sounds often consist of broadband noise.

In this paper, this sound source identification method is further developed and tested for recognition of specific sound sources in the real, complex environment, which have multiple sound sources mixed together, with a particular focus on construction noise sources.

3 METHODS

3.1 Overall sound source recognition method

For the real-world environmental sounds, in this paper, before the feature extraction and classification stages, a sound signal is first decomposed into successive frames of short duration. It is assumed that in the short duration of time (frame), the sound events remain relatively constant. Then the feature extraction and classification are applied based on each frame. The procedure of the whole sound source recognition method is shown in Figure 1.

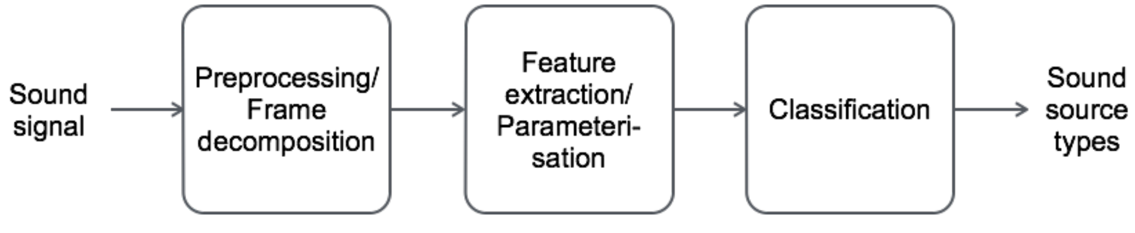
Sound signal
Preprocessing/ Frame decomposition
Feature extraction/ Parameteri-sation
Classification
Sound source types

Figure 1. Sound source recognition method

Here, the frame length of 1 minute with half overlap is used. The duration or length of frame is determined by the acoustic characteristics of the sound types to be identified, by examining the variation of spectrum with time of the sounds, as discussed following in Section 4.1. The 1-min duration is generally long enough to cover the dynamic characteristics of the sound events, and also shorter than the duration of each type of sound event.

3.2 Sound Recording

The primary sound data for this research were recorded by Southdowns Company, provided by Costain LBS redevelopment project. The station remains operational during the construction. The sound database forms a digital archive of soundscape in the construction area, which includes the sound pressure levels and sound recordings at some particular receivers made with professional microphones (CBA Spindlewood Ltd) and recorders, monitored all day long. The sound recordings have frequency responses from about 10 to 2800Hz, mono channel, and were stored in the digital format of AAC, at the sample rate of 22,050 Hz.

From the large database of recordings, in order to search for the appropriate indicators of feature extraction, 21 1-min recording segments, which represent the typical sound sources in that area, are selected and clipped as samples for the analyses. The main sound types include demolition, piling, vibration machine, public address system (PA system) of the railway station, airplane, and background noise. In each recording segment, only one dominant sound type can be heard.

To develop the classification method and test the recognition performance, 10 approximately 1-hour recordings are selected as samples from the database. Each of the recordings is then decomposed by successive frames (frame length of 1min and half overlap) for further processing and recognition. There are 118 frames in each recording. In total, 1179 frames are examined (one audio file is faulty).

3.3 Sound type identification by human's listening

To know the sound source types of recordings and compare them with the automatically identified types, the sound types in each 1-min recording segment and in each frame of the 1-hour recordings were identified by one of the authors through listening. The sound recordings were reproduced with Sennheiser HD 558 headphones, from a stationary computer. The sounds were presented using ArtemiS software package[13], and could be replayed as many times as needed while identifying the sound event types. Within a frame, the most dominant and arresting sound types were identified (maximum 3 types).

4 SOUND SOURCE RECOGNITION

4.1 Indicators for feature extraction

To look for the indicators of feature extraction, the variation of spectra of the 21 1-min recording segments are analysed using ArtemiS[13]; parts of the results are shown in Figure 1. By analysing the spectra with time, some initial results of the characteristics of the different sound types are drawn. For example, in Figure 1 (a), the alarm sound of a reversing construction vehicle have distinct pure tone component, indicated by the horizontal discontinuous line in the frequency range between 1k and 2k Hz. Demolition sounds are characterised by sudden changes in intensity, which happen simultaneously in all frequencies, indicated by the vertical lines in the spectrum. The sound of PA system, in Figure 1 (b), has its particular pattern in spectrum with time around the frequency range of 1k Hz, the intensity changes asynchronously at different frequencies in this range.

To reflect the respective characteristics of the different sound types, a number of indicators are then selected/developed. A series of indicators are examined, which include sound pressure level (SPL), A-weighted SPL, and psychoacoustic parameters, such as loudness, sharpness, tonality, pitch, and rhythm, calculated using ArtemiS[13] or Matlab with MIRtoolbox[14]. The results show that among the indicators, the acoustic/psychoacoustic parameters may not show the differences of the characteristics of the different sound types, but some specific indicators do, such as peak frequency in spectrum, event (peak) density and average value of peaks in spectral flux[15], and periodicity. For example, only reversing alarm sounds have distinct peak in spectrum at about 1.2k to 1.4k Hz. Demolition sounds have high average values of peaks in spectral flux, whereas PA system and background noise have relatively low average values. More detailed results are available upon request.

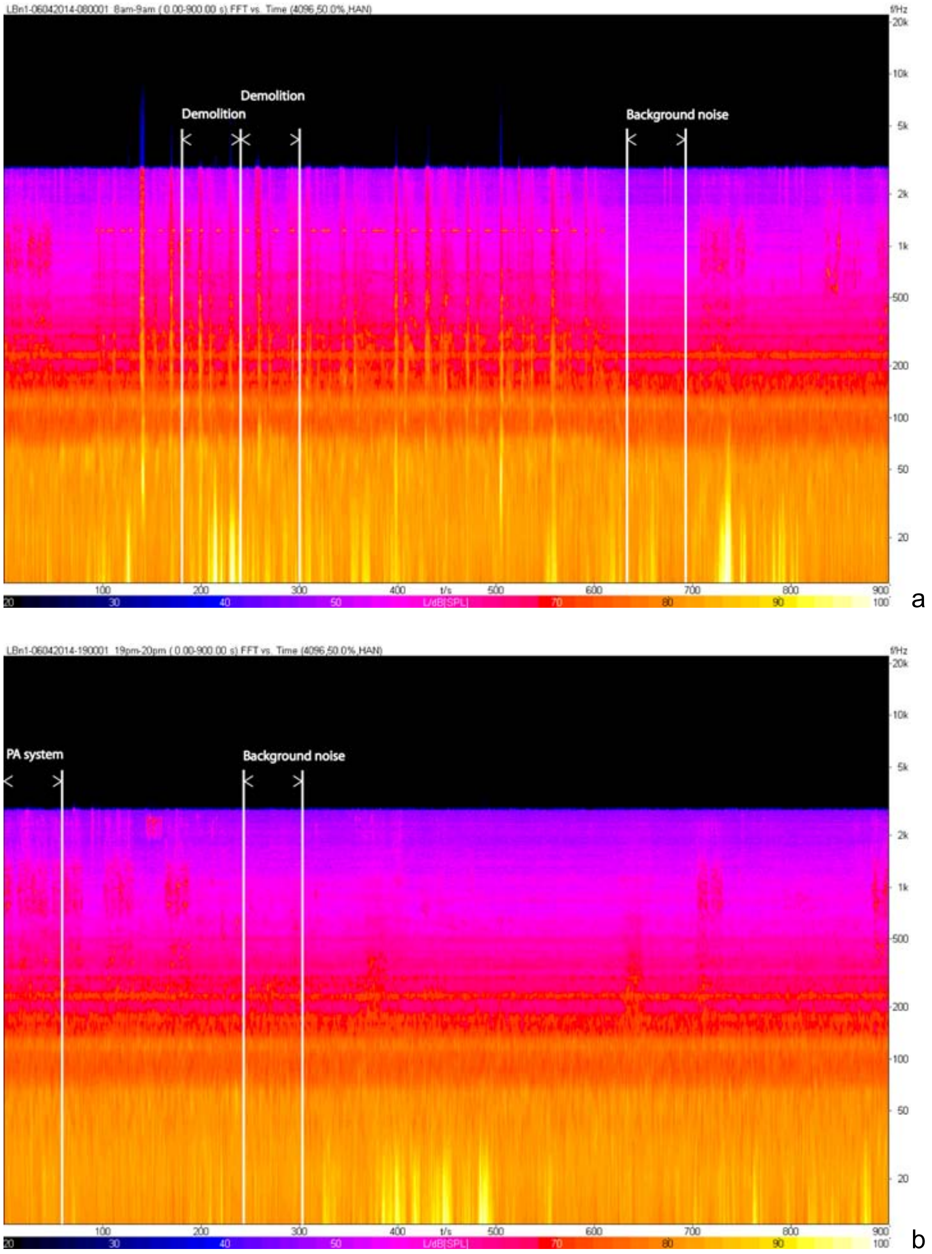

Figure 2. Spectra vs. time of the sound recordings. Each figure presents 15 min in duration of a recording, in which the 1-min recording segments are at the indicated areas between the two vertical white lines.

4.2 Automatic classification

The method for the classification stage is developed and tested using large size of samples. The 10 1-hour recordings are used. The indicators developed above are calculated for each frame of the recordings. Since multiple sound source types can be identified for each frame (that is, each frame is not restrict to one type of sound), each of the sound types to be identified is estimated respectively whether or not happened in each frame. To identify each type of sound, one or more indicators are used. Instead of machine learning models used in the previous research, which are computationally expensive, simple evaluation of the value ranges of the results of the indicators is used for the automatic recognition of sound types.

With these methods, the sound types identified in each sample frame are compared with those identified by human's listening. The numbers and percentages of correctly identified sample frames are shown in Table 1. Here, the accuracies of two sound types are shown, i.e., reversing alarm of construction vehicles and sound of demolition, both of which are arresting and annoying noises in construction sites. For each sound type, the table presents both the numbers/percentages of correctly identified frames containing a particular sound type and of frames not containing that sound type. Overall, it can be seen that the accuracy of recognition is generally high, which is 90% for identifying reversing alarm and 78% for demolition sound.

There might be a number of possible causes for the amount of recognition error. First, a very few sound events, which are not to be identified in this paper, might have similar characteristics to the above sound types in the indicators used for processing. For example, beep sound from a vehicle and reversing alarm both have prominent peaks/peak in spectrum, although they can be distinguished with additional indicators. Second, the characteristics of a sound type to be identified in the indicators might become unobvious and tend to be masked in some frames, since its loudness may be relatively small compared to the other sound types or background noise, or its duration is short in certain frame. Third, the error might also come from the judgment of human's listening. As only the most dominant and arresting sound types (maximum three) are identified through listening in a frame, the judgment may be uncertain sometimes.

Table 1. Recognition accuracies of reversing alarm and demolition sound

Human's listening	Signal processing		
	Correct	Total	Percentage
Non reversing alarm	855	942	90.8%
Reversing alarm	200	237	84.4%
Total	1055	1179	89.5%
Non demolition	745	983	75.8%
Demolition	169	196	86.2%
Total	914	1179	77.5%

5 DISCUSSIONS AND CONCLUSIONS

In this research, a number of typical construction noise sources have been identified from the real, complex environmental sound. This research used a number of methods different with those in the previous studies of environmental sound recognition. First, prior to the feature extraction and classification stages, without a source separation algorithm, the paper processed the environmental sounds directly by decomposing signal into successive short duration frames, and identified the sound source types within each frame, based on the assumption that sound events remain relatively constant in a short duration of time. Second, for the feature extraction, different with previous methods that mainly used MFCCs as indicators, the paper applied a range of specific acoustic/psychoacoustic and music related indicators, since MFCCs may be more suitable for speech and music that often consist of harmonic tonal components rather than environmental sounds, which often consist of broadband noise. Third, for the classification stage, simple evaluation of the value ranges of the result of the indicators was used in this paper, which is much more computationally inexpensive than machine learning algorithms that have been frequently used for recognition tasks, and thus more feasible for large-scale industry applications.

Using the methods, the accuracy of the recognition of a number of typical construction noise sources is high, based on the case study site of the construction project of LBS redevelopment. The accuracy is 90% for reversing alarm of construction vehicles and 78% for demolition sound, even though the quality of the sound recordings used here is not very high. The results show the possibility of the methods in automatic identification of sound source types from overall urban background noise, and also the potential for practical applications.

The accuracy of the sound source recognition can be further increased by using more indicators for the feature extraction, narrowing the evaluation conditions for the classification stage, or using machine learning algorithms if needed. Larger sample size used for developing the indicators and classification method can also increase the accuracy. In the current research project, more sound source types are being identified.

It is expected that the technique of automatic recognition of environmental sounds developed in the research would help address the existing noise problems for noise control, further benefit general areas in noise monitoring/mapping, and have more applications in construction sectors and beyond.

6 ACKNOWLEDGEMENTS

This research was supported by Innovation, Impact and Knowledge Exchange (IIKE) Collaborative R&D Award scheme at the University of Sheffield, funded by the Engineering and Physical Sciences Research Council (EPSRC). In-kind support was also obtained from Costain LBS project. Special thanks go to the Centre for Smart Infrastructure and Construction at University of Cambridge and Southwark Council for the valuable discussions of industry applications.

7 REFERENCES

1. R.M. Schafer, The Tuning of the World, Knopf. (1977).
2. D. Dubois, 'Categories as acts of meaning: The case of categories in olfaction and audition', Cognitive Science Quarterly 1, 35-68. (2000).
3. J. Kang, Urban Sound Environment, Taylor & Francis incorporating Spon. (2006).
4. M. Yang and J. Kang, Automatic identification of environmental sounds in soundscape, Proc. 42nd Inter-noise, Innsbruck (2013).
5. M. Yang, Natural and Urban Sounds in Soundscapes, PhD thesis, The University of Sheffield. (2013).
6. M. Cowling and R. Sitte, 'Comparison of techniques for environmental sound recognition', Pattern Recognition Letters 24, 2895-2907. (2003).
7. O. Bunting, J. Stammers, D. Chesmore, O. Bouzid, G.Y. Tian, C. Karatsovis and S. Dyne, Instrument for soundscape recognition, identification and evaluation (ISRIE): Technology and practical uses, Proc. Euronoise, Edinburgh (2009).
8. J.D. Krijnders, M.E. Niessen and T.C. Andringa, 'Sound event recognition through expectancy-based evaluation ofsignal-driven hypotheses', Pattern Recognition Letters 31, 1552-1559. (2010).

9. J.-J. Aucouturier, B. Defreville and F. Pachet, 'The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music', J. Acoust. Soc. Am. 122, 881-891. (2007).
10. J.-J. Aucouturier and B. Defreville, Sounds like a park: A computational technique to recognize soundscapes holistically, without source identification, Proc. 19th ICA, Madrid (2007).
11. M. Yang and J. Kang, Applicability and application of music features in soundscape, Proc. AIA-DAGA, Merano (2013).
12. B.P. Bogert, M.J.R. Healy and J.W. Tukey, The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking, Proc. the Symposium on Time Series Analysis, 209-243. (1963).
13. HEAD acoustics. GmbH, http://www.head-acoustics.de/eng/nvh_artemis.htm/ (Last accessed on 20 Jul 2013).
14. O. Lartillot and P. Toiviainen, A matlab toolbox for musical feature extraction from audio, Proc. 10th DAFx, 237-244. Bordeaux (2007).
15. J. Laroche, 'Efficient tempo and beat tracking in audio recordings', Journal of the Audio Engineering Society 51, 226-233. (2004).