Environmental Sound Recognition by Measuring Significant Changes in the Spectral Entropy

Jessica Beltrán-Márquez[1], Edgar Chávez[2], and Jesús Favela[1]

[1] CICESE, Mexico
{jbeltran,favela}@cicese.mx
[2] Universidad Michoacana, Mexico
elchavez@umich.mx

Abstract. Automatic identification of activities can be used to provide information to caregivers of persons with dementia for identifying assistance needs. Environmental audio provides significant and representative information of the context, making microphones a choice to identify activities automatically. However, in real situations, the audio captured by microphones comes from overlapping sound sources, making its identification a challenge for audio analysis and retrieval. In this paper we propose a succinct representation of the signal by measuring the multi-band spectral entropy of the signal frame by frame, followed by a cosine transform and binary codification, we call this the Cosine Multi-Band Spectral Entropy Signature (CMBSES). To test our proposal, we created a database of a mix-up of triples from a collection of nine environmental sounds in four different signal-to-noise ratios (SNR). We codified both the original sounds and the triples and then searched all the original sounds in the mix-up collection. To establish a ground truth we also tested the same database with 48 people of assorted ages. Our feature extraction outperforms the state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) and it also surpass humans in the experiment.

1 Introduction

Caring for a Person with Dementia (PwD) is a demanding task, often performed by a close relative that faces stress and burnout providing assistance 24/7. Context-Aware technologies can be designed to assist PwD and their caregivers [1]. Providing caregivers with information of the activities performed by the PwD would allow them to opportunistically identify when the elder might need assistance, for example detecting if the PwD is at risk, or if she is doing something unusual [2].

Research on automatic activity estimation has been conducted using a variety of sensors, such as accelerometers, video cameras, and readers of Radio Frequency Identification (RFID) tags. While microphones and video cameras are the most ubiquitous [3], microphones have the advantage of capturing information in all directions and are robust to changes in position and orientation, thus making data collection less intrusive. Audio also provides significant and

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2012, LNCS 7329, pp. 334–343, 2012.

highly representative information of context and requires only a microphone, which is already available in any mobile phone [4]. As a downside, audio cannot be used with activities that do not produce characteristic sounds, and also noise is always present in real situations thus making audio difficult to analyze.

Some authors in audio-based activity recognition use categories of high level semantic scenes ("meeting","library",etc.), while other approaches take the categories as sound events which can be defined as structurally meaningful units coming from a single source (e.g "footsteps" "door") [5]. Most attempts classify a small set of sounds in relatively limited and predefined contexts of interest. It is still an open problem the consideration of the idiosyncrasy of specific applications and the profile of the target users [6].

Narrowing the universe of scenes and audio events likely to occur may help to improve the recognition. Sound events collaborate to form auditory scenes and a combination of them can be used to infer a high level semantic scene, this combination can be adapted to specific users. Similarly, knowledge of the current auditory scene can help to disambiguate the possible multiple interpretations of sounds [7].

Automatic audio analysis in real situations is not an easy task because the audio captured by microphones contains a mixture of different sources. Humans are able to follow one sound source, for example, when several persons are talking we are still able to attend to one speaker [8]. But this ability is still challenging for automatic methods.

In this paper we present a method to identify sound events unprocessed. The aim is addressing the problem of activity recognition in mobile and real environments, hence we use monophonic audio as the one captured by a single microphone, such as the ones included in mobile phones. It is worth mentioning that audio identification can be used also in other applications beside activity recognition, for example for automatic annotation and segmentation of audio streams. The method used in this study consists in a classification based on features extracted using a binary signature formed with the entropy of the signal in 24 Barks sub-bands. The representation of the binary signature is very compact and allows signature comparison to be performed efficiently.

The rest of the paper is organized as follow: Section 2 presents the previous work, in section 3 we explain the process used for feature extraction, in section 4 we describe how the database used for the experiments was constructed, section 1 describes the experiment conducted and the results, section 6 presents the conclusions of the paper and provides directions for future work.

2 Previous Work

The research in activity recognition using auditory information usually performs the process shown in figure 1 below. The signal acquisition is done through microphones with a proper election of a sampling frequency and bit resolution. The pre-processing includes basic operations like filtering or smoothing. The audio feature extraction aims to find an adequate representation of the auditory

signals, the representation should be specific enough to differentiate between environmental sounds and general enough to allow variations of the same sounds to be identified as one. Finally a classifier is used to obtain the categories of the input audio.

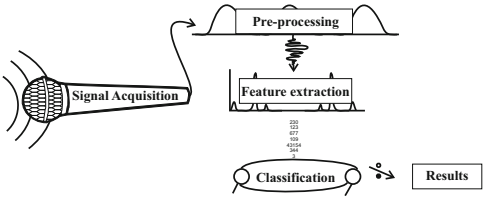

Fig. 1. Stages of the process of auditory recognition

Audio features of diverse nature have been tried in audio identification tasks. The Mel Frequency Cepstral Coefficients (MFCC) are the most popular features in audio information analysis and retrieval[9]. This feature was firstly studied for automatic Speech Recognition, and also have been used for scene or activity recognition. The MFCC represents the spectral envelope of the signal in each frame. In the classification front we find a great diversity also, the most used are Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM).

Below we briefly discuss some studies aiming at inferring activities. It is worth noticing that they can be found in very diverse areas in the scientific literature. One such area is *activity recognition* (AR), the second one corresponds to *computer scene recognition* (CASR) or *computer auditory scene analysis* (CASA). CASR studies aim at recognizing scenarios (restaurant, street, etc.) which can be used to infer activities. AR, on the other hand, focus on the final activities or events (hammering, jogging, etc). CASR studies generally use one microphone and take the audio input as an holistic signal while AR studies usually use more than one microphone and also include another sensors (like accelerometers) which are strategically located.

AR is explored in [10], where a linear discriminant analysis is performed on the FFT of the signal. The classification is done by measuring the euclidean distance to the mean of each class which is considered as a representative. They combine the signal analysis with accelerometers for arm motion to classify. They get from 85% (using an screw) to 100% classification in most of the activities. In [11] they use frequency domain features and a GMM as a classifier with the following results as recall percentage, brush 87.75%, wash 97.7%, shave 96.29%, electric brushing 86.30%, electric shaving 89.85%, and Other Activities = 89.28%. In [12] they use time and frequency features to create a 20 dimensional vector, and then to train an HMM, the recognition rate of sounds from ambients like Restaurant, Street, Lecture, Conversation and Other, they obtain 88.6 and improve to 94.4 using Linear Discriminant Analysis (LDA). They also show other results using more sensors. In [13] they use MFCC with a HMM classifier to get

the following results in accuracy percent, street (traffic) 93%, bus 81%, building site 100%, office 100%, lecture 100%, car (city) 100%, shopping mall 89%, street (people) 100%, supermarket 100%, laundrette 90%, car (highway) 98% and train 99%. One interesting approach is [14] where they use a matching pursuit algorithm and the MFCC, they also try GMM together with the K-Nearest Neighbor (KNN) classifier. The average accuracy rate is about 83.94 using GMM and 77.3 using KNN for ambients like Inside restaurants, playground, street (three types), train passing, inside vehicles, inside casinos, nature-daytime, nature-nighttime, ocean waves, running water/stream/river, raining/shower and thundering. The final example is [15] where they use MFCC plus time and frequency features with a Bayes Classifier and HMM, they obtain the following (% precision-% recall) pairs, walking 93-53, driving cars=100-100, riding elevators=78-80, riding a bus=25-90. This approach uses a hierarchal schema that consists in three coarse categories (music, voice, and ambient sounds) and finer intra categories, finally they take in consideration the design for mobile phones and proposes personalized aggregation of classes.

So far, none of the studies mentioned address the problem of environmental sound analysis when sounds overlap in a mixture of signals. Blind source separation allows the identification of different sources, for example through the use of *independent component analysis* (ICA), but it requires a fixed number of sensors corresponding to the sources to be identified. In this work we considered the audio as coming from a single source hence we cannot use this technique.

A recent approach [16] addresses the identification of mixture overlapped signals problem by doing preprocessing of the sound with source separation through an unsupervised non-negative matrix factorization (NMF) to produce separated tracks. Then a feature extraction step is performed in the tracks by the use of MFCC, and finally classification is done with a 3 state HMM. The reported results show a 52.6 of F_{score} used in the overall context experiment, where $F_{score} = 2\frac{precision \times recall}{precision + recall}$

The feature used in this paper is a modification of the Multi Band Spectral Entropy Signature (MBSES) which has been used in music information retrieval proving robustness to noise, equalization and loudness. In this feature, the entropy is calculated for every of 24 Bark bands per frame and then converted to a binary signature [17]. Below we provide a more detailed explanation of the feature.

3 Feature Extraction: MFCC and MBSES

Due to space restrictions we describe very briefly how to compute the MFCC and the MBSES. More details are found in the papers [18] for the MFCC and [17] for the MBSES.

Mel Frequency Cepstrum Coefficients (MFCC). We followed the process described in [18] to calculate MFCC. We used a pre-emphasis computed with the filter $H(z) = 1 - az^{-1}$, with $a = 0.95$. After the pre-emphasis the signal was

windowed in frames of 8192 samples with a Hamming window and an overlap of 50%. Every frame was transformed to the frequency domain through the N-point Fast Fourier Transform (FFT). Then we applied a 24 triangular band-pass filter-bank in Mel Scale from 0Hz to the Nyquist frequency. As a final step we obtained the Discrete Time Cosine Transform (DCT) of the log of the energy of each band.

The Multiband Spectral Entropy Signature (MBSES). To compute the MBSES the signal is framed in 8192 samples with an overlap of 50%. After this, a Hanning window is applied to every frame and then the N-point FFT is computed. The resulting frequency frames are then split critical bands 1 to 24 according to the Bark scale. For every critical band the spectral entropy is determined using the equation $H = \ln(2\pi e) + \frac{1}{2}\ln(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2)$. With σ_{xx} and σ_{yy} are the variances of the real and imaginary part, and σ_{xy} is the covariance between the real and the imaginary part of the FFT coefficients in the corresponding bands. With this process we obtain the analog of the spectrogram which computes the amount of energy in time and frequency; the *entropygram* gives the amount of information along the time for every critical band in the Bark scale. An illustration of what is obtained with the features is observed in figure 3 for the MFCC and the MBSES.

The entropygram obtained can be made more robust by a binarization, which eliminates a curious shift in the values of the entropy. This is achieved taking the sign of the time derivative of the *entropygram* as indicated in equation 1 where the bit corresponding to band b and frame n (i.e $bit(n, b)$) is determined with the sign of the difference of the *entropygram*'s entries H(n,b) and $H(n-1, b)$.

$$bit(n,b) = \begin{cases} 1 \; if \; H(n,b) - H(n,b-1) > 0 \\ 0 \; if \; H(n,b) - H(n,b-1) \leq 0 \end{cases} \tag{1}$$

Once binarized, the sound is a string of 24 bit symbols. The length of the string depends on the duration of the audio. This also allows to use Hamming distance to compare signatures. The binary version can be seen in figure 3.

3.1 The *MBSES

The derivative in the binarization process above introduces noise to the signal because only the sign is taken into account. This implies that any change in the slope, regardless of its size, equally contributes in the binarization. We will introduce a third symbol, the $\star$, to mark a non significant change in the slope, one that should be ignored. To this end we compute the derivative using central differences and then we apply the arctan$(\cdot)$ to the numerical derivative to obtain the slope in radians. If the absolute value of the slope pass a fixed threshold then we take the sign into account, if not, we put an $\star$. If s denote the slope, then the bit extraction follows as equation 2.

$$bit(n,b) = \begin{cases} \star & if \; s > \alpha \\ sgn(s) & \text{in other case} \end{cases} \tag{2}$$

An asterisk in a signature indicates indetermination when signatures are compared. This lead to a new distance explained in figure 2. An illustration of the signature is in figure 3 where the gray colored squares are the asterisks. Please notice that this binarization can be applied to plain MBSES or the CMBSES defined below.

B \ A	0	1	⋆
0	0	1	0
1	1	0	0
⋆	0	0	0

Fig. 2. The Hamming distance with asterisks

3.2 The CMBSES

Here the idea is to compute the MBSES as usual, and then compute a discrete cosine transform of the 24 entropy values in each frame. This is similar to the last step of the MFCC but instead of using the log of the energy, we used the entropy for this representation.

4 Database

We collected nine audio segments representing various types of sound sources from `http://www.freesound.org` and with 44100 Hz of sampling frequency and 16 bits depth (baby crying, keys, siren sound, bird singing, tooth brushing, music with voice, music without voice, male voice, and female voice). All sounds were cut to have duration of three seconds. We created a database by mixing the nine original sounds. First we formed the dataset "mixture_A"; this was obtained by mixing all the combinations of pairs of sounds with a 0dB Signal to Noise Ratio (SNR). The formula to calculate a SNR between a signal and noise is shown in equation 3, where P_{signal} means the power of the signal and P_{noise} the power of the noise.

$$SNR = 10 \log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{3}$$

After the previous procedure, we made four mixtures between the combination of two of the nine original sounds and all the elements from the "mixture_A". We avoided repetitions of sounds in a single mixture. These mixtures were obtained with four different SNR values (3.4dB, 5dB, 10dB and 20dB) where the nine original sounds were taken as the signal and the elements of mixture_A as the noise. With this procedure, four data sets of 252 elements were obtained for each SNR value. In each data set every original sound contributes as a signal or as a noise in 84 mixtures.

Figure 3 show the features MFCC, MBES, CMBES, Binary MBSES, Binary CMBSES and Binary C*MBSES from one individual sound (baby crying). In (b) we show the features for a mixture of the three sounds with 20dB and in (c) for 3.4dB of SNR. Here the baby crying is the signal while the bird signing and instrumental music are the noise. Notice the resemblance of the mixtures with the original sound, particularly with the C*MBSES.

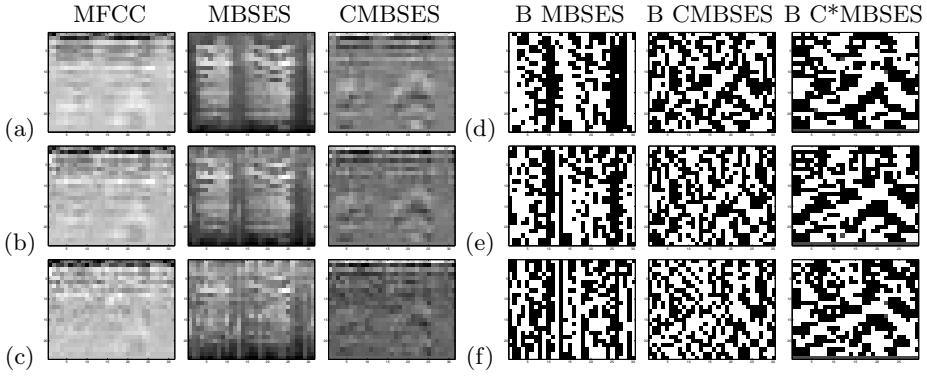

Fig. 3. Features MFCC, MBSES and CMBSES, and the binary versions B MBSES, B CMBSES and B C*MBSES of a *Crying baby* (a), and mixtures of sounds which include *Crying baby* in a 20dB (b) and 3.4dB (c). Please notice that in the binary signatures the shape of (a) preserves approximately in (c) and (d), this may not be apparent with the other features.

5 Experiments

We conducted an experiment comparing the five different features. This experiment consisted in identifying the occurrences of a given sound in the database, for this reason, the nine original sounds were compared with each SNR dataset by using a corresponding distance. The features used were: MFCC, MBSES, binary MBSES, binary CMBESES and binary C*MBESES. The corresponding distance to compare two sounds are the Euclidean distance for MFCC, a shifted euclidean distance for the MBSES, hamming distance for the next two and the distance mentioned in the table 2 for the binary C*MBSES.

To establish a ground truth we asked 48 subjects between 21-30 years old with headphones to hear the 21 mixtures of sounds from a web page that we created for the experiment (can be checked in `http://sound.natix.org`). Each subject heard different sounds that were randomly assigned from the complete database, the mixtures of sound could be of any of the four SNR dataset. Also, it was not assigned the same mixture of sound with a different SNR value to the same subject. The volunteers heard first the nine original sounds. After that we asked them to hear the mixtures of sounds and indicate with check buttons which of the nine original sounds they thought were part of the mixture. Subjects were

not informed of the number of sounds that were included in the mixtures and they were able to hear the mixtures as much as they wanted.

Figure 1 shows the recall obtained with every feature. We searched every sound in all the mixtures to obtain the results. As can be seen in figure 1 the results are better with low values of SNR. This happens because a 3.4dB mixture of sounds implies all the formant signals can be heard at once, like in a cocktail party. For the 20dB case, the opposite happens, because the sounds contributing as noise in the mixture are barely heard. It is also important to mention that humans heard several times every mixture. On average they heard 2.52 times the 3.4dB mixtures, 2.47 the 5dB, 2.61 the 20dB and 3.53 times the 20dB mixtures.

Table 1. The results of our experiment. Rows are different sounds as Baby Crying (i), Bird singing (ii), Keys (iii), Siren sound (iv), Tooth brushing (v), Music with voice (vi), Instrumental music (vii), Male voice (viii), Female voice (ix). We tested the following features (all with the same classifier) MFCC (a), MBSES (b), Binary MBSES (c), Binary CMBSES (d) and Binary C*MBSES (e). The (f) column is the average for the 48 volunteers. The four tables correspond to the different signal to noise ratio in the mixture.

	3.5 Db Mix						5 Db Mix					
	(a)	(b)	(c)	(d)	(e)	(f)	(a)	(b)	(c)	(d)	(e)	(f)
(i)	32.14	8.33	96.43	**100**	**100**	96.42	32.14	13.10	94.05	**100**	**100**	98.08
(ii)	51.19	64.29	96.43	**100**	**100**	96.42	50	59.52	94.05	**100**	**100**	97.61
(iii)	53.57	96.43	**97.62**	85.71	92.86	95.23	52.38	95.24	96.43	86.90	88.10	**100**
(iv)	2.38	78.57	76.19	82.14	86.90	**90.47**	3.57	75	73.81	78.57	80.95	**91.66**
(v)	**100**	**100**	**100**	**100**	**100**	97.61	**100**	**100**	**100**	**100**	**100**	98.80
(vi)	35.71	88.10	86.90	**100**	**100**	92.85	35.71	88.10	85.71	**100**	**100**	90.47
(vii)	**100**	98.81	98.81	**100**	**100**	94.04	**100**	98.81	98.81	**100**	**100**	96.42
(viii)	40.48	23.81	**100**	**100**	97.62	97.61	46.43	27.38	98.81	**100**	96.43	95.23
(ix)	95.24	58.33	**100**	**100**	98.81	97.61	92.86	55.95	**100**	**100**	**100**	96.42
Average	56.75	68.52	94.71	96.43	**97.35**	95.37	57.01	68.12	93.52	**96.16**	**96.16**	**96.16**

	10 Db Mix						20 Db Mix					
	(a)	(b)	(c)	(d)	(e)	(f)	(a)	(b)	(c)	(d)	(e)	(f)
(i)	35.71	29.76	89.29	**100**	**100**	95.23	39.29	33.33	71.43	92.86	86.90	**94.04**
(ii)	41.67	47.62	91.67	**100**	**100**	98.80	35.71	35.71	85.71	**96.43**	91.67	89.28
(iii)	46.43	85.71	89.29	86.90	85.71	**95.23**	41.67	76.19	83.33	83.33	65.48	**89.28**
(iv)	16.67	64.29	55.95	60.71	64.29	**73.80**	33.33	**66.67**	38.10	47.62	52.38	46.42
(v)	98.81	**100**	**100**	**100**	100	96.42	72.61	**100**	96.43	90.48	97.62	86.90
(vi)	32.14	83.33	84.52	**100**	97.62	94.04	35.71	54.76	72.62	78.57	**90.48**	72.61
(vii)	**100**	98.81	91.67	**100**	95.24	90.47	**100**	85.71	88.10	94.05	85.71	83.33
(viii)	50	34.52	90.48	92.86	94.05	**96.42**	48.81	33.33	80.95	**84.52**	77.38	77.38
(ix)	88.10	50	**100**	97.62	94.05	95.23	72.62	36.90	**86.90**	85.71	77.38	82.14
Average	56.61	66.01	88.10	**93.12**	92.33	92.85	53.30	58.07	78.17	83.73	**80.56**	80.15

6 Conclusion and Future Work

In this paper we classified mixed ambient sounds using a new feature, which consist in a binary signature formed with the entropy of the signal in 24 Bark sub-bands. The representation of the binary signature is very compact and allows signature comparison to be performed efficiently. We compared with the Mel Frequency Cepstral Coefficients (MFCC), which is currently the most used feature for audio-based activity recognition and surpassed the recognition capabilities of the MFCC. Furthermore, we also tested our signature against humans in an assorted collection of volunteers listening to the same collection of sounds and performing the same identification task. Our approach matches the performance of the volunteers (surpassing it by a small fraction), even if they had the opportunity to hear the mixtures several times. The results give us the hint of going in the right direction with our approach. A database of mixed sounds from everyday events was created using four different SNR values. To evaluate the features, every original sound was compared with the database aiming at finding those mixtures that contain it. In real life the sounds produced by different activities are overlapped hence it is important for automatic recognition to be able to identify all the sounds present in a segment of audio. This approach shows a feature which is capable of identifying sounds when they are not dominant in a signal which is an important step for recognize activities in real scenarios.

We are currently working in the problem of matching sounds of different sizes, or classes of sounds. To this end we are exploring approximate string matching techniques and partially observable Markov decision process.

Acknowledgements. This work was partially supported by a grant from the Alzheimer's Association (ETAC-10-173237) and by CONACYT trough a scholarship provided to the first author.

References

1. Rialle, V., Ollivet, C., Guigui, C., Hervé, C.: What do family caregivers of alzheimer's disease patients desire in smart home technologies? CoRR abs/0904.0437 (2009)
2. Morris, M., Lundell, J., Dishman, E., Needham, B.: New Perspectives on Ubiquitous Computing from Ethnographic Study of Elders with Cognitive Decline. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 227–242. Springer, Heidelberg (2003)
3. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. Comm. Mag. 48, 140–150 (2010)
4. Potamitis, I., Ganchev, T.: Generalized recognition of sound events: Approaches and applications, pp. 41–79 (2008)
5. Wichern, G., Xue, J., Thornburg, H., Mechtley, B., Spanias, A.: Segmentation, indexing, and retrieval for environmental and natural sounds. Trans. Audio, Speech and Lang. Proc. 18, 688–707 (2010)

6. Handte, M., Iqbal, U., Apolinarski, W., Marrón, P.J.: Challenges in ubiquitous context recognition with personal mobile devices. In: Proceedings of the 4th ACM International Workshop on Context-Awareness for Self-Managing Systems, CASEMANS 2010, pp. 6:40–6:45. ACM, New York (2010)
7. Niessen, M.E., van Maanen, L., Andringa, T.C.: Disambiguating sounds through context. In: Proceedings of the 2008 IEEE International Conference on Semantic Computing, pp. 88–95. IEEE Computer Society, Washington, DC (2008)
8. Bronkhorst, A.W.: The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acustica United with Acustica, 117–128 (January 2000)
9. Mitrovic, D., Zeppelzauer, M., Breiteneder, C.: Features for content-based audio retrieval. Advances in Computers 78, 71–150 (2010)
10. Ward, J.A., Lukowicz, P., Troster, G., Starner, T.E.: Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1553–1567 (2006)
11. Min, C.H., Ince, N.F., Tewfik, A.H.: Number Eusipco. In: Early Morning Activity Detection Using Acoustics and Wearable Wireless Sensors (2008)
12. Kern, N., Schiele, B., Schmidt, A.: Recognizing context for annotating a live life recording. Personal Ubiquitous Comput. 11, 251–263 (2007)
13. Ma, L., Milner, B., Smith, D.: Acoustic environment classification. ACM Trans. Speech Lang. Process. 3, 1–22 (2006)
14. Chu, S., Narayanan, S., Kuo, C.C.J.: Environmental sound recognition with time-frequency audio features. Trans. Audio, Speech and Lang. Proc. 17, 1142–1158 (2009)
15. Lu, H., Pan, W., Lane, N.D., Choudhury, T., Campbell, A.T.: Soundsense: scalable sound sensing for people-centric applications on mobile phones. In: Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services, MobiSys 2009, pp. 165–178. ACM, New York (2009)
16. Heittola, T., Mesaros, A., Virtanen, T., Eronen, A.: Sound event detection in multisource environments using source separation. In: Workshop on Machine Listening in Multisource Environments, pp. 36–40 (2011), `http://spandh.dcs.shef.ac.uk/projects/chime/workshop/`
17. Camarena-Ibarrola, A., Chávez, E., Tellez, E.S.: Robust Radio Broadcast Monitoring Using a Multi-Band Spectral Entropy Signature. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 587–594. Springer, Heidelberg (2009)
18. Sigurdsson, S., Petersen, K.B.,, T.L.S.: Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In: ISMIR, pp. 286–289 (2006)