# Water flow detection from a wearable device with a new feature, the spectral cover

Patrice Guyot, Julien Pinquier, Régine André-Obrecht
SAMoVA team - IRIT - University of Toulouse - France
{guyot,pinquier,obrecht}@irit.fr

## Abstract

*This paper presents a new system for water flow detection on real life recordings, and its application to medical context. The recognition system is based on an original feature for sound event detection in real life. This feature, called "spectral cover" shows an interesting behaviour to recognize water flow in a noisy environment. An experiment is realized with more than 7 hours of videos recorded by a wearable device. The system based on thresholds obtains good results for the water flow event recognition (F-measure of 66%). A comparison with classical approaches using MFCC or low levels descriptors and GMM classifiers is done to attest the good performance of our system. Adding the spectral cover to low levels descriptors also improve their performance and confirms the interest to this feature.*

## 1. Introduction

Despite significant progress in audio information retrieval this last decades, identifying environmental sound events in everyday life is still a challenging task. Indeed, the amount of information holding in an acoustic scene can be very important. For example, sound of cooking utensils may lead to infer someone's cooking activity; vehicle passing sound may lead to infer a nearby road. That's why audio can be sometimes more accurate than video to recognize some events: as a gunshot in a crowd or the use of tap water. Thus, applications of Audio Event Detection (AED) are numerous, for instance humanoid robots audition [21], sport (ball hit detection [27]), surveillance systems [4], more recently life-logging [1] and activity detection for older adults [24].

Even if the feasibility of the AED in real life is increasing, most of audio segmentation and classification methods are used on clean corpus like extracts of radio recordings [7], films, or studio produced music tunes [5]. In this case, the use of signature [20] can be an efficient solution to detect audio events. Nevertheless, a lot of acoustic features and retrieval methods which can be efficient in this particular context become obsolete in the everyday noisy life, due to the amount of learning data which would have been necessary to describe all acoustic cases [16].

A typical difficulty of identifying predetermined audio events in everyday life recordings comes from the possible overlapping of many different sound events. Indeed, this task can be seen as computational auditory scene analysis, where lots of different environmental sounds produce a mixture which must be analysed to obtain an auditory description of the environment for the listener [25].

The unknown presence of speech and various noises make this task very hard, even sometimes by human. For example, noises from indoor could comes from doors, electric devices or air-conditioning. Those from outdoor could be cars, planes, birds or rain (see [22] for a study in detail).

Another problem comes from the acoustic variability of sounds which could be labelled as the same audio event. Indeed, lot of objects which produce the same sound at a semantic level can produce very different acoustic sounds depending on their type or their material. For instance, if we focused on home environmental sounds, a door bell could differ a lot among different houses. The presence of different type of reverberation in houses may increase this difficulty.

Last but not least, the variability of the recording conditions like the choice of the microphone and it position in the space can deeply modify the acoustic mixture [23]. Hence we can say that acoustic event detection in real life is a non-deterministic problem.

We present in this paper a robust audio water flow recognition system, from a wearable device recording, for activities indexing. This paper is organized as follows. Section 2 presents the AED systems and gives a review of previous works in water flow detection. Section 3 explains the IMMED project and the aim to build a water flow recognition system. We propose our water flow recognition system in section 4. Section 5 describes experiments and compari-

son with a classical recognition system.

## 2 Previous works

The AED task purpose is to detect and recognize a closed set of pre-defined acoustic events in audio data. The most classical systems compute acoustic features to train some classifiers. The features can be extracted from the temporal domain (energy, zero-crossing rate, etc.) or from the spectral domain (spectral centroïd, spectral flux, spectral roll-off, etc). Some Low-Level Descriptors (LLD) have been standardized in the MPEG-7 norm. A large set of features have been described in [18].

If the Mel-Frequency Cepstral Coefficients (MFCC) are a very often used set of features in all types of audio recognition tasks and particularly speech [8], their efficiency for noisy sound recognition is discussed in several works in comparison to LLD [19, 12]. Even if the potential space of extracted features is theoretically huge [17], most of the research works on acoustic events are not focused on specific features.

If different works have been done in environmental sound event detection [1, 16], classification of environmental sound events for instance in a kitchen [13], or context recognition [3], only few studies deal specifically with water flow detection. These works are motivated by daily living activities recognition, underlying the hypothesis that several home human activities can be retrieved with water use analysis.

Some studies use sensors on the water pipes to provide an activity detection system. In [6], four microphones are installed in the basement, on hot water, cold water and drain pipes. This low cost system allows the recognition of eight daily living activities, with using the zero crossing rate (ZCR) and root mean square of the recording samples. In [9], Ibarz use a sensor attached to a water tap. The system detects if water is flowing in the pipe, but also quantifying the flow that allows to compute the amount of water used. The decision system is trained with Fast Fourrier Transform (FFT) coefficients and MFCC. The classification is perform by a k-Nearest Neighbor (kNN) system.

Two other studies use recordings from microphones to detect the Activities of Daily Living (ADL). Chen proposes in [2] to detect and recognize different bathroom activities with a single microphone close to the washing basin. The aim of this project is to monitor the bathroom ADL and increase understanding of personal hygiene behavioural problems of dementia patients. The system uses a classical approach: MFCC and Hidden Markov Models (HMM).

In [24], the purpose is water flow detection in a hand-washing task with a fixed camera placed above the sink. Taati make use of audio and video features to detect the water flow. This system aims to be used within a system that provides reminding prompts for Alzheimer's sufferers. Taati compute a signal to noise ratio, the ZCR, the spectral centroïd, the spectral roll-off at 85%, the spectral flux and MFCC. Different classifiers have been tested with quite closed results.

## 3. Activities of daily living with a wearable device

We are working on the IMMED project [15], which proposes to help the doctors in their diagnosis with a video monitoring system. Indeed, with a longer lifetime expectancy, there is nowadays a real challenge in helping the elderly population to keep their autonomy as long as possible. The IMMED project follows the "aging in place" approach wich reduce the healthcare costs and allows the elders to be independent and socially connected. This approach is limited by the autonomy of the patients and their ability to relize the activities of daily living. The project uses a video acquisition wearable device that the voluntary patient wears at home while he executes ADL. With providing videos of the people in age of dementia in their activities of daily living, the goal of the project is to assist the doctors to measure autonomy decline as part of the diagnosis process. The videos are added to the traditional survey based on interview of the patient and the relatives which take part in the doctor diagnosis.

As the doctors will not watch hours of videos, an automatic system of activities indexing is developed to ameliorate the navigation inside the videos. One of the project's purpose is to create an automatic indexing system to segment the video in ADL. By this way, the doctors could directly watch specific activities to evaluate the patient autonomy. Thus, an original interface permits to access the videos.

The final decision system uses video and audio cues like water flow events in a hierarchical HMM to take a decision about activities of the patients [11]. Focusing on the audio part, we noticed that water flow recognition can help to index several human activities like *cooking*, or *do the dishes*. Moreover water sounds are more universal that sounds coming from other activities like *make coffee* which depend strongly of the houses and the object implies in.

In comparison to the other water flow detection studies (describes in section 2), the aim of the IMMED project is to observe the patient in his personal home with the less intrusive approach as possible. In this condition the use of sensors, placed on the basement of a house or on water taps, is inconceivable. The other studies which deal with water flow measurement use microphone in a fixed place. Consequently the machine learning methods can be efficient thanks to a certain homogeneity of the training data. By contrast, our wearable device implies an important variabil-

ity of sounds due to different recording places and various activities executed by the device holder. First, each recording corresponds to a different home. Secondly, the device holder could be indoor, outdoor, and move from one room to another. Furthermore, the device holder can execute different activities, which imply so many different sounds. Hence these heterogeneous data makes very difficult the creation of a robust system. As the training data should be insufficient, we decide to focus on robust features.

## 4 Water flow detection

### 4.1 Classical features

A water flow sound usually presents a very noisy spectrum, which can be compared to a high colored noise. The main difficulty of our corpus is that water noises are overlapped with other noises and with speech. We can see on figure 1, an extract of 5 minutes composed by a long water flow event: a speech part overlaps the end of the water flow event and its amplitude is very high in comparison to the other parts. Lot of noises are present during all of the excerpt, mainly noises from the kitchen activities.

We have tested different low level descriptors on this excerpt: temporal features (energy, energy per band and ZCR) and spectral features (centroïd, spread, skewness, kurtosis, roll-off, flux, variation coefficient, and flatness).

According to the literature, variation coefficient is a good noise indicator [26], as well as spectral flatness [10].

All these features are computed on a 80 ms hamming window, and a smoothing is applied with a 4 s median filter. Figure 1 shows the most representative results: the zero crossing rate, the spectral centroïd and the spectral flatness: the curves of the ZCR and the spectral centroïd are impacted by the presence of the low frequencies of speech. The spectral flatness has a more interesting behaviour, but it is quite sensible to lots of other noises and silence.

### 4.2 Spectral cover

We introduce a new feature that we call *spectral cover* because it reacts to large spectral band sounds like high colored noises, even in presence of speech.

The spectral cover is given by:

$$SP = \frac{\sum_i (ampl(w_i) * w_i)^2}{[\sum_i ampl(w_i)]^\gamma}, 1 \leq \gamma \leq 2 \qquad (1)$$

where $w_i$ represent the frequencies and $ampl(w_i)$ their amplitudes, from the Fourrier transform .

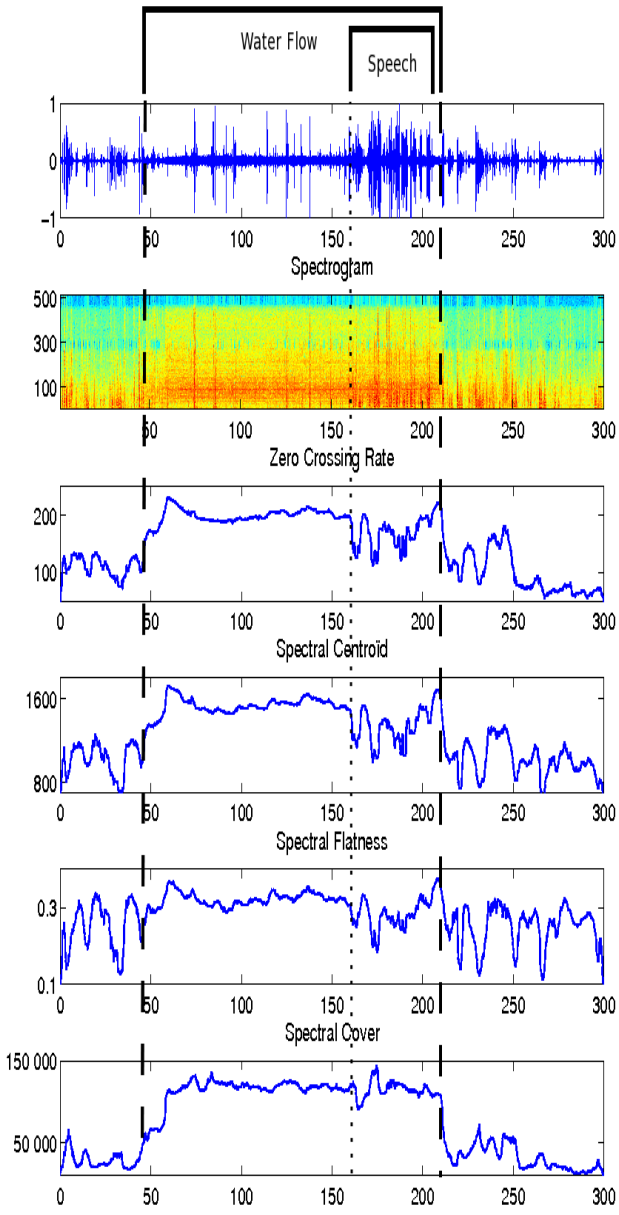


**Figure 1. Acoustic features comparison on a water flow event**

If we compare the spectral cover with the well-known spectral centroïd given by:

$$\mu = \frac{\sum_i ampl(w_i) * w_i}{\sum_i ampl(w_i)} \qquad (2)$$

An important difference is the introduction of a power in the numerator: it may boost the high frequencies, in com-

parison of the spectral centroïd or the spectral spread. An other difference, the power $\gamma$, makes our feature sensitive to the absolute signal level, unlike spectral centroïd. The $\gamma$ parameter allow to tune the sensibility of the feature to the signal level.

We can see on figure 1 that the spectral cover stays almost constant during the water flow event even in the presence of voice. It becomes lower outside the water flow sound event, that's why on this extract we can separate easily the water flow event with a simple threshold.

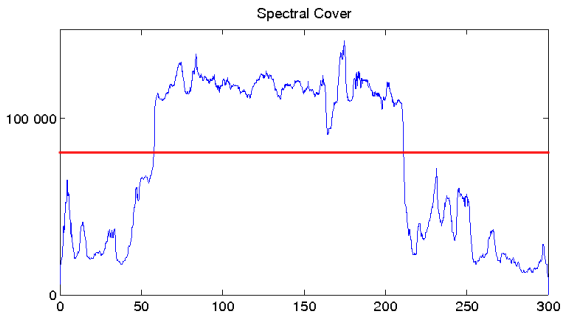


**Figure 2. Threshold on the spectral cover**

### 4.3 The water flow recognition system

Some experiments show the spectral cover is sensible to other audio events like the noise of a vacuum cleaner), so the system we propose operates in two steps. After computing the spectral cover on the signal, a first segmentation is performed to extract a family of sound events, the candidate sounds. We identify water flow events among this family in a second step. The system ends with a classical post-processing step. More precisely:

- We first compute a short-time Fourrier transform with a hamming window of 1024 samples (80 miliseconds) and an overlap of 50%. Then the spectral cover is computed on each frame. The minimum of the spectral cover values is extracted on a 2s window; this window is considered as a candidate sound if the value exceeds a threshold $T_1$ . With concatenating the juxtaposed values, we obtain a first selection, each segment is considered as a candidate sound or not.

- In the second step, we consider each candidate segment. If 85% of the frames included in the candidate segment, have a spectral cover value superior to a second threshold $T_2$, the segment is rejected as water flow; if not, it is accepted. The rejected segment correspond to sounds event like vacuum cleaner sounds which implies high levels of spectral cover.

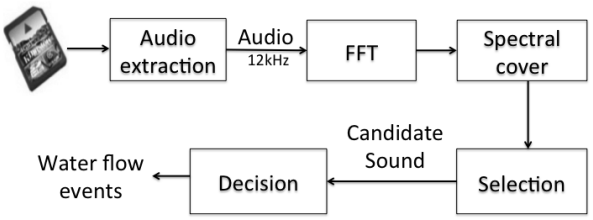


**Figure 3. System diagram**

- The post-processing step consists in removing all sound events smaller than 3 seconds.

## 5 Experiments

### 5.1 Corpus description

Our corpus is provided by the IMMED project: 20 videos of patients making activities of daily living in their own personal residence, recording with a camera (HD GoPro Fisheye) fixed on their shoulder. The videos are made in the presence of a psychologist who specifies a scenario of activities to realize. The records are from twenty minutes to two hours of duration. After extracting the audio part from the video, the audio sample rate is 12kHz . The listening reveals lots of environment noises, like opening and closing of doors, footsteps, kitchen utensils, and rubbing on the recording device. Some particular sounds like water flow or ring phone give obvious clues on the patient activity. Nevertheless, as the patient and the psychologist are in constant discussion, and due to the position of the microphone, speech is very present and can be loud in comparison to others sounds.

We annotated manually the videos in water flow events. We consider as water flow event each sound part containing at least one of the two sound events: the "fissing" sound created by the faucet, and the "splash" sound created by water impacts and bubbles. There is about 7 hours 30 minutes of videos with 85 water flow events for a total duration of 25 minutes.

### 5.2 Results

We fixed ours threshold values with one of the 20 videos. During its 39 minutes duration , 13 water flow events are present for a total duration of 3 minutes. The values are : $T_1 = 30000$, $T_2 = 80000$, and $\gamma = 1.5$

The results are evaluated by 4 classical metrics used in the french campaign ESTER2 [7]: Error-rate, Precision, Recall, and F-measure. The results of our system are presented in the table 1. Our system obtain 66% of F-measure. Among the missed events, some of them are not clearly audible or too short. The other missed events are only composed by little splash of water without a clearly audible fissing water flow sound. This kind of sound is never recognized by the system. To precise the nature of the false alarms, we analyse the system output on half of our corpus (4 hours of video): 22 false alarms are revealed and they are mainly due to the presence of speech and noise. Among them, we note speech overlaps (7 errors, 6 in the same file), manipulation noises like plastic bag or cookery (6 errors), television (3 errors), laughers (4 errors), manipulation of the camera (2 errors).

### 5.3 Comparison with MFFC & LLD in a GMM

We propose a comparative experiment between our system and a classical approach. This approach is based on a GMM classifier, and three different sets of features have been tested: MFCC, Low Level Descriptors (LLD), and LLD with our spectral cover (LLD+SC). MFCC and LLD are computed with the Yaafe audio extractor [14]. MFCC are composed of 24 coefficients from bands representing spectrum between 20Hz and 6kHz. To obtain performing LLD, a lot of feature combinations have been tested. Among them, the best results are obtained with energy, spectral flatness per band (per log-spaced band of 1/4 octave) and spectral shape statistics (including centroïd, skewness, and kurtosis). Several numbers of gaussians have been tested, and finally we have kept 4 Gaussians. As in our system, we have removed sound events smaller than three seconds. The standard leave-one-out testing protocol was used: one video is discarded from the database to train the GMM classifier with the 19 remaining videos. The excluded video is used as the test data.

Results are presented in the table 1. The results obtain with the GMM approach are not so good. We noticed that precision and recall are very unbalanced. That's could be the results of a specific learning data: the number of frames in the two classes non-water and water are quite unbalanced. Moreover, the water flow data are very heterogeneous, two much for a GMM of 4 components. As one could have predict, LLD gives better results than MFCC, and confirm some previous studies [19]. Adding spectral cover to LLD improves the performance; our spectral cover appears complementary to the chosen LLD. In comparison to these classical approaches, our system presents better performance. Moreover the computational cost of our system is insignificant compared with those of a GMM approach.

**Table 1. Comparative results**

| | System | GMM | | |
|---|---|---|---|---|
| | | MFCC | LLD | LLD+SC |
| Error rate | 0.04 | 0.09 | 0.08 | 0.07 |
| Precision | 0.54 | 0.35 | 0.39 | 0.44 |
| Recall | 0.83 | 0.87 | 0.91 | 0.88 |
| **F-measure** | **0.66** | **0.45** | **0.50** | **0.53** |

## 6 Conclusion

In this paper, we presented the problem of acoustic water flow detection in a specific and adverse environment: the audio signal is recorded by a wearable device inside personal home. Due to heterogeneous data and noisy environment, it was necessary to specify a new feature, called "Spectral Cover". A system based on robust thresholds was build to detect these water flow events. This system was assessed on IMMED corpus: more than 7 hours of real life records. The results (66% of F-measure and an error rate of 5%) are very satisfying regards to task difficulty.

Moreover, our system was compared to classical state-of-the-art audio recognition systems (GMM with a leave-one-out protocol). In conclusion, adding the spectral cover to classical low level descriptors in a GMM allows to increase the performance, but this system isn't better than our initial proposition. This confirm the interest and pertinence of choosing this feature to detect water flow sounds.

Furthermore, we have noticed that the spectral cover is also suitable to detect other sound events. Indeed, among the candidate segments, we have detected cell phone rings and vacuum cleaner use and this allows us to imagine future applications.

## References


[1] M. Al Masum Shaikh, M. Molla, and K. Hirose. Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense. In *Computer and Information Technology. ICCIT.* IEEE, 2008.

[2] J. Chen, A. Kam, J. Zhang, N. Liu, and L. Shue. Bathroom activity monitoring based on sound. *Pervasive Computing, PERVASIVE*, 2005.

[3] S. Chu, S. Narayanan, and C. Kuo. Environmental sound recognition with time-frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[4] C. Clavel, T. Ehrette, and G. Richard. Events detection for an audio-based surveillance system. In *International Conference on Multimedia and Expo, ICME.)*. IEEE, 2005.

[5] J. Downie. Music information retrieval. *Annual review of information science and technology*, 2003.

[6] J. Fogarty, C. Au, and S. Hudson. Sensing from the basement: a feasibility study of unobtrusive and low-cost home

activity recognition. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, 2006.

[7] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, and G. Gravier. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] J. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 2002.

[9] A. Ibarz, G. Bauer, R. Casas, A. Marco, and P. Lukowicz. Design and evaluation of a sound based water flow measurement system. *Smart Sensing and Context*, 2008.

[10] J. Johnston. Transform coding of audio signals using perceptual noise criteria. *Selected Areas in Communications*, 1988.

[11] S. Karaman, J. Benois-Pineau, R. Mégret, J. Pinquier, Y. Gaëstel, and J. Dartigues. Activities of daily living indexing by hierarchical hmm for dementia diagnostics. In *Content-Based Multimedia Indexing, CBMI*. IEEE, 2011.

[12] H. Kim, J. Burred, and T. Sikora. How efficient is mpeg-7 for general sound recognition. In *Proceedings of AES 25th International Conference*, 2005.

[13] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel. Temporal ica for classification of acoustic events in a kitchen environment. In *Proceedings of the INTERSPEECH.*, 2005.

[14] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR.*, 2010.

[15] R. Mégret, V. Dovgalecs, H. Wannous, S. Karaman, J. Benois-Pineau, E. El Khoury, J. Pinquier, P. Joly, R. André-Obrecht, and Y. Gaëstel. The immed project: wearable video monitoring of people with age dementia. In *Proceedings of the international conference on Multimedia*. ACM, 2010.

[16] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen. Acoustic event detection in real-life recordings. In *18th European Signal Processing Conference*, 2010.

[17] F. Pachet and P. Roy. Exploring billions of audio features. In *Content-Based Multimedia Indexing, CBMI*. IEEE, 2007.

[18] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 2004.

[19] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa. Computational auditory scene recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. IEEE, 1993.

[20] J. Pinquier and R. Andre-Obrecht. Jingle detection and identification in audio documents. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. IEEE, 2004.

[21] Y. Sasaki, M. Kaneyoshi, S. Kagami, H. Mizoguchi, and T. Enomoto. Daily sound recognition using pitch-cluster-maps for mobile robot audition. In *International Conference on Intelligent Robots and Systems, 2009. IROS*. IEEE.

[22] R. Schafer. *The soundscape: Our sonic environment and the tuning of the world*. Destiny Books, Rochester, 1994.

[23] J. Smolders, T. Claes, G. Sablon, and D. Van Compernolle. On the importance of the microphone position for speech recognition in the car. In *Acoustics, Speech, and Signal Processing, ICASSP*. IEEE, 1994.

[24] B. Taati, J. Snoek, D. Giesbrecht, and A. Mihailidis. Water flow detection in a handwashing task. In *Canadian Conference on Computer and Robot Vision (CRV), 2010*. IEEE, 2010.

[25] D. Wang and G. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press, 2006.

[26] D. Wilson and J. Wayman. Signal detector employing mean energy and variance of energy content comparison for noise detection, June 21 1994. US Patent 5,323,337.

[27] B. Zhang, W. Dou, and L. Chen. Ball hit detection in table tennis games based on audio analysis. In *Pattern Recognition, 2006. ICPR*. IEEE, 2006.