MADRID
inter.noise 2019
June 16 - 19
NOISE CONTROL FOR A BETTER ENVIRONMENT

Drip detection through its acoustic signature with variable background noise

Sobreira Seoane, Manuel A. [1]
AtlantTTic Research Center fort Telecommunication Technologies
E.E. Telecomunicación, Campus Universitario. E 36310-Vigo, Spain

ABSTRACT

In this paper, the problem of detection of an specific event– a drip–in the domestic acoustic scene is addressed. The detection of noises generated by water can be of great interest in domestic scenes because it can help to prevent domestic floods. In the case of drip sounds, it is quite difficult to get real sounds covering the different sound qualities that different drops may have. Their sound depends on many factors as their size, the kind of surface they hit, etc. In order to approach real life as much as possible, a database including real and synthesized sounds have been created. A training set has been set up using the speed of variation of the MFCC, the kurtosis and the probability density function of the high frequencies as features to train a SVM classifier.

Keywords: Automatic Detection, SVM, Acoustic Features
I-INCE Classification of Subject Number: 74

1. INTRODUCTION

In the last years the automatic detection of acoustic events is showing and increasing interest and activity. New applications related to the smart acoustic sensing of industrial, urban or domestic environments are arising. As examples, the well known MFCC (Mel Frequency Cepstral Coefficients) and a SVM (Support Vector Machine) are used to detect multiple sounds in domestic environments with the purpose of remote monitoring of elderly persons (presented by Alsina et al [1]). Sharma et al [2] [3], analyze different acoustic features, as pitch variation or the spectral envelop (formants), to detect the causes of a baby crying. Different features based on the processing of the normalized spectrogram, as the spectrogram entropy, are used by Jian [4].
The classification of multiple acoustic events in real life is a challenging and difficult task. We refer as *acoustic event* a specific sound produced by a sound source (a baby crying, a dripping tap, a glass break, a door opens). These acoustic events have a short duration and they are embedded in a specific *acoustic scenario* which contains a mixture

[1] msobre@gts.uvigo.es

of all the typical sounds coming from different sound sources. A domestic *sound* or *acoustic scene* might contain sounds as music, speech, a dishwasher humming, a vacuum cleaner noise, etc. In [5], the strategies and methods to follow in order to design a system which is able to detect and classify events in complex scenarios are described. In this paper, the problem of detection of a specific event – a drip– in the domestic acoustic scene is addressed. In order to design a simple system, the need of possible source separation, in order to extract the event of interest from the background noise is avoided, focusing on the extraction of features able to detect the specific event on background noise. The detection of noises generated by water can be of great interest in domestic scenes because it can help to prevent domestic floods. The detection of a drip can help to detect a problem at its early stage.
In the section 2, the method to generate the training database is addressed. In the case of drip sounds, it is quite difficult to get real sounds covering the different sound qualities that different drops may have. The timbre of a drip sound depends on the size of the drops, their speed when they hit a surface, the king of surface they hit or if they fall into water. Trying to approach the real situation as much as possible, recorded and synthesized drop sounds have been mixed with different background noises to obtain a database. The section 3 specifies the acoustic features used to train a linear SVM classifier. In the section 4, the results of different performed on the system are presented and section 5 presents the conclusions.

2. GENERATION OF THE TRAINING DATABASE

In order to set up an acoustic events classifier, a balanced database with a representative set of signals of the events of interests and “no events“ (or background noises) is needed. In the case of a drip detection, it is not easy to get a good set of drip sounds in different conditions. For the purpose of this investigation, three sources of sounds have been handled:

1. Recorded sounds of dripping water on different surfaces (as a hard floor, an aluminum sink, or a bath tube) and different background noises of the domestic sound scenario. In order to introduce variability in the database, the audio signals have been recorded using different devices:
   - A mobile phone with the built in microphone.

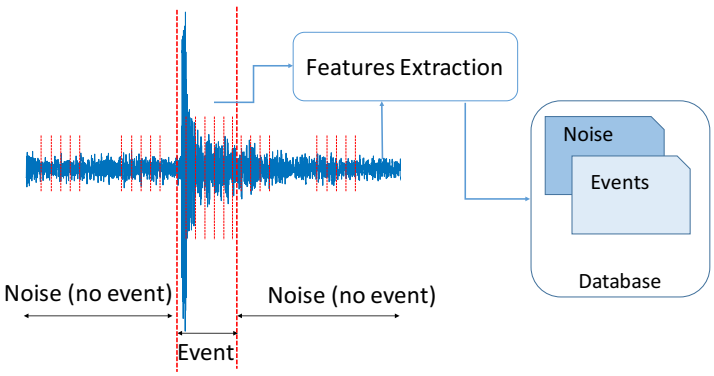
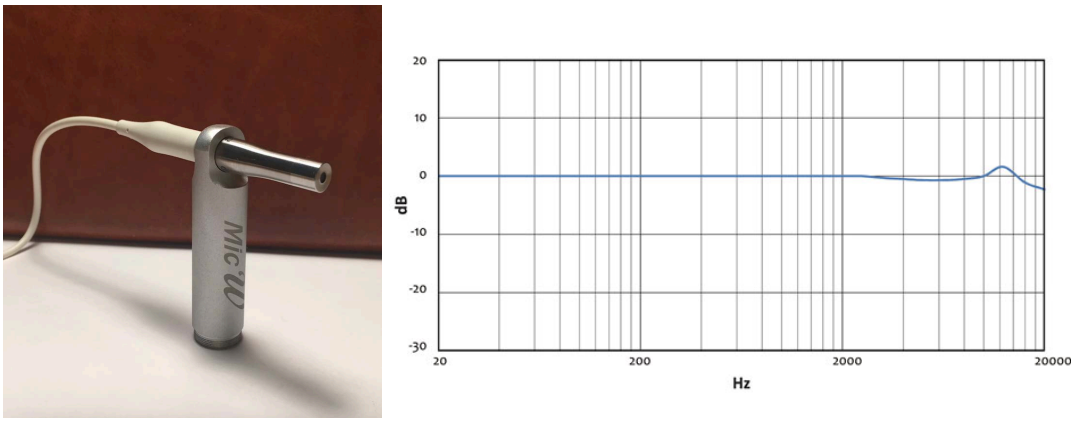


Figure 1: "Frame by frame" feature extraction approach



(a) i436 Microphone (b) Frequency response

Figure 2: i 436 Omnidirectional microphone. Sensitivity, S=44dB (6.3mV/Pa) and S/N>63 dB (Source: manufacturer specifications, http://www.mic-w.com/)

   - A mobile phone with an external microphone: Mic W i 436, a class 2 ominidirectional, electrect condenser microphone (figure 2).
   - An Edirol stereo R-09 voice recorder, with a built in stereo condenser microphone.

2. Sounds from the household section of the BBC database [6], as water droplets, vacuum cleaners, kettles, etc.

3. Synthesized dripping sounds: sound synthesis models have been generated to obtain sounds similar to a water drops falling over different kind of surfaces . The synthesized sounds have been mixed with different background noises.

It has been chosen to work on a 'frame by frame" approach, i.e., the audio dataset is split into small frames and the feature extraction is performed on each frame, see the figure 1. Each frame is labeled as "background noise" (no drip sound in the frame) or "event" (drip noise in the frame). As the aim of the research is dealing with the detection of short transient sounds, the duration of each frame is $t_f = 50$ ms. The sampling frequency of the audio dataset is $f_s = 48000$ Hz. As it can be seen in the figure 3, the duration of the signal of a drop is around 0.2 s. With the selected frame length, each 0.2 s event contains 4 frames. It should be notice that for low Signal to Noise ratios some of the frames can be masked due to the influence of the background noise.

2.1. Expansion of the database set using physical models

As previously mentioned, the purpose of this research is the detection of the sound produced when a drop hits a surface, either water or a hard surface. The perceived sound quality (timbre) of the radiated sound will depend on the characteristics of the drop (their size and speed) and the surface (size and material). In order to take into account the changes of the characteristics of the sound of the dripping water depending on the nature

of the surface and environmental conditions, its is needed to expand the recorded database using a model which can produce those changes that has not been possible to register in the recordings. Two models have been considered: the generation of the sound of a water drop that hits a rigid surface and a drip on water.

2.1.1 Drops falling on a surface:

In this case, it is assumed that the perceived/registered sound is mainly due to the radiation of the surface that has been hit by a drop. After being hit by a short duration of, it could be assumed that the sound is produced by the radiation of a set of acoustic decays. A modified model similar to the one used by Vörlander [7] has been chosen:

$$p(t) = \sum_{i=1}^{N} A_i \sin(2\pi f_i + \phi_i) \cdot w(t) \cdot e^{-\alpha t} \tag{1}$$

The Equation 1 describes a set of decaying finite cosines to simulate the short duration of the drop. The length of the signals are controlled by means of the time window $w(t)$, and chosen empirically, just by hearing that the effect is close to a real drop falling on a surface.f_i are the resonance frequencies of the surface. The size and speed of the drops will affect the amplitude of the sound radiated, while the size and material of the surface will determine de decaying resonance frequencies and the damping factor. To simulate that the database of drop sounds should have a wide number of possibilities, the resonances frequenciesf_i, the damping factor α, the phases ϕ_i and the amplitudes A_i are chosen randomly to simulate that a drop can hit any kind of surface.

2.1.2 Drops falling on water

As it is described by Guyot et al [8] and Moss and Hengching Yeh [9], the perceived sound of a drop is due to the radiation of the bubbles that are formed inside the water, just after the droplet hits the water surface. Moss [9] proposes several models that deal with the generation of the sound radiated by bubbles with different degrees of complexity. For the purpose of this job, the simple model for spherical bubbles have been chosen. The sound radiated by a spherical bubble can be calculated from the Equation 2 ,

$$p(t) = \epsilon r_o \sin(2\pi f(t) \cdot t) \cdot e^{-\beta_o t}, \quad \epsilon \in [0.01, 0.1], \tag{2}$$

where ϵ is a tunable parameter that allows to set the initial excitation of the bubbles, r_o is the radius of the bubble in meters, and:

– $f(t)$ is a time dependent frequency which depends on the main resonance of the bubble, f_o.

$$f(t) = f_o(1 + \xi\beta_o t), \tag{3}$$

with ξ a parameter with helps to adjust the effect of rising in pitch of the sound of a drop, taking $\xi \approx 0.1$ as its optimum value. The resonance frequency of the bubble is calculated as suggested by van den Doel [10]:

$$f_o = \frac{3}{r_o}, \tag{4}$$

– $\beta_o = \pi\ f_o \delta_{tot}$ is the attenuation factor of the decaying exponential and δ_{tot} the total damping of the bubble, that can be calculated as [10]:

$$\delta_{tot} = \frac{0.13}{r_o} + 0.0072\ r_o^{-3/2}. \tag{5}$$

From the last set of equations it becomes clear that the main parameter leading to the calculation of the radiated sound pressure of a drop falling on a water surface is the bubble size. The sound pressure, as shown in Equation 2, depends on the resonance frequency of the bubble and a damping factor and both terms, accordign to Equation 4 and Equation 5 can be estimated from the bubble radius r_o. In order to expand the database, a large set of drops has been generated taking the size of the drop as a random parameter. The radius of the drop is generated following a uniform distribution between 0.5 mm and 10 mm.

2.2. The generation of the training and test sets

The training data set generated tries to match real conditions as much as possible. The set of real drip sounds have been expanded with synthetic sounds randomly generated. A set of 16000 frames of drips sound have been created: 1000 frames comes from real sounds and 15000 have been generates. As the frame size is $t_f = 50ms$ this means that a set with 800 seconds of drip sound have been created. These sounds have been mixed with different background noises: domestic noises obtained from the BBC database [6] (vacuum cleaners, kettles, shavers, etc), speech and different kind of music covering different styles (pop, jazz, classic, etc). By mixing random drops with different background noises, a balanced training sets has been created trying to reproduce the variability of real life conditions. Each training set has 1000 drops (which means around 4000 frames belonging to the class “Drop”) and 1000 samples of background noises taken randomly from the background noises database. Therefore, the training signals last 66 seconds each.

3. ACOUSTIC FEATURES

One of the most critical aspects of audio detection and classification is the selection of a set of features that are relevant, they do not introduce redundancy and the distance the distance between classes is maximized [11] [12]. With the increasing interest on the detection of acoustic events, there is an extensive description of different acoustic features throughout the bibliography [13]. One of the more common approaches is the use of feature extraction methods that have been primary used for speech signals. The Mel-frequency cepstral coefficients (MFCC) and some other features based on the Mel scale (as the logarithmic Mel-Filter bank coefficients – FBANK) are of common use in the detection of acoustic events [14], although it is well known that they are quite sensitive to background noise [15]. The following features have been considered for the purpose of this job:

– High pass probability density function. The figure 3 shows the signal of a water droplet, its spectrogram and the spectrogram of the signal of a dripping tap on a sink . The persistence spectra of a drip sound on a continuous background noise and background noise are also shown. The persistence spectrum shows the probability density function for each frequency bin, i.e. the percentage of the time that a given

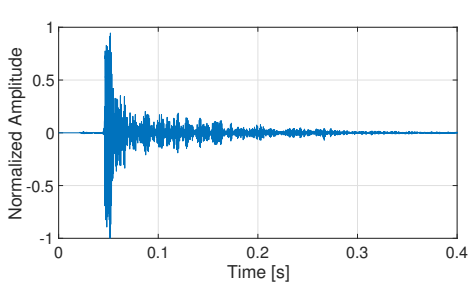
frequency is present in a signal. In the figure 4 (d) the PDF of a background noise signal frame to the PDF of a signal frame containing a drip sound on the same background noise are compared. It can be noticed how the probability density of the high frequencies increases. As most of background noises have higher energy at low frequencies, it is expected that this feature can be quite robust against most of the most common domestic background noises.

– Kurtosis: for a given set of samples of a signal (a frame), the kurtosis is calculated as the fourth standardized central moment:

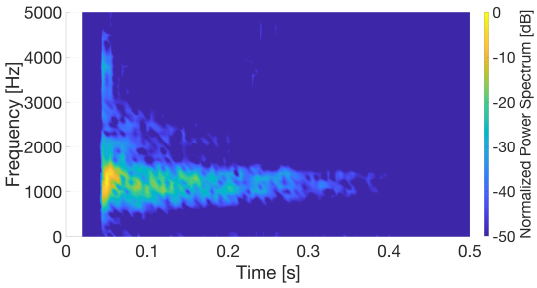
$$k = \frac{E(x-\mu)^4}{\sigma^4}, \tag{6}$$

where μ is the mean and σ the standard deviation of the set of samples. k=3 for normal distributed samples, while it takes hight values when transients (outliers) are present in the sample. The kurtosis is quite sensitive to impulses even with very low Signal to Noise ratio [16]. The figure 4 (a) shows the kurtosis for 30 seconds of signal of a dripping tap on water with a background noise. It can be noticed how the kurtosis becomes high in the presence of a the transient signal (drip sound).

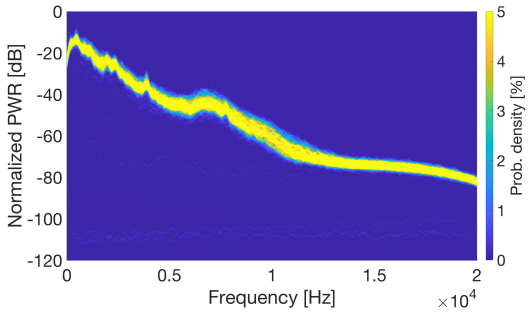
– Δ_{MFCC}: The speed of variation of the MFCC coefficients inside a frame have been also considered as an useful feature to detect the drip sound. The figure 4 (c) shows the Δ_{MFCC} for the example mentioned. It can be noticed how the peaks of the



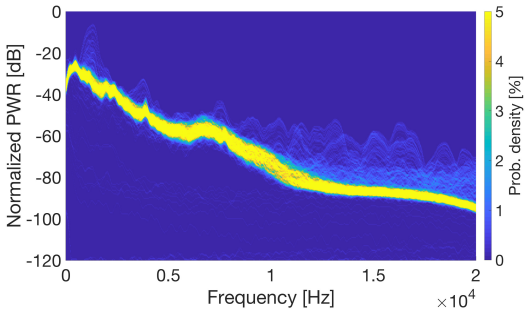
(a) Signal of a single drop



(b) Spectrogram (single drop)

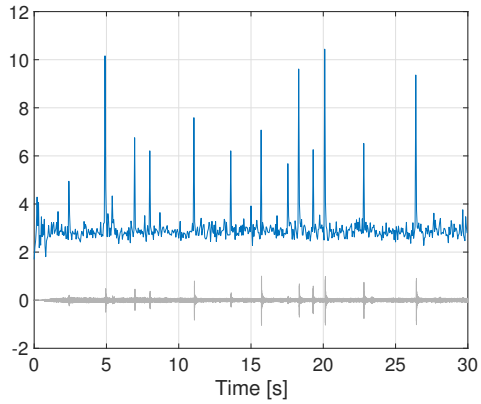


(c) Persistence spectrum, background noise

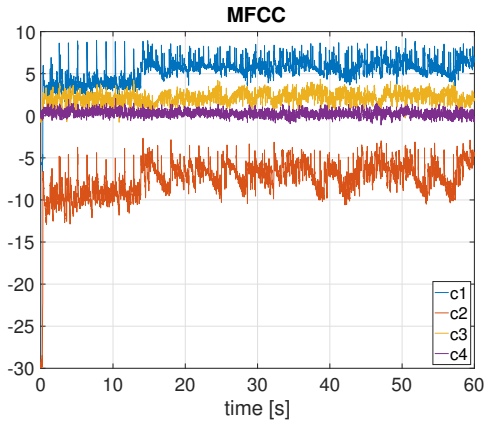


(d) Persistence spectrum, drops on background noise

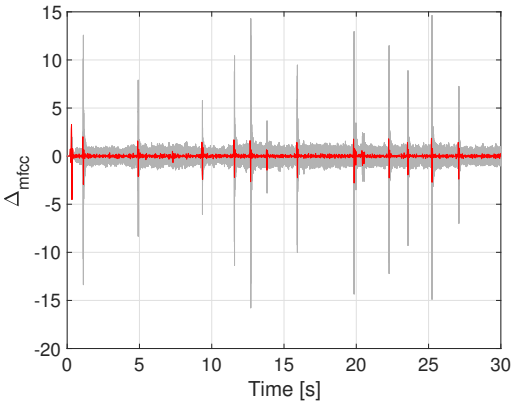
Figure 3: Signal and spectrogram of a dripping tap on water: (a) and (b). (c) Persistence spectrum of a continuous background noise (c) and the persistence spectrum of a drop on a continuous background noise (d).



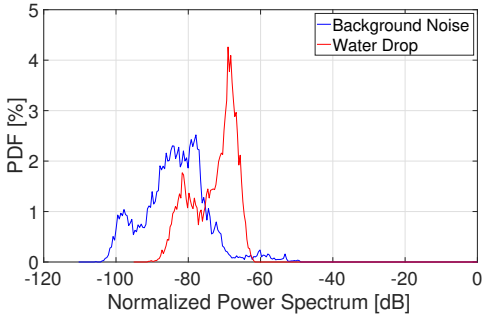
(a) Kurtosis.



(b) MFCC



(c) Speed of variation of the MFCC (in red), and the signal (grey)



(d) High Pass Probability Density Function

Figure 4: Acoustic features of a 30 seconds of signal of a dripping tap on water with background noise.(vacuum cleaner).

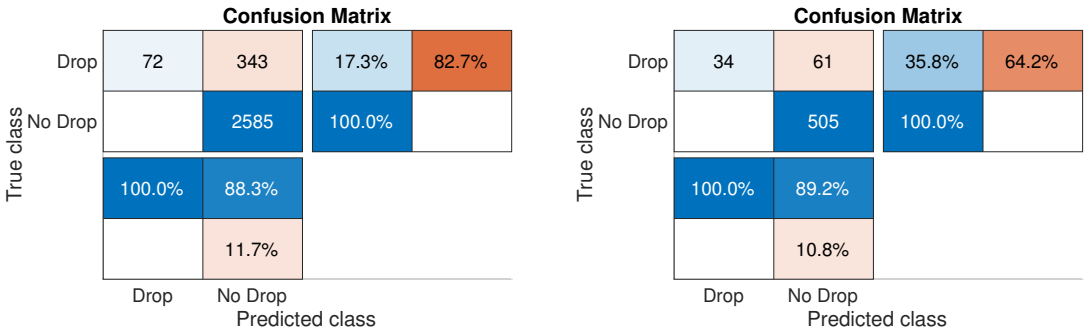
coefficients matches the frames where the event is present. For a given sample n, the Δ_{MFCC} are calculated as:

$$\Delta_{MFCC}[n] = MFCC[n] - MFCC[n-1]. \tag{7}$$

4. RESULTS AND DISCUSSION

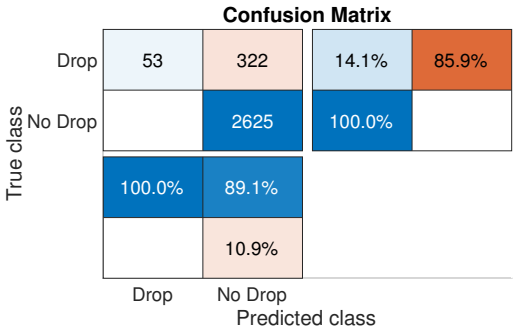
Once selected the acoustic features to be used, a training set has been created to train a linear SVM (Support Vector Machine) classifier [17] [18]. Different tests have been performed to check the performance of the detection:

- A set of test using samples of background noise and drips of the recorded databases not included in the training set .
- Test on real conditions.

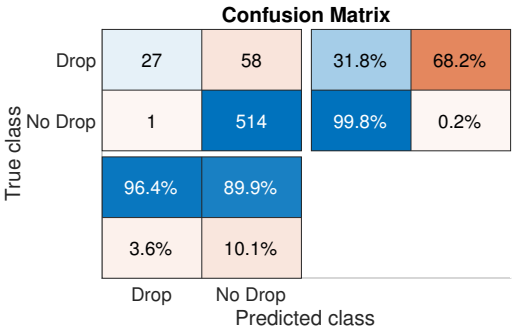


(a) Confusion Matrix, t_s=10 ms

(b) Confusion Matrix, t_s=50 ms



(c) Confusion Matrix, t_s=10 ms



(d) Confusion Matrix, t_s=50 ms

Figure 5: Confusion Matrixes for classification tests on the a subset of the prerecorded database: (a) and (b) continuous backgroung noise (vacuum cleaner). (c) and (d) background music.

4.1. Subset of the recorded database

The first set of tests have been performed using signals created from a subset of the recorded database that has been reserved as "test set", i.e. none of the events or the background noises used to create the test signals have not been used during the training stage. The figure 5 shows the confusion matrixes obtained in the cases:

a) Random set of drops (a mixture of recorded and synthesized drops) on a continuous background noise (vacuum cleaner). A SVM trained with features extracted with a frame length t_f = 10 ms is used.

b) The same test signal, but in this case the frame length is set to t_f = 50 ms.

c) Random set of drops on background music with t_f = 10 ms.

d) Random set of drops on background music with t_f = 50 ms.

The results shown in the figure 5 are obtained for a test signal of 30 seconds, which means a total 3000 frames in the case of a frame length $t_f = 10ms$ or 600 frames in the case of selecting a frame length of $t_f = 50ms$:

– For a frame length $t_f = 10ms$, the 17.3 % of the frames of the class "Drop" are correctly detected in the case of continuous background noise. In case of background music, the percentage of true positives is 14.1 % of the "Drops" are correctly detected. There is not false positives in both cases.

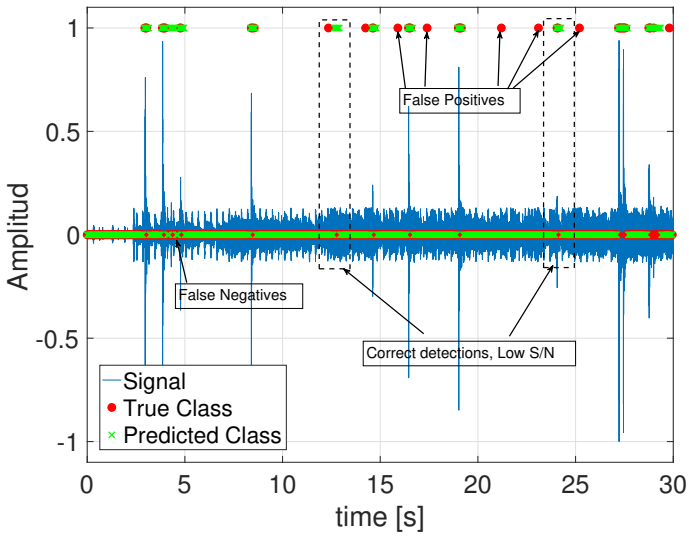


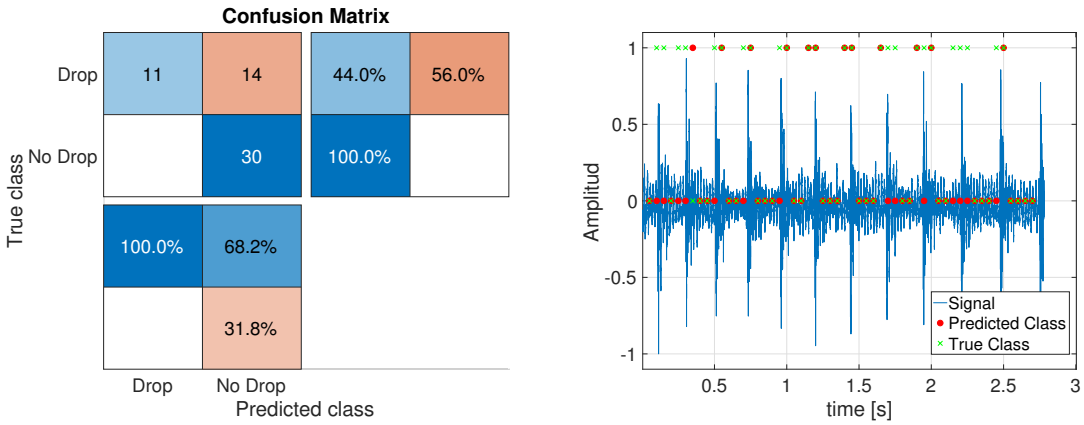
Figure 6: Detection of drip sounds on background music.

– If the frame length is $t_f = 50ms$, the percentage of true positives increases up to 35.8 % (continuous background noise) and 31.8 (background music), although in the case of background music there is a 3.6 % of false positives.

– The figure 6 shows an example of the results in the case of detection of drip sounds with background music. It can be noticed that there is not bursts of false positives: they are only isolated single frames wrongly classified as "drop". In a post-processing stage the decision of "event" could be taken only when a minimum number of frames have been detected as "event" within a time interval.

– The system shows the possibility to detect drops even for low signal to noise ratio.

4.2. Test on real conditions



(a) Confusion Matrix

(b) Results of the detection

Figure 7: Confusion Matrix for the real test: dripping tap on an aluminum sink with continuous background noise (vacuum cleaner). S/N=10 dB.

The system has been tested with a recording of a dripping tap on an aluminum sink with a background noise (a vacuum cleaner). The averaged Signal to Noise ratio during the recording was 10 dB. The figure 7 shows the results obtained. From a total of 25 frames expected as "Drop", only 11 have been classified as "Drops", being classified as "No Drop" (or Background Noise) 14 frames. The real test does not show any false positive, i.e. none of the frames expected as "No Drop" has been detected as "Drop". If it is taken into a count that a "frame by frame" detection is performed and each drop contains 3 or 4, frames it could be enough to detect correctly 1 over 3 or 4 frames of a "event". This means that an event could be classified as "drop" if at least 25 % of the frames within a time interval of ti=0.2 s (duration of a frame) have been classified as "Drop". It can be noticed in the example shown in the figure 7 (b) how at least one of the frames is correctly detected in 10 of the 12 drops recorded during the 3 seconds of the signal shown. This means that 83 % of the drops have been properly detected.

5. CONCLUSIONS

In this paper it has been evaluated the possibility of detection of drip sounds on real conditions, which means variable background noise. A training database has been created, using synthetic drip sounds with real recordings which have been mixed with different domestic background noises, as vacuum cleaners, washing machines, shavers, speech and different kinds of music. A training data set has been created from the data base, using the kurtosis, the probability density function and the speed of variation of the MFCC as features to train a SVM classifier. The table 1 shows a summary of the measurement of the classifier quality performance:

- Accuracy: measures the percentage of frames correctly classified over the total number of predictions.
- Recall: measures the percentage of events correctly classified (true positives) over the total number of expected events (i.e. true positives + false negatives).
- Specificity: measures the percentage of no-event frames (true negatives) over the total number of expected negatives (true negatives+false positives).
- Precision: measures the fraction of true positives over the total of positives expected.
- FMesure or F-Score: it is a weighted harmonic mean of the precision and recall:

$$F - Score = \frac{2precision \times recall}{precision + recall} \tag{8}$$

As it can be seen in the table 1, the classifier has a high precision and low recall, which means that even though a significant number of frames that should be classified as belonging to the "event" class, they are misclassified, those detected are correct with a high probability. The high specificity means that the probability of false alarm is really low. All the test performed show promising results, showing that it is possible to implement a drip detection through its acoustic signature.

Table 1: Classification measures from tests performed over the database set (samples with continuous background noise and background music), and a test performed on a recording on real life conditions.

	Continuous Noise	Background Music	Real Test
Accuracy (%)	93.55	90.33	72.73
Recall (%)	59.55	41.18	44.00
Specificity (%)	99.41	98.45	96.67
Precision(%)	94.64	81.40	91.67
FMeasure	0.73	0.63	59.46

6. REFERENCES

[1] Rosa M. Alsina Pagès, Joan Navarro, Francesc Alías, and Marcos Hervás. homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. *Sensors (Basel)*, 17, April 2017.

[2] Shivan Sharma et al. An intelligent system for infant cry detection and information in real time. In *Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017.

[3] Shubham Asthana, Naman Varma, and Vinay Mittal. Preliminary analysis of causes of infant cry. In *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 000468–000473, 12 2014.

[4] Jiaxing Ye, Takumi Kobayashi, and Masahiro Murakawa. Urban sound event classification based on local and global features aggregation. *Applied Acoustics*, 2017.

[5] Dan Ellis Tuomas Virtanen, Mark D. Plumbey, editor. *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[6] BBC sound effects. http://bbcsfx.acropolis.org.uk/

[7] M. Kob and M. Vörlander. Band filters and short reverberation times. *Acustica united with Acta Acustica*, 86:350–357, 2000.

[8] Patrice Guyot. Julien Pinquier, Régine André-Obre. Water sound recognition based on physical models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP 2013*, pages 793–797, Vancouver, May 2003.

[9] William Moss and Hengchin Yeh. Sounding liquids:automatic sound synthesis from fluid simulation. In *ACM Transactions on Graphics*, volume 28, December 2009.

[10] Kees van der Doel. Physically-based models for liquid sounds. In *Proceedings of ICAD 04-Tenth Meeting of the International Conference on Auditory Display*, Sydney, Australia, July 2004.

[11] Iryna Skrypnyk. Irrelevant features, class separability, and complexity of classification problems. In *IEEE 23rd International Conference on Tools with Artificial Intelligence*, Boca Raton, Fl. USA, November 2011.

[12] Iman Khosravi, Abdolreza Safari, and Saeid Homayouni. Msmd: maximum separability and minimum dependency feature selection for cropland classification from optical and radar data. *International Journal of Remote Sensing*, 39(8):2159–2176, 2018.

[13] Xiaodan Zhuang et al. Feature analysis and selection for acoustic event detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas,NV, USA, April 2008.

[14] Eva Vozáriková, Jozef Juhár, and Anton Čižmár. Acoustic events detection using mfcc and mpeg-7 descriptors. In Czyżewski A. Dziech A., editor, *Multimedia Communications, Services and Security. MCSS 2011. Communications in Computer and Information Science,*, volume 149. Springer, Berlin, Heidelberg, 2011.

[15] Courtenay V. Cotton and Daniel P. W. Ellis. Spectral vs. spectro-temporal features for acoustic event detection. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, November 2011.

[16] Cédric Gervaise, A Barazzutti, Sylvain Busson, Y Simard, and N Roy. Automatic detection of bioacoustics impulses based on kurtosis under weak signal to noise ratio. *Applied Acoustics*, 71:1020–1026, 11 2010.

[17] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, June 1999.

[18] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 13(2), march 2002.

[19] T. G. Leighton. *The acoustic buble*. Academic Press, 1994.

[20] B. W. Shuller. *Intelligent Audio Analysis*. Springer, 2013.