SOUND CLASSIFICATION IN INDOOR ENVIRONMENT THANKS TO BELIEF FUNCTIONS

Quentin Labourey[*†], *Denis Pellerin*[*], *Michele Rombaut*[*], *Olivier Aycard*[†], *Catherine Garbay*[†]

[*] Univ. Grenoble Alpes
GIPSA-Lab, F-38000 Grenoble, France

[†] Univ. Grenoble Alpes
LIG, F-38000 Grenoble, France

ABSTRACT

Sounds provide substantial information on human activities in an indoor environment, such as an apartment or a house, but it is a difficult task to classify them, mainly due to the variability and the diversity of realization of sounds in those environments. In this paper, sounds are considered as a class of information, to be mixed with other modalities (video in particular) in the design of ambient monitoring systems. As a consequence, we propose a classification scheme aimed at (i) exploiting the specificities of this modality with respect to others and (ii) leaving doubtful events for further analysis, so that the risk of errors is overall minimized. A dedicated taxonomy together with belief functions are proposed in this respect. Belief functions are an adapted way to face the variability of sounds, as they are able to quantify their impossibility to classify the signals when it differs too much from what is known by creating class of doubt. The algorithm is tested on a dataset composed of real-life signals.

***Index Terms*—** Sound classification, Indoor sounds, Belief functions, Features selection, Reject class

1. INTRODUCTION

Intelligent systems equipped with sensors are starting to appear in our homes: home automation, monitoring systems, companion robots, intelligent toys are becoming accessible to everyone. In order to get information about human activity, sensors (typically cameras and microphones) provide substantial amounts of data, but they all require adapted perception algorithms to interact with the user. In this paper, sounds are considered as a source of information, to be mixed with other modalities (video in particular) in the design of ambient monitoring systems (figure 1). Extensive works have been performed as regards the visual modality in indoor environment, but sound classification in those environments is still an open and complex problem. Sounds can be a very good indicator of the content of a scene, or at least give hints as to what the objects of interest in the scene are.

However, the variety of sounds encountered in real-life scenarios, as well as the small quantity of available labelled data to learn from, often lead to a restrained framework of classification: discrimination between speech and music, music genre differentiation, speaker recognition, footstep detection, etc... The variety of sounds also makes it difficult to propose a hard classification method as a sound coming from the same source can change depending on the situation. Moreover, in case of an interactive or reactive system, misclassification can have serious negative consequences, particularly in the case of monitoring systems.

As a consequence, we propose a classification scheme grounded on two pillars. Firstly, sound should be exploited as a modality that brings specific information, whose role is to complement others. Therefore, there is no need to ground the classification scheme on a wide taxonomy, accounting for any event in the monitored scene. Secondly, the risk of error should be minimized, in order that information provided by this modality may be considered as reliable. Hard classification is then to be avoided. Rather, doubtful events may be classified as such in a ”doubt”/”unknown” class and left for further investigation, under complementary modalities. A dedicated taxonomy together with belief functions are proposed in this respect.

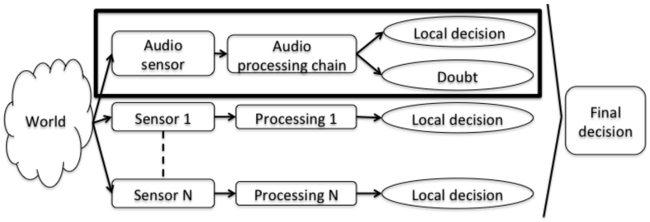

Fig. 1. General framework of an ambient monitoring system: the proposed system is in the bold rectangle

Section 2 briefly presents previous works on sound classification in different frameworks, section 3 presents the chosen taxonomy. Sections 4 describes the studied features and the method used for the selection of pertinent features. Section 5 describes how the doubt is implemented in the classification scheme thanks to a K-Nearest Neighbours algorithm in the context of belief functions. The dataset and the results of the classification on real-life signals are presented in sections 6 and 7.

2. PREVIOUS WORKS

Different approaches have been used for sound classification, although the global scheme is oftentimes similar: find a description space that separates the different classes at best and apply classification or separation in that space. Most of the time, finding the right description space to describe the signal at best means extracting features describing the most discriminant characteristics from the signals. Numerous types of classification algorithms can be used.

The literature on sound classification is diverse, in different fields:

- Speech/music discrimination or musical genre discrimination have been extensively explored with various methods, some examples of which are: [1] and [2] worked on the discrimination of a music dataset into different genre, using GMM and KNN for classification on various classical features. [3] worked on the differentiation between speech and music, introducing the uncertainty inside his classification with the use of the C-means algorithm. However this uncertainty was not used in the decision step. A more complete and recent review of speech/music discrimination methods can be found in [4].
- Indoor sound classification: The literature is less diverse on the topic, as it is fairly new, however some works have been performed, mainly on recognizing sound-events or particular activities that can be heard in a home: [5] worked on sound-event detection thanks to a humanoid robot with a method based on vector quantization with a very precise taxonomy on particular activities, [6] performed an experiment inside an automated home, where people were supposed to accomplish daily activities during a time and the sound was acquired. An event-detection system followed by a rather simple sound classification was executed, then automatic speech recognition was performed on that database, however the dataset itself is quite complex. Various other works on specific problems exist such as elder fall detection [7], footstep classification [8], but the present work tries to define a more general framework.

3. SEMANTIC GRANULARITY OF CLASSES

Most of the works described in the previous section are based on a semantically precise taxonomy [9], which can be hindering in some cases:

- Having a lot of precise classes means having to learn the different classes on precise objects of a defined environment. This would cause problems for adapting the system to a new environment.
- The diversity of indoor sounds makes it difficult to chose a granularity when semantic classes are chosen, and even inside a semantic class, the realizations can be diverse, e.g. should we make the difference between a door slamming and a window slamming, or footsteps on a wooden floor and footsteps on tiles ?
- It is important for an intelligent system to make as few errors as possible, particularly in case of monitoring. If two classes are close in terms of features but semantically very different (e.g. a person falling versus a door slammed), it might be better to consider that they are the same class to further investigate it with other sensors such as cameras or depth-sensors.

From these considerations, we decided to consider the following classes. These classes are meant to share a common fate as regards the overall goal of our design, being (i) solely identified on the basis of sound information, and (ii) discriminated with a rather high confidence:

- **Speech:** In indoor environment, speech is the most important indicator of a person being in presence. It is important for monitoring systems to be able to detect if a person is speaking or not.
- **Music:** Music is a very common sound that can be heard in indoor environment: radio, television, music is almost everywhere in our everyday life.
- **Short indoor sounds (impact):** It is difficult to separate the sound of a door slamming, a person falling, the impact of something thrown on a table, etc... However each time an impact is heard, it is important to be able to classify it as it can mean a problem.
- **Long indoor sounds:** This class regroups all sounds with sparse energy that are longer than an impact: steps, knocking, clothes friction, etc..
- **Doubt**: As said, the variability and the diversity of sounds that can be heard in indoor environments makes it important to be able to classify an extract as an unknown sound. The source of this type of sounds will require investigation by other sensors.

4. FEATURES SELECTION

To classify the extracts at best, it is necessary to chose adapted features. This section briefly describes the features that were studied for this work, and the way the signal is finally represented before classification, by selecting the features that separate the classes best individually. The features that were chosen in the final selection are marked with a ”*” sign. A more thorough description of the features can be found in [3].

The process of feature extraction is the following one: a sliding hamming window W is applied on the whole signal with overlapping. At each window position, features are extracted. As this represents a lot of data, those features are statistically aggregated to obtain one feature vector per signal, which is the final representation.

4.1. Classic sound features

Features in sound analysis are classically divided into 3 broad categories: temporal, spectral and transform features.

In temporal features, one of the most classic is the root mean square energy (E), which represents the quantity of energy associated with a portion of the signal. As such this feature does not give much information about the content of the signal but enables the computation of Low Energy frames (LE)* which is the fraction of frames which energy is lower than a portion of the average root mean square energy across all frames. This feature is efficient to detect the temporal sparsity of energy across the signal. More precisely it can easily discriminate signals with an energy temporally focused at one instant. Another frequently used temporal feature is the Zero Crossing Rate (ZCR), which has been extensively used in the speech/music discrimination framework. It counts the number of time the signal changes its sign inside an analysis window.

In spectral features, the spectral flux (SF)* represents the local temporal variations of the spectrum, while central frequency and bandwidth describe the repartition of the energy in a window of analysis. Spectral Rollof is defined as the quantity of the spectrum containing a defined percentage α of the power spectrum. Most of the time the percentage is chosen between 80% and 95%.

As for transform features, Mel-Frequency Cepstrum Coefficients (MFCC)* are amongst the most commonly used features in speech recognition and discrimination. They are obtained by mapping Fourier coefficients on the Mel-scale thanks to triangular overlapping windows. As for [3], only the first 13 MFCCs are extracted as the first coefficients are known to be the most significant.

4.2. Normalization and final representation

To compare the different features, the euclidian distance is used. As the features do not have the same dynamic, it is important to normalize each feature across the whole dataset to avoid favoring a feature with a bigger dynamic during the distance computations.

Each of the cited features are extracted for every window of analysis, which means that for a 10 second signal, with a 20 ms window analysis W, a total of 950 features is extracted from one signal. To obtain a representation of lesser dimension of the signal, we chose to compute for each feature the mean and variance, which leaves us with a representation of the signal by a feature vector s_i of size N = 40.

4.3. Feature selection

Extracting a lot of features does not mean obtaining better results in the classification. As we are going to base the classification on the euclidean distance between signals in the feature space, we can have a rough idea of how individual features will separate the different classes side-by-side, by computing the mean distance between each pair of class for that features only, and taking into account intra-class mean distance, to choose feature that represents the classes in the most compact way. That is why the following expression is computed for each feature and each pair of classes:

$$D_{C_i,C_j} = \frac{D_{inter}(C_i,C_j)}{D_{intra}(C_i) + D_{intra}(C_j)} \quad (1)$$

where $D_{inter}(C_i,C_j)$ is the mean distance between all pairs of classes C_i and C_j for the considered feature. Let C_i contain N_1 examples x_i of the considered feature, and C_j contain N_2 examples y_j:

$$D_{inter} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} |x_i - y_j| \quad (2)$$

D_{intra} is the same distance within one class. This enables us to see which features separates the classes best. The higher the distance, the more distinct the classes in the considered feature domain.

An arbitrary number of four features were chosen for the classification amongst those that yielded the higher distance between classes: two MFFC variances for MFCC number 2 and number 4, the spectral flux variance, and a very selective LE selecting the number of frames with a root mean square energy higher than 10 times the average value. It appears that the features that discriminate the classes at best individually are either transform domain features (MFCCs have shown great results in speech and music discrimination in the past), or representative of the temporal repartition of energy in the signal. It makes sense, since our classes are partly based on the sparsity of energy inside the signal. Spectral feature that describes that signal at a precise time like bandwidth or central frequency, might not suffice, as a class can be quite diverse spectrum-wise, while features that describes the evolution of the spectrum in time, such as the spectral flux might be better. Those four features are well adapted to our proposed taxonomy.

5. CLASSIFICATION SCHEME

As it was stated before, it is important to have as few misclassification as possible: if a person falling is classified as a person speaking by mistake, the system cannot be considered efficient.

The classification algorithm used for this work has to take into account the uncertainty of classification, which means ”hard” classification must be ruled out. Moreover, it is important to quantify the ignorance we have about the class of the considered extract: if the system is unable to classify the sound heard, information can be relayed to sensors that might be able to understand the nature or the source of the sound.

Belief functions are particularly adapted to that kind of classification [10]: they enable the classification system to quantify its ignorance about the considered extract. The classification used here is the belief K-Nearest neighbors: part of the dataset is used as a training set (labelled data), and the rest is used as the validation set.

For each signal s to classify amongst the class C = $\{\{C_1\}\cdot\cdot\cdot\{C_M\}\}$, the class of its K nearest neighbors x$_i$ with $i \in \{1, .., K\}$ is observed. For the ith neighbor of class C$_q$, the basic probability assignment on the subsets of C is built as follows:

$$m^{s,i}(\{C_q\}) = \alpha \exp^{-\gamma d^i} \tag{3}$$

$$m^{s,i}(C) = 1 - \alpha \exp^{-\gamma d^i} \tag{4}$$

$$m^{s,i}(\bar{C_q}) = 0 \tag{5}$$

where γ is a constant enabling to control the decrease in belief in class C$_q$ with respect to the distance d to the neighbor x$_i$. Here, γ is set to the mean euclidean distance within the class C$_q$. α is a constant enabling the doubt even if the extract to classify is placed at the exact same place as labelled data (here 0.95). m(C_q) represents the belief that is committed to C_q, and m(C) is the belief committed to ignorance (any class is possible). $m^{s,i}(\bar{C_q})$ represents the belief committed to any class other than C_q.

The belief masses of neighbors x$_i$ of the same class C$_q$ are then aggregated in one unique belief mass m$_q^s$:

$$m_q^s(\{C_q\}) = 1 - \prod_{x_i \in C_q} \left(1 - \alpha \exp^{\gamma d^i}\right) \tag{6}$$

$$m_q^s(C) = \prod_{x_i \in C_q} \left(1 - \alpha \exp^{\gamma d^i}\right) \tag{7}$$

Finally, all belief masses of all neighbors are aggregated in on final belief mass m^s thanks to a Dempster combination (as shown in [10]):

$$m^s(\{C_q\}) = \frac{m_q^s(\{C_q\}) \prod_{r \neq q} m_r^s(C)}{\eta} \tag{8}$$

$$m^s(C) = \frac{\prod_{q=1}^{M} m_q^s(C)}{\eta} \tag{9}$$

where η is a normalization constant:

$$\eta = \sum_{q=1}^{M} \left[\prod_{r \neq q} m_r^s(C)\right] + (1 - M) \prod_{q=1}^{M} m_q^s \tag{10}$$

The value 1-η represents the conflict: the higher the value of η, the more the belief mass cannot be attributed to any class (ignorance included).

At the end of this process, a unique belief function is obtained, representing for each class the belief for the extract to belong to that class. A decision can be taken as to which label L is assigned to the considered extract. Several types of decision exist in the literature, here the attributed class is the class with the highest belief mass.

$$L = \max_i (m^s(C_i)), C_i \subset C \tag{11}$$

It also means that if m(C) is the maximum, the extract will be classified as unknown (doubt), which turns the ignorance into a class itself.

6. DATASET AND PARAMETER SELECTIONS

The accessibility to indoor sound data acquisition is still complicated and the datasets are often quite complex. In this paper, the signals composing the database are taken from different datasets:

- Speech: 27 extracts of 10 seconds of speech were extracted, mainly from the GTZAN database [1] which regroups music and speech examples from various radio broadcasts and TV show, with different levels of background noise.
- Music: 30 extracts were also extracted from the GTZAN database.
- Impacts: To obtain sound of impacts of everyday life, we exploited the Sweet-home corpus created by Vacher et al. [6]: in this corpus, several participants were put inside the automated house DOMUS and asked to accomplish daily routine during 2 hours such as cooking, manipulating objects, sleeping, etc..., which contains impact-like sounds. The impacts extracted include: doors and windows slamming, impact of objects on walls and table, etc.. 16 sounds were extracted in total from the datasets, all 10 seconds-long.
- Long sounds: This class is composed of a variety of sounds extracted partly from the Sweet-home corpus and partly from other sounds found on the internet. It is composed by sounds of footsteps, door knocking, dishes making etc.. The duration of those signals vary from 2 to 13 seconds.

Each signal of the database is sampled at 22.5 kHz, and is processed with a 20 ms sliding window W with a 10 ms between two successive windows. For the experiments, the number of observed neighbors was set to 3, to enable us to decrease the size of the training set as much as possible.

7. RESULTS

A cross-validation method is applied to test the method on the dataset: training set and validation set are extracted randomly 1000 times from the dataset with a defined number of training samples. The separation in sets can be seen in table 1. A classification is considered an error if the class attributed to an extract is not the right one. The classification is not considered as an error if the system is unable to classify the extract.

The result of this classification can be seen in table 2 (line 1). The mean error (mean-number of misclassified signal) is 4.5 (less than 10%). The mean number of signals classified as unknown is 4.1, which brings the total of non-correctly classified signals with our method to 8.6. It is to be noted that the misclassified signals are equally spread amongst the classes, which would indicate that the features do not favor a particular class. A good portion of the doubt (2 signals) belongs to the class ”Long sounds” class, which makes sense as it is also the most complex class, while the rest of the doubt is divided almost equally amongst the other classes.

Class	Nb of training extracts	Nb of test extracts
Speech	11	16
Music	11	19
Impact	5	11
Long Sounds	5	12
Total	**32**	**58**

Table 1. Separation of each class into 2 different sets

Those results were compared to those given by a classic SVM with the same features (line 2 of table 2). The SVM gives out a higher number of hard misclassification, but a lower overall number of non-correctly classified extracts. The misclassified signals are higher in two classes: music and impact. It seems only logical that the added number of non-correctly classified extracts is higher in the case of the belief-KNN as the SVM defines a separating hyperplane for the data, and hard classifies the data according to that hyperplane. The belief-KNN would tend to classify an extract which is close to labelled extracts of various classes as unknown (doubt class). The only class were there is a loss in classification compared with the SVM is the speech class. This difference cannot be easily explained, although it can easily be corrected by taking other modalities into account (especially vision).

	Speech	Music	Impact	Long	Doubt	Total
KNN	1.4	1.3	0.7	1.1	4.1	8.6
SVM	0.6	2.5	2	1.1		6.3

Table 2. Results (mean number of misclassified signals)

8. CONCLUSION

The proposed classification method for indoor sounds signals obtains good classification performance, even though comparison with state-of-the art can be difficult. This method enables the non-classification of extracts in case the system does not reach a certain degree of confidence in one particular class, which makes it adapted to multi-sensor systems. Decreasing the semantic content of classes, making them more specifically adapted to the information conveyed by the sound modality enables the system to classify sounds that share a common fate .

9. ACKNOWLEDGMENTS

The authors would like to thank Vacher et al. [6] for the given access the Sweet-home corpus.
This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025).

REFERENCES

[1] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] T. Langlois and G. Marques, “Automatic music genre classification using a hierarchical clustering and a language model approach,” in *First International Conference on Advances in Multimedia*, 2009, pp. 188–193.

[3] M. Haque and J. Kim, “An analysis of content-based classification of audio signals using a fuzzy c-means algorithm,” *Multimedia Tools and Applications*, vol. 63, no. 1, pp. 77–92, 2013.

[4] M. Velayatipour and M. Mosleh, “A review on speech-music discrimination methods,” *International Journal of Computer Science and Network Solution*, vol. 2, no. 2, pp. 67–78, 2014.

[5] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, “Sound-event recognition with a companion humanoid,” in *12th IEEE-RAS International Conference on Humanoid Robots*, Nov. 2012, pp. 104–111.

[6] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, “The Sweet-Home speech and multimodal corpus for home automation interaction,” in *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, pp. 4499–4506.

[7] Y. Zigel, D. Litvak, and I. Gannot, “A method for automatic fall detection of elderly people using floor vibrations and sound: Proof of concept on human mimicking doll falls,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 12, pp. 2858–2867, Dec. 2009.

[8] A. Itai and H. Yasukawa, “Footstep classification using simple speech recognition technique,” in *IEEE International Symposium on Circuits and Systems*, May 2008, pp. 3234–3237.

[9] G. Song, D. Pellerin, and L. Granjon, “Different types of sounds influence gaze differently in videos,” *Journal of Eye Movement Research*, vol. 6, no. 4, pp. 1–13, Oct. 2013.

[10] T. Denoeux, “A k-nearest neighbor classification rule based on dempster-shafer theory,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, May 1995.