Automatic sounds clustering approach based on a likelihood measure computation

Maxime ROBIN*†, Grégoire NICOLLE‡, Alexandre ROTA‡
*Sorbonne University, Université de technologie de Compiègne, CNRS,
UMR 7338 Biomechanics and Bioengineering, Centre de recherche Royallieu - CS 60 319 - 60 203 Compiègne cedex
†KRG Corporate
Email : maxime.robin@openmailbox.org
‡E.S.M.E. Sudria
Email : greg.nicolle@gmail.com
Email : alex0rota@gmail.com

***Abstract*—This paper presents an automated sound clustering method. This technique aims to classify sounds in a non-supervised way. It helps for instance to gain time to label sounds for sounds database creation. The input is a cloud of sounds and the proposed algorithm regroup these sounds by similarity. These groups are created according to sound descriptors likelihood, without analyzing the content of the sound.**

***Keywords*—Sounds clustering, auto-clustering, artificial intelligence, clustering, non supervised, pre-classification.**

I. INTRODUCTION

A. Context

Creating a sound database is a long and fastidious process. This task is manual and is tedious to automatize. Many Artificial Intelligence systems work in a non supervised way, like neural networks. Neural networks depend a lot on their learning base (supervised or not). The neural network will provide wrong outputs if its learning is not advanced enough. Here we propose a new approach to create sound classes by resemblance. Our goal is not to determine what is the actual content of one sound, but to create an algorithm that determines the similarities between sounds and creates classes.

B. Survey : How do humans differentiate sounds ?

Starting from the cocktail party versus radio or TV sound problem from speech analysis, we performed a survey containing such sounds. This survey contains three different lengths of sounds : 2, 5 and 10 seconds. We postulated that we would get a higher number of good answers for 10 second sounds.

Table I
SURVEY RESULTS

	Good answers (%)	**Bad answers (%)**	**Did not answer (%)**
2s files (140 anwsers)	87.14	9.29	3.57
5s files (160 anwsers)	74.38	19.38	6.24
10s files (80 anwsers)	76.25	11.25	12.5

Based on the results (Table I) people give better answers for short files. The longer the files are the more misakes we get. We can also notice that the error rate is substantial.

II. APPROACH

We want to create clusters from disparate sounds (Figure 1). Naming and identifying sounds is not in the scope of this paper. Our aim is to group sounds naively by similarity without identifying them explicitly. Creating more groups than needed is not an issue as the aim is to minimize errors in clustering.

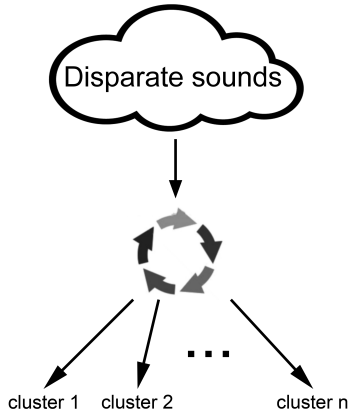

Figure 1. From unsorted sounds to clustered sounds

The algorithm splits each sound in small parts, which are qualified by 52 parameters. Then, for each parts, we determine the resemblance with all other parts on our database. This is indicated by $\varrho(\omega_i,\ \Omega_n)$ with ω_i the part actually measured and Ω_n all others parts of sounds excluding the actual sound. Tje likelihood measure (6) is based on the euclidean distance between each parameters of one window and each parameters of the other windows.

III. ALGORITHM AND IMPLEMENTATION

A. Suggested algorithm

We suggest an algorithm (Figure 2) in five steps. The last two steps are run for each sound, included linked sounds. Reason being links are not necessarily reciprocal.

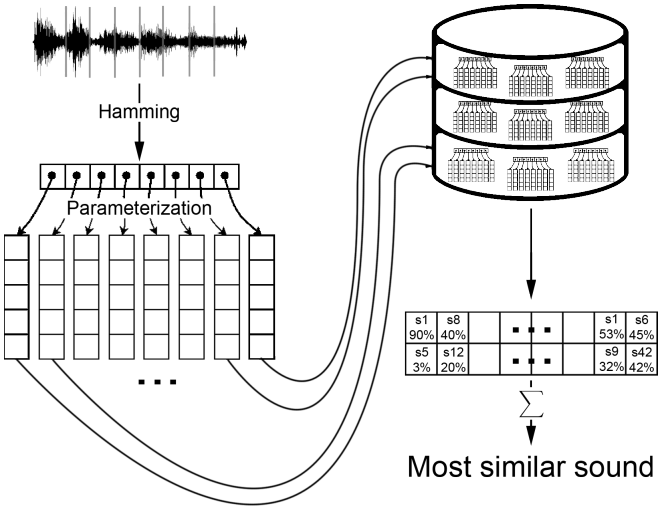

Figure 2. Proposed algorithm

B. Signal pre-processing

We decided to choose Hamming windowing (1) to split our signals into 20 ms time frames because rectangular windowing causes more side effects.

$$f_v(n) \triangleq \begin{cases} \alpha - (1-\alpha)\cos(\frac{2\pi n}{N}), & \text{if } 0 \leqslant n < N \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

In our case, we want to differentiate each sound. To dintinguish the kind of sound, music, voice, etc. is not mandatory. We chose hamming windowing which has a good resolution of close harmonic separation for this reason. We used 52 parameters to characterize each window.
These parameters belong to three categories : temporal, frequency and cepstral descriptors.

1) Temporal descriptors: We use Zero crossing rate (ZCR) and total energy as temporal parameters. ZCR (2) represents the main frequency of the signal, and is the number of times the signal inverts sign.

$$ZCR(t) = \frac{1}{2L}\sum_{\tau=1}^{L} |sgn(s_t(\tau)) - sgn(s_t(\tau-1))| \quad \text{with, } sgn(s_t(\tau)) = \begin{cases} 1, & \text{if } s_t(\tau) \geqslant 0 \\ -1, & \text{if } s_t(\tau) < 0 \end{cases} \quad (2)$$

Total signal energy (3) helps to identify : stationary, silence, temporary redundant signals.

$$e(t) = \frac{1}{N_t}\sum_{i=0}^{N_t-1} s_t^2(i) \quad (3)$$

2) Frequency descriptors: We use the spectral roll-off point (SRF) and Spectral Centroid (SC). These two parameters are used to determine the shape of the signal. Spectral Roll-off point (4) has a cutoff frequency where 95% of frequency based energy is contained.

$$SRF(t) = max\left\{ K \backslash \sum_{i=0}^{K} |S_t(f_i)|^2 < TH \sum_{i=0}^{N_t} |S_t(f_i)|^2 \right\} \quad (4)$$

The spectral centroid (5) is the frequency center of gravity for a signal. It is processed as the weighted average of frequencies in the signal.

$$SC(t) = \frac{\sum_{i=1}^{N_t} f_i S_t(f_i)}{\sum_{i=1}^{N_t} S_t(f_i)} \quad (5)$$

3) Cepstral descriptors: Mel-frequency cepstral coefficients (MFCC) are the coefficients that make up a mel-frequency cepstrum (MFC). MFC is a representation of a short-term power spectrum of a sound. We use 16 coefficients as parameters for one Hamming window.
We use derivative (Δ) and second derivative ($\Delta\Delta$) coefficients to keep a temporal factor. We can see the influence of one part of the signal on his neighbor. Δ and $\Delta\Delta$ coefficients are the derivative of each MFCC coefficient of the signal, this adds $16 + 16$ parameters for each time frame.

C. Proposed likelihood measure

All these descriptors create a vector characterizing each window (β) of a sound. For each window (ω_i), we define the similarity of all the others windows of each sound (Ω_n). The equation (6) represents the resemblance (ϱ) of a window ω_i facing another one ω_j (with $\omega_j \in \Omega_n$). This resemblance is defined here as the euclidean distance between the two vectors of the two windows. Each vector parameters is noted α_x.

$$\varrho(\omega_i,\ \omega_j) = \sum_{\alpha=0}^{N} dist(\omega_{i_\alpha}, \omega_{j_\alpha}) \quad \text{with, } dist(\omega_{i_\alpha}, \omega_{j_\alpha}) = \sqrt{\sum_{x=1}^{\eta} (i_{\alpha_x} - j_{\alpha_x})^2} \quad (6)$$

Using the k-Nearest Neighbors (k-NN) algorithm we decided to keep the two best candidates for each window. Summing these distances (7) returns the most similar sound to the actual sound.

$$\varrho(s_i, s_j) = \sum_{\beta=0}^{M} k\text{-}nn(\varrho(\omega_\beta,\ \omega_j)) \quad (7)$$

The most similar sound is then linked to the actual sound. Since the most similar sound of one sound is not necessarily the same as the sound which is linked to it, a cluster can be created.

IV. Results

For the evaluations we used sound databases and gave these to our system. For each test shown below you can find the number of sounds, in parenthesis. Each sound duration is between 4 and 10 seconds. The k-means is not relevant in a realistic environment since you have to fix the number of output clusters. And our system works independently from any input database.
As we can see on table II, the sounds are perfectly clustered, despite

an apparent similarity between the sounds. The table III shows that if we unbalance the number of sounds of each cluster the error rate increases. Table IV shows our algorithm with more inputs. The error rate is about twice lower than the one survey (table I) shows.
The suggested algorithm does not provide reliable results for sounds, laugh, female cry, scream, because there are clustering errors between human sounds. But if we regroup laugh, female cry and scream under the label *human sounds* the error rate is acceptable (about 5%). The proposed algorithm distinguishes human sounds from other sounds. But not humans sounds between them.
Table VI presents our algorithm tested on all files of the database (table V). The first panel corresponds to each class versus the others. The second panel is the same panel but regroups human sounds, eliminating the errors within the human sounds.

Table II
GLASS BREAKING (15) VS DISHES (15)

	Output Clusters	Error
K-means	2 (fixed)	20%
Proposed Algorithm	5	0%

Table III
GLASS BREAKING (15) VS DISHES(48)

	Output Clusters	Error
K-means	2 (fixed)	23.8%
Proposed Algorithm	5	3.17%

Table IV
DOOR CLAPPING (114) VS DISHES (96) VS LAUGH (99)

	Output Clusters	Error
K-means	4 (fixed)	25.7%
Proposed Algorithm	62	4.53%

Table V
FULL DATABASE

Label	Occurrences
Breahlessness	8
Brushteeth	5
Burp	16
Door knock	8
Cry (female)	17
Glass breaking	15
Laugh	49
Scream	48
Sneeze	35
Snore	24
Wipe	12
Yawn	26
Dishes	96
Door clapping	114

Table VI
ALL DATABASE TESTS

	Output Clusters	Error
Panel 1	73	26.43%
Panel 2	73	9.51%

V. CONCLUSIONS AND PERSPECTIVES

We suggested an algorithm that allows to cluster sounds without any learning and without analysis of the sound. This algorithm has an error rate of approximately 5%. But the errors are actually propagated on the grouping part of the algorithm. A correcting algorithm aiming to break false links between two sounds would greatly improve the error rate.

ACKNOWLEDGMENTS

We would like to thank D. ISTRATE, J. BOUDY, M. DOMINJON and A. BAUDON for their precious assistance in writing this paper.

REFERENCES

[1] D. Istrate, *Détection et Reconnaissance des Sons pour la Surveillance Médical*, 2003.
[2] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze, "Sélection de descripteurs audio pour la classification des sons environnementaux avec des svms mono-classe," *Colloque GRETSI, Troyes*, 11-14 septembre 2007.
[3] J. George, A. Cyril, B. I. Koshy, and L. Mary, "Exploring sound signature for vehicle detection and classification using ann," *International Journal on Soft Computing (IJSC) Vol.4, No.2*, May 2013.
[4] A. Aljaafreh and L. Dong, "an evaluation of feature extraction methods for vehicle classification based on acoustic signals," *International Conference on Networking, Sensing andControl (ICNSC)*, 2010.
[5] O.Zammit, X. Descombes, and J. Zerubia, "Apprentissage non supervisé des svm par un algorithme des k-moyennes entropique pour la détection de zones brûlées," *Colloque GRETSI, 11-14 septembre 2007, Troyes*, 11-14 septembre 2007.
[6] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," *AAAI-94 workshop on knowledge discovery in databases, volume 2*, 1994.
[7] P. Esling and C. Agon, "Multiobjective time series matching for audio classification and retrieval," *IEEE Transactions on Speech Audio and Language Processing 2012*, 2012.
[8] A. Minard, N. Misdariis, O. Houix, and P. Susini, "Categorisation de sons environnementaux sur la base de profils morphologiques," *10eme Congres Francais d'Acoustique*, 2010.
[9] S. Allegro, M. Büchler, and S. Launer, "Automatic sound classification inspired by auditory scene analysis," in *Eurospeech, Aalborg, Denmark*, 2001.
[10] R. L. Cannon, J. V. Dave, and J. C. Bezdek, "Efficient implementation of the fuzzy c-means clustering algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 2, pp. 248–255, 1986.
[11] L. Chen, S. Gunduz, and M. T. Ozsu, "Mixed type audio classification with support vector machine," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 781–784.
[12] M. Vacher, D. Istrate, L. Besacier, J.-F. Serignat, and E. Castelli, "Sound detection and classification for medical telesurvey," in *2nd Conference on Biomedical Engineering*, 2004, pp. 395–398.