AUDIO FEATURES FOR NOISY SOUND SEGMENTATION

Pierre Hanna, Nicolas Louis, Myriam Desainte-Catherine, Jenny Benois-Pineau
SCRIME - LaBRI
Université de Bordeaux 1
F-33405 Talence Cedex, France

ABSTRACT

Automatic audio classification usually considers sounds as music, speech, silence or noise, but works about the noise class are rare. Audio features are generally specific to speech or music signals. In this paper, we present a new audio feature sets that lead to the definition of four classes: colored, pseudo-periodic, impulsive and sinusoids within noises. This classification relies on works about the perception of noises. This audio feature set is experimented for noisy sound segmentation. Noise-to-noise transitions are characterized by means of statistical decision model based on Bayesian framework. This statistical method has been trained and experimented both on synthetic and real audio corpus. Using proposed feature set increases the discriminant power of Bayesian decision approach compared to a usual feature set.

1. INTRODUCTION

Advances in consumer home devices and broadcast technologies permit individuals to enjoy a large amount of audio/visual (A/V) content. To manage such wide quantity of incoming data, users need automatic techniques. In the last decade, different authors such as [1, 2] have proposed methods for structuring A/V content which are based on audiovisual descriptors. In parallel, a significant progress has been made in automatic audio classification for various application areas [3, 4]. Namely, a large amount of work addresses the problem of classification of audio into music, speech, silence and noise. Amongst those four classes, the noise class is the most complex and unstructured one. Still few research is devoted to the analysis of noise in audio.

In this paper, we are interested in a characterization of noisy sounds and its application to the problem of detecting noise-to-noise transitions in sound tracks. The work is based on the assumption that audio classification in speech, music, noise and silence has already been done. In fact, audio tracks in broadcast A/V programs or A/V works contain various noise segments and transitions between them. Locating those transitions is essential in the field of A/V content structuring. Thus, we are firstly interested in audio features which can characterize noisy sounds in the best way. Then, an adequate method for noise transition detection can be proposed, which we formulate as a *semi-blind segmentation* problem based on selected features. The paper is organized as follows. The choice of audio features and characterization of noisy sounds is presented in section 2. The segmentation method is introduced in section 3. Some results are presented in section 4. Conclusion and perspectives are given in section 5.

2. CLASSES OF NOISY SOUNDS AND AUDIO FEATURES

According to our previous work [5], noisy sounds can be classified into four classes. This classification is based on perceptual properties. In this section, we focus on these classes and propose the choice of relevant features for their identification.

2.1. Colored Noise

The first category of noisy sounds is intuitively the category composed of sounds that can perfectly be synthesized by filtering white noise: all the perceptual properties of these sounds are assumed to be contained in the short-time spectral envelopes [6]. The name of *Colored Noises* is due to the analogy with the color of light. The examples of such sounds are numerous: sounds from seashore, wind, breathing, etc...

The main characteristic of colored noises is the envelopes of their short-time spectra. It is useful to determine one or a few parameters for describing this property. A few features have been shown as useful for speech signal or musical sounds. For example the Spectral Roll-Off [7] or the spectral centroid are particularly useful to discriminate voice from unvoiced music. However, the application of these features to noisy sounds is less precise. We propose here to apply the research results of Goodwin about the residual modeling of sounds [8]. This work relies on the noise model of perception which states that a broadband noise is correctly represented by the time-varying energy in Equivalent Rectangular Bands (ERBs). However we propose to adapt this method by choosing the Bark scale instead of the ERBs, because the number of bark bands is smaller than the number of ERBs: this

choice implies a fewer number of features. Short-time spectral envelopes of noisy sounds are represented by the short-time energies within each Bark band. Therefore, one feature (composed of 26 values) characterizes the color of the noisy sounds. Variations of only one value indicate a modification of the color, that may be perceived, since each feature is related to the perception.

2.2. Pseudo-periodic Noises

Several natural noisy sounds are characterized by the pitch that can be perceived, for example machine noises, insect flies, scratching noise, etc... The pitch may have different strengths, and several pitches can be perceived at the same time. This property may have different explanations: the spectral envelope of sounds is composed of a frequency band of high energy [9], the noisy sounds are mixes of several sounds, the noisy sounds are considered as a sum of a few sinusoids that imply noises with perceived pitch [5] or the noisy sounds are assumed as rippled noise [10]. The two first cases respectively are presented in the sections 2.1 and 2.4. In this section, we propose two features which may characterize the class of noisy sounds that is described by the two other cases.

The feature we define has to describe the perceived pitch of a sound in two ways: the strength of the pitch and its value. A few methods have been proposed in psychoacoustics to measure the pitch strength of rippled noise [10]. The method we choose relies on the autocorrelation function Γ, and more precisely on the ratio (denoted AR) of the second maximum of the autocorrelation function (denoted $\Gamma(\tau)$) to its first value $\Gamma(0)$ (total energy of the signal). This feature is already known as a technique for the segmentation of speech into voiced and unvoiced parts, for the pitch estimation of a harmonic sound and for speech recognition [11]. Nevertheless it appears to be very useful to characterize noisy sounds.

Furthermore, two pseudo-periodic noises may have one similar autocorrelation ratio without being the same sound. That's why we propose a second feature to represent pseudo-periodic noises: the estimation of the period $p = \frac{\tau}{R}$ (R sample rate).

2.3. Impulsive Noise

Several natural noisy sounds are composed of periodic or aperiodic impulses. For example, one can think about applauses, walking steps, rain drops, etc... They are referred as *impulsive noise* [12]. Frequency of pulses composing this type of sounds has to be lower than approximately 20Hz. Otherwise this frequency is detected by the hearing system as a pitch.

The pulses contained in impulsive noises are similar to the transients (attacks) of instrumental sounds. Some methods for the detection of the transients have been developed relying on the variations of energy or on the zero-crossing rate (ZCR). However, we think that these methods can hardly lead to the definition of features for the characterization of impulsive noises. Indeed, the presence of energy in high frequencies may indicate pulses but also just the level of noise.

We propose to study the distribution of samples as explained in [5]. Properties about this distribution can be quantified by the kurtosis. A local pulse, characterized by a sharp probability density function, induces a high value of kurtosis. A kurtosis value is affected to each frame. A threshold has to be chosen in order to define frames that are assumed to contain one pulse. Each kurtosis value greater than the threshold is considered as an impulsive frame. The more important and audible the pulse is, the higher the kurtosis value is. Therefore, the kurtosis value is not only an indicator of the presence of pulses, but also a feature that describes the nature of the pulse.

We think that it is also important to be able to discriminate impulsive noises that do not differ by the nature of the pulses, but by their periodicity. That's why we propose to complete the kurtosis value with the periodicity of the pulses composing impulsive noises. This periodicity is null if only one pulse has been detected.

2.4. Sinusoids within Noise

Classification systems usually consider natural sounds as music sound, speech sound or noise. However, real world noises can rarely be assumed as pure noises, because they may be nothing but mixes of several sounds from different sources, that may be harmonic.

Here we consider the real world sounds which are assumed as being mixes of several sound sources. If one or some of these sources are harmonic or pseudo-harmonic and if the noise level is not too high, these harmonic sources can be perceived: they are thus important perceptual characteristics of such sounds. The examples of real world sounds of this class are numerous: street sounds-capes with horns, wind in trees with singing birds, seashore with seagulls, etc...

Natural noises are represented by short time amplitude spectra that are composed of peaks which correspond to sinusoids contained in the sound. The feature associated to this class of noisy sounds is simply the number of sinusoids. Several analysis methods have been proposed in the context of sound analysis/synthesis. An original method has been proposed in [5]. This technique is accurate with noisy sounds and it is independent from the general volume of the sound. It is based on the statistical analysis of the intensity fluctuations. A measure for each bin of the amplitude spectra is computed and a chosen threshold permits to define the number of bins that correspond to sinusoids. Therefore, a number of sinusoids is associated to each analysis frame. We consider the analyzed sound as a mix of noise and harmonic sounds if this number of sinusoids is greater than a chosen threshold.

3. STATISTICAL SEGMENTATION OF NOISE TRANSITIONS

After the study of noise signal features we address here the problem of detection of transitions between different

noisy sounds in a time-varying audio signal. This transition can happen both between noises of the same class and between noises from different classes as described in section 2. In any case we propose a blind segmentation scheme which consists in the following. For a pair of consecutive temporal windows on a temporal noise signal the problem is to check if the boundary between the windows corresponds to a change from one sound to another or the sound is continuous. This segmentation can be called "semi-blind". In fact we will use the descriptors which characterize noisy sounds the best (versus speech and music descriptors), but still realize a blind segmentation approach similarly to [1]. We formulate the segmentation problem in a general Bayesian framework. In our problem, the stochastic variable $x = (x_1, ..., x_m)^T$ represents a vector of audio features with m the total number of features measured along the time and two hypotheses are considered, $H1$ - "the absence of audio variation at time $t0$" and $H2$ - "audio variation at time $t0$". We will suppose that $H1$ and $H2$ form the partition of the space of hypotheses, that is $Pr(H1) + Pr(H2) = 1$ (Obviously, from the sense of our problem $Pr(H1 \times H2) = 0$). From the well-known Bayes formula the following may be induced:

$$\begin{cases} Pr(H1/x) = \frac{f(x/H1).Pr(H1)}{f(x)} \\ Pr(H2/x) = \frac{f(x/H2).Pr(H2)}{f(x)} \end{cases} \quad (1)$$

with $f(x)$ and $f(x/Hk)$ representing the probability density function and the conditional probability density respectively. We will suppose Gaussian distributions $N_0(\mu_0, \Sigma_0)$ associated to $H1$ for the interval $[t0 - n, t0 + n]$, $N_1(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$ associated to $H2$ for the intervals $[t0 - n, t0]$ and $[t0 + 1, t0 + n]$ respectively with n a parameter. Then a change of audio stream at time $t0$ can be expressed in terms of the likelihood ratio as: $\frac{L1}{L2} = \frac{A}{B}$ where $A = \prod_{t=t_0-n}^{t_0+n} \frac{1}{(2\pi)^{\frac{m}{2}} \times \sqrt{det(\Sigma_0)}} \times e^{-\frac{1}{2}(X_t^T \times \Sigma_0^{-1} \times X_t)}$

$$B = \prod_{t=t_0-n}^{t_0-1} \frac{1}{(2\pi)^{\frac{m}{2}} \times \sqrt{det(\Sigma_1)}} \times e^{-\frac{1}{2}(X_t^T \times \Sigma_1^{-1} \times X_t)}$$

and

$$\times \prod_{t=t_0}^{t_0+n} \frac{1}{(2\pi)^{\frac{m}{2}} \times \sqrt{det(\Sigma_2)}} \times e^{-\frac{1}{2}(X_t^T \times \Sigma_2^{-1} \times X_t)}$$

with $Lj = f(x/Hj) \times Pr(Hj)$ the likelihood function of the hypothesis Hj, $j = 1, 2$ and Σ_k the covariance matrix, $k = 0, 1, 2$ and X a centered measurement vector.

Following the usual development, that is taking the logarithm of the likelihood ratio, we obtain:

$$\begin{array}{l} M = \frac{n}{2}\left[ln(det(\Sigma_1)) + ln(det(\Sigma_2))\right] - n \times ln(det(\Sigma_0)) \\ + \frac{1}{2}[\sum_{t=t_0-n}^{t_0} (X_t^T \times \Sigma_1^{-1} \times X_t) \\ + \sum_{t=t_0+1}^{t_0+n} (X_t^T \times \Sigma_2^{-1} \times X_t) \\ - \sum_{t=t_0-n}^{t_0+n} (X_t^T \times \Sigma_0^{-1} \times X_t)] \\ M \begin{array}{l} < \\ > \end{array} 2 \times n \times ln(P/(1-P)) \begin{array}{l} => H2 \\ => H1 \end{array} \end{array} \quad (2)$$

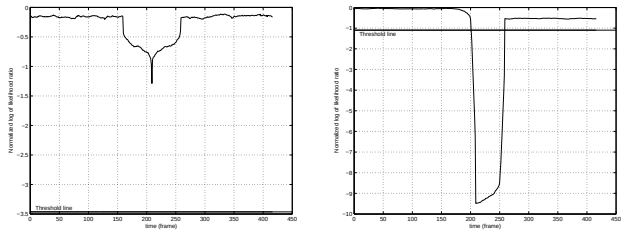
Here, P is the probability of the hypothesis $H2$ - noise transition. The segmentation method is thus as follows. Feature vectors x will be measured in two consecutive sliding windows. The decision on change will be made according to (2). In the next section the results on the use of various noise descriptors for semi-blind segmentation are presented.

4. EXPERIMENTS

Here we compare two sets of descriptors for noise segmentation using statistical method from Section 3. The first set of descriptors is constituted of Spectral Centroid, Spectral Flow, Roll-Off, Zero-Crossing Rate and Mel Frequency Spectral Coefficients (only the 13 first coefficients) [7]. The second one, which we propose, consists in the energy distribution amongst the Bark's Bands, Kurtosis, Period of Kurtosis, Auto-Correlation Ratio (ACR), Period of the ACR and Number of Sinusoids [7, 5].

In order to determine the relative discriminative power of the two sets of descriptors we conducted the experiments both on synthetic noise test data set and on excerpts from audio tracks of real-world broadcast corpus. The synthetic corpus was composed of impulsive, periodic, sinusoidal and colored noises generated with controlled parameters by the noise generation tool developed in [5]. Various combinations of noises inside the same class, described in section 2 such as impulsive, colored, etc ..., were tested. All combinations of noises from different classes following each other along the time were produced as well and submitted to the change detector. The real noise corpus contained a limited number (15) of transitions. The likelihood ratio normalized by the cardinal of feature set was computed for both sets of descriptors. The results of this comparison are shown in Figure 1. These results were obtained on a synthetic noisy sound corpus. Figure 1a presents the normalized log likelihood ratio for the descriptors of the group 1. Results for the proposed group 2 are shown in figure 1b. It can be seen that the group 2 exhibits much stronger minimum of the normalized log likelihood ratio in the case of noise change. This situation is typical. In table 1 we also show the differences between the global minimum and the closest minimum of normalized log likelihood ratio curve. The first column in table 1 contains the type of transition, e.g. i−i means that the transition is observed in ground truth between two impulsive noises. The second column contains the sound feature changed in the synthetic sound generated. The inter class transitions are notated with the key-word "Class". The last two columns contain the absolute difference between the global minimum and the closest minimum of the normalized log likelihood ratio. It can be stated that this difference in case of the second group of features is stronger in its absolute value. Therefore the discriminative power of the second group of descriptors is stronger.

As it can be seen from (2) the decision on a noise change is based on a probability dependent threshold and on the size of measurement window. It can be seen from the table 1 that the variability of the gap between the "change" minimum and the closest minimum is rather strong. This makes us conclude, that the threshold should be adaptive and we dynamically train it on the first measured windows supposing the continuity of noise. With this assumption



a) b)

Figure 1. Normalized log likelihood ratio for groups of audio features: a) First group, b) Second proposed group

Transitions	Differences	Gr. 1	Gr. 2
i1−i2	Period	0.50	4.10
i1−i3	Peaks' Magnitude	0.32	0.40
i−p	Class	0.50	19.00
i−s	Class	0.30	4.20
i−c	Class	0.35	3.40
p−s	Class	1.17	7.80
p1−p2	Magnitude of ACR	10.50	18.65
p1−p3	Period	10.00	18.80
p−c	Class	0.52	19.35
s1−s2	Magnitude of sin	1.30	4.85
s1−s3	Number of sin	0.60	2.55
s−c	Class	0.55	4.50
c−c	Color	0.68	3.86

Table 1. Comparison of discriminative power of two groups of descriptors: (i− impulsive, p− pseudo-periodic, c− colored noise and s− sinusoidal)

and on the proposed second descriptor set, the semi-blind segmentation method performs well on a limited real test data set we have. With dynamic training of a threshold we obtain a recall figure of 86.67% and the same precision of 86.67% with regard to the ground truth on a real broadcast audio noise transitions using the second group of descriptors.

5. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a new set of features to characterize noisy sounds. We proposed a semi-blind segmentation of noise transition based on a classical Bayesian approach. We have also shown that noise-to-noise transition detection can be improved using relevant features. For that purpose, our comparison between classical and proposed features illustrates that the discriminant power of the statistical segmentation rule has been considerably increased using our proposed feature set both on synthetic sound and real sound corpus.

However, on the one hand, more tests on real audio data set have to be done to validate the robustness of our decision method and to consolidate our choice of feature set proposed. On the other hand, conscious on the classical chicken-and-egg problem, we are nevertheless interested in a more extensive study of features appropriated to the classes of noisy sounds described in this paper, for change detection.

Acknowledgments

This research was carried out in the context of the SCRIME [1] project which is funded by the DMDTS of the French Culture Ministry, the Aquitaine Regional Council, the General Council of the Gironde Department and IDDAC of the Gironde Department.
The Broadcast Corpus we used for experiments has been partially supplied in the framework of a previous bilateral research contract with Philips Research NL.

6. REFERENCES

[1] E. Kijak, G. Gravier, L. Oisel, and P.Gros, "Audiovisual integration for tennis broadcast structuring," *Proceedings of CBMI'03, Rennes, France*, 2003.

[2] L. Chaisorn, T.-S. Chua, C.-K. Koh, Y. Zhao, H. Xu, and H. Feng, "A two-level multi-model approach for story segmentation of large news video corpus," *TRECVID WORKSHOP 2003*, 2003.

[3] M. F. McKinney and J. Breebaart, "Features fro audio and music classification," *Proc. of ISMIR'03*.

[4] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of ICASSP'97*, Munich, Germany.

[5] P. Hanna, "Modélisation statistique de sons bruités: étude de la densité spectrale, analyse, transformation musicale et synthèse," Ph.D. dissertation, LaBRI, Université Bordeaux I, 2003.

[6] X. Serra and J. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[7] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *CUIDADO Project Report*.

[8] M. Goodwin, "Residual modeling in music analysis-synthesis," *Proc. of ICASSP'96, Atlanta, GA, USA*.

[9] E. Zwicker and H. Fastl, *Psychoacoustics: facts and models*. Springer, 1999.

[10] W. Yost, "Pitch of iterated rippled noise," *Journal of Acoustical Society of America*, vol. 100, no. 1, pp. 3329–3335, 1996.

[11] A. Zolnay, R. Schulter, and H. Ney, "Extraction methods of voicing feature for robust speech recognition," in *Proc. of EuroSpeech'03*, vol. 1, Geneva, Switzerland, pp. 497–500.

[12] W. Hartmann, *Signals, Sound, and Sensation*. Modern Acoustics and Signal Processing, 1997.

[1] Studio de Cr´eation et de Recherche en Informatique et Musique ´electroacoustique