CLASSIFICATION OF AUDIO SCENES USING NARROW-BAND AUTOCORRELATION FEATURES

Xavier Valero, Francesc Alías

Grup de Recerca en Tecnologies Mèdia
La Salle - Universitat Ramon Llull, Barcelona, Catalonia, Spain
email: xvalero@salle.url.edu, falias@salle.url.edu, web: www.salle.url.edu

ABSTRACT

Multiple single sound events of very different characteristics might coincide in a given space and time, thus composing complex audio scenes. In that context, defining signal features capable of effectively analyzing the holistic audio scenes is a challenging task. This paper introduces a set of features that consider the temporal, spectral and perceptual characteristics of the audio scene signals. Specifically, the features are obtained from the autocorrelation function of band-pass signals computed after applying a Mel filter bank. The so-called Narrow-Band Autocorrelation (NB-ACF) features are compared to state-of-the-art signal features on a corpus of 4 hours composed of 15 audio scenes. Regardless of the learning algorithm employed, the NB-ACF attains the highest averaged recognition rates: 2.3 % higher than Mel Frequency Cepstral Coefficients and 5.6 % higher than Discrete Wavelet Coefficients.

Index Terms— Audio classification, feature extraction, autocorrelation function, environmental sound recognition, narrow-band signal analysis.

1. INTRODUCTION

Audio scene recognition aims at automatically identifying scenes or environments taking the audio as the main information source. We are talking about an emerging technology that might be applied in several fields. For instance, it may be used to enhance the robustness of speech processing algorithms in adverse noise conditions. In this context, the algorithm could be dynamically adapted to the given noise conditions by identifying the surrounding acoustic environment [1].

Hearing aid devices can be also improved thanks to audio scene recognition technology. In that context, the technology would allow the automatic adaptation of the device to the characteristics of the recognized acoustic situation, i.e., volume and equalization filter variations in noisy environments, quiet environments, in presence of music, etc. [2].

In robotics, generally the visual data is employed to make the robot interact with the environment. However, in absence of light, the robot totally loses the input information source. In order to reduce the dependency with the visual data, the audio data may be considered, representing a complementary source of information that, in addition, requires a lower computational processing cost than the visual signals [3]. Closely related, in a multimedia domain, the identification of the acoustic context may allow the portable devices (e.g., mobile phones) changing their working settings without human intervention according to the surrounding environment [4], e.g., by turning off the volume in a library or switching to hands-free mode inside a car but not in similar situations such as inside a bus or a train.

It should be noted that audio scenes, unlike speech or music, are unstructured audio signals (i.e., they lack of semantics). In addition, they might be composed of multiple environmental sound sources coinciding in space and time. Thus, robust signal features are needed in order to take into account all the details of such complex audio signals. An interesting comparison of signal features for audio context recognition was performed in [4]. They tested up to 11 common signal features: Mel Frequency Cepstral Coefficients (MFCC), Sub-Band Energy Ratio, Linear Predictive Coefficients and other low-level parameters such as Spectral Centroid, Zero Crossing Rate (ZCR) or Short Time Energy (STE). Tested on a corpus composed of 24 different soundscapes, the MFCC in conjunction with a Gaussian Mixture Model obtained the best performance.

However, as stated in [3], MFCC (like other traditional spectral-based features) might fail in describing noise-like signals with strong temporal domain signatures, such as insects chirping or rain sound. In [5], the importance of the temporal aspects of the acoustic signal is also highlighted. Therefore, signal features that take into account both spectral and temporal information should be employed for addressing the problem at hand [5].

This paper proposes a signal parameterization that, besides combining the spectro-temporal information needed to effectively analyze the complex audio scenes, also takes into account the human perception of sound. The proposed

parameterization is based on the autocorrelation function analysis of a set of narrow band-signals, a technique which has previously been used for sound mixture segregation [6], [7]. Unlike previous works, rather than using the whole autocorrelation, we propose a parameterization of this function using a set of perceptually motivated features.

The rest of the paper is organized as follows. Section 2 introduces the theory of the proposed signal features. Section 3 describes the implementation details, together with the machine learning algorithms employed to perform the audio scene classification. Section 4 presents the experimental evaluation carried out and Section 5 shows and discusses the obtained results. Finally, Section 6 draws up the conclusions and future work lines.

2. NARROW-BAND AUTOCORRELATION SIGNAL FEATURES

The proposed audio signal analysis first modifies the signal spectrum by means of an A-weighting filter in order to model the spectral response the human auditory system. Then, a bank of band-pass filters decomposes the broad-band signal $x(t)$ into a set of N narrow-band signals $y_j(t)$ (1):

$$y_j(t) = DFT^{-1}[X(f)W_A(f)H_j(f)] \quad (1)$$

where $X(f)$ is the DFT of the broad-band signal, $W_A(f)$ is the A-weighting filter and $H_j(f)$ is the band-pass filter centered at the $f_{o,j}$ frequency. Specifically, the filter bank is composed of N triangular filters $H_j(f)$ that follow the Mel scale, which is a perceptual scale of pitches judged by human listeners [8]. A 1000 Hz tone, with a level 40dB above the listener's threshold is defined as having a pitch of 1000 Mels. Below 1000 Hz, the Mel scale is approximately linear, whereby above the 1000 Hz the Mel scale is non-linear and follows a logarithmic pattern. Then, the frequency used in the Mel band-pass filters f_{Mel} is given by [8]:

$$f_{Mel} = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (2)$$

Next, the autocorrelation function (ACF) of each narrow-band signal $y_j(t)$ is computed. In addition, we define the band-normalized ACF of the signal filtered in the jth band as z_j (3). The narrow-band ACF is normalized by the total energy of the weighted signal $x_A(t)$. This process allows removing the dependency of the narrow-band analysis with the energy of the signal frame.

$$z_j(\tau) = \frac{1}{E(t)\,T}\int_0^T y_j(t)\,y_j(t+\tau)\,dt$$

$$E(t) = \frac{1}{T}\int_0^T x_A(t)^2\,dt \quad (3)$$

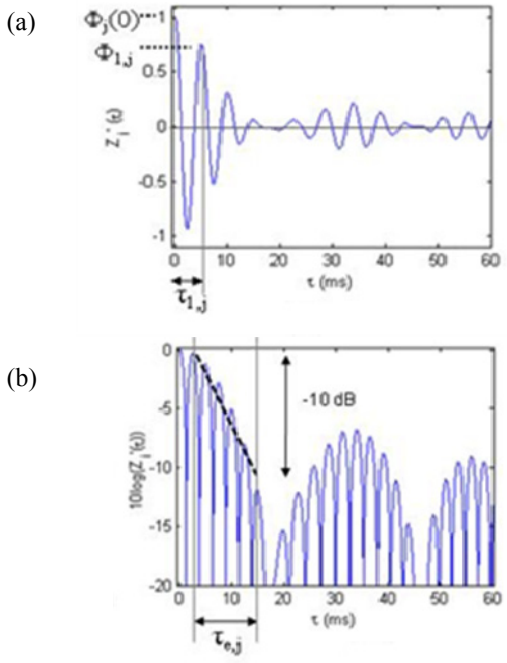

Fig. 1. Parameters extracted from the signal's Autocorrelation function. (a) Energy at the origin of the delay $\Phi_j(0)$, delay $\tau_{1,j}$ and amplitude $\Phi_{1,j}$: of the first ACF peak. (b) Effective envelope duration of the normalized ACF $\tau_{e,j}$.

where, hereafter, T is the interval in which the signal is integrated, τ represents the time delay and $x_A(t)$ is the weighted signal before applying the Mel filter bank.

At that point, the ACF is parameterized by extracting the four parameters proposed in [9] (see Fig.1). According to the author, the primary human auditory sensations (i.e., loudness, pitch, timbre and duration sensation) can be described by means of these four parameters. However, we have to redefine each parameter to cope with the narrow-band signal analysis context, as it is described in the following paragraphs.

- $\Phi_j(0)$: energy obtained from the ACF at the origin of the delay (4). It is related to the loudness or perceived sound pressure level of the sound signal at the jth band.

$$\Phi_j(0) = \frac{1}{T}\int_0^T y_j(t)^2\,dt \quad (4)$$

- $\tau_{1,j}$: delay of the first peak that can be found in the normalized ACF. This parameter is related to the dominant frequency of the narrow-band signal $y_j(t)$. It is computed as the delay of the largest $\Phi_j(\tau)$, starting from the first z_j zero crossing, denoted as T_K (5).

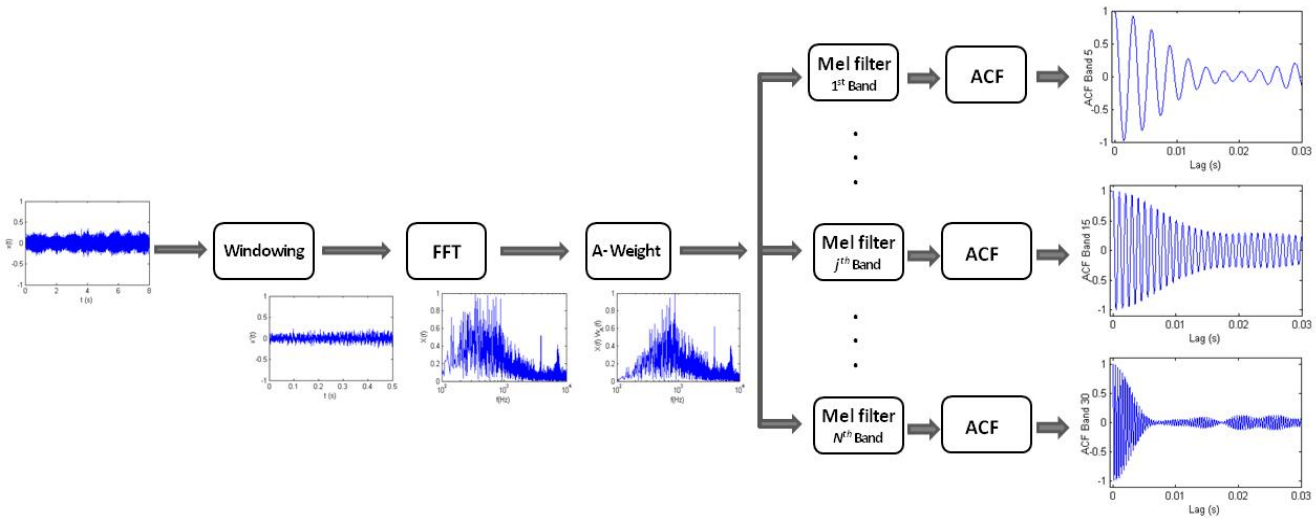

Fig. 2. Block diagram of the computation of the NB ACF features.

$$\tau_{1,j} = \arg\max_{\tau}\left\{ \frac{1}{E(t)(T-T_K)}\int_{T_K}^{T} y_j(t)\, y_j(t+\tau)\, dt \right\} \quad (5)$$

- $\Phi_{1,j}$: amplitude of the first peak (see (6)). It represents the strength of the dominant frequency $\tau_{1,j}$. High values of $\Phi_{1,j}$ are attributed to high-pitched signals, whereby low values of $\Phi_{1,j}$ are attributed to low-pitched signals.

$$\Phi_{1,j} = \max\left\{ \frac{1}{E(t)\,(T-T_K)}\int_{T_K}^{T} y_j(t)\, y_j(t+\tau)\, dt \right\} \quad (6)$$

- $\tau_{e,j}$: effective duration of the envelope of the normalized ACF. It is defined by the time that takes the $10\log(z_j(\tau))$ to decay 10 dB from its maximum value, and represents a repetitive feature within the sound signal. Thus, it is related to the reverberation of the signal $y_j(t)$. The algorithm to calculate $\tau_{e,j}$ performs a linear regression of the major peaks found in $10\log(z_j(\tau))$ and computes the equivalent decay time.

Finally, the parameters from the *N* narrow-band signals are merged into a unique signal feature vector (7).

$$\text{NB}-\text{ACF} = \{\Phi_1(0), \Phi_{1,1}, \tau_{1,1}, \tau_{e,1}, \ldots\ldots \Phi_N(0), \Phi_{1,N}, \tau_{1,N}, \tau_{e,N}\} \quad (7)$$

3. AUDIO SCENCE RECOGNITION SYSTEM

As a first step, the audio signal is framed employing a *hamming* window of 500 ms with 100 ms step, following the recommendation of [10] for the ACF signal analysis of environmental sounds. Note that the window used is larger than the typical length in spectral analysis [4], since the computation of ACF parameters requires a higher number of signal samples in order to achieve good resolution. Then, after applying an A-weighting filter, the framed signal is passed through a Mel filter bank that splits the signal into *N*=48 different narrow band signals. This value is obtained after adapting the Mel filter bank typically used in speech [11] to environmental sound analysis, whose bandwidth of interest is wider (in this work, it was set from 20Hz to 10 KHz). The ACF and the four parameters described in Section 2 are computed on each narrow band signal. Subsequently, Principal Component Analysis (PCA) is applied in order to reduce the dimensionality of the signal feature vector thus compacting the information [12].

Two different machine learning techniques have been selected to carry out the audio scene learning process: the K-Nearest Neighbor (KNN) and the Support Vector Machine (SVM). In this work, the SVM used a Gaussian radial basis function kernel, as in [13], and followed a traditional *one versus all* classification scheme. Both the number of neighbors from the KNN (*K=3*) and the sigma and C values from the SVM (1.3 and 1, respectively) were empirically selected so as to maximize the classification accuracy.

4. EXPERIMENTAL EVALUATION

4.1. Sound database

The corpus is composed of both self-recorded audio samples and audio samples extracted from a common sound library [14]. A total of 15 audio scenes from indoor and outdoor environments divided into five categories were considered (see Table 1). Each scene is represented by a set of minimum 150 samples and maximum of 300 samples, lasting 4s each one. In turn, each set was recorded in several locations (between 3 and 8, depending on the scene), so as

to guarantee certain data variability. The total corpus size is 3500 samples, which is equivalent to nearly 4 hours of audio data.

Category	Name	Samples
Outdoors-Natural	Seaside	251
	Countryside	150
Outdoors-City	Traffic	253
	Pedestrian	227
	Park	200
Indoor-leisure	Library	173
	Restaurant	194
	Stadium	296
Indoor-work environment	Classroom	200
	Office	288
	Factory	250
Indoor-means of transport	Station	198
	Inside car	300
	Inside bus	284
	Inside train	236
TOTAL		3500

Table 1. Audio scene corpus employed in the experiments.

4.2. Experimental setup

The proposed features are compared against other state-of-the-art signal parameterization techniques, covering time-domain (a combination of STE and ZCR), frequency-domain (MFCC) and time-frequency domain (Discrete Wavelet Coefficients (DWC)) signal features. The 4s audio sample was windowed with *hamming* windows of 30 ms long and an overlapping of 15 ms, as in [4]. In the case of MFCC, a vector of 13 coefficients (including the 0^{th} Cepstrum) was taken, whereby the DWC employed a "Daubechies" mother function with four vanishing moments [13]. The signal features (i.e. either MFCC, DWC or STE+ZCR) from all the 30ms frames contained in a 4s long sample are merged into a unique vector. Subsequently, PCA is applied by selecting the number of components that maximize the recognition performance of each signal feature (between 6 and 18, depending on the signal feature and on the machine learning technique)

The experiments consist in carrying out the classification of the corpus audio scene samples (one decision is taken for the whole 4s samples) following a 4-fold cross validation scheme [3]. Experiments are independently run considering the two machine learning techniques: KNN and SVM. Finally, a class-based analysis is carried out in order to analyze the classification performance achieved in every audio scene by the different signal features of the comparison.

5. RESULTS

5.1. General classification

The averaged classification rates attained when combining each signal feature with each machine learning technique are shown in Table 2. So far, the combination of STE and ZCR yield very poor accuracies, showing classification rates lower than 50% in combination with both classifiers. When employing the KNN algorithm, MFCC and DWC yield similar performances, whereas in combination with SVM, the MFCC attains an averaged classification rate 5.4 points higher than DWC. Nevertheless, both are outperformed by the proposed NB-ACF, which yields a classification rate of the 90% and a 91% when combined with the KNN and SVM, respectively.

Feature	KNN	SVM
NB-ACF	90.0 ± 0.9	91.0 ± 1.2
MFCC	87.4 ± 1.1	89.0 ± 1.3
DWC	86.2 ± 1.0	83.6 ± 1.1
STE+ZCR	47.5 ± 1.7	41.4 ± 3.0

Table 2. Mean and standard deviation (in %) of the classification rates attained by the combination of each signal feature with each machine learning algorithm.

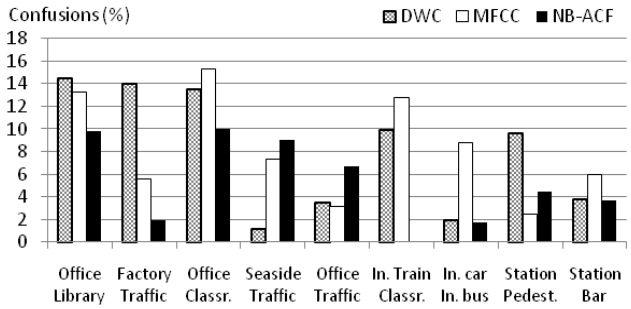

Fig. 3. Most frequent confusions obtained by the KNN classifier in combination with the DWC, MFCC and NB-ACF signal features.

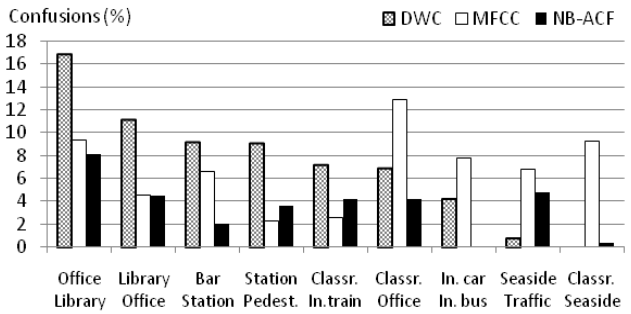

Fig. 4. Most frequent confusions obtained by the SVM classifier in combination with the DWC, MFCC and NB-ACF signal features.

5.2. Class-based analysis

The classification rate on each audio scene has been calculated in order to come up with a detailed comparison of the accuracy obtained when using DWC, MFCC and the proposed NB-ACF signal features (STE+ZCR are not included given the poor performances observed in Table 3). The comparison is based on the calculation of the confusion matrix when employing each signal feature in combination with KNN or SVM classifier. Due to space restrictions, the information of the confusion matrix is summarized by taking the three pairs of classes that showed the highest confusion rates for every signal feature.

As it can be observed from Fig 3, when employing the KNN classifier, the NB-ACF reduces significantly the confusion rates in the most critical cases (those that showed larger misclassifications when employing DWC and MFCC, i.e. *office-library* and *office-classroom*). Indeed, all the confusion rates yielded by NB-ACF remain below the 10%. It should also be noted the dramatic reduction of the confusions between *inside train* and *classroom*, and between *inside car* and *inside bus*.

Likewise, in combination with the SVM classifier, the highest misclassifications shown by NB-ACF are significantly reduced, remaining below the 8% (see Fig. 4). Specifically, the confusions between *classroom* and both *office* and *seaside* classes decreased nearly a 9% when employing NB-ACF instead of MFCC. Also the major misclassifications obtained when using DWC (i.e., *office-library*, *library-office* and *bar-station*) showed a significant reduction (6.5% to 9%) when using NB-ACF.

The superiority of NB-ACF in front of both DWC and MFCC turns out to be especially important in audio scenes such as *office*, *library* or *classroom*. These audio scenes usually present a high randomness on single sound events occurrence (e.g., people talking, door closing, phone ringing, etc.) and they are more difficult to recognize, as already noticed in previous works [12].

6. CONCLUSIONS

This paper has introduced a set of signal features for the classification of audio scenes that take into account their temporal, spectral and perceptual characteristics by analyzing the autocorrelation function of narrow-band signals (NB-ACF). In the experiments the NB-ACF descriptors have been compared to representative time domain (STE+ZCR), frequency domain (MFCC) and time-frequency domain (DWC) signal features. The classification rates attained by the STE+ZCR have shown that time-domain characteristics of the audio signals are not sufficient by themselves for conducting audio scene classification, whereby the spectral-domain characteristics (in this work, MFCC) attained a good performance. However, among all the tested signal features, the proposed NB-ACF, which is a spectro-temporal feature with perceptual basis, yielded the best classification rates for the corpus at hand.

The detailed analysis of the confusion matrices shows a reduction of misclassifications when employing NB-ACF, effect that is particularly interesting on those audio scenes where changes of sound events over time are more pronounced, such as *office, library* or *classroom*. Our future work will be focused on extending the corpus and testing the proposed features in different locations, as well as adapting the technique to real time applications.

7. REFERENCES

[1] F. Beritelli, R. Gasso, "A pattern recognition system for environmental sound classification based on MFCC and Neural Networks", in *Proc. IEEE International Congress on Signal Proc. And Communication Systems*, 2008.

[2] M. Büchler, S. Allegro, S. Launer, N. Dillier, "Sound classification in hearing aids Inspired by Auditory Scene Analysis", in *EURASIP Journal on Applied Signal Processing,* vol. 2005, January 2005.

[3] S. Chu, S. Narayanan, C.-C. Jay Kuo, "Environmental sound recognition with time-frequency audio features", in *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, no. 6, pp. 1142-1158, August 2009.

[4] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," in *IEEE Trans. Audio, Speech, Lang. Process*., vol. 14, no. 1, pp. 321–329, Jan. 2006.

[5] M. U. B. Altaf, B-H. Juang, "Audio signal classification with temporal envelopes", in *Proc. ICASSP,* 2011.

[6] M. Slaney, "The history and future of CASA", Chap. 13 of "Speech separation by humans and machines", Pierre Divenyi (Ed.), Kluwer Academic Publishers, pp. 199-211, 2005.

[7] R.O. Duda, R.F. Lyon, M. Slaney, "Correlograms and the Separation of Sounds", in *Proc. Asilomar Conf. On Signals, Systems and Computers*, 1990.

[8] S. Stevens, J. Volkmann, E. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch", in *Journal of the Acoustical Society of America* vol.8, pp.185–190, 1937.

[9] Y. Ando, "A theory of primary sensations and spatial sensations measuring environmental noise", in *Journal of Sound and Vibration*, vol. 241, no. 1, pp. 3-18, 2001.

[10] K. Fujii, Y. Soeta, Y. Ando, "Acoustical properties of aircraft noise measured by temporal and spatial factors", in *Journal of Sound and Vibration*, vol. 241, no. 1, pp. 69-78, 2001.

[11] M. Slaney, "Auditory Toolbox. Version 2", Technical Report #1998-010, Interval Research Corporation, 1998.

[12] X. Valero, P. Farré, F. Alías, "Comparison of machine learning techniques for automatic soundscape recognition", in *Proc. Forum Acusticum*, 2011.

[13] A. Rabaoui, M. Davy, S. Rossignaol, N. Ellouze, "Using one-class SVMs and Wavelets for audio surveillance", in *IEEE Trans. Information Forensics and Security,* vol. 3, no. 4, pp. 763-775, December 2008.

[14] "The Freesound Project," [Online]. Available: http://www.freesound.org/.