Feature Analysis for Audio Classification

Gaston Bengolea[1], Daniel Acevedo[1], Martín Rais[2,⋆], and Marta Mejail[1]

[1] Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires
gastonbengolea@gmail.com,
{dacevedo,marta}@dc.uba.ar
[2] Dpt. Matemàtiques i Informàtica / CMLA
Universitat de les Illes Balears / ENS Cachan
Spain, France
martin.rais@cmla-ens.cachan.fr

Abstract. In this work we analyze and implement several audio features. We emphasize our analysis on the ZCR feature and propose a modification making it more robust when signals are near zero. They are all used to discriminate the following audio classes: music, speech, environmental sound. An SVM classifier is used as a classification tool, which has proven to be efficient for audio classification. By means of a selection heuristic we draw conclusions of how they may be combined for fast classification.

1 Introduction

The analysis of audio features is an important task when an automatic audio classifier is being developed. In this work we aim at classifying audio signals according to a predefined audio category. This corresponds to the audio content analysis (ACA) field of study. The objective of ACA is the extraction of information from audio signals such as music recordings or any type of specific audio type that is stored on digital media. The information to be extracted is expected to allow a meaningful description or explanation of the raw audio data, which will lead to a more convenient processing. This processing may include automatic organization (tagging) of audio content in large databases as well as search and retrieve audio files with specific characteristics in such databases. Also, this processing may conduct to a more specialized task for a specific type of audio. For instance, in case of music recordings, applications range from tempo and key analysis -ultimately leading to the complete transcription of recordings into a score-like format- over the analysis of artists' performances of specific pieces of music [1], to transcribing news only segments [3], detecting commercials in TV broadcast program [4], transcribing lecture presentations [8], etc.

A common taxonomy of audio classes generally considers speech, music and environmental sound, although some other works include a mix of these classes

⋆ During this work, Martin Rais had a fellowship of the Ministerio de Economia y Competividad (Spain), reference BES-2012-057113, for the realization of his Ph.D.

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 239–246, 2014.

or other subclasses. For example, Chen et al. departed the audio data into five types: music, speech, environmental sound, speech with music background, and environmental sound with music background [2]; Zhang parsed audio data into silence, speech, harmonic environmental sound, music, song, speech with music background (speech+music), environmental sound with music background, non-harmonic sound, etc. [12]. Once each of these classes are established in the audio signal, several other applications arise. For instance, in case of speech, the speech activity detection (SAD) has applications in a variety of contexts such as speech coding, automatic speech recognition (ASR), speaker and language identification, and speech enhancement.

Audio classification is generally based on features estimated over short time audio samples, followed by a state-of-the-art classifier. Each of these features represent some particular characteristic which make them more suitable to detect certain types of audio that are present in the audio clip. A well-known feature called Zero Crossing Rate (ZCR) gives a rough estimate of the spectral properties of audio signal and it is related with its noisiness; generally, voiced audio clips have much smaller ZCR that unvoiced clips making it suitable for speech discrimination. In this work, we analyze the ZCR and propose a modification making it more robust when signals are near zero. One of the firsts approaches by Sanunders used this feature and the short time energy to classify radio program into speech and music [10]. Other work by Panagiotakis used only energy and frequency features to discriminate these two classes [7].

There are several more audio features to consider. In this work we analyse High Zero Crossing Rate Ratio, Spectral Flux, Low Short-Time Energy Ratio, Noise Frame Ratio and Band Periodicity audio features. We use them to discriminate the following predefined audio classes: music, speech, environmental sound. An SVM classifier is used as a classification tool, which has proven to be efficient for audio classification [5]. By means of a selection heuristic we draw conclusions of how they may be combined for fast classification.

2 Audio Features

In order to compute the features, we have an audio clip x which has been chopped into N consecutive frames per second, having each frame L samples (see Fig. 1). We will refer to x_n to the n-th frame and $x_n(l)$ to the l-th sample within the n-th frame, for $0 \leq n \leq N-1$ and $0 \leq l \leq L-1$. For audio classification, based on the work of [6], the input signal is downsampled to 8000Hz (samples per second), and $N = 40$ frames per second having a total of $L = 200$ samples per frame. Then, for each second of the audio, several features have been implemented and evaluated and a support vector machine classifier for each type is employed to detect if the second has content related to the type.

High Zero Crossing Rate Ratio (HZCRR). HZCRR is defined as the ratio of the number of frames whose Zero Crossing Rate (ZCR) are above 1.5-fold average zero-crossing rate in an 1-second window [6]. The ZCR is defined as the

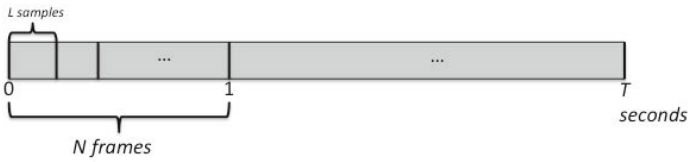


Fig. 1. Sketch of a T-seconds signal x partitioned into N frames per second and L samples per frame

ratio of the number of times a signal crosses the x-axis and is an approximate measure of noisiness and has proven to be a discriminative feature for audio signals.

$$ZCR(x_n) = \frac{1}{2L}\sum_{l=1}^{L-1} |sgn(x_n(l)) - sgn(x_n(l-1))| \tag{1}$$

where

$$sgn(x) = \begin{cases} -1 & \textbf{if } x < 0 \\ 0 & \textbf{if } x = 0 \\ 1 & \textbf{if } x > 0 \end{cases} \tag{2}$$

After evaluating this feature, we detected that for some audios, the zero crossing rates were unreasonably high. This was because the signal oscillated when close to zero. To fix this, a thresholded version of the ZCR, the TZCR feature is proposed. The idea is to divide the space in three distinct non-overlapping areas: the zero area, delimited by $[-t, t]$, the positive values higher than the threshold t, and the negative values lower than $-t$. The TZCR feature is then defined by

$$TZCR(x_n) = \frac{1}{2L}\sum_{l=1}^{L-1} TZC(x_n(l)) \tag{3}$$

where

$$TZC(x_n(l)) = \begin{cases} |sgn(x_n(l)) - sgn(x_n(l-1))| & \textbf{if } |x_n(l)| > t \textbf{ and } |x_n(l-1)| > t \\ 1 & \textbf{if } |x_n(l)| > t \textbf{ and } |x_n(l-1)| \leq t \\ 1 & \textbf{if } |x_n(l)| \leq t \textbf{ and } |x_n(l-1)| > t \\ 0 & \textbf{otherwise} \end{cases} \tag{4}$$

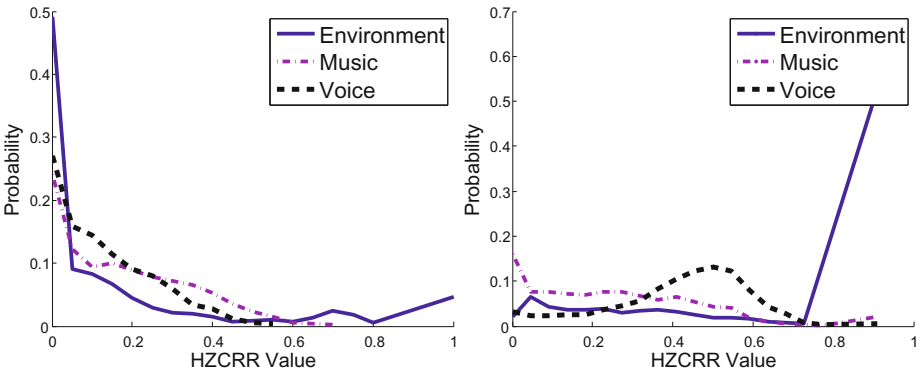
As in the original Zero Crossing metric, when the discrete function x_n goes from negative to positive, it accounts for 2 ZC, and when a consecutive pair of values goes to zero coming from something different, it accounts for 1 ZC. Our thresholded version keeps the same definition, however the 'zero' is now a region that covers the range $[-t, t]$. Finally, the HZCRR feature becomes

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1} sgn(TZCR(x_n) - 1.5 \cdot \overline{TZCR}) + 1 \tag{5}$$

where

$$\overline{TZCR} = \frac{1}{N} \sum_{n=0}^{N-1} TZCR(x_n) \tag{6}$$

Fig. 2 shows the histograms of the values for this feature, the first using the original ZCR and the second using the proposed TZCR. Under the original formulation, it is easily perceptible how the discrimination of the audio types is not clear and the three curves look similar. This does not occur under the proposed formulation where if the HZCRR value is between 0 and 0.25, the analyzed second is probably music, if the value lies between 0.4 and 0.7 there is a high probability the input signal is voice, and values higher than 0.75, we are clearly dealing with an environmental sound. In the non-defined intervals, this new feature may not be discriminative enough and other features have to be used.



(a) Using ZCR (eq. 1) (b) Using TZCR (eq. 3)

Fig. 2. Comparison of histograms of HZCRR values for three different audio classes: music, voice and environment

Spectral Flux (SF). The spectral flux [9] measures the spectrum fluctuations between two consecutive audio frames. It is defined as

$$SF_n(x) = \sum_{k=1}^{L-1} |X_n(k) - X_{n-1}(k)| \tag{7}$$

where X_n is the Discrete Fourier Transform of the n-th audio frame x_n. The Spectral Flux SF feature estimated in a 1-second window is defined as the average of the SF_n's:

$$SF = \frac{1}{N-1} \sum_{n=1}^{N-1} SF_n(x) \tag{8}$$

Low Short-Time Energy Ratio (LSTER). LSTER is defined as the ratio of the number of frames whose short-time energy are less than 0.5 time of average short-time energy in a 1-sec window.

$$LSTER = \frac{1}{2N} \sum_{n=0}^{N-1} sgn\left(\frac{\overline{STE}}{2} - STE(x_n)\right) + 1 \tag{9}$$

where

$$STE(x_n) = \frac{1}{L} \sum_{l=0}^{L-1} x_n^2(l), \qquad \overline{STE} = \frac{1}{N} \sum_{n=0}^{N-1} STE(x_n) \tag{10}$$

Noise Frame Ratio (NFR). Let x_n be a frame, $0 \leq n \leq N-1$, and let

$$\hat{A}_n(m) = \frac{A_n(m)}{A_n(0)} = \frac{\sum_{l=0}^{L-1-m} x_n(l)x_n(l+m)}{\sum_{l=0}^{L-1} x_n^2(l)} \tag{11}$$

be the normalised autocorrelation sequence of the frame x_n. We consider this frame x_n is a noise frame NF_n if $\max_m \hat{A}_n(m) < Th$. Finally, we define the Noise Frame Ratio

$$NFR = \frac{\#NF_n}{N} \tag{12}$$

Band Periodicity (BP). We define a subband x^{band} as the audio sequence containing the frequency range $[F_1, F_2]$ of the frequencies in x. In this work we considered four subbands in the following ranges: $[500, 1000]$ Hz, $[1000, 2000]$ Hz, $[2000, 3000]$ Hz, and $[3000, 4000]$ Hz. The periodicity property of x^{band} is derived by subband correlation analysis and is represented by the maximum local peak of the normalized correlation function. The normalized correlation function $r_{band,n}$ for the n-th frame is calculated as

$$r_{band,n}(k) = \frac{\sum_{l=0}^{L-1} x_n^{band}(l-k)\, x_n^{band}(l)}{\sqrt{\sum_{l=0}^{L-1} (x_n^{band}(l-k))^2}\sqrt{\sum_{l=0}^{L-1} (x_n^{band}(l))^2}}, \qquad k = 0, \ldots, L-1$$

where $x_n^{band}(l)$ refers to values from the current frame when $l \geq 0$; if $l \leq -1$ then we refer to values in the previous frame $x_{n-1}^{band}(l)$. Then, the band periodicity in a 1-second window for each subband is estimated as

$$BP_{band} = \frac{1}{N} \sum_{n=0}^{N-1} r_{band,n}(k_p)$$

where k_p is the index of the maximum local peak: $k_p = \arg\max_k r_{band,n}(k)$.

3 Classification and Results

A training set consisting of around 86 minutes (206640 frames), formed by 1714 seconds of music, 1736 seconds of environment, and 1716 seconds of voice was manually labeled. For each audio type (music, voice and environment), a separate labeling of the training set was performed indicating if there was presence of that audio type (a binary decision) on every 1-second segment. Then, once features were calculated for each 1-second audio segment, they are grouped together and used to train three Support Vector Machine classifiers [11]. We used the libSVM library[1] and a radial basis function as the kernel. To optimize classification, a 5-fold cross-validation procedure is performed varying the cost parameter C and the γ parameter of the radial kernel. Note that for the development of the results, when the BP feature is mentioned, it means that all four subbands (features BP_1, BP_2, BP_3 and BP_4) are used.

The test set used is formed by 550 frames of voice, 583 frames of music and, 630 of environment sound; precision, recall and accuracy metrics have been used to evaluate the algorithm.

In table 1, results for each SVM are shown. It should be noted that using a single SVM to separate between classes obtains excellent results. By using a multi-SVM schema, the possible outcomes increase dramatically. However, it is remarkable how even after using multiple classifiers to detect the audio classes, the proposed method is able to achieve results over 85%, which allows to successfully perform multi-class classification. We have performed an analysis considering all the combinations of features, having $\binom{5}{k}$ combinations, for $k = 1, \ldots, 5$. Each of these combinations was used to train and test the SVM classifier (obtaining a confusion matrix for each test and the corresponding rates).

Table 1. Precision, recall and accuracy rates

	Precision	Recall	Accuracy
Voice	0.8935	0.8606	0.9114
Music	0.9200	0.8470	0.9103
Environment	0.9838	0.9560	0.9787

In Table 2 we summarize these results, showing the best selection of features with respect to precision, recall and accuracy. This gives us an idea of which are the most discriminative features for each audio class (voice, music and environment) and sets an heuristic for selecting features that is depicted in the following paragraph. The last column shows the results using all the features.

We observe that, when using two features, HZCRR and BP achieve an accuracy rate near 90% and all the metrics are high for classes voice and environment. These two features are present for all the best selections of k features, for $k \geq 2$. In the case of the environment class, adding any number of features to these two

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

achieves a negligible increase for both precision and recall. Then, this suggests that HZCRR and BP are sufficient for classifying a test frame as environment or voice class. Since the computation of features is the most time-consuming task, we consider that adding both the SF and the LSTER features improves music classification (i.e., reducing the false positive rate for these classes). We note also that adding the NFR feature does not increase performance significatively, and thus, its usage is not recommended.

Table 2. Each column shows the best selection of features with respect to precision, recall and accuracy for each class

	1 feature	2 features	3 features	4 features	5 features
VOICE	[BP] Precision: 0.991 Recall: 0.952 Accuracy: 0.944	[HZCRR,BP] Precision: 0.991 Recall: 0.981 Accuracy: 0.973	[HZCRR, SF, BP] Precision: 0.993 [HZCRR, LSTER,BP] Recall: 0.986 Accuracy: 0.978	[LSTER,HZCRR, SF,BP] Precision: 0.995 Recall: 0.986 Accuracy: 0.981	[NFR, LSTER,HZCRR, SF,BP] Precision: 0.994 Recall: 0.987 Accuracy: 0.981
MUSIC	[BP] Precision: 0.872 [SF] Recall: 0.781 [BP] Accuracy: 0.747	[HZCRR,BP] Precision: 0.951 Recall: 0.915 Accuracy: 0.874	[HZCRR, SF,BP] Precision: 0.959 [LSTER, SF, BP] Recall: 0.933 [HZCRR, SF,BP] Accuracy: 0.895	[HZCRR,LSTER,SF,BP] Precision: 0.968 [HZCRR,SF,BP,NFR] Recall: 0.939 [HZCRR,LSTER,SF,BP] Accuracy: 0.910	[NFR, LSTER,HZCRR, SF,BP] Precision: 0.937 Recall: 0.888 Accuracy: 0.910
ENVIRONMENT	[BP] Precision: 0.943 Recall: 0.881 Accuracy: 0.944	[HZCRR, BP] Precision: 0.991 Recall: 0.981 Accuracy: 0.973	[HZCRR, BP, SF] Precision:0.993 [HZCRR, LSTER, BP], Recall: 0.986 Accuracy: 0.978	[LSTER,HZCRR,BP, SF] Precision: 0.995 Recall: 0.986 Accuracy: 0.981	[NFR, LSTER,HZCRR, SF,BP] Precision: 0.994 Recall: 0.987 Accuracy: 0.981

3.1 TZCR Results

In order to evaluate the proposed HZCRR feature using TZCR values, an SVM was trained using only the HZCRR feature and evaluated for each audio type. The threshold was empirically set to 0.1 which offered the best results. Table 3 shows the improvement over the original formulation. The F-Measure metric (also known as F_1 score) is defined as $F_1 = 2 \cdot (precision \cdot recall)/(precision + recall)$ and can be interpreted as a weighted average between the precision and the recall.

Table 3. Evaluation of both variants of the HZCRR feature for all audio types

	ZCR Voice	TZCR Voice	ZCR Music	TZCR Music	ZCR Env.	TZCR Env.
Recall	63.39 %	**82.22 %**	72.12 %	**85.12 %**	67.73 %	**82.57 %**
Precision	**100 %**	86.01 %	**88.56 %**	84.24 %	**99.59 %**	87.44 %
Accuracy	63.39 %	**77.91 %**	65.98 %	**73.43 %**	67.54 %	**81.50 %**
F-Measure	77.59 %	**87.58 %**	79.50 %	**84.68 %**	80.62 %	**89.81 %**

4 Conclusions and Future Work

In this work we have analysed several audio features for classification of audio clips according to predefined classes. We have emphasized our analysis on the ZCR feature detecting that by using the original definition, it yielded high values when not expected. For that, we have introduced a modification making it more robust as the signal approaches to zero. In future work, we plan to apply this improved feature in the wavelet domain. At each step of the wavelet transform, an approximation and details of the original signal are computed. After several steps, an approximation at different resolution levels may be obtained and we expect to achieve better classification rates estimating the HZCRR to these approximation coefficients.

The analysis performed in this paper has allowed us to infer an heuristic for selection of the best features that are more suitable for classification of specific types of audio. This heuristic saves computational times since not all of the features are necessary to estimate.

References

1. Chai, W.: Semantic segmentation and summarization of music: methods based on tonality and recurrent structure. IEEE Signal Proc. Mag. 23(2), 124–132 (2006)
2. Chen, S.L., Gunduz, Ozsu, M.T.: Mixed type audio classification with support vector machine. In: IEEE International Conference on Multimedia and Expo, pp. 781–784 (July 2006)
3. Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C.: Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech and Audio Processing 12(4), 401–408 (2004)
4. Johnson, S.E., Woodland, P.C.: A method for direct audio search with applications to indexing and retrieval. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2000, vol. 3, pp. 1427–1430 (2000)
5. Z., S., Lu, H.-J.Z.L., Li: Content-based audio segmentation using support vector machines. In: IEEE International Conference on Multimedia and Expo, ICME 2001, pp. 749–752 (August 2001)
6. Lu, L., Zhang, H.-J., Jiang, H.: Content analysis for audio classification and segmentation. IEEE Trans. on Speech and Audio Processing 10(7), 504–516 (2002)
7. Panagiotakis, C., Tziritas, G.: A speech/music discriminator based on rms and zero-crossings. IEEE Transactions on Multimedia 7(1), 155–166 (2005)
8. Park, A., Hazen, T.J., Glass, J.R.: Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In: IEEE Int'l Conf. on Acoustics, Speech, and Signal Proc. (2005)
9. Sadjadi, S., Hansen, J.: Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Proc. Letters 20(3), 197–200 (2013)
10. Saunders, J.: Real-time discrimination of broadcast speech/music. In: IEEE Int'l Conf. on Acoustics, Speech, and Signal Proc., vol. 2, pp. 993–996 (1996)
11. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc, New York (1995)
12. Zhang, C.-C.J.T., Kuo: Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on Speech and Audio Processing 9(4), 441–457 (2001)