Radon-based Audio Classification Features

Author
Gonzalez, Ruben

Published
2012

Conference Title
IEEE International Conference on Multimedia & Expo (ICME2012)

DOI
https://doi.org/10.1109/ICME.2012.155

Copyright Statement
Copyright 2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Downloaded from
http://hdl.handle.net/10072/50681
Radon-based Audio Classification Features

Ruben Gonzalez
Institute for Integrated and Intelligent Systems
Griffith University, Gold Coast Campus, QLD, Australia,

Abstract—This paper presents novel features for audio classification based on the Radon transform. The performance of these features are evaluated against the widely accepted MFCC based feature sets for a wide range of audio data sets.

Keywords: Audio classification, content-based retrieval, indexing, spectral features, MFCC, Radon Transform, machine learning, k-NN classification, musical instruments, frog calls, insect sounds, speech and music discrimination.

I. INTRODUCTION

Content-based retrieval is fundamentally a two-step task; salient features are first extracted from the data, which are then used for classification via machine learning approaches. Effective audio content-based retrieval requires a robust feature set that can capture salient information in audio signals across a wide variety of audio classes. It also requires a robust method for classifying the audio based on the selected feature set. The most common classification methods used for this audio class recognition include Gaussian Mixture Models (GMM), K-Nearest Neighbour (k-NN), Neural Networks (NN), support vector machines (SVM), and Hidden Markov Models (HMM).

The choice of classification method has been shown to be largely insignificant. Arias [1] compared GMM and SVM to classify four audio classes (speech, music, applause, laughter) using features consisting of MFCC features plus energy and their derivatives and found that performance was relatively comparable. Chu et al., [2] investigated the problem of determining between five different classes of environmental sounds; Hallway, Café, Lobby, Elevator and Sidewalk using k-NN, GMM and SV classifiers. While all three classifiers performed within 3% of each other it was observed that, “the KNN classifier works well overall, outperforming GMM and is roughly 1000 times faster than SVM.” Peltonen [6] also found that overall k-NN performed better than GMM. Liu and Wan [7] found k-NN to outperform GMM in all cases. Lefèvre [3] also observed, “that the k-NN estimator outperforms the GMM estimator in identification tasks.”

For reasons of efficiency and effectiveness, rather than operating directly on raw audio data, classification methods operate on an abstraction of the audio data expressed as a small feature set. The size of this feature set is known as its dimensionality. The objective in selecting these features is to isolate properties of the underlying audio data that have statistical or psychological salience. Sometimes, as in the case of the MFCC, they also attempt to mimic physiological processes. The disadvantage of selecting any particular set of features is that whatever information is not encapsulated by them is automatically discarded. Hence, while choice of classification method is relatively unimportant this is not the case with the choice of features for audio classification.

A wide variety of audio features, extracted from signals in either the temporal or frequency domains have been presented in the literature. Of these, the Mel-Frequency Cepstral features (MFCC), which are frequency transformed and logarithmically scaled, appear to be universally recognised as the most generally effective. The MFCC has been shown to outperform the MPEG7 features [4]. McKinney and Breebaart [5] evaluated four different feature sets comprising of the highest ranked individual features obtained from one of four methods including; (i) spectral and temporal features; (ii) MFCC, (iii) psychoacoustic features (roughness, loudness, sharpness, etc) and (iv) features derived from the temporal envelopes from an auditory filterbank. An iterative ranking process ensured that each feature set was evaluated for its optimum performance. While they found that for music genre recognition the temporal envelope features performed the best, for other classes of audio MFCC and spectral features performed better.

Peltonen [6] individually evaluated eleven types of time domain and frequency domain features including MFCC, band-energy ratios, LP-cepstra and LPC for a total of 26 different acoustic scenes using both kNN and GMM classifiers. He found that overall MFCC features outperformed the other features. Liu and Wan [7] evaluated a total of fifty-eight different temporal and spectral features to classify twenty-three different musical instruments into five different classes (brass, keyboard, percussion, string and woodwind) using both k-NN and GMM. They found that temporal features alone gave the lowest performance, followed by spectral features alone and then MFCC features alone. The best performance was achieved when temporal, spectral and MFCC features were combined, followed closely by spectral and MFCC features combined.

Both spectral and MFCC features are extracted from the Short Time Fourier Transform (STFT) that only captures frequency information from the signal at particular instances in time. Due to the comparatively short analysis windows used, the STFT on its own is sensitive to noise and unable to capture transient data in the signal. These problems may be overcome by training the classifier using a large number of feature vectors extracting from the data, increasing training time and memory resource requirements. This paper proposes an alternate approach based on forming feature vectors that summarise the time-frequency distribution of the energy in the signal. The proposed feature set, based on the Radon transform is shown...
to provide superior performance compared to MFCC based feature sets across a range of diverse datasets.

II. SPECTRAL-RADON TRANSFORM

The Radon transform, \( R(\rho, \theta) \) of a continuous 2D function \( g(x, y) \) in the Euclidean plane \( \mathbb{R}^2 \) is a function defined on the space of straight lines \( L \) by the line integral along each such line. Let each said line be represented by the parameters \((\rho, \theta)\) where \( \rho \) is the perpendicular distance of the line from a given point in the plane and \( \theta \) is the orientation of the line, as depicted in Figure 1.

These lines can be described parametrically in terms of \( s \) extending from \(-\infty \) to \(+\infty \) as:

\[
\begin{align*}
x(s) &= s \sin \theta + \rho \cos \alpha \\
y(s) &= s \cos \theta + \rho \sin \alpha
\end{align*}
\]

Since \( \alpha \) and \( \theta \) are related \((\alpha = 90 - \theta)\) these equations may be rewritten as:

\[
\begin{align*}
x(s) &= s \sin \theta + \rho \sin \theta \\
y(s) &= s \cos \theta + \rho \cos \theta
\end{align*}
\]

The integral along each line can now be defined as:

\[
R(\rho, \tau) = \int_{-\infty}^{\infty} g(x(s), y(s)) \, ds
\]

On substitution of (1) into (2) this becomes:

\[
R(\rho, \theta) = \int_{-\infty}^{\infty} g(s \sin \theta + \rho \sin \theta, s \cos \theta + \rho \cos \theta) \, ds
\]

The Radon transform is applied to the time-frequency energy distribution of the audio signal. This is calculated as a simple spectrogram \( X(\tau, \omega) \) defined as the contiguous set of the magnitudes of STFTs of the signal:

\[
X(\tau, \omega) = |\mathcal{F}\{z(t)w(t-\tau)\}|^2
\]

Where \( z(t) \) is the input signal, \( w(t) \) is a window function, and \( \tau \) is the time index. Letting \( \tau = x \) and \( \omega = y \), the Spectral-Radon Transform (SRT) may be defined as:

\[
SRT(\rho, \theta) = R[X(\tau, \omega)]
\]

This SRT representation has interesting properties. Frequency partials in the spectrum become individual traces that have three main characteristics. The vertical location of the tip of each trace is related to the average frequency of the partial. The minimum thickness of a trace is a function of the amount of frequency modulation or noise in the partial. The maximum thickness of traces elsewhere indicates the length of partials. The location of the minimum thickness along a trace indicates the spectral gradient of the partial. Zero gradients result in the minima being at the tips of a trace. Increasingly positive slopes have focus approaching the centre of the trace from the left hand side while increasingly negative slopes the focus approaches the centre from the right hand side. These characteristics are depicted in Figures 3 – 5 for various combinations of frequencies, noise and spectral gradients. On the left of each image is the spectrogram from which the SRT traces are obtained. In practice most natural sounds consist of numerous partials with different characteristics.
III. AUDIO FEATURE EXTRACTION

The performance of two Radon transform derived feature sets was evaluated. The first set (SRT1) consisted of the principle components of the SRT. Commonly either principle component analysis (PCA) or its equivalent Karhunen-Loeve (KL) transform can be used to obtain these components. In practice the KL transform is often approximated by means of the Discrete Cosine Transform (DCT). In this case a 2D-DCT is used as per (7) and the first few coefficients used to form the feature vectors.

\[ D2(i,j) = \sum_{\theta} \sum_{\rho} SRT(\rho, \theta) \cos(\omega \cdot i) \cos(\rho \cdot j) \]  

(7)

Instead of attempting to capture the principle components of the entire SRT surface, the second feature set (SRT2) focused on capturing the principle frequency and temporal characteristics of the traces. This consisted of the principle components of two vectors formed by first projecting each row of the SRT onto the \( \rho \) axis and each column of the SRT onto the \( \theta \) axis as follows:

\[ f_1(\rho) = \sum_{\theta} SRT(\rho, \theta) \cos(\omega \cdot i) \]  
\[ f_2(\theta) = \sum_{\rho} SRT(\rho, \theta) \cos(\rho \cdot j) \]  

(7)

The principle components were obtained from the 1D-DCT of each vector as in (8) for \( f_1(\rho) \) and similarly for \( f_2(\theta) \):

\[ DCT(i) = \sum_{\rho} f_1(\rho) \cos(\omega \cdot i) \]  
\[ DCT(\theta) = \sum_{\theta} f_2(\theta) \cos(\rho \cdot j) \]  

(8)

These two Radon derived feature sets were compared against three others. The baseline feature set is comprised of only MFCC features. An enhanced MFCC (MFCC+) feature set adds spectral and temporal features to the baseline features. Both of these other feature sets rely on short time Fourier transforms, and since the SRT based features are derived from a time-frequency distribution of the signal a third feature set was investigated to isolate the performance contribution of the SRT from the effects of the time distribution of the spectra. This other feature set consists of the principle components of the spectrogram. Accordingly the Principle Spectral-Temporal Coefficients (PSTC) feature set obtained by taking the first few coefficients of a 2D-DCT of the spectrogram provides a statistically optimum representation of the spectrum for any given number of coefficients.

\[ PSTC(i,j) = \sum_{\theta} \sum_{\rho} X(t, \omega) \cos(\omega \cdot i) \cos(\rho \cdot j) \]  

(9)

A. MFCC Features

Mel-frequency cepstral coefficients are calculated from the short-term Fourier Transform as the cepstrum of the mel-warped spectrum. The frequencies of the Fourier coefficients are remapped onto the mel scale using relationship (10) and octave-wide, triangular overlapping windows. Finally the cepstrum (11) is obtained from the remapped coefficients \( m(n) \).

\[ Mel(f) = 2595 \log_{10}(1 + f/100) \]  
\[ c(k) = DCT \left\{ \log(\text{DFT}(m(n))) \right\} \]  

(10)  
(11)

The enhanced MFCC+ dataset includes four common temporal features and six spectral features. The four temporal features are the Zero Crossing Rate (ZCR), the root-mean-square (RMS) value, short-term energy (E), and energy flux (F). These are defined as follows:

\[ ZCR = \sum_{k=2}^{K} |\text{sgn}(x(k)) - \text{sgn}(x(k-1))| \]
\[ \text{sgn}(n) = \begin{cases} 1, & n > 0 \\ 0, & n = 0 \\ -1, & n < 0 \end{cases} \]
\[ E = \frac{1}{K} \sum_{k=1}^{K} |X(k)|^2 \]
\[ F = E(n) - E(n-1) \]

The six spectral features used are the signal bandwidth (BW), spectral centroid (SC), and pitch (P) by means of subharmonic summation [8], pitch and harmonicity via Bregman’s method [9] and the skew, which is the percentage of energy in the pitch relative to the harmonic partials.

\[ BW = \sqrt{\sum_{k=1}^{K} (k - SC)^2 |X(k)|^2 / \left(\sum_{k=1}^{K} |X(k)|^2\right)} \]
\[ SC = \left(\sum_{k=1}^{K} k |X(k)|^2 / \left(\sum_{k=1}^{K} |X(k)|^2\right)\right) \]
\[ P = f : f >= 0 \land \forall g >= 0, H(f) >= H(g), H(f) = \sum_{k=1}^{K} h_k X(k \cdot f) \]

IV. DATA SETS

Five widely differing datasets were used to evaluate the performance of SRT features.

The ‘Four Audio’ dataset consisted of 415 separate recordings of speech (71), music (197), applause (85) and laughter (62). These were obtained from various sources including recordings, live performances, and broadcast media. These were all of 2.5 seconds in duration and sampled at 44.1kHz and 16 bits.
The ‘Frog Calls’ dataset consisted of recordings of 72 different species of native Australian frog calls. They were sampled at 22.05kHz and 16 bits and were each of 50 seconds in duration.

The ‘Musical Instruments’ dataset consisted of 1195 recordings of 97 different musical instruments. These were categorised into one of twelve different classes: piano, harpsichord, plucked string, bowed string, woodwind, brass, organ, tympani, metallic tuned percussive, wooden tuned percussive, non-tuned percussive, and others. These were all sampled at 44.1kHz and 16 bits and were each of one second in duration.

The ‘Environmental’ sounds dataset consisted of 205 recordings of a 20 classes of environmental sounds including: sirens, chimes, music, insect noises, ambient outdoor (wind, rain etc), storm, thunder, ambient office sounds, animal sounds, screams, laughter, car engines, traffic, power tools, explosions and gunshots. These were sampled at 11 kHz and 16 bits and were each of 500 milliseconds in duration.

The ‘Genres’ dataset consisted of 1887 recordings of nine different classes of contemporary music including: rap, pop, rhythm and blues, alternative, electric, folk, jazz, rock, and blues. Each was 10 seconds in duration.

The ‘Birds’ dataset consisted of recordings of 657 different species of 24 different families of birds. Each was 10 seconds in duration.

V. EXPERIMENTS

For each given dataset, feature vectors of varying size ranging from 8 to 96 dimensions were formed as the first N features from each of the five feature sets. These vectors were then evaluated using a k-NN classifier (k=1) using ten-fold cross validation.

Since most of the features used in the experiments were obtained via the short-time Fourier transform it was necessary to first determine the optimal window size. To do this the classifier was trained with features extracted from Fourier spectra at various window sizes. This was performed for all data sets and all feature sets. For almost all the datasets the best results were obtained using the largest analysis window size, being 1024 samples. For the MFCC based features using the ‘Instruments’ and ‘Environmental’ datasets the best performance was achieved using 512 and 128 samples respectively. In the case of the PSTC features, an analysis window size of 256 samples was found optimal.

The reason for this is that the number of vectors required for training puts downward pressure on the number of samples available from which to form each training vector. This results in a tradeoff between forming features that provide better spectral or temporal resolution. In the case of the Instruments and Environmental datasets, increased temporal resolution provided better performance.

To ensure that the classifier was equally trained in all cases, for any given dataset exactly the same number of samples were used in forming the training vectors for all of the feature sets. To ensure that performance of the feature sets was independent of the number of training vectors, the optimal number of training vectors for each feature set and data set was first determined. This was undertaken by extracting multiple vectors sequentially with no overlap from each recording. In all cases the best performance was obtained when the classifier was trained with the maximum amount of training vectors that could be extracted from the datasets.

For the ‘Four Audio’ dataset Table 1 and in Fig. 7 show the best performance of 5% error was obtained with the SRT2 features and the worst at 25% error using the SRT1 features. The PSTC features provided the next best performance with 10% error, followed by MFCC+ features at 15% error and then standard MFCC features that achieved 16% error at higher dimensions. While the MFCC and MFCC+ feature sets required using 100 training vectors from each recording, the PSTC and SRT1 required 20 training vectors, each obtained from 100ms of audio data while the SRT2 required a single training vector obtained from 2 seconds from each recording.

<table>
<thead>
<tr>
<th>Features</th>
<th>Dimensions (Feature set size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>8    16   18    26   32   42   64</td>
</tr>
<tr>
<td>MFCC+</td>
<td>0.307 0.223 0.220 0.205 0.164</td>
</tr>
<tr>
<td>PSTC</td>
<td>0.135 0.121 0.100 0.094</td>
</tr>
<tr>
<td>SRT1</td>
<td>0.410 .339 .255</td>
</tr>
<tr>
<td>SRT2</td>
<td>0.079 0.053 0.055</td>
</tr>
</tbody>
</table>

Fig. 7. Normalised Error Rate for the Four Audio Dataset
Table 2. Normalised Error Rates for the Frog Call Dataset

<table>
<thead>
<tr>
<th>Dimensions (Feature set size)</th>
<th>Features</th>
<th>8</th>
<th>16</th>
<th>18</th>
<th>26</th>
<th>32</th>
<th>42</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>M F C C</td>
<td>0.350</td>
<td>0.214</td>
<td>0.176</td>
<td>0.171</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M F C C +</td>
<td>0.505</td>
<td>0.280</td>
<td>0.222</td>
<td>0.187</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P S T C</td>
<td>0.207</td>
<td>0.141</td>
<td>0.094</td>
<td>0.084</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S R T 1</td>
<td>0.78</td>
<td>0.71</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S R T 2</td>
<td>0.24</td>
<td>0.15</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The minimum error with the Environmental sounds dataset was 25% obtained with the SRT2 feature set which narrowly outperformed the M F C C + features. These were followed by the P S T C and standard M F C C and finally S R T 1 features as shown in Table 4 and in Fig. 10.

Table 4. Normalised Error Rate for the Environmental Dataset

<table>
<thead>
<tr>
<th>Dimensions (Feature set size)</th>
<th>Features</th>
<th>8</th>
<th>16</th>
<th>18</th>
<th>26</th>
<th>32</th>
<th>42</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>M F C C</td>
<td>0.412</td>
<td>0.363</td>
<td>0.349</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M F C C +</td>
<td>0.453</td>
<td>0.342</td>
<td>0.282</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P S T C</td>
<td>0.446</td>
<td>0.369</td>
<td>0.307</td>
<td>0.332</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S R T 1</td>
<td>0.69</td>
<td>0.60</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S R T 2</td>
<td>0.26</td>
<td>0.280</td>
<td>0.256</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the case of the Music Genre dataset the best performance was again obtained using the S R T 2 features that achieved a minimum error rate of 57% while the S R T 1 features again performed poorly as shown in Table 5 and in Fig. 11.

Table 5. Normalised Error Rates for the Musical Genre Dataset

<table>
<thead>
<tr>
<th>Dimensions (Feature set size)</th>
<th>Features</th>
<th>8</th>
<th>16</th>
<th>18</th>
<th>26</th>
<th>32</th>
<th>42</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>M F C C</td>
<td>0.750</td>
<td>0.707</td>
<td>0.674</td>
<td>0.648</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M F C C +</td>
<td>0.733</td>
<td>0.679</td>
<td>0.646</td>
<td>0.621</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P S T C</td>
<td>0.631</td>
<td>0.578</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S R T 1</td>
<td>0.76</td>
<td>0.78</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S R T 2</td>
<td>0.63</td>
<td>0.577</td>
<td>0.573</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In common with the Frog call dataset the PSTC features also outperformed the SRT2 features in the bird call dataset achieving a 42% compared to the SRT2’s 46% as shown in Table 6 and in Fig. 12. For low dimensional vectors the MFCC feature set initially performed better than MFCC+ but fell behind it as dimensionality increased. The lowest error rates achieved by the MFCC and MFCC+ feature sets were obtained with 250 training vectors and 1024 sample window. This gave an error rate of 49% for MFCC+ and 50% for the standard MFCC feature set.

Table 6. Normalised Error Rates for the Bird Call Dataset

<table>
<thead>
<tr>
<th>Dimensions (Feature set size)</th>
<th>8</th>
<th>16</th>
<th>26</th>
<th>32</th>
<th>42</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.678</td>
<td>0.498</td>
<td>0.529</td>
<td>0.523</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC+</td>
<td>0.716</td>
<td>0.576</td>
<td>0.518</td>
<td>0.485</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSTC</td>
<td>0.590</td>
<td>0.507</td>
<td>0.434</td>
<td>0.425</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRT1</td>
<td>0.85</td>
<td>0.84</td>
<td>0.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRT2</td>
<td>0.52</td>
<td>0.468</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 11. Normalised Error Rate for Music Genre Dataset

Fig. 12. Normalised Error Rate for the Bird Call Dataset

VI. CONCLUSIONS

The spectral-Radon transform has applications for audio indexing, content-based retrieval, acoustic recognition and even acoustic visualization. This paper has presented SRT derived features for audio classification that provided lower error rates than MFCC based classification features on all six distinct datasets evaluated. As the salient information is captured by the detailed changes in the characteristics of individual SRT traces simply using the first few principal components of the SRT surface as features does not work. Yet isolating the specific frequency and temporal characteristics of SRT traces yields superior performance.

While these SRT features do not always provide the highest performance when compared to feature sets that may be more attuned to the particularities of a given data set, SRT features provided better overall performance across all data sets. The improved performance is not just due to the features capturing temporal variation of the signal as the SRT features also outperformed the spectrogram principle component features in four of the six datasets.

An additional advantage of the SRT features is that the number of training vectors and associated memory resources are significantly reduced over MFCC based feature sets. Accordingly SRT based features are shown to be a useful tool for audio classification, indexing and content based retrieval.

REFERENCES