Better than MFCC Audio Classification Features

Ruben Gonzalez

Institute for Intelligent Integrated Systems, Griffith University,
School of Information and Communication Technology,
Gold Coast, Australia.

Abstract. Mel-Frequency Ceptral Coeffienents (MFCCs) are generally the features of choice for both audio classification and content-based retrieval due to their proven performance. This paper presents alternate feature sets that not only consistently outperform MFCC features but are simpler to calculate.

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

1 Introduction

Content-based retrieval is fundamentally a two-step task; salient features are first extracted from the data, which can be then used for class recognition 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 8 MFCC features plus energy and their derivatives and found that performance was relatively comparable. Chu et.al., [2] investigated the problem of correctly determining between one of five different classes of environmental sounds; Hallway, Café, Lobby, Elevator and Sidewalk using k-NN, GMM and SVN 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 capture properties of the underlying audio data that have statistical or psychological saliency. Sometimes as in the case of the MFCC, they also attempt to mimic psychological processes. The disadvantage of selecting any particular set of features is that any information that is not represented 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 features have been presented in the literature, being extracted from audio signals in either the temporal or frequency domains. 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 nine 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. This iterative ranking process ensured that each feature set was evaluated for its possible 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, percussions, 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 a combination of spectral and MFCC features.

This paper presents three alternate feature sets to the MFCC that are less computationally complex and superior in performance across a range of diverse datasets.

2 Audio Feature Extraction

This paper evaluates the performance of five different feature sets. The baseline feature set is comprised on only MFCC features. An enhanced MFCC (MFCC+) feature set adds spectral and temporal features to the MFCC. The three proposed feature sets include the Principle Spectral Coefficients (PSC), the Principle Cepstral Coefficients (PCC) and the Principle Spectral-Temporal Coefficients (PSTC).

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 (1) and octave-wide, triangular overlapping windows. Finally the cepstrum (2) is obtained using the remapped coefficients m(n).

$$Mel(f) = 2595\log_{10}(1 + f/100) \tag{1}$$

$$c(k) = DCT\left\{ \log\left|DFT\{m(n)\}\right| \right\} \tag{2}$$

The enhanced MFCC+ dataset includes four common temporal features and six spectral features. The four c 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:

$$ZCR = \sum_{k=2}^{K} \left|\mathrm{sgn}(x(k)) - \mathrm{sgn}(x(k-1))\right| \qquad \mathrm{sgn}(n) = \begin{cases} 1, n > 0 \\ 0, n = 0 \\ -1, n < 0 \end{cases} \tag{3}$$

$$E = \frac{1}{K}\left(\sum_{k=1}^{K} |X(k)|^2 \right) \tag{4}$$

$$F = E(n) - E(n\text{-}1) \tag{5}$$

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.

$$BW = \sqrt{\left(\sum_{k=1}^{K} (k - SC)^2 |X(k)|^2 \right) \Big/ \left(\sum_{k=1}^{K} |X(k)|^2 \right)} \tag{6}$$

$$SC = \left(\sum_{k=1}^{K} k \times |X(k)|^2 \right) \Big/ \left(\sum_{k=1}^{K} |X(k)|^2 \right) \tag{7}$$

$$P = f : f >= 0 \wedge \forall g >= 0, H(f) >= H(g);$$

$$H(f) = \sum_{k=1}^{K} h_k X(k \cdot f) \tag{8}$$

Rather than just arbitrarily selecting features based on a subjective notion of saliency, it is possible from a statistical perspective to identify the principle components in any given data set. These principle components are guaranteed to optimally represent the underlying data for any number of components. 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). A statistically optimum feature set of the Principle Spectral Components (PSCs) can accordingly be obtained by taking the first few DCT coefficients of the spectrum obtained via a short-time Fourier transform (where || represents the complex magnitude):

$$PSC(k) = DCT\left\{ \left| DFT\{x(n)\} \right| \right\} \tag{9}$$

A variation on the PSC that provides more even scaling of the feature set by whitening the spectrum is to use the Principle Cepstral Components (PCCs):

$$PCC(k) = DCT\left\{ \log\left| DFT\{x(n)\} \right| \right\} \tag{10}$$

As the PSC and PCC methods are formed from a one-dimensional spectrum they are unable to capture the evolution of the spectrum over time. The temporal characteristics of sounds are known to be important for identifying some classes of audio. Accordingly the PSTC feature set captures the principle information contained in the time-frequency distribution of energy in the audio signals. This feature set is obtained by taking a two-dimensional Discrete Cosine Transform (DCT) of the audio signal's spectrogram.

3 Data Sets

To evaluate the performance of the proposed audio classification features five widely differing datasets were used.

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 1629 recordings of 74 different species of native Australian frog calls [10]. They were sampled at 22.05kHz and 16 bits and were each of 250 milliseconds in duration.

The 'Insect' dataset consisted of recordings of the sounds made by 381 different species of insects and were categorised according to the four following families: Katydid, Cricket, Cicada, and others. These were all 5 seconds in duration and sampled at 44.1kHz and 16 bits.

The 'Musical Instruments' dataset consisted of 1345 recordings of 97 different musical instruments [11]. These were categories 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 500 milliseconds 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.

4 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 the Four Audio, Frog Call, and Insect datasets the best results for all feature sets were obtained using the largest window size, being 1024 samples. With the exception of the PSTC features, the best performance for all feature sets using the Instruments and Environmental datasets was achieved using a 512-sample window.

In the case of the PSTC features, the analysis window size was 256 samples for the Instruments dataset and 128 samples for the Environmental sounds dataset. 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, which was the same for all feature sets with the exception of the PSTC features.

For the ‘Four Audio’ dataset the best performance was obtained using 100 training vectors for each recording for all feature sets except the PSTC feature set where 20 training vectors were used. The results shown in **Table 1** and in **Fig. 1** show the normalized classification error rate for each of the feature sets for each size of vector evaluated. For this dataset the PSTC features clearly outperformed all the others with the PCC features providing the second best performance. At lot dimensions the PSC features performed only marginally better than the MFCC but approached the PCC

features at higher dimensions when enough coefficients were used. The MFCC features alone had the worst overall performance. The MFCC+ enhanced features provided moderate performance at low dimensions that did not improve much at higher dimensions.

Due to the small amount of data available for analysis in the Frog Call dataset only five training vectors could be extracted from each recording while using a window size of 1024 samples. In the case of the PSTC features only a single train vector could be extracted from the data and this seriously impeded the performance of the PSTC feature set. The results are shown in **Table 2** and **Fig. 2**. While PSTC features outperformed the MFCC and MFCC+ features, it fell behind the PSC and PCC features. The PCC features again outperformed the PSC features at all dimensions. Notably, the MFCC features outperformed the enhanced MFCC+ features at all but the highest dimensions.

The best performance for the Insect sounds dataset was obtained using 50 training vectors (20 in the case of PSTC features) and a window size of 1024 samples. This dataset provided similar results to the Frog Call dataset as demonstrated by the results shown in **Table 3** and in **Fig. 3**. The PTFC features again provided the best overall performance at all dimensions. As is to be expected the MFCC+ performed better than MFCC except at low dimensions, but both of these again demonstrated the worst performance. The PSC and PCC features provided very similar results except that the PSC features became less competitive at higher dimensions.

The Instrument dataset results departed markedly from those of the other datasets. The best performance in all cases was obtained using 50 training vectors and a 512 sample window for all feature sets except the PSTC features which made use of 5 training vectors and a 256 sample window. The enhanced MFCC+ narrowly outperformed the PSTC features as shown in **Table 4** and in **Fig. 4**. Notably the PCC features performed even worse than the standard MFCC features. Yet the PSC features managed to provide better performance than MFCC for all dimensions. To evaluate the contribution of the spectral and temporal features in the MFCC+ dataset to the performance, these were added to the PSC data set to form an enhanced PSC+ dataset. These enhanced PSC+ features were able to approach but not exceed the performance of the MFCC+ features.

The best performance with the Environmental sounds dataset was obtained using 10 vectors and a 512 sample window in all cases except the PSTC features which used 5 training vectors and as 128 sample window. The PSC features in this case provided the best performance as is shown in **Table 5** and in **Fig. 5**. The PCC features only managed to surpass the performance of PSC at high dimensions. Notably the PTFC features provided moderate performance better than MFCC but worse than MFCC+ at low to mid dimensions but deteriorating at higher dimensions. It is unclear if this was due to having insufficient samples to obtain enough training vectors at a high enough frequency resolution, or some other intrinsic characteristic of the dataset.

Table 1. Feature Set Normalised Error Rate for the Four Audio Dataset

	Dimensions (Feature set size)								
Features	8	16	18	26	32	42	64	74	96
PSC	0.297	0.219			0.183		0.129		0.121
PCC	0.245	0.170			0.144		0.119		0.116
MFCC	0.307	0.223	0.220		0.205		0.164		
MFCC+			0.215	0.168		0.161		0.145	
PSTC	0.135	0.121			0.100		0.094		0. 094

Table 2. Feature Set Normalised Error Rate for the Frog Call Dataset

	Dimensions (Feature set size)								
Features	8	16	18	26	32	42	64	74	96
PSC	0.296	0.191			0.142		0.131		0.145
PCC	0.192	0.114			0.095		0.107		0.130
MFCC	0.283	0.226			0.195		0.214		
MFCC+			0.322	0.258		0.214		0.214	
PSTC	0.281	0.228			0.181		0.161		0.168

Table 3. Feature Set Normalised Error Rate for the Insect Dataset

	Dimensions (Feature set size)								
Features	8	16	18	26	32	42	64	74	96
PSC	0.046	0.021			0.014		0.021		0.022
PCC	0.049	0.022			0.016		0.015		0.015
MFCC	0.084	0.051			0.039		0.040		
MFCC+			0.049	0.032		0.026		0.028	
PSTC	0.018	0.013			0.008		0.009		0.014

Table 4. Feature Set Normalised Error Rate for the Musical Instrument Dataset

	Dimensions (Feature set size)								
Features	8	16	18	26	32	42	64	74	96
PSC	0.599	0.491			0.379		0.281		0.233
PCC	0.646	0.581			0.507		0.439		0.407
MFCC	0.642	0.531			0.423		0.325		
MFCC+	0.451		0.373	0.326		0.286		0.250	
PSTC	0.480	0.402			0.327		0.293		0.308
PSC+			0.400	0.354		0.303		0.254	

Table 5. Feature Set Normalised Error Rate for the Environmental Sounds Dataset

	Dimensions (Feature set size)								
Features	8	16	18	26	32	42	64	74	96
PSC	0.326	0.252			0.213		0.214		0.232
PCC	0.348	0.274			0.240		0.211		0.219
MFCC	0.412	0.363			0.349		0.300		
MFCC+	0.453		0.342	0.282		0.280		0.249	
PSTC	0.446	0.369			0.307		0.332		0.355

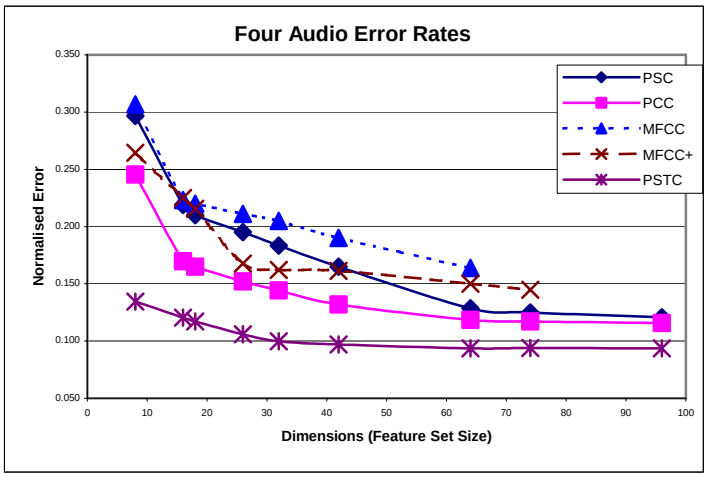

Fig. 1. Feature Set Normalised Error Rate for the Four Audio Dataset

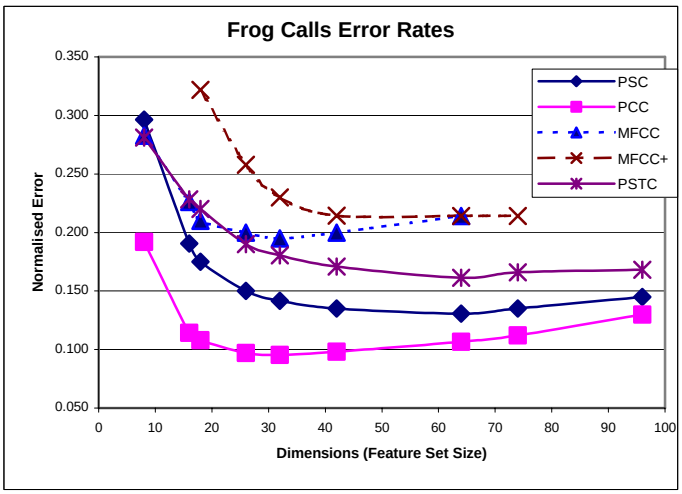

Fig. 2. Feature Set Normalised Error Rate for the Frog Call Dataset

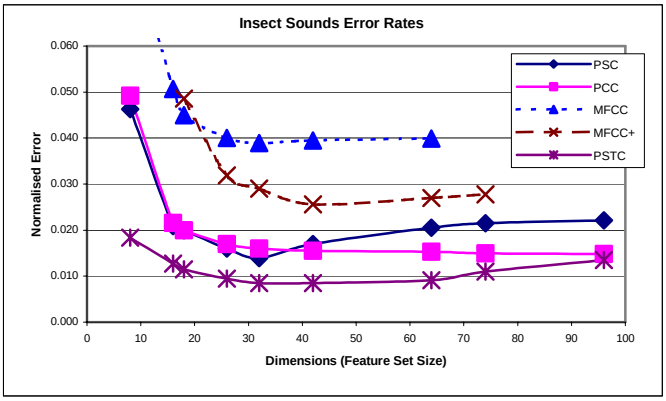

Fig. 3. Feature Set Normalised Error Rate for the Insect Dataset

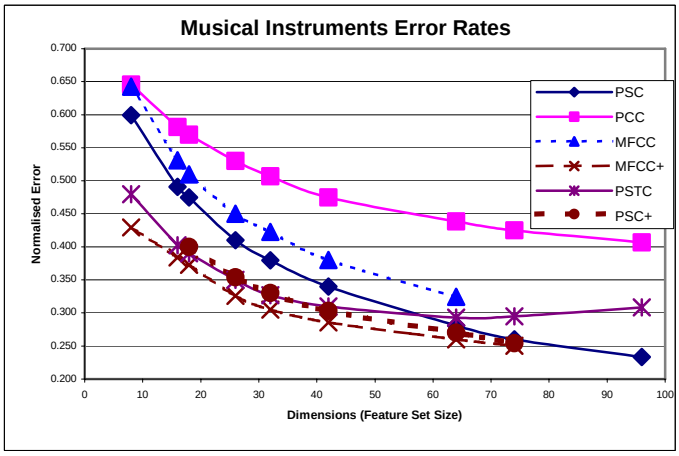

Fig. 4. Feature Set Normalised Error Rate for the Musical Instrument Dataset

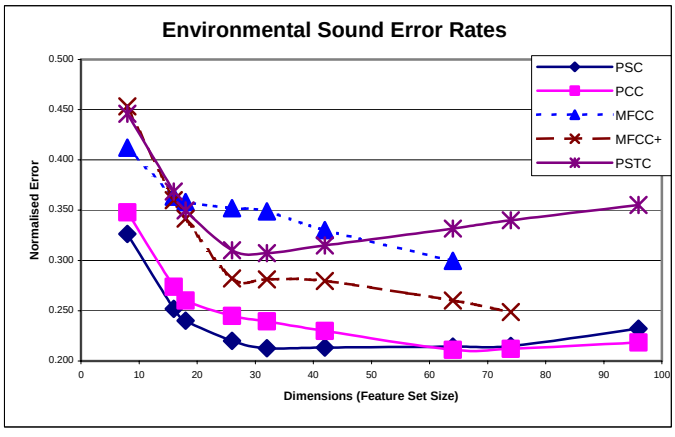

Fig. 5. Feature Set Normalised Error Rate for the Environmental Sounds Dataset

5 Conclusions

This paper has presented three new feature sets for audio classification based on principle spectral components for audio indexing, content based retrieval and acoustic recognition that provide improved performance over feature sets comprised of only MFCC and MFCC plus spectral and temporal features in most cases. When sufficient data is available the PSTC features provide the best overall performance by a factor of 50-100%. When insufficient data is available either the PSC or PCC feature sets generally provide better performance than MFCC and MFCC enhanced feature sets. These have the added benefit of requiring less processing to obtain than MFCC based feature sets.

6 References

[1] José Anibal Arias, Julien Pinquier and Régine André-Obrecht, *"Evaluation Of Classification Techniques For Audio Indexing,"* Proceedings of 13th European Signal Processing Conference, September 4-8, 2005. EUSIPCO'2005, Antalya, Turkey.

[2] S. Chu, S. Narayanan, C.-C. Jay Kuo, and Maja J. Mataric. *"Where am i? scene recognition for mobile robots using audio features"*. In Proc. of ICME, Toronto, Canada, July 2006.

[3] Lefèvre F., "A Confidence Measure based on the K-nn Probability Estimator", International Conference on Spoken Language Processing, Beijing, 2000

[4] Kim, H-G., Moreau, N., Sikora., "Audio Classification Based on MPEG-7 Spectral Basis Representations" IEEE Trans. On Circuits And Systems For Video Technology,Vol.14,No.5, May 2004.

[5] M.F. McKinney, J. Breebaart. "*Features for audio and music classification.*" In Proc. of the Intern. Conf. on Music Information Retrieval (ISMIR 2004), pp. 151-158, Plymouth MA, 2004.

[6] Peltonen, V. Tuomi, J. Klapuri, A. Huopaniemi, J. Sorsa, T., "*Computational auditory scene recognition*", Proceeding of. International Conference on Acoustics, Speech, and Signal Processing, 2002. (ICASSP '02). May 13-17, 2002, Orlando, FL, USA, vol.2, pp:1941-1944.

[7] Mingchun Liu and Chunru Wan. 2001. "Feature selection for automatic classification of musical instrument sounds." In Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries (JCDL '01). ACM, New York, NY, USA, 247-248.

[8] D.J.Hermes, "Measurement of pitch by subharmonic summation" J. Acoust. Soc. Am. Volume 83, Issue 1, pp. 257-264 (January 1988)

[9] Bregman, Albert S., Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, Massachusetts: The MIT Press, 1990 (hardcover)/1994 (paperback).

[10] D.Stewart, "Australian Frog Calls - Subtropical East", [Audio Recording]

[11] McGill University Master Samples, 1993 "Vol 1: Classical Sounds" [Audio Recording]