July 2003

Performance of MPEG-7 low level audio descriptors with compressed data

Jason Lukasiak
University of Wollongong, jl01@uow.edu.au

David A. Stirling
University of Wollongong, stirling@uow.edu.au

N. Harders
University of Wollongong

S. Perrow
University of Wollongong

Recommended Citation

Lukasiak, Jason; Stirling, David A.; Harders, N.; and Perrow, S.: Performance of MPEG-7 low level audio descriptors with compressed data 2003.
https://ro.uow.edu.au/infopapers/114

Performance of MPEG-7 low level audio descriptors with compressed data

Abstract

This paper presents a detailed analysis of lossy compression effects on a set of the MPEG-7 low-level audio descriptors. The analysis results show that lossy compression has a detrimental effect on the integrity of practical search and retrieval schemes that utilize the low level audio descriptors. Methods are then proposed to reduce the detrimental effects of compression in searching schemes. These proposed methods include multi-frame searching and machine learning derived prediction. The proposed mechanisms greatly reduce the effect of compression on the set of MPEG-7 descriptors; however, future scope is identified to develop new audio descriptors that account for compression effects in their structure.

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was published as: Lukasiak, J, Stirling, D, Harders, N & Perrow, S, Performance of MPEG-7 low level audio descriptors with compressed data, Proceedings International Conference on Multimedia and Expo, 6-9 July 2003, vol 3, 273-276. Copyright IEEE 2003.

PERFORMANCE OF MPEG-7 LOW LEVEL AUDIO DESCRIPTORS WITH COMPRESSED DATA

J. Lukasiak, D. Stirling, N. Harders, S. Perrow

TITR, University of Wollongong

Wollongong, NSW, Australia, 2522

ABSTRACT

This paper presents a detailed analysis of lossy compression effects on a set of the MPEG-7 low-level audio descriptors. The analysis results show that lossy compression has a detrimental effect on the integrity of practical search and retrieval schemes that utilize the low level audio descriptors. Methods are then proposed to reduce the detrimental effects of compression in searching schemes. These proposed methods include multi-frame searching and machine learning derived prediction. The proposed mechanisms greatly reduce the effect of compression on the set of MPEG-7 descriptors, however, future scope is identified to develop new audio descriptors that account for compression effects in their structure.

1. INTRODUCTION

With the ever increasing volume of Multi-Media (MM) data available via shared networks, such as the Internet or even large organizational intranets, meaningful and efficient storage, retrieval, archiving and filtering of the available MM data is becoming increasingly difficult. Current text based search and retrieval schemes rely on meaningful textual tags being associated with every MM item. Such text based methods are limited in usefulness by the quality (content) of the tag used. For example, it may be possible to find a particular MM item via Authors name or title, but it is extremely unlikely that content specific features such as colour, melody or frequency structure would be identifiable using a text based tag. Overcoming this limitation is the realm of the new MPEG-7 standard [1]. This standard provides a structured framework for describing MM content in a platform independent environment [1,2].

At its lowest level, the MPEG-7 standard specifies a set of low-level descriptors that are calculated directly from the MM content [2]. Examples of the low-level descriptors generated are Color space [2] and audio spectral envelope [3]. A full description of the MPEG-7 Audio standard can be found in [3] and an excellent overview of the entire standard in [2].

Having descriptors associated with MM data that describe the actual content of the data, provides the potential for powerful manipulation of content supply and consumption. The manipulation could involve finding all MM items in a database that have a blue background or selecting the Audio segments that represent specific sources (such as dogs barking) [4][7].

Whilst the proposed MPEG-7 standard offers a powerful new scheme for control of MM data, there are a number of issues that may limit practical application of the standard. Examples of such limitations are the complexity involved and the integrity of the descriptors in compressed environments. As a substantial amount of MM data is compressed before storage using lossy compression schemes that remove the perceptually redundant information from the signal, such as MP3 for audio and JPEG for images, the effects of compression on the integrity of the descriptors is extremely important for practical applications. For example, can we find an MP3 audio file (or an audio segment previously compressed by MP3) that matches our target song, using the low-level descriptors generated from a CD?

The effect of compression on the MPEG-7 low-level audio descriptors is the focus of this paper. A thorough analysis of the effect of compression on five of the seventeen low-level audio descriptors is presented in section 2. These five descriptors were selected due to space constraints in presenting results for the full set of descriptors, and also, due to these descriptors being frame based (as opposed to entire file based). The combination of the prescribed descriptors also presents a compact description of the underlying audio data.

Once identified, some initial methods for reducing the effects of compression are detailed in Section 3. Finally the major points are summarized in Section 4.

2. EFFECT OF COMPRESSION ON LOW-LEVEL DESCRIPTORS

The five audio low-level descriptors selected for the analysis were; Audio Power (AP), Audio Waveform (AW), Audio Spectrum Envelope (ASE), Audio Spectrum Centroid (ASC) and Audio Spectral Spread (ASS) [3].

To determine the effect of compression on the audio low-level descriptors, two well-known audio compression algorithms, MP3 [5] and WMA [6], were used to compress and uncompress 90 16-bit 44.1kHz sampled audio files. Each of the audio files was of approximately 10 seconds duration, with 50 files representing instrumental only signals and 40 representing combinational signals such as pop music.

The MP3 encoder was operated at 128, 160 and 192 kb/s and the WMA coder was operated at both 128 and 160 kb/s.

The MPEG-7 low level audio descriptors defined above, were then calculated for both the files which had been compressed/uncompressed and the original files. A frame size of 20 ms was used to calculate the descriptors

0-7803-7965-9/03/$17.00 ©2003 IEEE

2.1 Objective Measures

To give an objective indication of the effect of compression on the descriptors, the Segmented Signal to Noise Ratio (SegSNR) was used. This was calculated as:

$$SegSNR = \frac{1}{N}\sum_{x=0}^{N-1} 10\log_{10}\left(\frac{\sum_{n=1}^{r} x_n^2}{\sum_{n=1}^{r}(x_n-\overline{x}_n)^2}\right) \qquad (1)$$

where x_n is the descriptor from the origirial file, . $\overline{x_n}$ is the descriptor from the compressed file, r is the dimension of the descriptor per frame and N is the number of frames in the file. The SegSNR for each descriptor and compression configuration was calculated for each file, with the results summarized in Section 2.3.

2.2 Practical Measures

Determining the effect of compression noise on the descriptors requires not only objective measures, but also analysis of the performance in a practical searching scheme. To this end, a simple searching scheme was utilized that attempts to locate a specific frame in the compressed file, using the descriptors generated for the target frame from uncompressed data. This is a simplified version of search/retrieval schemes outlined in [7]. The searching scheme selects the frame in the compressed file that minimizes the Mean Squared Error (MSE) defined as:

$$MSE = \frac{1}{r}\sum_{n=1}^{r}(x_n-\overline{x}_n)^2 \qquad (2)$$

where x_n, $\overline{x_n}$ and r are as defined for (1). It should be noted that when the descriptors are generated from original uncompressed data, the above scheme returns the correct frame for all individual (and combinations of) descriptors.

2.3 Practical Results

The results for the average SegSNR across the 90 test files are shown in Table 1. The average percentage incorrect frames identified using (2) are shown in Table 2.

2.3.1 Results for AP

The AP describes the instantaneous power of each input frame and consists of only a single scalar value per frame [3]. The average SegSNR values for the AP descriptor are shown in row 3 of Table 1 and the search mis-classification percentage for a subset of the files is shown in row 3 of Table 2.

The results in Table 1 indicate that the AP descriptor achieves a very high SegSNR value of over 40 dB for all of the compression schemes. This very high SegSNR would appear to indicate that compression has very little effect on the AP descriptor, however, the results in Table 2 conflict with this view. The results in Table 2 indicate that despite the very high SegSNR values reported in Table 1, the AP descriptor is a very unreliable search mechanism. This inability to adequately match the target frame is due to both the AP descriptor having very similar values across adjacent frames and the fact that logarithmic amplitude quantisation is employed in most audio encoders.

	MP3			WMA	
Bit rate in kb/s	128	160	192	128	160
AP	42.71	43.22	43.12	44.55	47.4
AW	37.33	40.33	41.44	39.55	42
ASC	44.63	47.72	48.55	47.77	50.53
ASS	47.43	49.15	49.87	50.77	53.69
ASE	35.24	39.05	40.36	37.58	40.35

Table 1: Average SegSNR values for dB

	MP3			WMA	
Bit rate in kb/s	128	160	192	128	160
AP	84.75	84.3	86	82.45	78.4
AW	56.25	46.4	41.3	47.8	38.55
ASC	86.55	82.5	81.15	83.15	82.5
ASS	90.55	86	89.8	85.85	83.35
ASE	1.325	0.475	0.59	1.34	1.04
ASS/ASC	34.6	26.6	22.2	23.85	16.45
ALL	9.05	4.72	2.55	4.15	2.25

Table 2: Percentage incorrect frames for searching

2.3.2 Results for AW

The AW descriptor provides a low resolution representation of the time domain envelope and consists of 2 scalar values (max and min) per frame [3]. The average SegSNR values are shown in row 4 of Table 1 and the search mis-classification percentage is shown in row 4 of Table 2.

The SegSNR results indicate that whilst the AW descriptor has slightly lower SegSNR values than those for AP, the values are still very high. The worst case is 37.33dB for MP3 at 128kb/s. The searching results indicate that despite the AW exhibiting lower SegSNR values than the AP descriptor, it provides a more reliable searching mechanism. This improvement is due to the fact that the AW descriptor has two values for each frame and the probability of two frames having very similar descriptors is reduced. However, the AW descriptor still finds incorrect frames on approximately 40 to 60 percent of occasions.

2.3.3 Results for ASC

The ASC descriptor represents the center of gravity of the frequency spectrum and is a single scalar value per frame that indicates the octave shift from 1 kHz of the centroid value [3]. The average SegSNR values are shown in row 5 of Table 1 and the search mis-classification percentage is shown in row 5 of Table 2.

The results indicate that in an objective sense, compression has very little effect on the ASC descriptors. This result is clearly evidenced by the smallest value for SegSNR being in excess of 40 dB. However, as for the AP descriptor, the average search results indicate that due to the relative stability of the centroid value across frames, the minor effects of compression evident in the SegSNR values are sufficient to cause the ASC descriptor to be unsuitable for frame identification in compressed environments.

2.3.4 Results for ASS

The ASS descriptor describes the RMS deviation from the centroid value (ASC) for a given frame and consists of a single scalar value per frame that represents the octave spread from the ASC value [3]. The average SegSNR values are shown in row 6 of Table 1 and the search mis-classification percentage is shown in row 6 of Table 2.

As for the ASC descriptor, the SegSNR values indicate that in an objective sense compression produces little degradation on the ASS values. Again however, the search results indicate that the ASS provides a very unreliable search mechanism in compressed environments. The descriptor often finds the incorrect frame in over 90% of instances.

2.3.5 Results for ASE

The ASE descriptor provides a representation of the power spectrum for each frame of the audio file and consists of a vector of values for each frame, with each vector component representing the magnitude of a particular frequency band [3]. The number of frequency bands (and hence the length of the ASE descriptor) is variable according to a predetermined set of user parameters. These parameters include loEdge, hiEdge and Resolution [3]; where loEdge represents the lowest edge (frequency) of the frequency bands, hiEdge represents the highest edge (frequency) of the frequency bands and resolution defines the width of the frequency bands (in octaves with respect to 1 kHz) between loEdge and hiEdge. The ASE also contains two additional values representing 0Hz-loEdge and hiEdge-Sampling_freq/2.

An important note in the standard [3] is that for fine resolutions (i.e. < ¼ octave) the window length (length of frame) restricts the minimum value for loEdge such that atleast 1 FFT frequency coefficient is present in each band. The result of this restriction is that for resolutions less than ¼ octave, the resultant ASE descriptor becomes biased. This bias is due to the fact that the value representing 0Hz-loEdge has many FFT coefficients lumped into it, and thus, becomes very large, whilst the neighboring bands contain only a single coefficient. The net result of this effect is that resolutions less than ¼ octave are not suitable for searching or identifying complex audio signals that have significant low frequency content (such as speech). To alleviate this problem the ASE descriptor generated for this work used a resolution of ¼ octave, which produced 32 frequency bands for each frame. The average SegSNR values for ASE using this configuration are shown in row 7 of Table 1 and the search mis-classification percentage is shown in row 7 of Table 2.

The values in Table 1 indicate that the ASE descriptor produces lower objective results in the presence of compression than the other spectral descriptors, ASS and ASC. This should be expected as the ASE produces much finer resolution than those other descriptors, and hence, the effects of removing masked components in the compression scheme produces more visible objective distortions. It is also clearly evident that as the bit rate of the compression schemes increases, the SegSNR also increases. This effect is due to the compression schemes using the additional bits available to better represent the spectral envelope of the signal.

The results in Table 2 indicate that the ASE produces fairly reliable search results for all compression schemes. This is obviously due to the fine resolution present in the descriptor and thus the likelihood of two similar frames existing (even in the presence of compression noise) is lower than for the more generic descriptors. However, the search still fails on approximately 1% of occasions for high compression rates.

2.3.6 Results for Combined Descriptor Searches

To ascertain if improved search results could be achieved by combining multiple descriptors together into meta-descriptors, we formed two meta-descriptors. The first of these meta-descriptors combined all five of the specified descriptors together and the second combined ASC and ASS to produce a compact representation of frequency content. The search results for these two meta-descriptors are shown in Table 2 rows 7 & 8 respectively.

Comparing the results for the meta-descriptor containing all of the descriptors to those for the ASE, indicates that including the additional descriptors into the search actually degrades the searching performance. This result indicates, that because the additional descriptors may have larger absolute values than the individual ASE components, the ambiguity introduced into these additional descriptors by compression is sufficient to degrade the unweighted search performance used in (2). Better results may be achieved by introducing a weighting function into (2).

When compared to the results for the descriptors ASC and ASS in isolation, the results for the combined ASS/ASC meta-descriptor indicate that combining the descriptors together reduces the incorrect results by between 50 and 66%. This improved result is encouraging and supports the finding that implementing more sophisticated search (weighted) mechanisms for meta-descriptors may improve performance in compressed environments.

2.3.7 Summary of Results

The results presented for all descriptors indicate that an objective measure of the effects of compression gives little indication of the actual performance degradation in a practical search situation. The results indicate that compression noise is a significant problem for practical applications of the low level audio descriptors.

It should be noted that on average, across all of the results presented in Section 2.3, the WMA coder had less effect on the descriptors than the MP3 encoder. This result is most likely attributed to the fact that the WMA algorithm is significantly more modern than MP3, and thus, exploits more sophisticated signal processing and psycho-acoustic techniques.

3. METHODS FOR IMPROVING PERFORMANCE IN COMPRESSED ENVIRONMENTS

A number of methods were trialed to improve the search performance in compressed environments. Initially linear signal processing techniques such as Vector classification [9], linear prediction and vector linear prediction [8] were employed.

However, due to the non-stationarity and complexity of the compression noise, these methods offered . little or no improvement.

To provide a more flexible modeling/prediction mechanism a number of alternative approaches, emanating from Machine Learning (ML) were considered. The practical focus of these endeavors is to detect useful, but often-implicit patterns in empirical data, and to further construct descriptive models of these. Whilst there are several plausible techniques that could be considered in improving the performance of the low level descriptors, the impact of each is governed primarily by how the various aspects and concepts are both represented and modeled. A useful overall guide to this technology is found in [10].

The approach taken in this paper is to learn rule-based predictive models, based on compression-effected descriptors that can essentially predict their compression free state. For this task we employed the ML algorithm Cubist [11] and the results for the all meta-descriptor are shown in Table 3. The results in Table 3 use MP3 128 kb/s compression and are for a set of test files that were not used in training the predictor. The all meta-descriptor was chosen, as the 37 by 37 dimension predictor presents the greatest challenge for the ML procedure.

Descriptor	ALL
non-predicted	4.1
ML: Cubist	3.2

Table 3: Comparison of search performance for ML prediction.

As can be seen in Table 3, the ML approach significantly improves the missed-framing rate by 0.9% compared to the simple non-predicted method. This improvement is readily explained by the nature of the machine learning approach generally. As such algorithms seek to learn specific concepts, they create and grow a model to fit or explain the training data. In contrast, often the reverse situation arises, where a fixed model is employed and data is fitted to it.

In an attempt to further improve searching performance we constructed an improved search algorithm that incorporates the previous, current and future frames (PCF) into the search procedure. This modification incorporates a larger time scale (adjacent frames) into the calculation and allows more accurate searching, as the evolution, pattern or trend is identified as opposed to a single sample value.

The results of using the PCF search on the five individual descriptors with 128kbps MP3 compression are shown in Table 4. Comparing the results in Table 4 with those in Table 2 clearly indicates that increasing the time scale of the search dramatically improves the search performance; an improvement from 84.75% to 9.8% mis-classified frame for the AP descriptor.

4. CONCLUSION

An extended analysis of lossy compression effects on a set of the MPEG-7 low level audio descriptors was conducted. This analysis exposed a distinct degradation in the performance of practical searching schemes when lossy compression has been used to modify the MM files.

	AP	AW	ASC	ASS	ASE
Average	9.8	4.8	11.9	15.7	0.4

Table 4: Percentage incorrect frames for the PCF search

Methods tested to reduce the effects of compression, indicate that prediction schemes based on machine learning offer the best performance in modeling the compression noise. Also, a more complex searching mechanism that incorporated the previous, current and future frames into the search criteria greatly reduced the number of incorrectly identified frames. The Authors fell that combining these two approaches could further improve searching reliability in compressed environments, however, the overall reliability still may be somewhat below that necessary for practical applications. The authors propose that future work should focus on developing new audio descriptors that incorporate the characteristics of compression, such as psycho-acoustic models, into their structure. The development of these descriptors should greatly improve reliability for compressed file searches.

5. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11/N4031, *Overview of the Mpeg-7 Standard (version 5)*, International Organisation for Standardisation, Singapore, March 2001.

[2] S. Chang, T. Sikora and A Puri, "Overview of the MPEG-7 Standard", IEEE Trans. On Circuits and Systems for Video Tech., Vol. 11, No.6, pp. 688-695, June 2001.

[3] ISO/IEC FDIS 15938-4, *Information Technology Multimedia Content Description Interface,Part 4: Audio*, International Organisation for Standardisation, Singapore, March 2001.

[4] M. Casey, "MPEG-7 Sound Recognition Tools" IEEE Trans. On Circuits and Systems for Video Tech., Vol. 11, No.6, pp. 737-747, June 2001.

[5] ISO/IEC JTCI/SC29, "Information Technology-Coding of Motion Pictures and associated audio for digital storage media upto about 1.5Mbit/s – IS 11172 (Part 3, Audio)", 1992

[6] Microsoft, "Windows Media Encoder", available at http://www.microsoft.com/windows/windowsmedia/WM7/encoder/whitepaper.asp, 15th of July 2002.

[7] S. Quackenbush, and A. Lindsay, "Overview of MPEG-7 audio", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11 Issue 6, pp. 725-729, June 2001 .

[8] M.Yong, G. Davidson, A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction", Proc. Of ICASSP, Vol. 1, pp.402-405, 1988.

[9] A. Gersho, R.M. Gray, *Vector quantisation and signal compression*, Kluwer Academic Publishers, 1992.

[10] I. H. Witten and E. Frank, *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, 2000.

[11] Cubist, (version 1.13), Rulequest Research, www.rulequest.com.