Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Musical Onset Detection using MPEG-7 Audio Descriptors

D. Smith (1), E. Cheng (2) and I. S. Burnett (2)

(1) CSIRO ICT Centre, Hobart Tasmania, Australia
(2) RMIT University, Melbourne Victoria, Australia

PACS: 43.75.Zz

ABSTRACT

An onset detection system that exploits MPEG-7 audio descriptors is proposed in this paper, with investigations into the feasibility of MPEG-7 based onset detection performed across a diverse database of music. Detection functions were developed from both individual MPEG-7 descriptors and combinations of descriptors (joint detection functions). The results indicated that individual descriptors could achieve respectable detection performance (maximum F-measure of 0.753) with basic waveform features. Average detection performance could be improved by up to 11.2%, however, when joint detection functions were comprised of diverse combinations of MPEG-7 descriptors. This may be attributed to the increased capability of detection functions, composed of different spectral and temporal features, in capturing the variation in onset characteristics from different musical styles. It is thus concluded that the proposed onset detection system could be plausibly integrated into an existing MPEG-7 audio analysis system with minimal computational overhead.

INTRODUCTION

Musical onset detection is the process of identifying the transient sections of a musical composition associated with the beginning of a note. Onset detection is an important preprocessing step in many advanced music processing tasks such as music editing, beat tracking, tempo estimation, music retrieval and automatic score transcription. The detection of musical onsets, however, is a non-trivial problem: musical signals are usually composed of complex mixtures of overlapping notes and instruments that can be pitched, non-pitched, harmonic, or inharmonic. For example, the onset of a pitched, harmonic instrument such as a violin exhibits vastly different characteristics to the onset of an unpitched, inharmonic instrument such as a snare drum.

A number of approaches have been introduced for the detection of musical onsets, and the typical paradigm is composed of two distinct stages: estimating detection functions and the onset detection (peak-picking) algorithm. The detection functions reduce the original signal into functions that are better suited for identifying onset transients, where peak-picking algorithms are then applied to identify the time points in the detection functions corresponding to musical onsets.

Onset detection algorithms have previously been developed using a number of detection functions that individually track changes in features extracted from the temporal and frequency signal domain, such as spectral difference [1][2], phase deviation [1][3], and energy [4]. However, previous approaches to onset detection are far from perfect: when detection functions only track one feature at a time, it is difficult to capture the variation of onsets in different types of music. For example, temporal features exhibit improved onset detection performance with inharmonic and unpitched percussion instruments, whilst spectral features are well suited to more pitched, harmonic instruments [1]. Thus, combinations of detection functions, such as energy and phase [5][6], have been proposed to improve upon using the detection functions alone. Hybrid frequency subband approaches have also investigated energy changes at high-frequencies augmented with low frequency spectral changes [7].

Given the varied temporal/spectral detection functions currently utilised for musical onset detection, this paper proposes and performs a preliminary study into the application of MPEG-7 audio descriptors [8][9] to form onset detection functions. MPEG-7 is an ISO standard for multimedia content description composed of temporal and spectral descriptors ranging from low level features (e.g., spectral centroid, log attack time) to more music specific features (e.g., audio signature, percussive instrument timbre) [10]. Such a range of temporal, spectral and musical descriptors is thus well suited to detect a variety of musical onsets.

Furthermore, as an international standard, MPEG-7 is utilised in various audio analysis applications such as audio classification/retrieval [11], audio sports event detection [12], general sound recognition [13], audio fingerprinting [14], analysis of environmental sounds [15], speaker recognition [16], query-by-humming [17], musical instrument classification [18] and analysis [19], and audio database management [20]. Consequently, with the breadth of applications already using MPEG-7 descriptors, the proposed onset detection approach could be integrated into an existing MPEG-7 audio analysis system to allow for audio analysis tasks to be comprehensively and efficiently performed in a unified framework.

In the remainder of this paper, background into the MPEG-7 audio description framework is presented. The proposed mu-

sical onset detection system that employs MPEG-7 audio descriptors is then described. Experimental results of this onset detection system are presented and discussed: in particular, the performance of the detection system with individual MPEG-7 descriptors and different subsets of the MPEG-7 descriptors were evaluated and compared. The paper concludes with indications into the future work to be performed

MPEG-7 AUDIO DESCRIPTORS

MPEG-7 is an international standard (ISO/IEC 15938) [9] for multimedia content description consisting of 12 Parts encompassing the description frameworks for audio, voice, video, images, graphs and 3D models. The audio description framework forms Part 4 of the standard (ISO/IEC 15938-4) [8][9], and includes a set of Low-Level Descriptors (LLD) that characterise the temporal and frequency signal features, in addition to high-level Description Schemes (DS) that are more application-specific e.g., audio signature or instrument timbre.

The following subset of the available MPEG-7 LLD and high-level DS were applied to musical onset detection in this paper, chosen due to their applicability to (monophonic) musical signal onsets (with abbreviations are denoted in brackets):

- AudioWaveform LLD (Maximum – AXV, Minimum – AMV);
- AudioPower LLD (AP);
- AudioSpectrumCentroid LLD (ASC);
- AudioSpectrumFlatness LLD (ASF);
- AudioSpectrumSpread LLD (ASS);
- AudioFundamentalFrequency LLD (F0);
- AudioSignature DS (Mean – ASM, Variance – ASV);
- InstrumentTimbre DS (HarmonicSpectralCentroid - HSC, HarmonicSpectralDeviation - HSD, HarmonicSpectralSpread - HSS).

AudioWaveform LLD

The AudioWaveform LLD describes the temporal signal waveform in an efficient manner, to enable a 'summary' display of an audio file. The LLD attributes utilised in the proposed musical onset detection system are the signal maximum (AXV) and minimum values (AMV) from each temporal frame.

AudioPower LLD

The AudioPower LLD describes the instantaneous power *P(t)* for the input signal *s(t),* temporally smoothed over each frame *w*: $P_w(t) = \left|S_w(t)\right|^2$.

AudioSpectrumCentroid LLD

The AudioSpectrumCentroid LLD describes the centre of gravity of the log-frequency power spectrum, indicating the dominance of low or high frequency content. The power spectrum coefficients are given by:

$$P_x(k) = \frac{1}{lw * NFFT}\left|X_w(k)\right|^2$$

where *lw* is the window length, *NFFT* the Fast Fourier Transform (FFT) length, and $X_w(k)$ the FFT coefficients of the w^{th} window in the k^{th} FFT frequency bin. As specified in the MPEG-7 standard, power spectrum coefficients below 62.5Hz are summed to give a modified power spectrum *P'(n)*:

$$bound = floor\left(\frac{62.5 \times NFFT}{FS}\right); P'_w(0) = \sum_{k=0}^{bound} P_w(k), f(0) = 31.25$$

$$P'_w(n) = P_w(n + bound), f(n) = (n + bound)\frac{FS}{NFFT}$$

where *FS* is the sampling frequency and $n = 1,..,\frac{NFFT}{2} - bound$. The spectrum centroid is then calculated as:

$$C_w = \frac{\sum_n \log_2(\frac{f(n)}{1000}) P'_w(n)}{\sum_n P'_w(n)}$$

AudioSpectrumFlatness LLD

The AudioSpectrumFlatness LLD describes how 'flat' the signal spectrum is, where noise-like signals tend to exhibit a more broadband 'flat' spectrum whilst tonal signals are more narrowband thus deviating from a flat spectral shape.

The AudioSpectrumFlatness LLD, hereby abbreviated as *SFM*, is calculated in frequency bands of logarithmic ¼ octave resolution as the ratio of the geometric to arithmetic mean of the power spectrum coefficients:

$$SFM_b = \frac{\sqrt[ih(b)-il(b)+1]{\prod_{i=il(b)}^{ih(b)} \left|X_w(i)\right|^2}}{\frac{1}{ih(b)-il(b)+1}\sum_{i=il(b)}^{ih(b)} \left|X_w(i)\right|^2}$$

where each frequency band *b* is bounded by the *il* and *ih* power spectrum coefficients.

AudioSpectrumSpread LLD

The AudioSpectrumSpread LLD describes the second moment of the log-frequency power spectrum. Calculated as the Root Mean Square (RMS) deviation of the log-frequency spectrum with respect to its centre of gravity, the AudioSpectrumSpread indicates how distributed the signal power spectrum is from the centroid, which can indicate the presence of narrowband or wideband signals. Hence, the AudioSpectrumSpread calculation is an extension to the AudioSpectrumCentroid descriptor:

$$S_w = \sqrt{\frac{\sum_n ((\log_2(\frac{f(n)}{1000}) - C_w)^2 P'_w(n))}{\sum_n P'_w(n)}}$$

AudioFundamentalFrequency LLD

The AudioFundamentalFrequency LLD is the estimated fundamental frequency (F0), where the estimation technique is not specified within the MPEG-7 standard and is thus open to implementation. In this paper, the technique based on the normalised auto-correlation is used, where the first maximum of the auto-correlation function indicates the fundamental period and hence F0.

Audio Signature DS

The Audio Signature DS is a compact representation intended for use in automatic audio signal identification. The AudioSpectralFlatness LLD is statistically temporally summarised across adjacent frames to form the signature, according to a decimation factor. However, in this paper, to ensure that a signature descriptor is obtained for each frame for onset de-

tection, the SpectralFlatness is summarised across frequency bands. That is, the statistical summarisation is a reduction in frequency, rather than temporal, resolution.

Instrument Timbre DS

The Instrument Timbre Description Scheme encompasses the HarmonicSpectralCentroid, HarmonicSpectralDeviation, HarmonicSpectralSpread, HarmonicSpectralVariation, LogAttackTime, SpectralCentroid, and TemporalCentroid descriptors. As these descriptors are defined to be calculated (or averaged) over the duration of the audio signal (which would not be useful for onset detection), this paper utilised the *instantaneous* HarmonicSpectralCentroid (HSC), HarmonicSpectralDeviation (HSD), and HarmonicSpectralSpread (HSS) 'pre-descriptor' values calculated for each analysis frame as part of the harmonic descriptor calculations. Due to the harmonic nature of these three instantaneous pre-descriptors, the frequency location and amplitude of harmonic peaks in the signal must firstly be identified; harmonic peaks are defined as spectral peaks that occur at multiples of F0, and can thus be identified from the windowed FFT of each analysis frame.

Instantaneous HarmonicSpectralCentroid (HSC): defined as the weighted mean of the harmonic spectral peaks, calculated for each frame *w* according to:

$$IHSC_w = \frac{\sum_{h=1}^{H} f_w(h) \cdot A_w(h)}{\sum_{h=1}^{H} A_w(h)}$$

where *A* and *f* denote the amplitude and frequency of the estimated harmonic peaks, respectively, and *H* is the total number of harmonics considered.

Instantaneous HarmonicSpectralDeviation (HSD): defined as the spectral deviation of the log-frequency spectrum amplitude from the spectral envelope:

$$IHSD_w = \frac{\sum_{h=1}^{H} |\log_{10}(A_w(h)) - \log_{10}(SE_w(h))|}{\sum_{h=1}^{H} \log_{10}(A_w(h))}$$

where *SE* denotes the spectral envelope around the h^{th} harmonic peak.

Instantaneous HarmonicSpectralSpread (HSS): defined as the weighted standard deviation of the harmonic spectral peaks, normalised by the Instantaneous HarmonicSpectralCentroid:

$$IHSS_w = \frac{1}{IHSC_w} \sqrt{\frac{\sum_{h=1}^{H} A_w^2(h) \cdot [f_w(h) - IHSC_w]^2}{\sum_{h=1}^{H} A_w^2(h)}}$$

PROPOSED ONSET DETECTION FUNCTIONS

Individual MPEG-7 Descriptor Functions

The twelve MPEG-7 descriptors explained in the previous section were computed using the MPEG-7 Audio Reference Software Toolkit (for Matlab) [21] to perform onset detection in the experiments of this paper. To conform to the MPEG-7 standard, this paper utilised window and hop sizes of 30ms and 10ms, respectively.

The detection functions $\phi(n)$ that were used to identify onsets were formed by computing the difference between adjacent windows of the MPEG-7 descriptors as:

$$\phi_i(n) = |(d_i\varphi(n) - \mu_i)/\sigma_i|$$

$$d_i\varphi(n) = \varphi_i(n) - \varphi_i(n-1)$$

where *i*=1 to 12, μ_i is the mean and σ_i is the standard deviation of the difference function $d_i\varphi(t)$ used for normalisation. A peak picking algorithm was then used to identify the windows in $\phi(n)$ that contained local maxima.

Onsets were estimated at the windows of local maxima of $\phi(n)$ that were greater than the adaptive threshold ($T_i(n_m)$). $T_i(n_m)$ is defined as:

$$T_i(n_m) = \mu_i + m_i(n_m)$$

$$m_i(n_m) = median\{[\phi_i(n_m - W/2)...\phi_i(n_m + W/2)]\}$$

where n_m represents the windows with local maxima in the function, W is the number of windows used to compute the median and μ_i is the mean of ϕ_i. This threshold is related to [1], however, we replace their constant term with μ_i as this eliminates the need to set any parameters and enables the threshold to account for variations between signals. If onset detection was being performed with the MPEG-7 descriptors in real time (as opposed to post-processing as considered here), the function mean in the threshold would have to be replaced with a running average of detection functions across signal windows.

Joint MPEG-7 Descriptor Functions

Detection functions combining multiple MPEG-7 descriptors were formed by obtaining individual MPEG-7 detection functions and their thresholds (using the methodology outlined above) and then computing the following:

$$\text{if } \phi_i(n_m) \geq T_i(n_m), \text{ then } v_i(n_m) = 1$$

$$\text{if } (\sum_{i=1}^{C} v_i(n)) > L, \quad \text{then } n \in \text{onset}$$

where *C* is the number of functions and *L* is the number of functions that must have an onset in order to classify the current window n as an onset. The vector $v_i(n)$ is initialised as a 1x*N* of zeros, where *N* is the number of windows in $\phi_i(n)$.

EXPERIMENTAL EVALUATION

Test Data

The data set used in this experiment was a database of music obtained from [1]. This database was chosen as it contained 1065 hand-annotated onsets from a range of commercial and non-commercial recordings of various musical styles. The data set consisted of 23 monophonic recordings sampled at 44.1kHz, ranging in duration from 1.3 seconds to 1 minute.

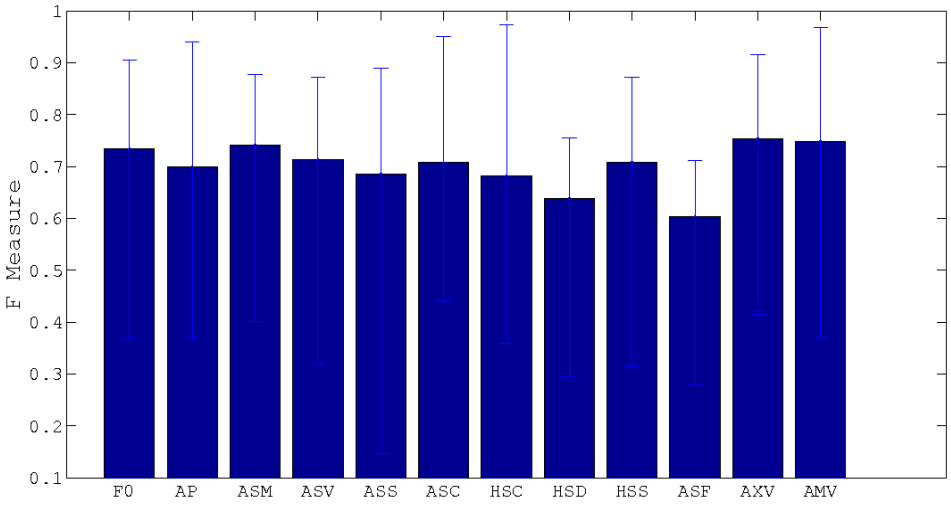

Figure 1. The average F-measure and 95% confidence interval across the data set for each of the 12 MPEG-7 descriptor detection functions.

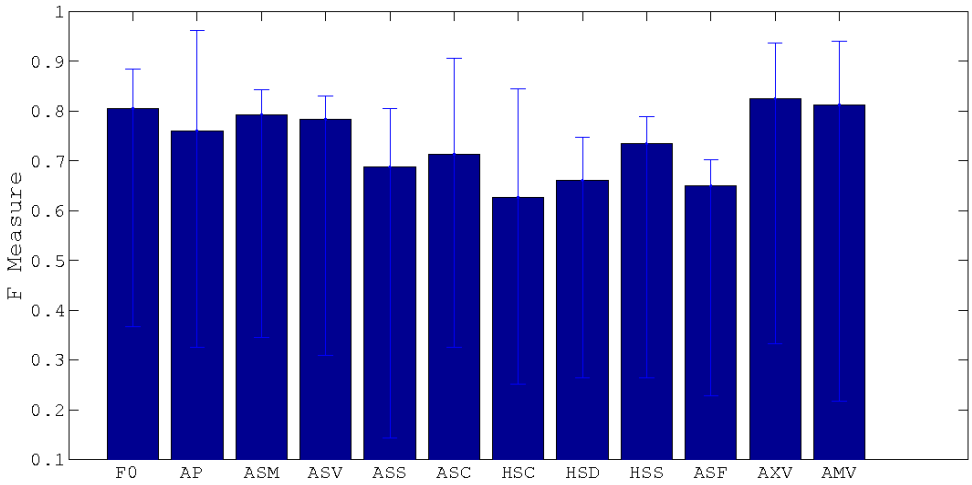

Figure 2. The average F-measure and 95% confidence intervals across each of the 12 MPEG-7 descriptor functions for the pitched percussive class of recordings.

For analysis of the proposed MPEG-7 detection functions, the data set was categorized into four musical classes: pitched percussive (pp - e.g., piano), non-pitched percussive (npp - e.g., snare drum), pitched non-percussive (pnp - e.g violin)., and complex (com – e.g., pop music).

Evaluation Metric

The F-measure was used to evaluate the onset detection performance in the experiments of this paper. The statistic can be defined as:

$$F = \frac{2 \cdot PC \cdot RC}{PC + RC}$$

where *PC* is the precision rate of the detection and *RC* is the recall rate of the onset estimates. These statistics are defined as:

$$RC = \frac{p}{a}; PC = \frac{p}{t}$$

such that *p* is the number of correctly identified onsets, *t* is the total number of estimated onsets and *a* is the true number of onsets. The F-measure can range between 0 and 1 with a value of 1 indicating that all onsets have been detected without false positives and a value of 0 indicating that there were not any correct onsets detected.

In this paper, an onset estimate was considered to be correctly detected when it was estimated within 50ms of the annotated onset positions of the data set. Further, a window was placed around the annotated onset positions to compensate for the error associated with the hand labelling of onsets [1].

RESULTS AND DISCUSSION

Individual MPEG-7 Descriptor Functions

Figure 1 shows the detection performance of the proposed MPEG-7 detection functions across the whole test dataset, indicating that all twelve proposed MPEG-7 detection functions achieved a moderate level of detection performance. The detection functions of the Audio Waveform descriptors, AMV and AXV, achieved the highest average F-measures of 0.748 and 0.753, respectively. The detection function of the ASF descriptor produced the lowest F-measure of 0.603. Although the average F-measures of the better performing descriptors were respectable, the spread of performance (wide confidence intervals) suggests that the variability of

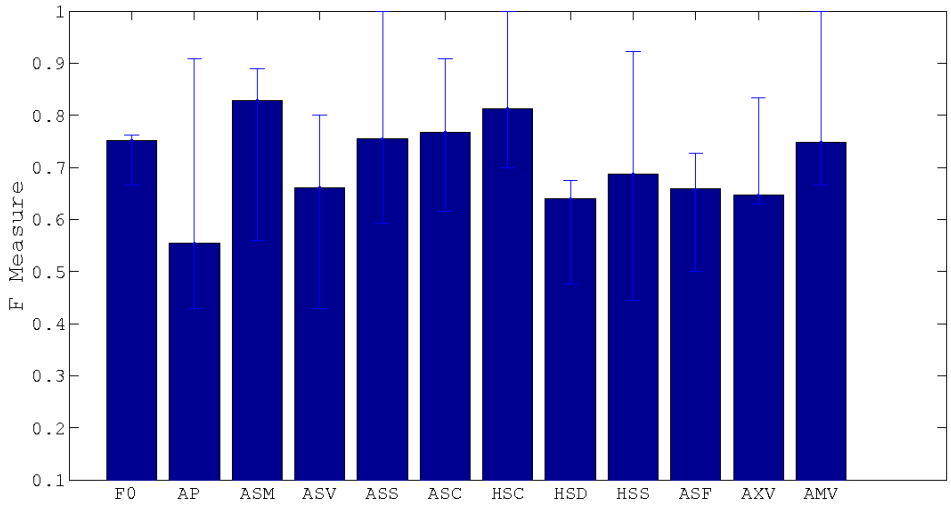

Figure 3. The average F-measure and 95% confidence intervals across each of the 12 MPEG-7 descriptor functions for the pitched non-percussive class of recordings.

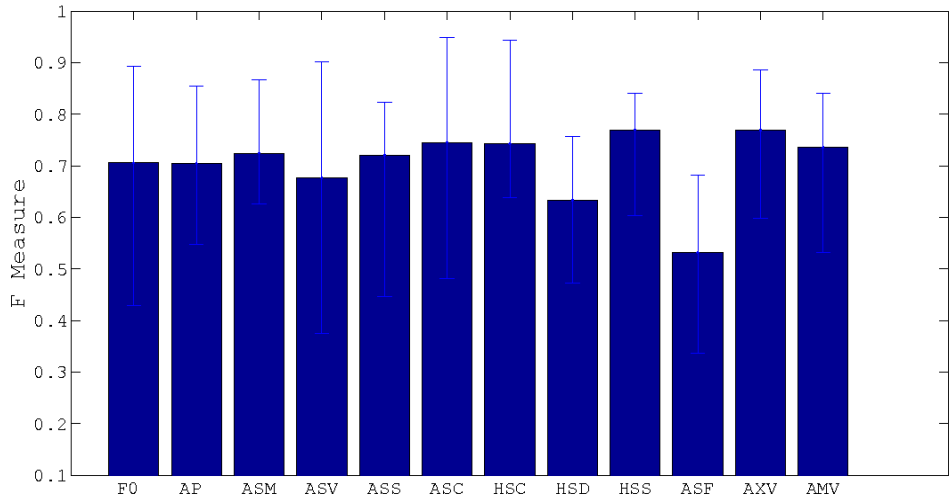

Figure 4. The average F-measure and 95% confidence intervals across each of the 12 MPEG-7 descriptor functions for the non-pitched percussive class of recordings.

onset detection accuracy would be considerable across different music styles.

Figure 2 shows the average F-measures of the pitched-percussive music class. The detectors of the AudioWaveform and F0 descriptors performed the best with average F-measures of 0.825 and 0.805, respectively. The strong performance of the F0 based descriptor was expected, given that the signals were pitched. The harmonic-based detection functions produced relatively low F-measures when compared to the spectral and energy based descriptor functions (i.e. AP, ASV, and ASM), despite the signals in this class having a harmonic structure that changes with note transitions. In fact, the HSC was the worst performing descriptor function with an F-measure of 0.628.

In the pitched-non-percussive class of music, the spectral based ASM and HSC descriptors were shown to produce the best performing detection functions in Figure 3, with average F-measures of 0.829 and 0.813, respectively. This was in contrast to the pitched-percussive class where the HSC descriptor was the worst performing function. The energy based descriptor AP provided the weakest detection performance with an F-measure of 0.555. This was not unexpected, given that note transitions of non-percussive instruments aren't associated with sudden changes in energy.

Figure 4 shows the average F-measures of the non-pitched-percussive music class. The HSS and AXV descriptors are shown to achieve the highest F-measures for this class of music with 0.770 and 0.769, respectively. Unexpectedly, it didn't appear that the HSS and the InstrumentTimbre harmonic-based descriptors in general, exhibited a drop-off in detection performance despite the non-harmonic structure of the signals. The ASF descriptor was the worst performing function with a poor F-measure of 0.531. We hypothesize that this is due to percussive non-pitched instruments commonly having a flat spectrum, which reduces the ability of the ASF function to discriminate between note transitions.

Figure 5 shows the average F-measures of the complex music class with all descriptor functions being consistently low with average values ranging between 0.600 for ASF and 0.682 for HSC. Furthermore, the confidence intervals across all descriptors were relatively small, indicating that the onset detection performance was poor across a significant proportion of the complex signals. This suggests that complex music is the most challenging class to perform onset detection for using a wide range of detection functions.

In general, from the results presented in Figures 1-5, the AudioWaveform descriptors were the best choice for percussive instruments that exhibited salient changes in energy at

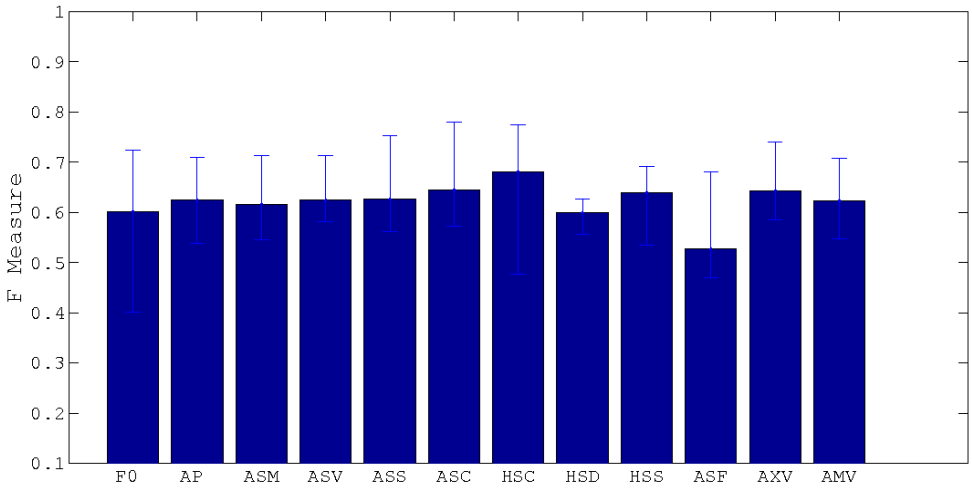

Figure 5. The average F-measure and 95% confidence intervals across each of the 12 MPEG-7 descriptor functions for the complex class of recordings.

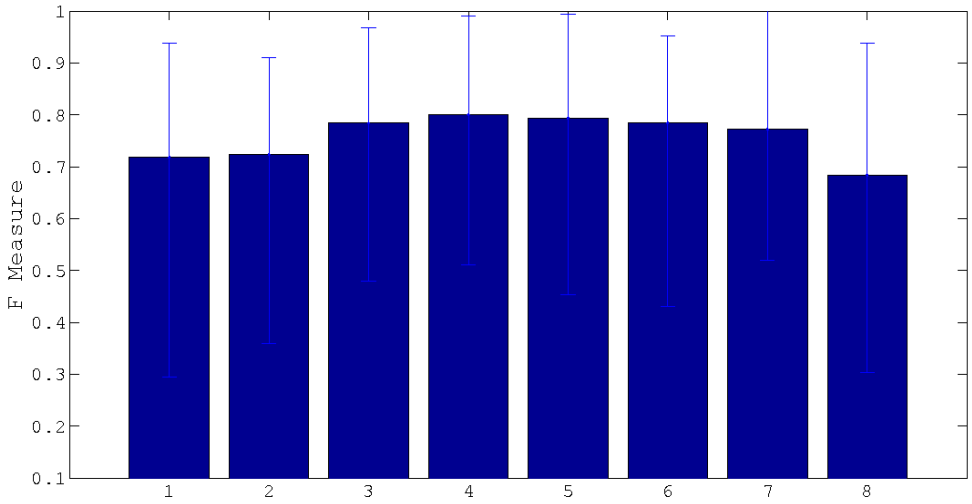

Figure 6. The average F-measure performance of 8 different joint MPEG-7 descriptor detection functions.

onsets, whether or not the instruments were pitched. These results illustrate that very simple statistics can be effective in performing onset detection across time, particularly when signals exhibit significant energy changes. For non-percussive signals, however, simple spectral-based descriptors, such as the harmonic spectral centroid and signature mean (i.e., average spectral flatness), were found to be the most appropriate.

Combinations of MPEG-7 Descriptor Functions

Eight joint detection functions were evaluated in this paper, composed of the following the constituent descriptors:

1. AXV and AMV
2. AXV, AXV and ASM
3. AXV, AXV, ASM and F0
4. AXV, AMV, F0, ASM and HSS
5. AXV, AMV, F0, ASM, ASV and HSS
6. All 12 of the MPEG-7 descriptors
7. AXV, HSS, F0, AP and ASM
8. AXV, HSC and ASM

The first five of these joint functions consisted of combinations of the best performing descriptors from Figure 1. Function 1 consisted of the two highest performing descriptors (AXV AMV); each of the next four functions then incorporated its predecessor's descriptors plus the next highest performing descriptor. The joint function 6 was comprised of all 12 of the MPEG-7 descriptor functions. Function 7 incorporated five different types of descriptors that possessed spectral, temporal, energy, harmonic and pitch based features. Function 8 was comprised of the best performing descriptor functions for each of the four sub-classes of music.

The onset detection performance of the detection functions of the joint descriptors are shown in Figure 6. Five of the eight joint detection functions improved the detection performance of the highest performing individual descriptor AMX by between 2.5% and 6.4%. There was also an average 16% reduction in the variance of the joint models when compared to the AMX descriptor. These results illustrate that using a combination of MPEG-7 descriptors can improve the onset detection performance across a database of varied musical styles, as multiple features of signals that are associated with different types of onsets are exploited.

A comparison of the joint detection functions in Figure 6 indicates that combinations of the two or three highest performing descriptors (functions 1 and 2) achieved lower detection performance than their underlying constituent descriptor functions. This could be because of the homogeneity of the features used in these particular joint functions: the features were either solely or predominately comprised of AMV and AMX, both of which are similar temporal-based features.

Combinations of four or more of the highest performing MPEG-7 descriptors functions (functions 3 to 5), however, were advantageous over individual descriptors. The joint function that provided the highest F measure of 0.801 was comprised of the top five performing descriptors AMX, AMV, F0, ASM and HSS. This combination offered an improvement in detection of between 6.4% and 11.2% over the individual descriptor functions at minimum. Figure 6 shows that there was a decline in the F-measure, however, as additional descriptors were incorporated into the joint functions. The average F-measure for the joint function consisting of all 12 descriptors was 0.784, suggesting that the inclusion of weaker performing descriptors into the joint detection function reduced the overall detection performance.

The average detection performance of function 7, which was comprised of a collection of five descriptors with unique features (i.e. spectral, power, pitch, harmonic and time features), was shown to be superior to its constituent descriptors with an F-measure of 0.772. Although this further confirms that feature diversity is advantageous, the detection performance of function 7 was lower than other joint functions that possessed more homogenous features. This result confirms that it is not only the diversity of the MPEG-7 features, but the performance of its underlying functions that contributes to the joint detection performance.

CONCLUSION

An onset detection system utilising MPEG-7 audio descriptors was proposed in this paper. The detection system was evaluated across a range of 12 individual MPEG-7 descriptor functions and 8 joint functions combining sets of MPEG-7 descriptors. Reasonable detection performance was obtained by using a number of different individual descriptors: basic temporal AudioWaveform descriptors achieved the highest average detection performance with an average F-measure of 0.753. It was also found that the detection performance of joint functions were superior to their underlying constituent descriptors with an advantage of at least 6.4%. Furthermore, the performance of the joint detection functions was more robust than constituent functions given a significant drop in performance variation between musical styles. The best performing joint detection functions were found to be composed of diverse combinations of MPEG-7 descriptors (i.e., spectral, temporal, harmonic and pitch based features) that individually achieved relatively good detection performance.

Results indicate that it would be possible for the proposed onset detection system to be efficiently integrated into an existing MPEG-7 audio analysis system with minimal computational overhead. The additional performance advantage gained by using sets of MPEG-7 descriptors would outweigh the additional complexity in forming the detection function in the system.

Future work will focus upon further improving detection performance, in particular, reducing the variance associated with the accuracy of onset detection across signals. One potential avenue will be to include high-level audio descriptor schemes of MPEG-7 into the detection system. This could include the Melody Contour Description Scheme that defines rhythmic and beat information. The major challenge to address with this approach is the integration of higher level information into the detection system, particularly given that these schemes often operate upon much longer temporal scales than the basic descriptors utilised in this paper. A further improvement could utilise the MPEG-7 descriptors to perform a pre-processing task to automatically classify the musical styles/genre and then deploy MPEG-7 descriptor detection functions appropriate to the identified musical style.

REFERENCES

1 J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. B. Sandler, “A Tutorial on Onset Detection in Music Signals,” *IEEE Trans. Speech and Audio Processing* **13** (5), 1035-1047 (2005).

2 P. Masri, “Computer Modeling of Sound for Transformation and Synthesis of Musical Signal,” *Ph.D. thesis, University of Bristol* (1996).

3 J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* **5**, 441-444 (2003).

4 F. Jaillet, X. Rodet, “Detection and modelling of fast attack transients,” *Proc. Int. Computer Music Conference (ICMC)* (2001).

5 J. Bello, C. Duxbury, M. Davies, and M. Sandler, “On the use of phase and energy for musical onset detection in the complex domain,” *IEEE Sig. Proc. Letters* **11** (6), 553–556 (2004).

6 C. Duxbury, J. Bello, M. Davies, and M. Sandler, “Complex Domain Onset Detection for Musical Signals,” *Proc. Int. Conf. Digital Audio Effects (DAFx)* (2003).

7 C. Duxbury, M. Sandler, and M. Davies, “A hybrid approach to musical note onset detection,” *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 33–38 (2002).

8 ISO/IEC 15938-4 (2002). Information Technology - Multimedia Content Description Interface – Part 4: Audio.

9 H. Kim, N. Moreau, T. Sikora, *MPEG-7 Audio and Beyond* (Wiley, West Sussex, 2005).

10 ISO/IEC 15938 (2002). Information Technology – Multimedia Content Description Interface.

11 H. Kim, N. Moreau, T. Sikora, “Audio classification based on MPEG-7 spectral basis representations,” *IEEE Trans. Circuits and Systems for Video Technology*. **14**, 716-725, (2004).

12 Z. Xiong, R. Radhakrishnan, A. Divakaran, T. Huang, “Audio-based highlights extraction from baseball, golf and soccer games in a unified framework,” *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP ‘03)* **5,** 628-631 (2003).

13 M. Casey, “MPEG-7 sound recognition tools,” *IEEE Trans. on Circuits and Systems for Video Technology* **11**, 737-747 (2001).

14 O. Hellmuth, E. Allamanche, M. Cremer, H. Grossmann, H. Holger, J. Herre, T. Kastner, “Using MPEG-7 Audio Fingerprinting in Real-World Applications,” *Proc. AES 115th Convention* (Oct. 2003)

15 D. Mitrovic, M. Zeppelzauer, H. Eidenberger, “Analysis of the Data Quality of Audio Descriptions of Environmental Sounds,” *Journal of Digital Information Management* **5** (2), 48-55 (April 2007).

16 H. Kim, N. Moreau, T. Sikora, “Speaker Recognition Using MPEG-7 Descriptors,” *Proc. Eurospeech* (2003).

17 J-M. Batke, G. Eisenberg, P. Weishaupt, T. Sikora, “A Query by Humming System using MPEG-7 Descriptors,” *Proc. AES 116th Convention* (May 2004).

18 P. Szczuko, P. Dalka, M. Dabrowski, B. Kostek, “MPEG-7-based Low-Level Descriptor Effectiveness in the Automatic Musical Sound Classification,” *Proc. AES 116th Convention* (May 2004).

19 E. Łukasik, “MPEG-7 Musical Instrument Timbre Descriptors Performance in Discriminating Violin Voices,” Proc. IEEE Workshop on Signal Processing, 87-90 (2004).

20 M. Jacob, “Managing Large Sound Databases using MPEG7,” *Proc. AES 25th Int. Conf: Metadata for Audio* (June 2004).

21 MPEG-7 Multimedia Software Resources Available: http://mpeg7.doc.gold.ac.uk/