DETECTING MUSIC IN AMBIENT AUDIO BY LONG-WINDOW AUTOCORRELATION

Keansub Lee and Daniel P. W. Ellis

LabROSA, Department of Electrical Engineering
Columbia University, New York, NY 10027 USA
{kslee, dpwe}@ee.columbia.edu

ABSTRACT

We address the problem of detecting music in the background of ambient real-world audio recordings such as the sound track of consumer-shot video. Such material may contain high levels of noises, and we seek to devise features that will reveal music content in such circumstances. Sustained, steady musical pitches show significant, structured autocorrelation at when calculated over windows of hundreds of milliseconds, where autocorrelation of aperiodic noise has become negligible at higher-lag points if a signal is whitened by LPC. Using such features, further compensated by their long-term average to remove the effect of stationary periodic noise, we produce GMM and SVM based classifiers with high performance compared with previous approaches, as verified on a corpus of real consumer video.

Index Terms— Speech analysis, Music, Acoustic signal detection, Correlation

1. INTRODUCTION

Short video clips are in some cases replacing still-image snapshots as a medium for the casual recording of daily life. While the thousands of digital photos in a typical user's collection already present a serious navigation and browsing challenge, video clips, which may not be well represented by a single thumbnail image, can make things still more difficult.

However, video clips contain much richer information than single images, and consequently present many new opportunities for the automatic extraction of information that can be used in intelligent browsing systems. We are particularly interested in exploiting the acoustic information – the soundtrack – that is available for video, and in seeing what useful information can be reliably extracted from these kinds of data.

One attribute that we see as both informative and useful to users, and at the same time technically feasible, is the detection of background music. For instance, if a user is searching for the video clip of a certain event, they are likely to able to remember (or guess) if there was music in the background, and thereby limit the scope of a search. In a manual labeling of a database of over 1000 video clips recorded by real users of current digital cameras (which include video capability), approximately 18% were found to include music – enough to be a generally-useful feature, while still retaining some discriminative power.

There has been a substantial amount of work relating to the detection of music in audio, or the discrimination between a few categories such as speech and music, including [1, 2, 3, 4]. However, the soundtrack of 'consumer video' has many characteristics that distinguish it from the broadcast audio that has most commonly been considered in this work: Casual recordings made with small, hand-held cameras will very often contain a great deal of spurious, non-stationary noise such as babble, crowd, traffic, or handling artifacts. This unpredictable noise can have a great impact on detection algorithms, particularly if they rely on the global characteristics of the signal (e.g. the broad spectral shape encoded by MFCC features) which may now be dominated by noise.

In trying to design robust features, we focus on the two key characteristics of music worldwide: Pitch and Rhythm. *Pitch* refers to the perceived musical notes that build up melodies and harmony, and is generally conveyed by locally-periodic signals (thus possessing a spectrum with harmonic peaks); musical instruments are usually designed to have relatively stable periods, and musical notes typically last for hundreds of milliseconds before the pitch is changed. *Rhythm* is the regular temporal structuring of note events giving rise to a sense of beat or pulse, usually at several hierarchically-related levels (beat, bar, etc.). While a given musical instance may lack clear pitch (e.g. percussion music) or a strong rhythm (e.g. an extremely 'romantic' piano style), it is difficult to imagine music possessing neither.

In the next section, we describe a music detection feature for detecting the stable periodicities of pitch that is robust to high levels of background noise. This feature, combined with a rhythm-detection feature based on the beat tracker of [5], is evaluated in section 3, and compared to previous music detection features. We discuss the results in section 4.

This work was supported by the NSF (grant IIS-0238301), the Eastman Kodak company, and EU project AMIDA (via ICSI).

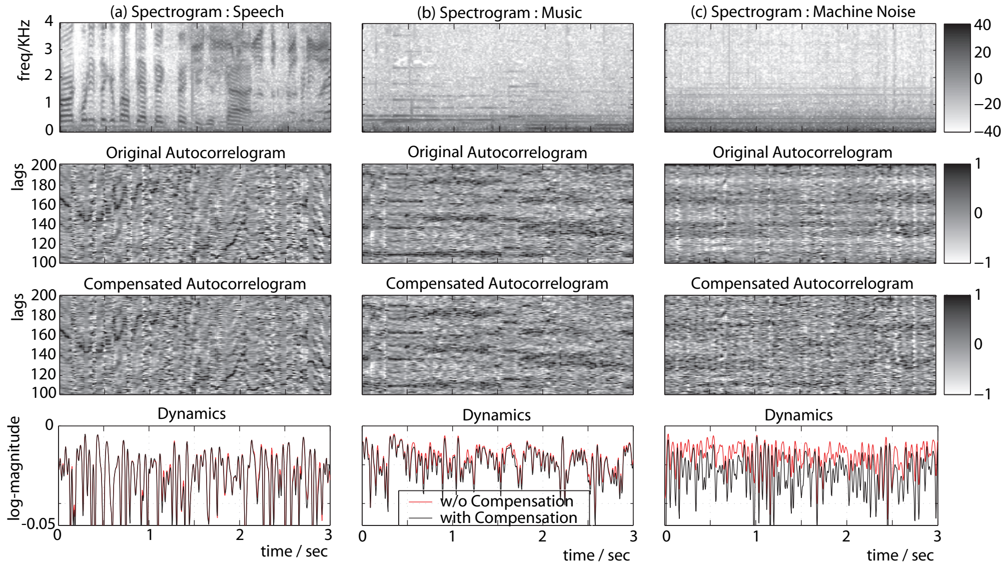

Fig. 1. Examples of noisy speech, music and machine sound from a consumer audio recording.

2. MUSICAL PITCH DETECTION

Our strategy for detecting musical pitches is to identify the autocorrelation function (ACF) peaks resulting from the periodic, pitched energy that are stationary for around 100..500 ms, but to exclude aperiodic noise and stationary periodicity arising from background noise. Whitening by Linear Predictive (LP) inverse filtering prior to ACF concentrates aperiodic noise energy around zero lag, so we use only higher-lag coefficients to avoid this energy. Calculating the ACF over 100 ms windows emphasizes periodicities stable on that time scale, but we then subtract the long-term average ACF to remove any stationary, periodic background. Finally, the stability (or dynamics) of pitch content is estimated by a feature composed of the cosine similarity between successive frames of the compensated ACF.

2.1. LPC Whitening and ACF

Mono input recordings are resampled to 16 kHz, and fit with a 12th order LPC model over 64 ms windows every 32 ms. Further processing is applied to residual of this modeling, which is a spectrally flat (whitened) version of the original signal, preserving any pitch-rate periodicity.The short-time ACF $r_{ee}(n,\tau)$ for each LPC residual envelope output $e(n)$ at a given time index n may be defined as:

$$r_{ee}(n,\tau) = \sum_{i=n+1}^{n+W} e(i)e(i+\tau) \tag{1}$$

where W is an integration window size, and $r_{ee}(n,\tau)$ is calculated over 100 ms windows every 5 ms for lag $\tau = 0 \ldots 200$ samples (i.e. up to 12.5 ms, for a lowest pitch of 80 Hz).

2.2. ACF Compensation

Assume that residual $e(n)$ consists of a clean musical signal $m(n)$ and a background aperiodic noise $a(n)$ and stationary periodic noise $b(n)$ i.e. $e(n) = m(n) + a(n) + b(n)$. If the noise $a(n)$ and $b(n)$ are zero-mean and uncorrelated with $m(n)$ and each other for large W, the ACF is given by:

$$r_{ee}(n,\tau) = r_{mm}(n,\tau) + r_{aa}(n,\tau) + r_{bb}(n,\tau) \tag{2}$$

To simplify notation, variables n and τ are henceforth dropped.

2.2.1. *Aperiodic Noise Suppression*

The effect of the LPC whitening is to concentrate the ACF of unstructured noise, r_{aa} at or close to the zero-lag bins. We can remove the influence of aperiodic noise from our features by using only the coefficients of r_{ee} for lag $\tau \geq \tau_1$ samples (i.e. in our system, $\tau \geq 100$).

$$r_{ee} = r_{mm} + r_{bb},\ for\ \tau \geq 100 \tag{3}$$

Once the low-lag region has been removed, ACF r_{ee} is normalized by its peak value $||r_{ee}||$ to lie in the range -1 to 1.

2.2.2. *Long-time Stationary Periodic Noise Suppression*

A common form of interference in environmental recordings is a stationary periodic noise such as the steady hum of a machine as shown in the third column of figure 1, resulting in

Table 1. Speech-music classification, with and without vocals (broadcast audio corpus, single Gaussian classifier with full covariance). Each value indicates how many of the 2.4 second segments out of a total of 120 are correctly classified as speech (first number) or music (second number). The best performance of each column is shown in bold. Features are described in the text.

Feature	Speech vs. Music w/ vocals	Speech vs. Music w/o vocals
Rth	96/120, 65/120	96/120, 62/120
mDyn	114/120, 99/120	114/120, 104/120
vDyn	89/120, 115/120	89/120, 116/120
4HzE	**106/120**, **118/120**	**106/120**, **120/120**
vFlux	106/120, 116/120	**106/120**, **120/120**
Rth+mDyn	111/120, 109/120	**111/120**, **114/120**
mDyn+vDyn	114/120, 101/120	114/120, 104/120
4HzE+vFlux	**104/120**, **118/120**	104/120, 120/120
Rth+mDyn+vDyn	112/120, 114/120	**112/120**, **117/120**
Rth+4HzE+vFlux	103/120, 119/120	103/120, 120/120
Rth+mDyn+4HzE	**108/120**, **119/120**	108/120, 120/120
Rth+mDyn+vFlux	108/120, 117/120	108/120, 120/120

ACF ridges that are not, in fact, related to music [6]. The ACF contribution of this noise r_{bb} will change very little with time, so it can be approximated as the long-time average of r_{ee} over M adjacent frames (covering around 10 second). We can estimate the autocorrelation of the music signal, $\hat{r}_{mm}$, as the difference between the local ACF and its long-term average,

$$\hat{r}_{mm} = r_{ee} - \gamma \cdot \hat{r}_{bb} = r_{ee} - \gamma \cdot avg\{r_{ee}\} \tag{4}$$

γ is a scaling term to accommodate the per-frame normalization of the high-lag ACF and is calculated as the best projection of the average onto the current frame:

$$\gamma = \frac{\sum_{\tau} r_{ee} \cdot avg\{r_{ee}\}}{\sum_{\tau} avg\{r_{ee}\}^2}, \; for \; \tau \geq 100 \tag{5}$$

This estimated music ACF $\hat{r}_{mm}$ is shown in the third row of figure 1.

2.3. Pitch Dynamics Estimation

The stability of pitch in time can be estimated by comparing temporally adjacent pairs of the estimated music ACFs:

$$\Upsilon(n) = S_{cos}\{\hat{r}_{mm}(n), \hat{r}_{mm}(n+1)\} \tag{6}$$

where S_{cos} is the cosine similarity (dot product divided by both magnitudes) between the two AC vectors. Υ is shown in the fourth row of figure 1. The sustained pitches of music result in flat pitch contours in the ACF, and values of Υ that approach 1, as shown in the second column of figure 1. By contrast, speech (column 1) has a constantly-changing pitch contour, resulting in a generally smaller Υ, and the initially larger Υ of stationary periodic noise from e.g. machine is attenuated by our algorithm (column 3).

3. EVALUATION

The pitch dynamics feature Υ was summarized by its mean (mDyn) and variance (vDyn) for the purpose of classifying clips. We compared these features with others that have been successfully used in music detection [2], namely the 4Hz Modulation Energy (4HzE), Variance of the spectral Flux (vFlux) and Rhythm (Rth) which we took as the largest peak value of normalized ACF of an 'onset strength' signal [5] over the tempo range (50-300 BPM).

Table 1 compares performance on a data set of random clips captured from broadcast radio, as used in [2]. The data was randomly divided into a 15 s segments, giving 120 for training and a 60 for testing (20 each of speech, music with vocals, and music without vocals). Classification was performed by a likelihood ratio test of single Gaussians fit to the training data. 4HzE and vFlux have the best performance among single features, but Rth + mDyn + vDyn has the best performance (by a small margin) in distinguishing speech from vocal-free music.

However, classification of clean broadcast audio is not the main goal of our current work. We also tested these features on the soundtracks of 1873 video clips from the YouTube [7], retured by consumer-relevant search terms such as 'animal', 'people', 'birthday', 'sports' and 'music', then filtered to retain only unedited, raw consumer video. Clips were manually sorted into 653 (34.9%) that contained music, and 1220 (65.1%) that did not. We labeled a clip as music if it included clearly-audible professional or quality amateur music (regardless of vocals or other instruments) throughout. These clips are recorded in a variety of locations such as home, street, park and restaurant, and frequently contain noise including background voices and many different types of a mechanical noise.

We used a 10 fold cross-validation to evaluate the performance in terms of the accuracy, d' (the equivalent separation of two normalized Gaussian distributions), and Average Precision (the average of the precision of the ordered returned list truncated at every true item). We compared two classifiers, a single Gaussian as above, and an SVM with an RBF kernel. At each fold, the classifier is trained on 40% of the data, tuned on 20%, and then are tested on the remaining 40% selected at random. For comparison,we also report the performance of the '1G+KL with MFCC' system from [8], which simply takes the mean and covariance matrix of MFCC features over the entire clip, and then uses an SVM classifier with a symmetrized Kullback-Leibler (KL) kernel.

As shown in table 2, the new mDyn feature is significantly

Table 2. Music/Non-music Classification Performance on YouTube consumer recordings. Each data point represents the mean and standard deviation of the clip-based performance over 10 cross-validated experiments. d' is a threshold-independent measure of the separation between two unit-variance Gaussian distributions. AP is the Average Precision over all relevant clips. The best performance of each column is shown in bold for the first three blocks.

	One Gaussian Classifier			SVM Classifier		
Features	Accuracy(%)	d'	AP(%)	Accuracy(%)	d'	AP(%)
Rth	$\mathbf{81.9 \pm 0.87}$	$\mathbf{1.85 \pm 0.06}$	$\mathbf{75.8 \pm 1.67}$	$\mathbf{82 \pm 0.87}$	$\mathbf{1.83 \pm 0.08}$	$\mathbf{80.9 \pm 1.02}$
mDyn	80.6 ± 0.73	1.67 ± 0.05	70.9 ± 2.08	81.1 ± 0.69	1.66 ± 0.06	77.5 ± 2.23
vDyn	63 ± 1.44	0.76 ± 0.08	47.7 ± 1.26	66.7 ± 1.37	0.57 ± 0.11	50.2 ± 2.47
4HzE	65.2 ± 1.08	0.87 ± 0.07	53.7 ± 1.62	68.6 ± 0.89	0.81 ± 0.07	53 ± 1.47
vFlux	61.5 ± 1.17	0.74 ± 0.09	52.4 ± 1.85	67.3 ± 1.36	0.68 ± 0.09	50.2 ± 2.68
Rth+mDyn	$\mathbf{86.9 \pm 0.84}$	$\mathbf{2.17 \pm 0.08}$	$\mathbf{86.1 \pm 0.94}$	$\mathbf{88.6 \pm 0.55}$	$\mathbf{2.36 \pm 0.07}$	$\mathbf{91.7 \pm 0.69}$
4HzE+vFlux	63.9 ± 1.39	0.79 ± 0.07	53.1 ± 2.3	68.5 ± 0.77	0.76 ± 0.07	51.9 ± 1.98
Rth+mDyn+vDyn	89.9 ± 0.67	2.49 ± 0.08	88.2 ± 1.17	90.7 ± 0.81	2.61 ± 0.1	$\mathbf{94.8 \pm 0.67}$
Rth+4HzE+vFlux	83 ± 1.32	1.9 ± 0.15	80.5 ± 1.8	85.1 ± 0.91	2.02 ± 0.08	84.7 ± 2
Rth+mDyn+4HzE	88.9 ± 1	2.4 ± 0.12	88 ± 1.16	90.6 ± 1.06	2.57 ± 0.13	94 ± 1.25
Rth+mDyn+vFlux	$\mathbf{90 \pm 0.72}$	$\mathbf{2.49 \pm 0.08}$	$\mathbf{89.3 \pm 1.25}$	$\mathbf{91.4 \pm 0.84}$	$\mathbf{2.67 \pm 0.1}$	93.8 ± 0.89
Rth+mDyn+vDyn+4HzE	90.6 ± 1.03	2.57 ± 0.12	89.3 ± 1.4	91.7 ± 1.02	2.72 ± 0.14	95.4 ± 0.76
Rth+mDyn+vDyn+vFlux	90.2 ± 0.71	2.52 ± 0.09	88.9 ± 0.97	91.3 ± 0.78	2.66 ± 0.1	94.8 ± 0.87
1G+KL with MFCC	N/A	N/A	N/A	80.2 ± 0.75	1.68 ± 0.007	80.4 ± 1.82

better than previous features 4HzE or vFlux, which are less able to detect music in the presence of highly-variable noise. The best 2 and 3 feature combinations are 'Rth + mDyn' and 'Rth + mDyn + vFlux' (which slightly outperforms 'Rth + mDyn + vDyn' on most metrics). This confirms the success of the pitch dynamics feature, Υ, in detecting music in noise. Matlab code for these features are available[1].

4. DISCUSSION AND CONCLUSIONS

An examination of misclassified clips revealed that many represent genuinely ambiguous cases, for instance weak or intermittent music, or partially-musical sounds such as a piano being struck by a baby. There were many examples of singing, such as birthday parties, that were not considered music by the annotators but were still detected. More clear-cut false alarms occurred with cheering, screaming, and some alarm sounds such as car horns and telephone rings.

In this paper, we have proposed a robust musical pitch detection algorithm for identifying the presence of music in noisy, highly-variable environmental recordings such as the soundtracks of consumer video recordings. We have introduced a new technique for estimating the dynamics of musical pitch and suppressing both aperiodic and stationary periodic noises in the autocorrelation domain. The performance of our proposed algorithm is significantly better than existing music detection features for the kinds of data we are addressing.

[1] http://www.ee.columbia.edu/~kslee/projects-music.html

5. REFERENCES

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. ICASSP*, 1996, pp. 993–996.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP*, 1997, pp. 1331–1334.

[3] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Tr. Speech and Audio Proc.*, pp. 441–457, 2001.

[4] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework.," *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.

[5] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.

[6] K. Lee and D. P. W. Ellis, "Voice activity detection in personal audio recordings using autocorrelogram compensation," in *Proc. Interspeech*, Pittsburgh, 2006, pp. 1970–1973.

[7] "YouTube - broadcast yourself," 2006, http://www.youtube.com/.

[8] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in *MIR workshop, ACM Multimedia*, Germany, Sep. 2007.