# Overlapped music segmentation using a new effective feature and random forests

**Duraid Y. Mohammed[1], Khamis A.Al-Karawi[2], Philip J. Duncan[3], Francis F. Li[4]**
[1]School of Engineering, Al-Iraqia University, Baghdad, Iraq
[2]Diyala University, Baquba, Kegubernuran Diyala, Iraq
[3,4]School of Computing, Science and Engineering, University of Salford, Salford, Greater Manchester, United Kingdom

**Article Info**

***Article history:***

Received Feb 2, 2019
Revised Apr 20, 2019
Accepted May 20, 2019

***Keywords:***

Audio content analysis
Audio indexing
Entropy
Music detection
Real world audio classification

## ABSTRACT

In the field of audio classification, audio signals may be broadly divided into three classes: speech, music and events. Most studies, however, neglect that real audio soundtracks can have any combination of these classes simultaneously. This can result in information loss, thus compromising the knowledge discovery. In this study, a novel feature, "Entrocy", is proposed for the detection of music in both pure form and overlapping with the other audio classes. Entrocy is defined as the variation of the information (or entropy) in an audio segment over time. Segments, which contain music, were found to have lower Entrocy since there are fewer abrupt changes over time. We have also compared Entrocy with existing music detection features and the entrocy showing a good performance

***Corresponding Author:***

Philip J. Duncan,
School of Computing, Science and Engineering,
University of Salford, Salford,
Greater Manchester, United Kingdom.
Email: P.J.Duncan@salford.ac.uk

## 1. INTRODUCTION

Automatic classification of real-world audio soundtracks into speech music and/or other events, particularly when overlaps take place, is a challenging problem. Classification is important in archive management, information mining from big data and many other applications. No previous study has been general enough to propose a universal system that will maximize information retrieval for further information mining [1]. Much of the previous research found in the literature focused on classifying soundtracks into speech or music, in a mutually exclusive manner. Automatic classification of overlapped soundtracks into speech and/or music, when the two sometimes overlap, is a particularly challenging problem, since overlapping can contaminate relevant characteristics and features, causing incorrect classification or information loss. This problem has been recognized by the Audio and Acoustics Signal Processing challenge (AASP), which is sponsored by the IEEE Signal Processing Society. One conclusion drawn from the AASP was that "the task of recognizing individual potentially overlapping sounds becomes significantly challenging and the performance of systems that are even prepared to deal with polyphonic content falls dramatically" [1]. Although it is well understood that the overlapping of the classes might mitigate the information retrieval system's performance, little research has been performed with the aim of addressing this problem.

To name some, Shokouhi et al. [2] proposed an overlapped speech detection algorithm with a presence of noise condition to estimate the likelihood of overlapping speech. Zhang [3] presented an approach to segmentation and annotation of audio-visual data, which uses three main categories: silence, with or without music components, and each of the last two categories further classified into more components. Four types of audio features were used, namely, the short-time energy function, the short-time average zero-crossing rate, the short-time fundamental frequency and the spectral peak tracks. The segmentation performance of the proposed method of music, speech with background music and sound effect with background music detection was up to 94.5%, 86% and 87.5% respectively.

Lee [4] has suggested 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 and the detection of music in ambient audio by using pitch. Lee has also reported improved music/speech discrimination results with vocals present using the proposed features when combined with rhythm and 4 Hz Modulation Energy (4HzE). The use of these features has achieved 91.4% accuracy where it is combined with a variance of spectral flux and rhythm. For speech/music discrimination, Sell et al. [5] has derived new features from chroma vector based on the musical tonality.

Tomonori [6] has investigated four sets of features for background music detection. The first set of empirical features (EMPR) included 4-Hz modulation energy, the percentage of low-energy frames, the spectral centroid, the spectral roll-off point (Spectral Roll-off point Frequency denotes the frequency below which 85% of the spectrum magnitude can be found. Most of the signal power is concentrated within a certain range of frequencies.), the spectral flux (spectral flux refers to how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame), and the zero-crossing rate. The second set is spectral power from the linear-scaled frequency (SPEC). The third set is spectral power features with the Mel-frequencies (SPMF). The fourth set is Mel-frequency cepstral coefficients MFCCs, and empirical features (EMPR). The experimental results show that the EMPR feature set provides the highest music detection accuracy with high music- speech mixing ratio, whereas the SPMF set reflects the highest results with the low music-speech ratio (-10 dB and -20dB). Tzanetakis [7] implemented a real-time speech/ music discrimination system with the following features: Spectral centroid, spectral flux, pitch, MFCC, linear prediction (LPC), zero crossings, RMS, and spectral roll-off.

However, the semantic review shows there is no single feature space, which promises to address the problem stated above of real world audio. Ravindran et al. explore the utilization of physiologically impressed auditory features with each physiologically motivated and applied mathematical audio classification techniques [8]. The authors have a tendency to use features derived from a biophysically defendable model of the early auditory system for audio categorization employing a neural network classifier. The authors have also employed a Gaussian-mixture-model (GMM)-based classifier for the aim of comparison and show that the neural-network-based performance is better. Further, the authors used features from a more advanced model of the sensory system and show that the features calculated from this model outperform the features extracted from the early modality stage. The features provide sensible classification performance with solely one-second information segments used for learning and testing phase. Bisot et al. [9] have suggested a supervised nonnegative matrix factorization (NMF) model for overlapping event detection in real audio soundtracks. The authors have begun by highlight the efficacy of non-Euclidean NMF to learn representations for distinguishing and classifying acoustic events in a multi-label setting. Thereafter, the authors suggested training a classifier and using NMF decomposition in a joint optimization problem. This has been done with a general β-divergence version of the nonnegative task-driven dictionary-learning model. An experimental appraisal is performed on the event set of the DCASE 2016 task3 challenge. The suggested supervised NMF-based system enhances performance over the baseline system.

Nam et al., have developed a method for perceiving sound resources in a mixture. The suggested method employed the regression technique, in which the author estimate the relative proportion of sound sources in the given mixture sound signal. Furthermore, the magnitudes of the sound resources were estimated without actually separating the sources. The basic strategy that utilised is based on supervised source separation using probabilistic latent component analysis (PLCA) [10]. A dictionary of the basic components has been estimated from the isolated training data of that mixture. Thereafter, a set of mixture weights of the relative proportion of each source has estimated from the given mixture source without actually separating the sources. The suggested approach has tested on a blend of 5 of sounds and reveals that it's convincing in precisely appraising the relative magnitudes of the sounds in the mixture.

The Audio and Acoustics Signal Processing challenge (AASP) [11], which is sponsored by IEEE signal processing society, provides a competition to classify real-world scene signals with overlap and without overlap issues, but the scope is limited to speech and other indoor/outdoor events; no music scores have been included. The objective was to detect prominent events and ignore other contents of the soundtracks. One of the AASP conclusions was “For the polyphonic case, the task of recognizing individual potentially overlapping sounds becomes significantly challenging and the performance of systems that are even prepared to deal with polyphonic content falls dramatically” [11]. In [12-13] a new technique has been proposed for non-exclusive classification through deploying a timestamp with three classifiers, each of which works as a sensor to detect its particular classes. Thus, the start and ending of each class can be determined even when overlapped. In this work, 'the classification decision is based on the variation behavior across a number of consecutive frames at the same time (visualization of sound)'. For example, an audio file comprising car engines, babble noise, cars moving, shut and open buses doors is more likely to be a bus station.

## 2. DATABASE

Three standard audio classes (speech and music and event sound) were considered in this work. The audio samples in the training dataset were collected from the GTZAN music/speech collection [14], AASP [6], DVD audio tracks, TV and BBC broadcast. The AASP challenge provided two datasets: one for scenes or soundscapes classification (SC) and the other one for event classification. The AASP data was gathered from 10 different places in the London area: Inside an office, park, quiet street, open market, restaurant, supermarket, tube train, tube station, bus and busy Street. The event dataset collected from inside the office is further divided into two sets: monophonic denoted as 'office live' (OL) and polyphonic denoted as 'office synthetic' (OS). Event types used were: alert (beep sound), clearing throat, cough, door slam, drawer, keyboard clicks, keys (keys put on table), knock (door knock), laughter, mouse click, page turn, pen drop, phone, printer, speech and switch. The speech and music from the GTZAN dataset were mixed with OL audio or pink noise to generate two different datasets, G1 and G2. G1 consisted of eight groups: pure-speech, pure-music, pure-ambient/event, speech over music (SM), speech over event (SE), music- event, speech (MSE) - music- event (ME) or silence. The second dataset G2 was generated by mixing speech and music with pink noise in 20 dB and 25 dB SNR for further validation. All the mixed classes were mixed using the same strategy, as explained in [13].

## 3. RATIONALE AND DEFINITION

Existing features for classification are predominantly established on artificially tailored non-overlapping audio clips. A real world audio soundtrack might include one of the segment types explained earlier. An alternative feature has therefore been proposed for improving the detection of the music occurrences, and has shown promising results for cases where the music was pure or overlapping with other classes. The essential idea in employing the frequency calculation as a feature is to indicate the amount of uncertainty across a number of consecutive frames. This has been done through a combination of entropy and frequency and therefore we call it *entrocy*. The assumption here as has mentioned before is 'the classification decision is based on the variation behavior across a number of consecutive frames at the same time (visualization of sound).

Entropy and frequency calculations have evolved separately. The theory of entropy was proposed for the first time by Shannon [15] to measure the degree of information in the signal through calculation of the probability density function (pdf) of each sample in the frame; the randomness distribution of the data is also represented. Entropy is employed in diverse classification problems and had been providing adequate results through the ability to detect the complexity of a signal. The range of its application varies from images, automatic speech recognition, health and ecology. It is also applied to [16] calculating the entropy of audio spectra for discriminating clean speech/noisy speech, and is proposed as a feature for robust automatic speech recognition (ASR). Misra [16] had shown that clean speech has a lower level of spectral entropy than noisy speech because the mean of abrupt changes is higher in a noisy environment. In [17], the same feature was applied on the sub-band of short-time Fourier transform (STFT) and combined with some of the MFCC bands to improve the result of the ASR in a noisy environment; this feature was called spectral entropy. The maxent technique refers to the maximum entropy model, proposed for the first time by [18] as a statistical model for natural language processing and has since been deployed in diverse applications.

## 4. EXPERIMENTAL METHOD

Our classification scheme started by resampling each audio file to a standard sample rate of 22.05 kHz, 16 bit resolution. Each audio file is segmented with a time window with size of 1102 samples and 50% overlap into a series of analytical frames. The overlap has been selected to trade off with the increasing frequency resolution (number of frequency bins). We then calculated the entropy for each frame as defined in the following equation.

$$H = -\frac{1}{\log_2(n)}\left[\sum_{i=1}^{n} p(xi) * \log_2 p(xi)\right] \tag{1}$$

where *p(xi)* is the value of the probability mass function (pmf) at *X = xi* and *n* represents the length of the frame. The probability mass function (pmf) of each sample in the frame can be defined as follows.

$$pr(x_i) = pr(X = x_i) = pr(\{s \in S : X(s) = x_i\}) \tag{2}$$

- Entropy is normalized to the logarithm of the frame length to eliminate the dependency on the frame's length. Therefore, the entropy value will be between 0, and 1 in the case of maximum randomness. Thus, the gradient descent runs much faster and converges in less iteration.
- The experimental results, as shown in the Table 1, indicate that most of the music genre frames have higher randomness (entropy) than the speech.

Table 1. Entropy values for speech and music frames

| Frame type | No. of frames | Mean of entropy | Std of entropy |
|---|---|---|---|
| Speech | 13959 | 0.527 | 0 |
| Music | 11964 | 0.642 | 0 |

- Then after, the calculated entropy vector H, which extracted from N frames, is segmented into 32 samples segments, for frequency calculation. A smaller window results in more compression of the frequency axis as Fs/window size becomes smaller. Consequently, a larger window gives higher resolution in the frequency domain. However, increasing the resolution is valid as long as the signal remains stationary.
- The segmentation is applied with a moving window 16 samples (50%) at a time on the calculated entropy vector (H).

$$E(k) = a(k)\sum_{i=0}^{n} x(i)Cos(\frac{(2i+1)\pi k}{2n}), k = 0,1,\ldots,n-1 \tag{3}$$

For spectral analysis purposes, each segment is multiplied by Hanning window. The Discrete Cosine Transform (DCT) is then calculated for each segment using (3). The DCT method is used to calculate the variance of each 32 adjacent entropy values.

where,

$$\mathrm{a(k)} = \begin{cases} \sqrt{1/N} & k = 0 \\ \sqrt{2/N} & k \neq 0 \end{cases} \tag{4}$$

The centre of gravity, which reflects the spectral shape of the $i^{th}$ entropy segment, is also calculated from the DCT coefficients and is defined as

$$Ci = \frac{\sum_{k=1}^{F} f(E(k))^2}{\sum_{k=1}^{F} (E(k))^2} \tag{5}$$

where, *F* is the number of DCT coefficients , *f* is the frequency centre and *E(k)* is the magnitude of the DCT coefficients calculated using (3).

The average of the segments here will not be equal to zero. Therefore, the first coefficient is ignored to eliminate the mean and the final output is represented by 16 coefficients (15- as frequency bins and one reflecting the centre of frequency gravity). Figure 1 shows the procedure for the entrocy feature calculation. The suggested method is simple and also computationally efficient. The entrocy calculation is done without any computationally expensive optimizations or sophisticated mathematical operations performed in any of the above calculation stages. The method is straightforward to understand and relates to audio content, complexity and homogeneity. Moreover, the construction is general and can be applied in most of the audio information retrieval studies such as music/speech discrimination, segmentation, music information retrieval or classification.

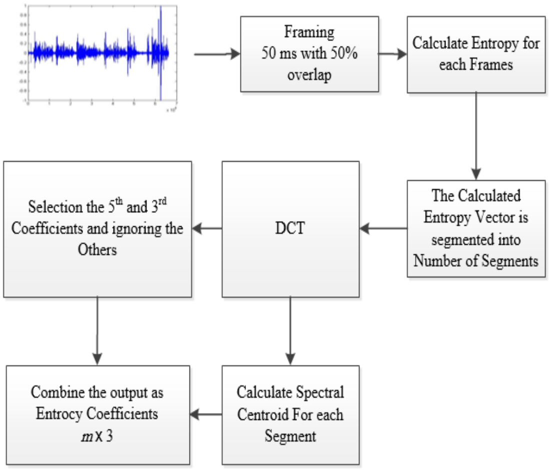
Framing
50 ms with 50% overlap
Calculate Entropy for each Frames
The Calculated Entropy Vector is segmented into Number of Segments
DCT
Selection the 5th and 3rd Coefficients and ignoring the Others
Combine the output as Entrocy Coefficients $m$ x 3
Calculate Spectral Centroid For each Segment

Figure 1. Entrocy calculation procedure

## 5. ENTROCY VALIDATION

For the purposes of Entrocy validation, an experiment to build a module for detection of music against all other classes has been carried out. The samples that contain music score (music or mix) will be labelled as 1 and the remaining samples will be labelled as 0. The machine learning method used is the Random Forests technique trained with varying numbers of trees; it has been established empirically that 1000 trees are an optimal size. Figure 2 shows a simple random forest, with 20 trees using the Entrocy feature.

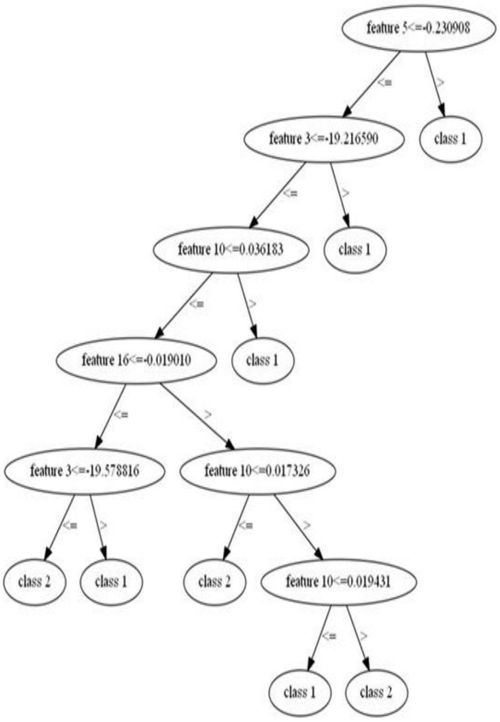

Figure 2. Random forest DTs, Class 1 classes with music, Class 2 without music, features number represents the Entrocy coefficients, and the value represents the threshold of hyper-planes respective to the feature axis

Class 1 represents music, class 2 represents other classes without music occurring; the features number represents the Entrocy coefficients, and the numerical value represents the threshold of hyperplanes with respect to that feature, which is determined by the Random Forest technique.

The basic concept of random forest trees is that they are working as collection of hyperplanes (threshold), where each threshold is orthogonal to the respective features axis [19]. Figure 3 shows the 2-Dimensional features (Entrocy coefficients 5 and 3) visualization where the fifth coefficient is localized as the X-axis and third coefficient as the Y-axis with their hyperplanes threshold.

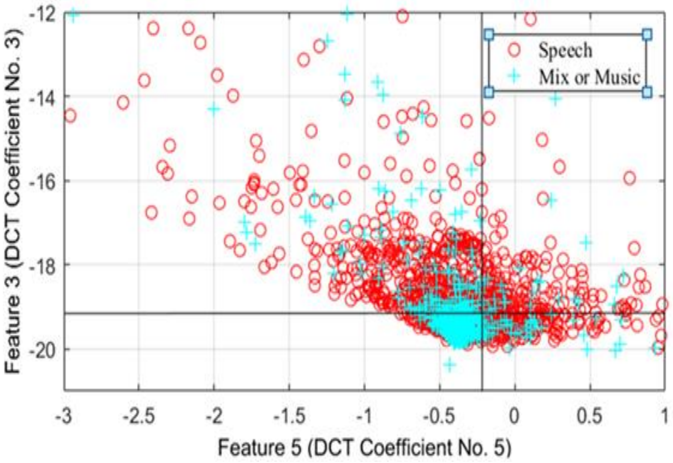

Figure 3. Feature Spaces with their respective thresholds from Figure 2

## 6. MACHINE LEARNING METHOD

As discussed in the preceding section, the purpose of this paper is to design feature based on transformation of raw audio data to categorize music occurrences?, motivated by the characteristics of data distribution, variation and direction. For the training, testing and validation a random forest technique is deployed. During the past decades, many segmentation and classification systems have been developed. The adoption and development of classification methods are typically application driven. Some examples of those methods are (Dhanalakshmi et al.) audio categorization using radial basis function neural networks (RBFNN), which is based on radial Basis Function (RBF), or distan ce function, as activation function for hidden layer’s neurons and SVM reliant on event type into six different categories (music, news, cartoon, movie, sports and advertisements), Hidden Markov Models and Gaussian Mixture Models and neural Network [20].

Random forest is considered to be the modern generation of the decision tree machine learning and it is used by many researchers recently, i.e. to improve the discrimination between speech and non-speech [21] and to decrease the storage of speech through saving only speech segments. The authors illustrated that random forest results were better than other decision tree algorithms. Also, smoothing is used over 5 segments to improve the results. Random forest was compared against the bagging and bootstrapping decision tree algorithms by [22] for classifying environmental audio into five different sets (bird, wind, rain, frog and thunder) based on the different size of training samples; the results show that the stability of random forest against the other two algorithms.

Random forest is proposed for the first time by Breiman [23]. Many classification trees are built on the same dataset, which randomly divides the data into both test sets (90%), and learning sets (10%). The decision trees are different from each other due to samples are dragging randomly (different bootstrap) from the training dataset for each tree, this process called “bagging” and considerings only a small random subset of the available data features as candidates for each split. This randomisation reduces the correlation between individual decision trees. In the testing phase, the test vector passes through all trees in the random forest, and each tree makes an independent decision. The final decision will be the class which has the most votes over all subtrees in the random forest. i.e, if the random forest consists of four trees used to classify an object into three classes, then the prediction will be for the class, which has the highest vote over all four trees. The trees are growing based on the following strategy.

To conclude, the construction of DTs starts at the root node then continuously partitions the feature space. A classification tree is established consuming a labelled data, $\{v = (x_i, y_i), i = 1 \cdots m\}$, where $x_i$ is the data samples and $y_i$ the respective class labels [24]:

- The Number of trees (B) is determined by the user at the beginning. Each tree is trained independently of the others with a randomly selected subset of the training set per tree.
- The Out-of-bag (OOB) technique is employed for selection of the training samples, the training data set is divided into three parts, two of which are used for tree construction with the ability to swap, and the last one is left "out of bag" for testing. It is possible that some of the samples might be replicated in more than one tree and other of the samples never picked, as shown in Figure 4. Each tree in the RFs randomly selects a set of features and training samples from the training vector. OOB is used to calculate both the estimated error of the forests and the variable importance. This property makes RFs more robust against noise and the dependency on dataset problem.
- A binary split function, $Sp = \{x, \theta_j\} \in \{0,1\}$ which is associated with each internal node, passes the patterns $x$ at either internal node j to the left or right child node based on the decision (0 or 1).

This procedure is repeated for all features to select the best one to split. Then parameters $\theta_j$ are optimised for all tree nodes during training, to select features with a higher information gain (impurity function). In general, splitting refers to determining which feature should be used and at which node by measuring impurity gain, and the optimal cut-point for that feature. The feature with the highest information gain value is selected to behave as a root node, which is used to split the dataset. In other words, features are organised based on the priority (importance) from the top down.

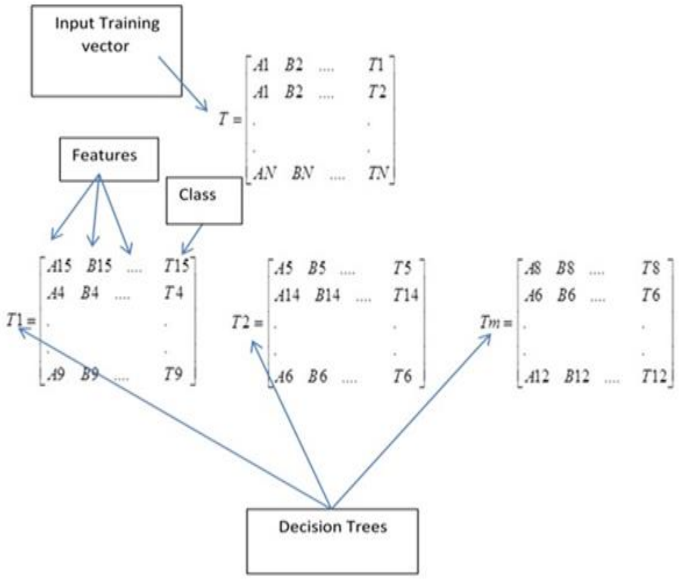

Figure 4. Subset samples selection (RF)

In this experiment, the goal as discussed for building a module for music against all other classes, the features vector will be (N*3) where N represents a number of samples and respective single class label. In the training phase, the samples which content on the music score will be labeled as 1 and the rest samples as 0. The random forest trained with 1000 trees as explained before.

## 7. RESULTS AND DISCUSSION

The results of the experimental method indicate that the pure music/pure speech discrimination was significant where Zero Crossing Ratio ZCR or entrocy is used with all samples selected from GTZAN dataset described above. Table 2 and Table 3 illustrate the results of the two classification scenarios. The error rate is defined as the ratio between the misclassified samples and a total number of samples. The music group included all samples with music background (Music (M), Speech over Music (SM), Speech over Music and

Event (SME) and Music over Event (ME)) and the group without music included all other samples. From Table 2, the results show that ZCR outperformed other features since speech samples have a higher rate of zero crossing than pure music. However, it has failed when used to discriminate between classes with and without music due to an under-fitting problem, which has led to non-convergence, while entrocy presents high performance in music/non-music discrimination.

Table 2. Speech/ Music discrimination Error Rate, the ratio between number of misclassified frames and total number of frames n each group

| Features | Pure Music | Pure Speech |
|---|---|---|
| ZCR | 5.33% | 3.75% |
| Entrocy | 8.60% | 10.48% |
| Spectral Entropy | 49.53% | 63.85% |
| Spectral Centroid | 60.84% | 42.13% |
| Roll_Off | 42.37% | 36.00% |
| Chroma | 29.12% | 59.18% |

Table 3. Music detection Error Rate, the ratio between number of misclassified frames and total number of frames in each group

| Features | with Music | without Music |
|---|---|---|
| ZCR | 25.87% | 82.25% |
| Entrocy | 12.26% | 17.46% |
| Spectral Entropy | 34.07% | 38.48% |
| Spectral Centroid | 41.40% | 44.46% |
| Roll_Off | 62.83% | 36.78% |
| Chroma | 61.31% | 37.62% |

Consequently, ZCR has been combined with entrocy to recognize samples with music from samples without music. It has achieved the best results when detecting the presence or absence of music with high accuracy. The classification performance of music, SM, ME, SME, Speech, Speech with Event background and Event was up to 94.19%, 95.33%, 94.31%, 81.68%, 67.82%, 86.34% and 95.90% respectively. Finally, pink noise was mixed with speech and music in two different signals to noise ratios (SNR) 20 and 25 dB in place of an event. Figure 5 shows the error rate respective to each mixing group. As noticed the results presented here may facilitate improvements in music detection when training the system on a specific class of things i.e. noise here are better than when using wide range of events.

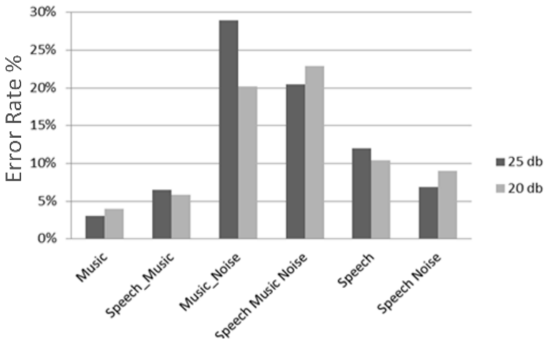

Figure 5. Entrocy Error Rate for music detection

## 8. CONCLUSION

A novel feature, ’entrocy’ is proposed in this paper for classification of overlapped audio comprising three main classes (speech, music and event). It reflects the class randomness level fluctuation over time. It is logical to identify the class based on estimating the homogeneity of changes through the frequency of the segments contents. This is the key to information preservation through detecting a particular class even when overlapped with other classes and will be the basis for non-exclusive segmentation highlighted in [25]. Music, speech and event sounds can be consistently detected and segmented through developing one module belonging to each class. For the music module, random forest machine learning using the entrocy feature has shown promising results. Event sound recognition is challenging since there is often no prior knowledge about how many different types of events there are likely to be, and which ones are of interest. Therefore, an ad hoc approach might be best for event processing based on the application due to the event sounds being an open set. When the (OL) dataset or pink noise is employed, the entrocy is effective for the detection of music occurrences.

We therefore conclude that entrocy has great potential as a feature for overlapped audio classification and has outperformed existing features in the overlapped condition. In addition, it provides a cost and time gain for the classification process due to achieving these results with only low dimensional features.It has also been observed that for detection of speech, music and event with convenient results a longer period of the frame might be beneficial; in this paper 50 ms has been used and the final decision was respective to [50ms*32=1.6 seconds]. This is not surprising since the human listener has difficulty differentiating isolated sounds having short duration. The experimental results show that the frames that involve music, even when overlapping with other audio classes, have lower entrocy than those with music absent. It worth noting that this feature will be further evaluated through application in different classification areas as future work.

## REFERENCES

[1] Giannoulis, D., et al. *Detection and classification of acoustic scenes and events: An IEEE AASP challenge*. in 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2013. IEEE.

[2] Shokouhi, N., et al. *Robust overlapped speech detection and its application in word-count estimation for prof-life-log data*. in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. IEEE.

[3] Zhang, T. and C.C.J. Kuo, Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 2001. 9(4): p. 441-457.

[4] Lee, K. and D.P.W. Ellis. *Detecting music in ambient audio by long-window autocorrelation*. in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. 2008.

[5] Sell, G. and P. Clark. Music tonality features for speech/music discrimination. in 2014 IEEE International Conference on Acoustics, *Speech and Signal Processing* (ICASSP). 2014.

[6] Tomonori, I., M. Ryo, and K. Kunio. *A background music detection method based on robust feature extraction*. in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. 2008.

[7] Tzanetakis, G. and P. Cook. *A framework for audio analysis based on classification and temporal segmentation.* in EUROMICRO Conference, 1999. Proceedings. 25th. 1999.

[8] Ravindran, S., K. Schlemmer, and D.V. Anderson, *A Physiologically Inspired Method for Audio Classification*. EURASIP Journal on Advances in Signal Processing, 2005. 2005(9): p. 561326.

[9] Bisot, V., S. Essid, and G. Richard. *Overlapping sound event detection with supervised Nonnegative Matrix Factorization*. in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017.

[10] Nam, J., G.J. Mysore, and P. Smaragdis. Sound Recognition in Mixtures. in LVA/ICA. 2012. Springer.

[11] Giannoulis, D., et al., *Detection and classification of acoustic scenes and events: An ieee aasp challenge*, in Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. 2013, IEEE. p. 1-4.

[12] Li, F.F. Nonexclusive audio segmentation and indexing as a pre-processor for audio information mining. in Image and Signal Processing (CISP), 2013 6th International Congress on. 2013.

[13] Mohammed, D.Y., et al., A *system for semantic information extraction from mixed soundtracks deploying MARSYAS framework, in Industrial Informatics (INDIN*), 2015 IEEE 13th International Conference on. 2015. p. 1084-1089.

[14] Tzanetakis, G. Music Analysis, Retrieval and Synthesis for Audio Signals. [cited 2014 20 Jan]; Available from: http://marsyas.info/about/overview/.

[15] Shannon, C.E., A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 1948. 5(1): p. 3-55.

[16] 16. Misra, H., et al. *Spectral entropy based feature for robust ASR*. in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. 2004. IEEE.

[17] Toh, A.M., R. Togneri, and S. Nordholm, *Spectral entropy as speech features for speech recognition*. Proceedings of PEECS, 2005. 1.

[18] Berger, A.L., V.J.D. Pietra, and S.A.D. Pietra, A maximum entropy approach to natural language processing. *Computational linguistics*, 1996. 22(1): p. 39-71.

[19] Kargupta, K., B.-H. Park, and H. Dutta, Orthogonal decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 2006. 18(8): p. 1028-1042.

[20] Dhanalakshmi, P., S. Palanivel, and V. Ramalingam, Classification of audio signals using SVM and RBFNN. Expert Systems with Applications, *ELSEVIER*, 2009. 36(3): p. 6069-6075.

[21] Thambi, S.V., et al. *Random forest algorithm for improving the performance of speech/non-speech detection*. in Computational Systems and Communications (ICCSC), 2014 First International Conference on. 2014.

[22] Zhang, Y., Selected Features for Classifying Environmental Audio Data with Random Forest. *The Open Automation and Control Systems Journa*l, 2015. 7(1).

[23] Breiman, L., Random Forests. Machine Learning, 2001. 45(1): p. 5-32.

[24] Hastie, T., et al., The elements of statistical learning: data mining, inference and prediction. second edition ed. *The Mathematical Intelligencer*. Vol. 27. 2005, California: springer. 83-85.

[25] Duncan, P.J., D.Y. Mohammed, and F.F. Li, Audio Information Mining–Pragmatic Review, Outlook, and a Universal Open Architecture, in *Audio Engineering Society Convention* 136. 2014, Audio Engineering Society: BERLIN.