Mobile Audio Intelligence: From Real Time Segmentation to Crowd Sourced Semantics

Lazaros Vrysis
Aristotle University of Thessaloniki
AUTh University Campus
Greece, 54124
lvrysis@auth.gr

Nikolaos Tsipas
Aristotle University of Thessaloniki
AUTh University Campus
Greece, 54124
nitsipas@auth.gr

Charalampos Dimoulas
Aristotle University of Thessaloniki
AUTh University Campus
Greece, 54124
babis@eng.auth.gr

George Papanikolaou
Aristotle University of Thessaloniki
AUTh University Campus
Greece, 54124
pap@eng.auth.gr

ABSTRACT
The task of general audio detection and segmentation based in means of machine learning is very popular and high-demanding procedure nowadays. Most relevant works in the last decade aim at modelling audio in order to conduct a semantics analysis and a high–level categorization. A generic strategy that would detect audio events as means of transitions from one audio state to another is considered interesting and would support whole classification workflow. This work investigates the possibilities in designing a robust bimodal segmentation algorithm for audio that would perform well in different conditions without relying on complicated machine learning schemes by minimizing prior knowledge for detection model, and thus, delivering consistent performance for any input signal and computing environment. Additionally, a modern user-generated content approach for populating and updating ground truth databases is presented. Both techniques are implemented and embedded as upgrades, in a mobile software environment for smartphones.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications - *Data mining*. H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing – *Methodologies and techniques, Modeling, Signal analysis, synthesis, and processing.* I.5.2 [**Pattern Recognition**] Design Methodology - *Classifier design and evaluation, Feature evaluation and selection, Pattern analysis.*

General Terms
Algorithm, measurement, experimentation, design.

Keywords
Audio, sound, events, detection, bimodal, segmentation, semantics, mobile, software, acoustics, measurements.

AM15, October 07-09, 2015, Thessaloniki, Greece
© 2015 ACM. ISBN 978-1-4503-3896-7/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2814895.2814906

1. INTRODUCTION
Audio Events Detection (AED) is a common task with many applications in the research fields of Automatic Speech Recognition (ASR), Music Information Retrieval (MIR) and Environmental Sound Recognition (ESR). For many years the scientific research has been concentrated to model human speech [1], resulting in the well-known text to speech and speech to text implementations that continuously tend to improve, leading to usable consumer products in the last decade. Audio content other than speech was out of the scope of the scientists but is trending lately [5]. Audio signals come from animals or the Mother Nature, music, noise produced from engines and various machines synthesize the whole audio scenery defining the meaning of a soundscape [11]. The effort to classify all these different sounds and detect audio-content changes is challenging.

Speech may be the most informative auditory information source for humans, music should follow, but other kinds of sounds also carry useful information that can contribute in a higher level of contextual analysis. Rapidly growing audiovisual content is the daily outcome from various computational environments, produced by single users or professionals, including enterprises such as television and radio stations, personal devices like smartphones or other systems such as surveillance systems. When coming up against huge amounts of content, machine-based automatic semantics annotation is necessary and a considerable convenience for managing and retrieving data.

Categorization of non-speech sounds also may help improve speech recognition performance, isolating target information from unwanted noise. Hierarchical classification schemes can be used with higher efficiency, transforming complex classification tasks into smaller and thus, easier to solve [15]. Finally, modern User Generated Content (UGC) approaches for collecting data can be very useful in order to speed up ground truth populating tasks.

1.1 Related Work
When comes to AED, a rich list of works can be found in the scientific community. Most of them set up a restricted content-basis experiment not completely solving a practical, every day problem. This is a reasonable choice, considering the huge difficulties for developing a generic classification scheme and mechanism, taking into account that classification challenges with more than a dozen of categories lead to poor performance results. It's

worth noting that many works are based on CLEAR (Classification of Events, Activities and Relationships) Framework in order to conduct an evaluation of their proposed algorithms using the corresponding dataset [16].

Zhuang et al [19] perform real-world acoustic events detection, working on a feature pool of Mel Frequency Cepstral Coefficients (MFCCs) and Log Frequency Filterbank Parameters (LFPs) supported by numerous classifiers, such as Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs), Support Vector Machine (SVM) and Gaussian Mixture Models (GMMs). Experiments demonstrate that the presented setup contributes toward improved performance, compared to previous best-performing approaches. Zhou and Zhuang [20]-[21] highlight the growing research interest in the AED field and note that there are key differences between speech and other context audio signals. They perform a feature selection process based on Adaboost [7] algorithm and compare classification results of different feature sets (MFCCs, Adaboost Selected, Delta Values, Acceleration Values) based on HMM AED system architecture. They conclude that using a complete set of features designed for speech recognition is not optimal, and a custom feature set could yield better performance. Elo et al [8] utilize MFCCs, Perceptual Linear Prediction (PLP) combined with Zero Crossing Rate (ZCR) values and SVM or HMMs as classification algorithms without demonstrating clear results. Atrey et al [1] propose a system that incorporates a hierarchical approach for audio event detection (categorizing talk, cry, knock, footsteps, walk, run, vocal and non-vocal events), a classifier based on GMMs and various feature sets - consisting of Linear Predictor Coefficients (LPCs), Linear Predictive Cepstral Coefficients (LPCCs, Log Frequency Cepstral Coefficients (LFCCs). Their outcome sums up that LPCs perform well for segmenting background and foreground audio activity, LFCC performs better in demarcating between the vocal and non-vocal events and LFCC as well as ZCR are good for classifying between door knock and footstep events. Kotsakis et al investigate various audio pattern classifiers in broadcast audio semantic analysis using supervised training [10]. Hierarchical and hybrid classification taxonomies are deployed facilitating efficient speaker recognition/identification, speech/music discrimination, and generally speech/non-speech detection-segmentation, leading to remarkable performance values.

A recent work of Dimoulas and Symeonidis [6] presents MAESTRO, a framework for automated multimedia files organization that embeds an interesting audio transition landmarking and bimodal segmentation technique, called Match-ware. Match-ware exploits wavelet-based signal transforms in combination with exponential moving average audio feature thresholding for audio transitions detections. Tsipas et al come up against similar task in musical pieces introducing Vector-Quantization is as an adaptive filtering mechanism for time-lag matrices while a structure-feature based self-similarity matrix is proposed for novelty detection [17]. The method is evaluated against state-of-the-art implementations showing performance improvements.

The current system has been implemented in order to be deployed as an additive utility to iSMAARTer "Intelligent Sound Measurement Audio Analysis & Recording Tool". iSMAARTer is a mobile software capable of performing audio measurements combined with embedded semantics processing and long-term intelligent analysis [18]. Strategies and algorithms keeping computational requirements low have been presented, while cloud-based services have been introduced, exploiting mobile devices' sensors for sound mapping tasks.

The rest of the paper is organized as follows: Section 2 presents the improvements of iSMAARTer towards audio semantic analysis; Section 3 introduces a proposed bi-modal segmentation algorithm while Section 4 describes the evaluation of this strategy. In Sections 5 and 6 conclusions and further work thoughts are included.

2. SYSTEM IMPLEMENTATION

2.1 Event Transition Detection

Along with the low-level acoustic measurement modes iSMARTER also provides long-term audio analysis capabilities, based on semantic audio processing concepts. This, higher-level, Long Term Analysis (LTA) module brings real time audio-pattern recognition visually resulting into an event detection markup timeline. Event detection resides on a pattern-based user-generated sound library.

The newly added segmentation approach comes to support the existing infrastructure without requiring prior knowledge or a pattern-matching table. Both algorithms run in cooperation with each other, aiming at improved performance. From a user's viewpoint, the LTA's event detection timeline has been reformed, including new events detection highlighting along with pattern matching and feature analysis output.

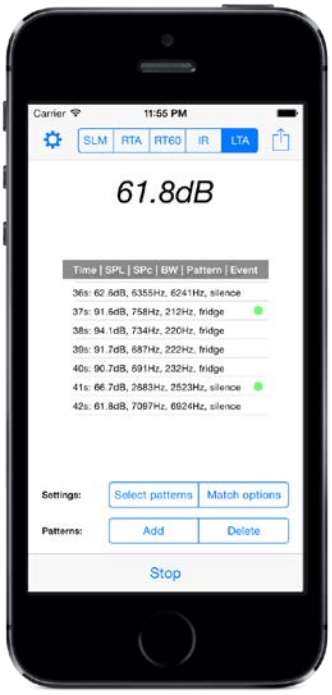

Figure 1. iSMAARTer Running Long-Term Analysis with Event Detection Enabled

2.2 User Generated Ground Truth Data

The software includes a cloud-based session manager that handles all the users' data, aiming at building user-generated, spatiotemporal digital maps used for storing measurements. In addition to this, users can manage their own sound library, used for real time audio events detection. This sound database handles raw audio data combined with its corresponding higher level semantics, in the form of the users' annotations.

Crowd sourced semantics capability enables users to share their pattern-storing matrix and get access to a worldwide cloud-based library. The benefit of this approach is twofold. Firstly, a single user can semantically analyze sounds without having to build up its own sound library gaining undeniable convenience. The application performs the specified task utilizing data that is already available, and thus, offering an improved user experience.

Secondly, this user-generated data gathering contributes in populating ground-truth data. The content is stored in iSMAARTer's cloud infrastructure and can be used in big-data analysis tasks, offering valuable services in the field of audio semantics analysis.

3. SEGMENTATION ALGORITHM

3.1 General Workflow

The workflow of the proposed algorithm is presented in Fig. 2. Raw input data is being separated into overlapping successive frames and a Hann window is applied. Each frame generates a baseline feature vector, its components are standardized and delta values are calculated (except from these that already carry transitional information, such as Spectral Flux). The latter undergoes a thresholding process and finally a frame is marked as a segmentation point or not if certain number of components exceed the above mentioned threshold.

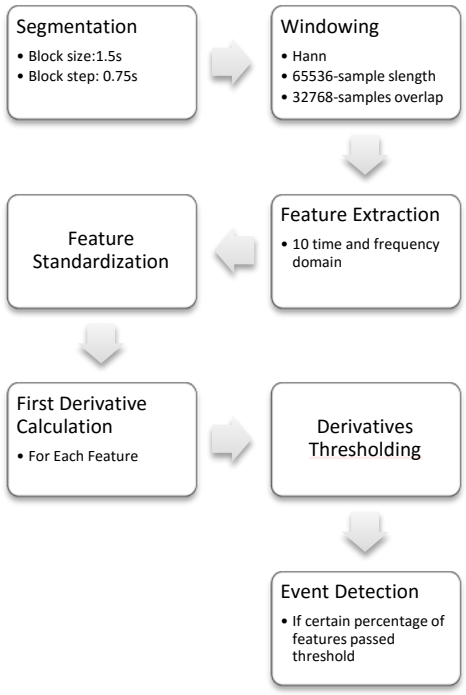

Figure 2. General Workflow of the Algorithm

3.2 Signal Segmentation

The input audio signal is being separated in blocks with a size of (W_S) *32768-65536 samples* (*~0.75-1.5s*, in respect to the sampling rate of the recordings) and a Hann window is being applied on each of them with overlapping (W_O). The choice of a large window introduces well-known time-domain fuzziness but was made in purpose and considering two facts. Firstly, a coarse segmentation is initially intended, not aiming at sub second accuracy and omitting events that change within brief time periods Secondly, a large window smooths out successive feature-vectors, helping the algorithm not to detect multiple false positive events. The overlapping is necessary so as not to lose information, and as a result, events between windowed frames. This setup can lead to a total maximum accuracy error of *+/- 372ms*.

3.3 Feature Processing

After signal windowing the feature extraction process follows. The final feature vector consists of time and frequency domain features; most of them are low lever MPEG7 descriptors. Raw feature values are standardized based on mean and standard deviation values that were derived for each feature analyzing the development dataset resulting in the vectors *F(t)*.

$$F(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ \dots \\ f_N(t) \end{bmatrix} \quad (1)$$

Standardized feature values are used to calculate the corresponding deltas *D(t)*.

$$D(t) = \begin{bmatrix} f_1(t) - f_1(t-1) \\ f_2(t) - f_2(t-1) \\ \dots \\ f_N(t) - f_N(t-1) \end{bmatrix} \quad (2)$$

The resulting data undergoes a thresholding process in order to end up to a binary output, getting a value of 1 if a vector component exceeds the threshold and 0 otherwise. The thresholding value T_1 is adjustable; an optimum value of the parameter was used throughout the testing and evaluation process

$$Q(t) = \begin{bmatrix} q_1(t) \\ q_2(t) \\ \dots \\ q_N(t) \end{bmatrix} \quad (3)$$

Where

$$q_{n,i} = \begin{cases} 1, f_{n,i} \geq T_1 \\ 0, f_{n,i} < T_1 \end{cases} \quad (4)$$

3.4 Event Detection

The feature processing procedure results in a feature vector for each audio frame, consisting of binary data. A sum of the vector components is performed as eq. 5 demonstrates. The system marks a specific frame as transition point or not by making use of an additional second thresholding parameter T_2 (eq. 6). T_2 can be set in the range *[0, 1]*. Parametric analysis for T_2, in order to optimize the algorithm was carried out. Increasing T_2, likewise T_1, makes the system stricter in detecting events.

$$SQ_n(t) = \frac{\sum_1^N q_i(t)}{10} \quad (5)$$

$$E_n = \begin{cases} YES, SQ_n \geq T_1 \\ NO, SQ_n < T_2 \end{cases} \quad (6)$$

3.5 Feature Selection

While developing the system, many features were calculated and evaluated, such as twelve MFCCs, RMS energy, Loudness, Temporal Max, Min, Mean, Skewness, Kurtosis, Crest Factor, Spectral Smoothness, Skewness, Kurtosis, Slope, but after a dimension reduction process the best performing were selected. In the sys-

tem's development phase Sonic Visualizer [4] was used for building the full feature set and Weka [9] for the feature ranking procedure. In order to evaluate the feature set, two different ranking algorithms were employed, the "InfoGainAttributeEval" and "OneRAttributeEval". The first one evaluates the importance of each attribute individually by estimating the information gain with respect to the class using entropy measures. The second algorithm is a simple classification algorithm that generates a one-level decision tree and accounts the number of corrected classified instances.

The VAMP Plugins that we used for extraction the audio features were Libxtract [3], MIR.EDU [13] and Queen Mary Plugin Set [12].

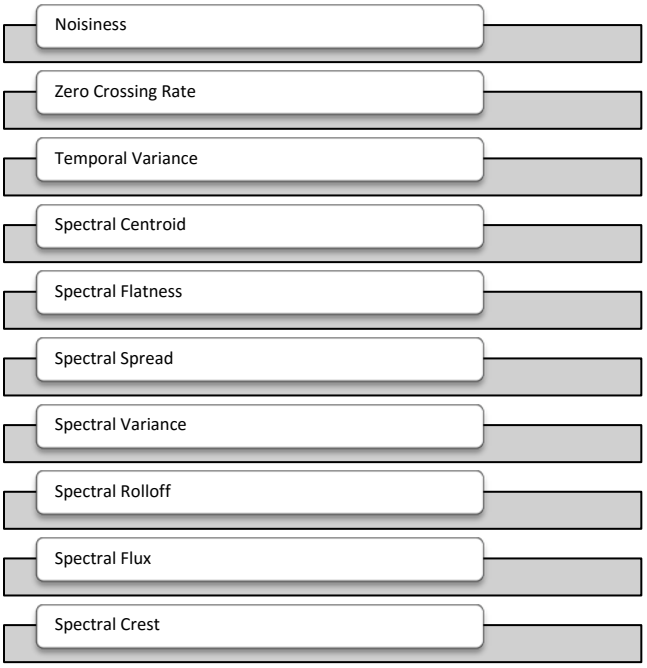

Figure 3. Selected Features

RMS, Loudness and other absolute power-based features were tested in detail so as to deliver, as much as possible, a generic approach independent from incoming sound pressure level and input gain of devices avoiding biasing of the algorithm to our specific dataset. Eventually, the feature selection process showed a weak coherence of this kind of features and audio transition points.

In order to highlight performance variance for different features, we ended up with three feature sets under testing. The first consists of the ten best performing feature, the second is a collection of 12 MFCCs and the third one a combination of both of them.

Table 1. Feature Sets

Feature Set	Transition Points
10S	10 Selected Features
MFCCs	12 MFCCs
MFCCs+10S	12 MFCCs + 10 Selected

4. EXPERIMENTAL RESULTS

In order to examine the efficiency of the proposed algorithm experiments were conducted. We created a specially designed audio data, annotated with ground truth information, used as an input to detection algorithm. Common evaluation metrics were employed so as to get a numeric measure of the method's performance. System's parameters were set as follows: W_S*=65536 samples, 50%* overlap with W_O*=32768 samples,* T_1*=1,* T_2*=25%.*

4.1 Dataset

The dataset that was used for the development and performance evaluation of the system is mainly hold audio data and is divided in two subsets. Both subsets carry uncompressed monophonic audio data with *44,100Hz, 16bit* sampling rate and bitrate characteristics. Recordings include a variety of audio content, combining speech, radio and TV broadcasts, music, environmental and household sounds, electronic devices' noise, ringing, etc.

The first one was used for deriving mean and standard deviation values for each feature so as to supply necessary data for the standardization process of the algorithm. It is a sixty minutes compilation of sounds collected from libraries and recordings that carried out for the purposes of the experiment.

The second was used for measuring the performance of the system and was annotated in order to highlight audio events in terms of transition points. Sonic Visualizer was used for this procedure. This dataset is split up in four subsets, with slightly different content:

- Subset #1 is a compilation of environmental sounds derived from sound libraries
- Subset #2 mostly consists of household sounds, speech, TV and radio recording
- Subset #3 contains audio content from an office working environment
- Subset #4 includes outdoor audio events in urban environment

Table 2. Evaluation Dataset

Dataset #	Transition Points
Set 1	15
Set 2	18
Set 3	20
Set 4	22
Overall	55

The dataset is available at: *research.playcompass.com/files/LVLib-1.zip*

4.2 Performance Measures

To evaluate the accuracy of the transition points detection, common performance measures for binary classification were employed. In particular, the traditional or balanced *f1-score* was used as the main performance indicator.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (7)$$

For the calculation of the precision and recall measures, a certain threshold *t* was defined and used to classify detected transition points as correct or incorrect. If the detected transition point time difference from a transition point in the ground truth data was less or equal to the predefined threshold then the detected transition points were considered correct. Common values for the parameter *t* in bibliography lie in a range between *0.5s* and *3s* [14].

$$\boldsymbol{precision} = \frac{\boldsymbol{\#\ of\ detected\ transition\ points\ within\ t}}{\boldsymbol{\#\ of\ detected\ transition\ points}} \quad (\mathbf{8})$$

$$recall = \frac{\#\ of\ detected\ transition\ points\ within\ t}{\#\ of\ ground\ truth\ transition\ points} \quad (9)$$

In our case, a threshold *t=1.5s* was selected for the evaluation as our optimization goal was more towards optimal segmentation and less towards sub second accuracy. Furthermore, the accuracy of the ground truth data can be controversial when sub second precision requirements are brought into the equation.

4.3 Results

The performance results for each feature set are presented through the following tables. Precision, Recall and F-Score values are demonstrated for each subset and overall values as well.

Table 4. Performance Results (10S)

Dataset #	Precision	Recall	F-Score
Set 1	100.0%	73.3%	84.6%
Set 2	68.1%	83.3%	75.0%
Set 3	91.6%	64.7%	75.8%
Set 4	90.9%	95.2%	93.1%
Overall	**85.1%**	**80.3%**	**82.6%**

Table 4 holds the results for the first feature set, with 10 selected features. We can notice decent performance and balanced precision and recall outputs.

Table 5. Performance Results (MFCCs)

Dataset #	Precision	Recall	F-Score
Set 1	92.3%	75.0%	82.8%
Set 2	28.3%	72.2%	40.6%
Set 3	85.7%	70.6%	77.4%
Set 4	61.5%	80.0%	69.6%
Overall	**53.5%**	**74.7%**	**62.4%**

Table 5 displays the same metrics for our second feature set, which was made up by 12 MFCCs. Overall performance is notably lower, specifically, precision metric for the second subset is remarkably low. After some examination of this behavior, we found out that the problem comes from the speech segments of the signal. MFCCs prove to be very sensitive when comes to speech, their delta values fluctuate considerably leading to many false positive transition point detections within a single segment of speech.

Last table presents results with MFCCs and 10S feature set combined. Before even examining the results, we do not easily expect improved values compared to these of the first feature set, due to the poor performance of the MFCCs. Indeed, performance lies between the previous ones.

Table 6. Performance Results (MFCCs+10S)

Dataset #	Precision	Recall	F-Score
Set 1	100.0%	73.3%	84.6%
Set 2	40.5%	83.3%	54.6%
Set 3	92.3%	70.6%	80.0%
Set 4	86.4%	90.5%	88.4%
Overall	**68.7%**	**80.3%**	**74.0%**

5. CONCLUSION

This paper mainly presented a real-time segmentation algorithm that can be implemented in numerous audio annotation and semantic analysis platforms. Experimental results show that such a generic methodology for audio segmentation can be useful and has potential to improve. The algorithm does not lie heavily on power based audio features, making it independent from audio input that varies in different devices and recordings. Feature sets tryouts and ranking showed that in some cases certain features may perform extremely well or remarkably low, like MFCCs in speech.

Additionally, a strategy for crowd-sourcing annotated audio samples gathering and, generally, user generated content turns workable for populating ground truth data and should increasingly be followed in similar implementations.

6. FURTHER WORK

Considering the experimental results, the procedure's methodology and the related works by others, it is clear that much further work and testing can be performed. Many more feature sets, such as LPCs, LPCCs, LFCCs, PLPs or extra individual features should be tried out and evaluated. The algorithm's transition point detection mechanism can be enhanced, combining Dimoulas and Symeonidis' Match-ware [6] strategy with the power-individual and standardization logic presented in this paper. Finally, segmentation process can be used in coordination with classification tasks boosting the performance for both.

7. REFERENCES

[1] Atrey, P. K., Maddage, N. C., & Kankanhalli, M. S. Audio based event detection for multimedia surveillance, in *IEEE Acoustics, Speech and Signal Processing*, 2006, IEEE.

[2] Avdelidis, K., Dimoulas, C., Kalliris, G., Papanikolaou, G. Adaptive phoneme alignment based on rough set theory, in *Rough Sets and Current Trends in Computing*, 2010, Springer Berlin Heidelberg, 100-109.

[3] Bullock, J. Libxtract: A lightweight library for audio feature extraction, in *International Computer Music Conference*, *43*, 2007.

[4] Cannam, C., Landone, C., Sandler, M. B., & Bello, J. P. The Sonic Visualiser: A Visualisation Platform for Semantic De-

scriptors from Musical Signals, in *International Society for Music Information Retrieval*, 2006, 324-327.

[5] Chu, S., Narayanan, S., Kuo, C. J.Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, *17* (6), 2009, IEEE, 1142-1158.

[6] Dimoulas, C., Symeonidis, A. Syncing Shared Multimedia through Audiovisual Bimodal Segmentation, *IEEE MultiMedia*, *22* (3), 2015, IEEE, DOI:10.1109/MMUL.2015.33 2015.

[7] Freund, Y., Schapire, R., Abe, N. A short introduction to boosting, *Journal-Japanese Society for Artificial Intelligence*, *14*, 1999, 771-780.

[8] Elo, J. P., Bugalho, M., Trancoso, I., Neto, J., Abad, A., Serralheiro, A. Non-speech audio event detection, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, IEEE, 1973-1976.

[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter*, *11* (1), 2009, 10-18.

[10] Kotsakis, R., Kalliris, G., Dimoulas, C. Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification. *Speech Communication*, *54* (6), 2012, 743-762.

[11] Niessen, M., Cance, C.,Dubois, D. Categories for soundscape: toward a hybrid classification, in *Inter-Noise and Noise-Con Congress and Conference*, 2010, Institute of Noise Control Engineering, 5816-5829.

[12] Queen Mary University of London. 2013. Vamp Plugins. Retrieved June 21, 2015, from http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html

[13] Salamon, J., Gómez, E. Mir.edu: An open-source library for teaching sound and music description, in *International Society for Music Information Retrieval*, (Tapei, Taiwan, 2014.

[14] Serra, J., Müller, M., Grosche, P., Arcos, J. L. Unsupervised detection of music boundaries by time series structure features, in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[15] Silla Jr, C. N., Freitas, A. A. A survey of hierarchical classification across different application domains, *Data Mining and Knowledge Discovery 22*, 2011, 31-72.

[16] Stiefelhagen, R., Bernardin, K., Bowers, R., Rose, R. T., Michel, M., & Garofolo, J. The CLEAR 2007 evaluation, in *Multimodal Technologies for Perception of Humans*, 2008, Springer Berlin Heidelberg, 3-34.

[17] Tsipas, N., Vrysis, L., Dimoulas, C. A., & Papanikolaou, G. Content-Based Music Structure Analysis Using Vector Quantization, in *Audio Engineering Society Convention 138*, 2015, Audio Engineering Society.

[18] Vrysis, L., Dimoulas, C. A., Kalliris, G. M., & Papanikolaou, G. Mobile Audio Measurements Platform: Toward Audio Semantic Intelligence into Ubiquitous Computing Environments, in *Audio Engineering Society Convention 134*, 2013, Audio Engineering Society.

[19] Zhuang, X., Zhou, X., Hasegawa-Johnson, M. A., Huang, T. S. Real-world acoustic event detection, *Pattern Recognition Letters*, *31*, 2010, 1543-1551.

[20] Zhuang, X., Zhou, X., Huang, T. S., Hasegawa-Johnson, M. Feature analysis and selection for acoustic event detection, in *IEEE International Conference In Acoustics, Speech and Signal Processing*, 2008, IEEE, 17-20.

[21] Zhou, X., Zhuang, X., Liu, M., Tang, H., Hasegawa-Johnson, M., Huang, T. HMM-based acoustic event detection with AdaBoost feature selection, in *Multimodal technologies for perception of humans*, 2008, Springer Berlin Heidelberg, 345-353.