Experiments in Automatic Genre Classification of Full-length Music Tracks using Audio Activity Rate

Shiva Sundaram and Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory (SAIL)
Dept. of Electrical Engineering-Systems, University of Southern California
3740 McClintock Ave, EEB 400, Los Angeles, CA 90089. USA.
Email: shiva.sundaram@usc.edu, shri@sipi.usc.edu

***Abstract*— The activity rate of an audio clip in terms of three defined attributes results in a generic, quantitative measure of various acoustic sources present in it. The objective of this work is to verify if the acoustic structure measured in terms of these three attributes can be used for genre classification of music tracks. For this, we experiment on classification of full-length music tracks by using a dynamic time warping approach for time-series similarity (derived from the activity rate measure) and also a Hidden Markov Model based classifier. The performance of directly using timbral (Mel-frequency Cepstral Coefficients) features is also presented. Using only the activity rate measure we obtain classification performance that is about 35% better than baseline chance and this compares well with other proposed systems that use musical information such as beat histogram or pitch based melody information.**

I. Introduction

Automatic genre classification is a subset of the larger automatic music information retrieval problem. Here, the retrieval system is based on identifying the type of a given song/musical piece in terms of its musical genre. The genre of a given song depends on a variety of factors such as the sequence of music/vocals, the instruments present, the beats, rhythm, melody etc. It is an overall high-level label used to describe artists' style, and to categorize and organize music. Genre labels of music content are completely based on subjective interpretation of a given musical piece and in most cases, it is predefined by the artists, directors/producers or even well-informed music enthusiasts. The resulting classes and musical types are highly overlapping and it remains a challenge to develop automatic genre classification systems.

The objective of this work is to perform classification of complex audio scenes using information that represents an underlying acoustic structure of the given audio track. In this approach, a song or a music piece is seen as a complex audio scene where a variety of acoustic sources come together. The underlying structure is derived using measures that are based on generic, domain-independent audio attributes. This attribute-based approach has already been successfully applied to segmenting music signals [1]. The results of our experiments show that it is possible to discern complex audio scenes (music genres in this case) using the analysis method presented here.

There are a variety of elaborate systems that identify the music genre of an unknown musical piece with satisfactory accuracy. They are usually domain dependent since they use information (such as beats or rhythms, similarity between segments etc) that is specific to the music domain. For example in [2], the authors investigate the performance of rhythmic content features such as beat histogram and pitch content for the classification task. In [4], the study compares the performance of short-time spectral features, medium time timbral features and long time features such as beat spectrum and beat histogram. In [6], the authors develop a classification system based on textual musical transcription derived through unsupervised segmentation of an available training set. The system proposed in [7] transforms a piece of music into a sequence of abstract acoustical events derived from the activation of a trained feed-forward neural network. This work is most similar to the one presented here, in that, the authors develop a way to derive an underlying *hidden* structure, and use that interpretation for the target classification task.

In contrast to the work discussed earlier, the approach proposed here is dependent on a set of time-varying quantitative metric of generic attributes known as event *activity rate*. This time-varying metric essentially brings out the structure in a song and in this work we test the usefulness of this measure in recognizing music genres. The premise for this approach is the human auditory system which is able to quantitatively measure the interaction of the events presented to it in a scene as a whole instead of identifying the individual sources [9]. We think that this approach offers the flexibility to be applied to a variety of audio processing problems and in this work we focus on music genre classification. Mainly, we assess the performance of the activity rate measure with two experiments: (1) using Dynamic-Time Warping (DTW) [10], we recognize the genre of a given full-length test track by comparing the similarity between the activity rate measure of this track with a set of previously organized tracks and finally labeling it with the genre that is closest to it. (2) By training a classifier that models the activity rate measure as observations of Hidden Markov Models (HMM). We also present the performance obtained by modeling the extracted Mel-frequency Cepstral Coefficients (MFCC) features directly as observations of HMMs.

Due to artistic styles, a song may have sections that sound like other genres and yet the overall song may be deemed to belong to just one particular genre. For example, the reader may be aware that frequently, tracks that belong to the genre *Metal* may have sections that sound like *Rock/Pop*. Hence, the experiments are performed on full-length tracks. Also, since it allows for similarity measure based on the structure of a

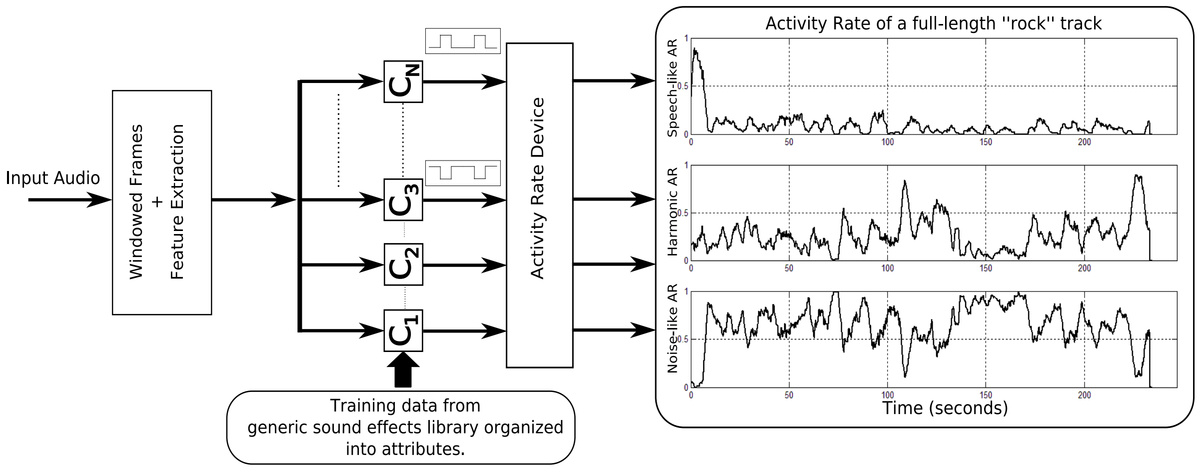

Fig. 1. Extraction of "Activity Rate" for a given clip in terms of *speech-like* (top), *harmonic* (middle), and *noise-like* (bottom) attributes.

whole song, this aspect is useful for classification using the activity rate measure with DTW.

In the next section, the details of the attributes and implementation for deriving the activity rate measure discussed. Then, the experiments performed to test the ideas discussed here are presented.

II. The Event Activity Rate Measure

As mentioned earlier, the experiments presented here are based on a time-varying quantitative metric that characterizes a given audio scene (a song in this case) in terms of three attributes defined here. The motivation for this stems from the fact that the human auditory system not only depends on identifying sources based on perceived signal level characteristics but also on the overall understanding, relevance and temporal/spatial placement of acoustic events. As originally suggested in [1], we focus on three types of attributes: *speech-like*, *harmonic* and *noise-like*. *Speech-like* mainly covers individual speech, conversations in a crowd, laughter, and human vocalizations. Examples of sounds with *noise-like* attributes are vacuum cleaners and car's engine. Other acoustic sources that have such *noise-like* characteristics are waves on a seashore, machine-shop tools, heavy rain, breathing sounds etc. Along the same lines, a wide variety of acoustic sources such as door bells, string musical instruments such as violin, guitar, (excluding percussion instruments), telephone ring-tones, sirens, beeps, pure tones etc. can be categorized under the group *harmonic* i.e. sources that are harmonically rich. These attributes were chosen because of three main reasons:

1) They are sufficiently distinct in terms of perception.
2) They are sufficiently generic and domain independent. Any given scene will have one or more of these acoustic events. For example, in a song, vocal sections are *speech-like* events and instruments such as a piano or a guitar are *harmonic* events.
3) They are also suitable for automatic processing techniques because they are accurately discernible using common acoustic features such as MFCCs [1].

By independently detecting the number of *speech-like*, *harmonic* and *noise-like* events in a given clip, it is possible to generate a 3 dimensional time-varying vector where each dimension corresponds to a quantitative measure of the individual categories for that particular instant. The next section describes measuring a given music clip in terms of this activity rate.

A. Event Activity Rate: Implementation

The complexity of a given clip can be described in terms of the different acoustic sources present in it. The activity rate is defined as the number of events (noise-like/speech-like/harmonic) detected per unit time of analysis. For example, in a song, for sections with just music, the harmonic activity rate is expected to be high whereas for sections with singing, the speech-like activity rate is expected to be high. Thus in effect, the activity rate provides an aggregate quantitative measure of interaction of the individual events. It can also be seen as a way to visualize the underlying structure of a given clip.

Fig. 1 depicts the process of extracting the activity rate for a given music clip. The first stage of the system is a feature extraction stage. It maps a given windowed audio frame of T_s duration (usually from 20 to 100 milliseconds.) to a point in a D dimensional feature space Ω. Here, the feature is a popular, perceptually-motivated 39 dimensional vector comprising of 13 Mel Frequency Cepstral Coefficients (MFCC), its delta(MFCC-D) and delta-delta (MFCC-DD). The features are relevant here because they model the front-end perception of the human auditory system and they have also been successfully applied in classification of general audio ([5] and references therein). The silence frames are treated separately using the root mean-squared energy (RMS) of the signal and they are not considered for the activity rate estimation.

The features extracted per frame are classified into one of speech-like, harmonic or noise-like attribute. We construct the identification of the attributes using a bank of classifiers (as detectors). Each classifier is trained to detect either *speech-*

like, *harmonic* or *noise-like* events. The time series of these classifier outputs are then assimilated to provide a final quantitative description in term of the three attributes. For the detection aspect, three 5-Nearest Neighbour (5-NN) classifiers were implemented in a one-against all scheme (illustrated as $C_1, C_2, \ldots, C_N$ in Fig. 1). The training data for these classifiers were obtained from 7.28 hours (approximately 2.42 hours for each attribute) of audio clips selected from the general BBC Sound Effects library [13]. The clips forming the training data of the classifiers were grouped according to the way they are interpreted after listening to them. For example, as mentioned earlier, the sound of a vacuum cleaner and the sound of waves on a seashore are both considered *noise-like*. Each of these sources are grouped based on these perceived attributes regardless of the linguistic label/class or their signal feature similarity/dissimilarity. Instead of the broad *speech-like*, *noise-like* and *harmonic* categories, further narrow, additional descriptions are also possible, however, in this study we restrict it to these three.

The output of the individual classifiers is an indicator function whose output is "1" if the j^{th} frame $w_k^j = 1$, i.e if it is classified as the corresponding k^{th} attribute. The activity rate is estimated by calculating the moving average of the classifier outputs over a defined observation time window. This can be mathematically expressed as:

$$r_m^k = \frac{1}{M_r}\left(\sum_{j=m-\frac{M_r}{2}}^{j=m+\frac{M_r}{2}} I\{w_{k,j} = 1\}\right),$$

$$\mathbf{r_m} = (r_m^1, r_m^2, \ldots, r_m^N)$$

here, w is the feature frame, r_m^k is the activity rate of the k^{th} attribute at the m^{th} time slice. M_r is the size of the window over which the activity rate is estimated and typically $T_s < M_r \times T_s$ and $M_r \approx 50$ to 100. $\mathbf{r_m}$ is the output of the Activity Rate Device which is a $N = 3$ dimensional signal. It takes continuous values in the interval $[0, 1]$. A value close to 0 indicates no activity and 1 indicates high activity for the corresponding attribute. An example output of a track from the "Rock/Pop" genre is also illustrated in Fig. 1. The activity rate can be seen as a low-dimensional representation of the underlying structure of a complex audio signal with multiple sources. It also shows the temporal evolution of the three attributes indicating the interaction of the different sources present in the audio track.

In this work we use the measured activity rate as "features" for 2 different classification approach designed to capture the temporal aspects of the activity rate signals: (1) a nearest-neighbour classifier using dynamic time-warping to measure similarity between a given test music track and a previously organized training set and (2) an HMM-based classifier that captures the temporal statistics of these measures as states and transition probabilities. We show that the time-varying structure represented by the activity rate signals is useful for genre classification of music tracks. Next the experiments performed are discussed.

TABLE I
DISTRIBUTION OF NUMBER OF SONGS IN EACH GENRE IN THE EXPERIMENTAL DATABASE

Genre	No. of songs
Classical	318
Ambient	199
Electronica	206
Metal	201
Rock-Pop	278
Total	1202

III. EXPERIMENTS

For the experiments we used 1202 full-length MPEG1-Layer 3 (mp3) tracks encoded at 128 kbps collected from the Magnatune database available online (*http://www.magnatune.com*). The distribution of number of songs in each class is shown in Table I. The mp3 codec is transparent at this bit-rate and all the songs were first converted into 44.1kHz one-channel audio before the feature extraction stage. We focus on classification into 5 genres: *Classical*, *Ambient*, *Rock-Pop*, *Metal*, and *Electronica*. The genre information for each track was obtained from the ID3 tag embedded in the mp3 format and information obtained from the website. We test the performance of the activity rate measure for the genre classification task of full-length tracks. Three experiments are designed for this purpose. The first two experiments deal with the performance of using only the activity rate measure as "features" for the classification task. In the third experiment, we also present the performance of an HMM-based classifier built using the MFCCs extracted from the full-length tracks. The details are given below:

*A. **Expt. 1**: Using Activity Rate and DTW*

The activity rate measure is a time-series of the interaction of the various attributes. Also, because there are inter-song artistic variations and duration differences we obtain a *distance* measure between the activity rate time-series of two songs using Dynamic Time Warping (DTW). DTW essentially aligns two time-series of unequal length by defining a warping function to map the samples of one time series onto the other under a set of constraints. The Euclidean distance between the time-aligned series (in this case, the activity rate measures) is calculated to be the measure of similarity between them. Since the activity rate is a time-series of the underlying structure of the song in terms of three attributes, this gives us a measure of similarity between the songs. Thus using a nearest-neighbour approach (9 nearest neighbours), by comparing a given test track with a previously organized training data, the genre of the given track can be identified.

The activity rate signals are slow-varying signals. Also, in this case, the overall temporal envelope of the signal is more important than the minor variations. Therefore, for this experiment, we perform a $4-$level asymmetric dyadic wavelet decomposition of the activity rate signals by decomposing only the low-frequency sub-band output from each level. The output after the $4-$levels of decomposition is used for computing the DTW based similarity measure described above.

B. *Expt. 2: Using Activity Rate in an HMM-based classifier*

Using the activity rate signals as features, it is possible to train a classifier to recognize the genre of a given test music track. In this experiment we train an HMM-based classifier using the Baum-Welch algorithm similar to the procedures in a continuous speech recognition task. By this we attempt to capture the genre-specific temporal statistics of the activity rate signals. For training, each track of the training set is split into ten segments with same genre labels. Again, similar to continuous speech recognition, using the trained HMMs, the final identification of a given test track was performed using Viterbi decoding. The decoding procedure results in multiple segments that are labeled with the genre with the maximum likelihood score. Then by summing the output likelihood scores of the individual segments for each genre, the label with maximum score is chosen to be the genre label for the whole song. Mathematically, this is expressed as:

$$\mathbf{L} = \underset{i}{\mathrm{argmax}} \left\{ \sum_{j=1}^{j=R_i} S_j^i \right\}.$$

In this equation S_j^i is the output likelihood score of the j^{th} segment that has been marked with i^{th} genre label. R_i is the number of segments that has been marked with the i^{th} genre label in a given track, and $\mathbf{L}$ is the overall label given to the track. Here, the number of states of the HMMs and the number of Gaussian mixtures for each state was determined experimentally, and the best performing states-mixtures combination was chosen.

C. *Expt. 3: Using MFCCs in an HMM-based classifier*

Here again, the training and test procedures are same as Experiment 2 described previously. Only, instead of using the activity rate measure, the extracted MFCCs are used directly as features. MFCC features represent the timbral qualities in music [3]. Along with the 13 MFCCs we also include the MFCC-D and MFCC-DD coefficients that also represent the inter-frame spectral changes. The features for these experiments were extracted every 50 milliseconds with a window size of 100 milliseconds. These values were experimentally determined to give the best classification performance.

The three experiments attempt the genre classification task at two different levels. Experiment 3 uses detailed signal level measures that represent spectral properties and inter-frame variations. Experiments 1 and 2 use higher-level information and represent a given music in terms of an aggregate measure in terms of the attributes. The MFCC features are directly used for classification in experiment 3. They are further processed into activity rate measure as shown in Fig. 1 and used for classification in experiments 1 and 2. It is important to point out that the activity rate measure derived in this work is based on a generic sound effects database [13] that is different from the database used for training the genre classifiers.

The results of these experiments are discussed next.

TABLE II
CONFUSION MATRIX FOR EXPT. 1: USING ACTIVITY RATE SIGNALS AND DTW FOR SIMILARITY MEASURE(USING 9-NEAREST NEIGHBOUR RULE).

class. as →	Classical	Ambient	Electronica	Metal	Rock-Pop
Classical	**97.03**	00.00	00.15	01.63	01.19
Ambient	29.00	**57.05**	00.94	02.98	10.03
Electronica	34.32	00.45	**07.13**	17.38	40.71
Metal	03.64	00.00	00.00	**74.49**	21.87
Rock-Pop	24.18	00.00	01.28	34.42	**40.11**
Chance	**20.00**	**18.96**	**19.79**	**20.36**	**20.89**

TABLE III
CONFUSION MATRIX FOR EXPT. 2: USING ACTIVITY RATE SIGNALS IN AN HMM-BASED CLASSIFIER.

class. as →	Classical	Ambient	Electronica	Metal	Rock-Pop
Classical	**71.07**	15.87	02.93	05.47	04.67
Ambient	24.15	**55.65**	04.94	10.17	05.08
Electronica	06.62	03.71	**40.00**	34.44	15.23
Metal	00.13	01.55	00.52	**91.33**	06.47
Rock-Pop	01.94	05.30	13.31	45.22	**34.24**
Chance	**20.00**	**18.96**	**19.79**	**20.36**	**20.89**

TABLE IV
CONFUSION MATRIX FOR EXPT. 3: USING MFCCS AS FEATURES IN AN HMM-BASED CLASSIFIER

class. as →	Classical	Ambient	Electronica	Metal	Rock-Pop
Classical	**99.33**	00.26	00.13	00.26	00.00
Ambient	09.89	**77.97**	04.10	02.68	05.37
Electronica	01.05	04.90	**82.11**	03.17	08.71
Metal	00.00	00.13	01.03	**90.03**	08.79
Rock-Pop	00.51	03.74	05.03	34.75	**55.94**
Chance	**20.00**	**18.96**	**19.79**	**20.36**	**20.89**

IV. RESULTS AND DISCUSSION

The classification performance for each experiment was estimated using the average of 10-fold cross-validation. For each cross-validation, 60% of the songs were randomly selected for the training set and remaining 40% was used for testing. The testing was done with approximately equal priors. For experiments 1 and 2 the MFCC features were extracted every 25 milliseconds (50 milliseconds feature window size) and $M_r = 50$ for estimating the activity rate. For experiment 2, a 10-state 16 Gaussian mixture left-to-right HMM topology was determined to have the best performance. For experiment 3, a 32 state 16 mixture Gaussian mixture left-to-right HMM topology was determined to have the best performance.

The results of experiments 1 and 2 that use the activity signals are listed in Table II, III. It can be seen that the classification performance is well above chance except the *Electronica* case. The two most confusing classes for the anomalous result for the *Electronica* case in experiment 1 are *Classical* and *Rock-Pop*. This could partly be due to the type of tracks present in the data whose acoustic structure are similar to either *Classical* songs or *Rock-Pop* songs. However, since this does not show up in the results of experiment 2 (using the same activity rate measure and a different classification approach), it could be attributed to certain issues with the similarity measure using the DTW algorithm (such as cases of non-zero mean in the signals [11]). Another observation is that the tracks that belong to *Metal* are mostly misclassified as *Rock-Pop* and vice versa. This can be attributed to the stylistic similarity between *Metal* and *Rock-Pop* tracks. This trend can be seen in the results of all the experiments shown in Table

II-IV. Another trend is the significantly better performance for the *Classical* and *Metal* classes. This essentially indicates that the content of the two genres and their structure are distinct compared to the other genre classes.

The MFCC results shown in IV are better than by just using activity rate signals. This can partly be because the activity rate experiments presented here are based on only three attributes. This can also be due to the fact that the timbral information represented by the MFCC features provide a more detailed information that is useful for genre classification task. However, a re-examination of the results by combining the classes *Classical* with *Ambient* and *Metal* with *Rock-Pop* (since they are mis-classified as the other) in experiments 1 and 3 indicate comparable performance with an average accuracy of 88.51% and 94.34% respectively. This good performance for this combined case as opposed to considering all the five genre classes can be a result of using only three attributes for the activity rate measure. As a part of our future work discussed in the next section, we would like to explore this aspect by including more attributes in the analysis.

While there are differences in data sets and the genre classes under examination, the results shown for the activity rate signals are similar to the trends observed in other systems such as [2], [3], [7]. These results of experiments 1 and 2 are also competitive with other systems that use both music-specific information [3] [4] and a purely statistical approach [8]. Thus they can appropriately capture information that is relevant to genre classification.

V. Conclusion and Future Work

In this work we present experimental results on genre classification of full-length music tracks using the activity rate of three attributes defined here. The three attributes *speech-like*, *harmonic*, and *noise-like*, are categories of acoustic sources that are based on how they are perceived and interpreted and not based on their linguistic labels. As mentioned earlier, The main motivation for this approach is the way the human auditory system performs most audio processing tasks with an overall understanding of a sound source and its properties without trying to actually recognize each and everyone of them. The activity rate measure is a quantitative metric that measures a given audio signal, resulting in a time-varying vector that represents an abstraction of the underlying acoustic structure in terms of these three attributes. Another motivation for applying this measure for genre classification task is to understand the way *non-musicians* hear music. It has been suggested in [12] that:

"*Non-musicians cannot recognize intervals, do not make categorical decisions about pitch, do not understand the functional properties which theoreticians impute to chords and tones, do not recognize common musical structures, and might not even maintain octave similarity*".

In this respect, the work presented here presents an alternative method for music analysis.

The main conclusion from our results is that it is indeed possible to recognize genre using the acoustic structure information represented by the event activity rate measure. This is feasible due to the distinct content (in terms of musical instruments and vocals), and the inter-genre styles which can be captured without actually identifying the individual sources. It is sufficient to just have an aggregate measure of the various attributes. The interesting aspect of the experiments is that the activity rate measure derived in this work is based on a *generic* sound effects database that is different from the database used for training the genre classifiers.

As a part of our future work, we would like to increase the number of attributes and also the dimensions of the activity rate signals and see if it further improves the classification accuracy. This would also be useful in recognizing more classes of music genres and developing a hierarchical scheme with better classification strategies as suggested in [2] and [3]. This would entail creating more attributes based on perception or resolving the existing attributes into smaller categories. One such idea is to resolve the *noise-like* category into *machine-generated* (such as engine noise, computer printers, fan noise etc.) and *natural* noise (such as waves on a seashore, wind noise, rainfall sounds). Other ideas include detecting the category of *impulsive* sounds. We would also like to apply the approach presented here to a wider information retrieval application such as video genre classification using audio.

References

[1] S. Sundaram and S. Narayanan, "An attribute-based approach to audio description applied to segmenting vocal sections in popular music songs", In proc. of the International Workshop on Multimedia Signal Processing (MMSP) 2006, pages 103-107, Victoria, Canada.

[2] G. Tzanetakis and P. Cook, "Music Genre Classification of Audio Signals", IEEE Trans. on Speech and Audio Processing,Vol. 10, No. 5, July 2002.

[3] N. Scaringella, G. Zoia and D. Mlynek, "Automatic Genre Classification of Music Content: A Survey", IEEE Signal Processing Magazine, pages 133-141, Vol. 23, Issue: 2.

[4] A. Meng, P. Ahrendt and J. Larsen, "Improving Music Genre Classification by Short-Time Feature Integration", In proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2005, pages 497-500, Vol. 5Philadelphia, USA.

[5] D. Li, I.k. Sethi,N. Dimitrova,T. McGee, "Classification of general audio data for content-based retrieval," in Pattern Recog. Letters, Vol.22, 533-544, 2001.

[6] J. Lee, and C. -H. Lee, "A Study on Music Genre Classification Based on Universal Acoustic Models", ISMIR 2006. 7th International Conference on Music Information Retrieval, 8 - 12 October 2006, Victoria, Canada.

[7] H. Soltau, T. Schultz, M. Westphal and A. Waibel,*"Recognition of Music Types"*, In proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1998 , pages 1137 - 1140 Vol. 2.

[8] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "Musical Genre Classification VIA Generalized Gaussian and Alpha-Stable Modeling", In proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2006. Vol.5, Toulouse, France.

[9] A. S. Bregman, "Auditory Scene Analysis: The Perceptual Organization of Sounds." London: The MIT Press, 1990.

[10] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson " Considerations in dynamic time warping algorithms for discrete word recognition", IEEE Trans. on Acoustics, Speech, and Signal Processing, pages 575-582, Vol. 26, 1978.

[11] E. J. Keogh and M. J. Pazzani, "Derivative Dynamic Time Warping", First SIAM International Conference on Data Mining 2001. Chicago IL.

[12] K.D. Martin, E.D. Scheirer, and B.L. Vercoe, "Music content analysis through models of audition," In Proc., ACM Multimedia '98 Workshop on Content Process of Music for Multimedia App., Bristol UK, (Sept.1998).

[13] "The BBC Sound Effects Library- Original Series." http://www.sound-ideas.com