International Journal of Advancements in Electronics and Electrical Engineering – IJAEEE
Volume 3 : Issue 1 *[ISSN 2319 – 7498]*
Publication Date : 09 January 2014

Performing Venue Information Retrieval from Recorded Soundtracks

-An acoustic feature extraction approach

Francis F. Li

***Abstract*—Keywords and descriptors are important metadata for multimedia content. Such metadata are associated with the programme, and are generated to facilitate indexing, search, clustering, archiving, semantic analysis and many other potential applications. Information about performing or recording venues provides a useful cue for content search and authentication, but is difficult to obtain form the media content. This paper proposes to determine the recording venues from extracted room acoustic features contained in the soundtracks. Room acoustic decay curves are obtained via maximum likelihood estimation from recording. The decay curve is a statistical description of the impulse response of a room, and provides a good discriminator for recording venues. Machine learning is then performed on the estimated decay curves to make the decision. This paper presents the rationale of the method, describes the algorithms and validates the method by simulations.**

***Keywords*—acoustic feature, recording venue, maximum liklihood estimation, machine learning, media content.**

I. Introduction

With the rapid growth of multimedia content on the Internet, ever increasing capacity of media archives, and the emerging of new media technologies such as multimedia content management systems, enhanced digital audio and video broadcasting and semantic web, methods to automate information extraction and generate metadata have received more and more attention in recent years to meet the demand of effective indexing and search of media contents. The standardization of Multimedia Contents Description Interface in MPEG-7 is an important milestone in the advancement of the technology, allowing the use of XML to store metadata and tag them alongside the actual media signals.

There exist a number of MPEG-7 encoders and automated metadata generation schemes from soundtracks. While the MPEG-7 employs up to17 temporal and spectral Low Level Descriptors (LLDs) to depict audio frames in great details for further analysis, the others takes an ad hoc approach to feature space selection [1]. At a high level, the selected LLDs or feature spaces are used to generate semantically meaningful metadata, i.e. keywords.

Francis F. Li
The University of Salford
UK
e-mail: f.f.li@salford.ac.uk

Figure 1 illustrates a typical structure of such high level metadata generators. On the first classification and segmentation layer, feature spaces are computed and processed to segment the soundtrack into speech, music or event sounds. They are time stamped and tagged for metadata generation and also sent to one of the three subsequent filtering and separation processing stages accordingly. The three different types of audio segments are cleaned using appropriate de-noising algorithms. For multiple talker speech signals, source separation may be performed where necessary and possible. Three dedicated audio recognition/classification sub-systems are used: an automated speech recognition (ASR) sub-system, a music information retrieval (MIR) system, and an event sound classification system. A final stage gathers information from previous stages, performs logical reasoning/inference and semantic analysis to generate metadata. Such a system seems sophisticated enough to generate semantic metadata about the media content, but it is difficult to acquire the information about acoustics of recording environment or performing venues.

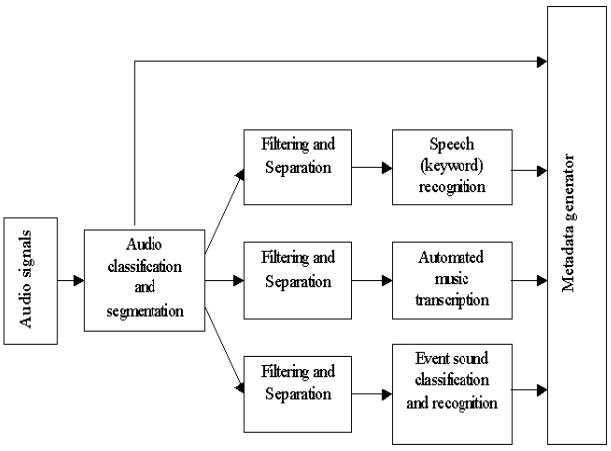

Figure 1. A typical high level metadata extraction system for audio tracks

Metadata that describe recording environment, either by specifying the venue or indicating acoustic conditions, e.g. in a large concert hall, with a long reverberation time or in a small recording studio with a very short reverberation time, are invaluable. It is also useful to be able to identify if a particular soundtrack is genuinely recorded or captured live in a particular venue or synthesized technically in a studio. Furthermore, features of acoustic environment can even be used to help authenticate recording work to some extent and therefore might be use for forensic purposes.

This paper addresses these demands and proposes the use of blind machine audition techniques for acoustic features

developed in the past few years [2-6] to determine acoustic feature from recorded soundtracks and provide extra information for additional metadata.

II. Rationale

Properties of sound propagation form sources to receiving positions in a recording space are described by acoustic transfer functions in the frequency domain or impulse responses in the time domain. The room impulse responses are determined by the geometrical shape and structure of the space, acoustic properties of interior surfaces, occupants (including furniture and audience) and source-receiver positions. Therefore the impulse responses are unique acoustic features of that space. Figure 2 sketches a typical impulse response in a concert hall.

Imagine that an impulsive sound excitation is applied in the room, shortly after the arrival of direct sound, discrete early reflections follow, and then an exponential decay process represents the reverberation [7].

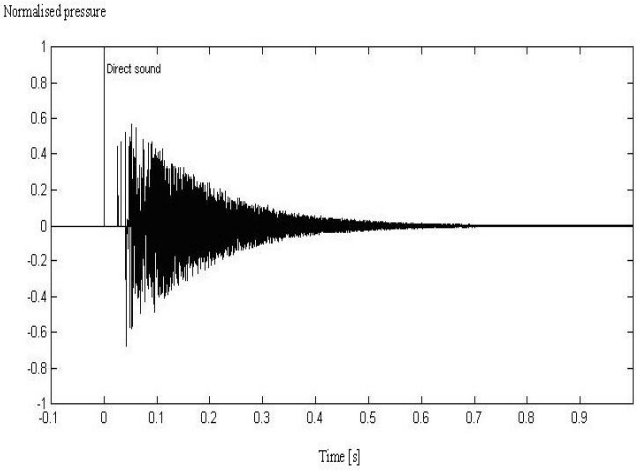

Figure 2. A typical impulse response of a concert hall

The room impulse response colours the sound sources, hence listeners perceive different acoustic effects in different venues. Such coloration is encoded in the recorded soundtracks by the convolution of the source and impulse response, since room is essentially a linear transmission system of sound.

$$r(t) = s(t) \otimes h(t) \qquad (1)$$

where *r(s), s(t) and h(t)* are received sound, source and room impulse response respectively. If impulse responses can be extracted from the recorded soundtracks and a large database of impulse responses from existing venues is made available, the recording venue can be determined from matching the impulse responses. Room impulse responses are non-minimum phase in nature. De-convolution to obtain the room impulse response is known to be an ill-posed inverse problem mathematically. Extracting the impulse responses from recorded soundtracks means that the source *s(t)* is not available. Blind de-convolution increases the level difficulty. No blind de-convolution or approximation methods developed so far can resolve the subtle details of room impulse responses necessary to effectively differentiate two different rooms, to the best of the author's knowledge. Room acoustic parameters such as reverberation parameters decay curves are statistical features of impulse responses [7]. Especially the decay curve $\tilde{h}^2(t)$, which describes how the energy level in the space reduces over time when a stationary excitation is stopped. It can be analytically calculated from the impulse response using the Schroeder backwards integration following

$$\tilde{h}^2(t) = \int_t^\infty h^2(t)dx = \int_0^\infty h^2(x)dx - \int_0^t h^2(x)dx \qquad (2)$$

In fact common monaural room acoustics parameters are evaluated from the decay curves. It is therefore postulated that room acoustic parameters and decay curves can be used as feature spaces to identify different recording spaces.

For the purpose of in-situ room acoustics measurements, a number of semi-blind and blind estimation methods for room acoustic parameters and decay curves were developed [2-6]. In particular maximum likelihood estimation method with a multi-section decay model has been proven adequate to obtain decay curves from received or recorded arbitrary sounds such as speech, music or event sound. With the estimated decay curves and the known decay curve of performance venues stored in a database, the recording venue can be determined by a machine learning scheme or simply using the Euclidean distance of the two with a suitably determined threshold.

III. Extracting Decay Curves from Soundtracks

Blind estimation of decay curve can be achieved using Maximum Likelihood Estimation (MLE) of decay phases found in speech, music or other sound signals based on a suitably chosen decay model [6]. In this study, decays found in signal envelopes are to be estimated by the MLE.

The MLE is a parametric estimation method. If there exists a parametric model for a statistical process, in the form of a probability density function *f*, then the probability that a particular set of parameters θ are the parameters that generated a set of observed data $x_1, x_2, \ldots x_n$, is known as the likelihood *L* denoted by

$$L(\theta) = f(x_1, x_2 \ldots x_n \mid \theta) \qquad (3)$$

In the MLE, an analytic model of an underlying process needs to be assumed first determined and then a likelihood function formulated. The parameters that result in a maximum

in the likelihood function are the most likely parameters that generated the observed set of data. Once the model is chosen and maximum likelihood function formulated, many existing optimization routines can be used to determine the parameter(s) by maximizing the $L(\theta)$. Let a room impulse response *h[n]* be modelled as a random Gaussian sequence *r[n]* modulated by a decaying envelope, *e[n]*.

$$h[n] = e[n]r[n] \tag{4}$$

where *n* is the sample number. The envelope is represented by a sum of exponentials:

$$e[n] = \sum_{k=1}^{M} \alpha_k a_k{}^n \tag{5}$$

where a_k represent decay rates, α_k are weighting factors and M is the number of decays. If two decay rates are chosen, it can be weighted by a single factor.

$$e[n] = \alpha a_1{}^n + (1-\alpha) a_2{}^n \tag{6}$$

where a_1 and a_2 represent the two decay rates and α is a weighting factor that changes the level of contribution from each individual decay. This enables the representation of an energy response with a non-uniform decay rate and by changing α the model can adapt to best fit the decay phases. More exponentials, as formulated in Equation 5, can be used to model the decay but at the cost of extra computational overhead when optimizing the likelihood function. However, as the purpose here is to identify the recording venues using acoustic features, the two decay rate model was previously found adequate for room acoustics modelling [6]. The likelihood function for the two decay rate model is formulated below: The likelihood of a sequence of independent, identically distributed, Gaussian variables occurring can be written as

$$L(r,\sigma,\mu) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(r[n]-\mu)^2}{2\sigma^2}\right)} \tag{7}$$

where μ is the mean and σ^2 the variance of the Gaussian process. The room impulse response model has no DC component, so $\mu=0$. For the decay phases found in reverberated sounds *s*, the envelope is of interest. Thus the probability of the sequence, which has a zero mean and is modulated by an envelope *e*, is given by

$$L(s;\sigma,e) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi e[n]}\sigma} e^{\left(-\frac{s[n]^2}{2e[n]^2\sigma^2}\right)} \tag{8}$$

rearranged to give:

$$L(s,\sigma,e) = e^{\left(\sum_{n=0}^{N-1}\frac{-s[n]^2}{2e[n]^2\sigma^2}\right)} \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \prod_{n=0}^{N-1} \frac{1}{e[n]} \tag{9}$$

The proposed decay model Equation 6 is substituted into Equation 9. It is more convenient to work with a logarithmic likelihood function, since the multiplication becomes summation. The log likelihood function becomes

$$\ln\{L(s,\sigma,a_1,a_2,\alpha)\} = -\sum_{n=0}^{N-1} \frac{\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right]^2 s[n]^2}{2\sigma^2} - N/2 \ln\left(2\pi\sigma^2\right) - \sum_{n=0}^{N-1} \ln\left[\alpha a_1{}^n + (1-\alpha)a_2{}^n\right] \tag{10}$$

Maximizing the log likelihood function with respect to the decay parameters α, a_1 and a_2 yields the most likely values for these parameters. This is achieved by minimizing the minus log-likelihood function. The Sequential Quadratic Programming (SQP) type of algorithm is found suitable for this application [8]. Once the parameters in Equation 6 are determined, the decay curve is obtained.

MLE is performed on the envelope of sound signals obtained via Hilbert Transform. A 0.5 second moving windows over typically a 60 second excerpt would be sufficient. Given the model described in Equation 3, the decay curve is completely determined by three parameters α, , a_1 and a_2.

IV. Machine Learning Approach to Venue Identification

A. *Octave band decay curves as a feature space*

As discussed in Section 2, decay curves of the recording spaces provide a good feature space to differentiate acoustics and hence the venue. Nonetheless, to further differentiate subtle discrepancies found in recording spaces, octave band features were used and found beneficial. The decay curves are estimated from recorded music signals in 5 octave bands, 250 Hz, 500 Hz, 1 kHz, 2 kHz and 4 kHz sub-bands respectively, yielding a feature space with 15 parameters (three parameters as described by Equation 6 were used for each sub-band). This is done by pre-filtering the soundtracks with octave band filters, and then performing the MLE algorithm.

B. *Artificial Neural Networks classifier*

A straightforward Euclidian distance between the MLE estimated decay curves and pre-measured and saved ones can be use as an index to quantify the level of the similarity of recording spaces. The distance provides an indication of how

likely the recording was taken in a particular space. A threshold can be determined for decision making. However, there are minor discrepancies in predicted decay curves due to different source (music) signals. In this study machine learning is used to statistically learn from examples and help better make the decision.

Twelve impulse responses acquired in different recording venues (in this study 12 different concert halls) are convolved with a variety excerpts from 20 anechoic music excerpts to generate a training/validation data set. The objective is to train the system to recognize sound tracks recorded in a particular hall from others. Six typical feed-forward artificial neural networks (ANNs) as depicted in Figure 3 are used, each is trained to detects one particular concert hall. The ANNs has two middle layers with sigmod functions and a bi-level output layer. The numbers of neurons are 15, 12, 6 and 1 on input layer, inner layers and output layer respectively. Typical training and validation regimes were followed. Results give 100% correct recognition of the 6 intended concert halls.

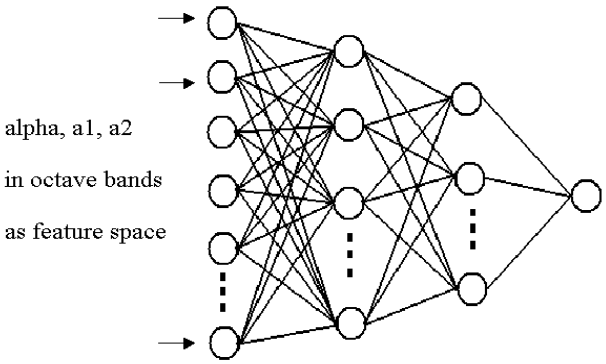

Figure 3. One of the 6 ANNs

Although the number of impulse responses used to train and validate the ANNs is not particularly large, it is worth mentioning that the method is still deemed to work very well. The 12 impulse responses used are all concert hall ones, with reverberation times circa 2 seconds. The results indicate that the use of octave band decay curve can differentiate subtle discrepancies of these concert halls, hence providing a good feature space for venue identification and authentication.

V. Discussion about Application Scenarios

Two major application scenarios are considered. The first is to determine the recording venue. The second is to indicate the type of recording spaces.

A. *A. Identify Recording venues*

As validated in section IV, a set of ANNs can be trained, with each recognize a particular venue but rejects others. This means that a fully working version of the algorithm will need to be excursively trained on all recording venues to be recognized. It works fine for venue authentication of a limited number of venues where previously recorded samples are available. One possible way to mitigate the need to collect a lot of information about recording venues and train a large number of ANNs, is to use the Euclidean distance of the proposed feature spaces as an index to determine how likely the recording is make in a particular hall.

B. *Identify the type of recording space*

The second application scenario is to determine what type of venue the recording was made in. This can be a relatively easier task. From the MLE estimation method, decay curves can be reconstructed according to Equation 6. The reverberation time (RT) can subsequently be calculated following the ISO 3382 standard, i.e. a line fitting to the logarithmic decay curve in the region from -5 to -35dB and then extrapolate line to determine the time that it takes for the energy to decay by 60 dB [9]. Based upon a knowledgebase of typical reverberation times for various possible recording spaces, the types of recording venue can be determined, e.g. small room if RT<0.6 s, large recording studio if RT is circa 1 s, or concert hall, if RT>1.5s.

C. *Compressed sound tracks*

It is known that audio classification algorithms often show performance degradation when compressed signals are presented to them. MPEG-2 Audio Layer III compressed signals down to 96 kbps have been used to test the training system, the 100% classification accuracy was maintained. This is not surprising, as the feature space is taken from the envelopes of the signals, a very low frequency statistical feature that is not affected by common audio compression.

VI. Concluding Remarks

Commonplace automatic metadata generation tools and systems do not typically extract information about acoustic conditions of recording venues. Nor would they generate recording venue related metadata from soundtracks. Room acoustics decay curves seem to offer a good feature space to differentiate different recording venues. A dual-decay model based maximum likelihood estimation can be used to adequately estimate room acoustics decay curves from recorded soundtracks. Combining the blind decay curve estimation and a machine learning algorithm trained on a database of acoustic features of various spaces, recording venues can be determined.

The results presented here are based on a pilot study with a small number of cases, more testing on a larger database and fine tuning of the algorithms are needed.

References

[1] H. Kim, N. Moreau and T. Sikora, MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval, Wiley, 2005.

[2] T. J. Cox, F. Li, and P. Darlington, "Extraction of room reverberation time from speech using artificial neural networks," Journal of AES, Vol. 49, No. 4, pp. 219-230, 2001.

[3] F. F. Li and T. J. Cox, "Speech transmission index from running speech: A neural network approach," Journal of Acoust. Soc. Am., Vol. 113, Issue 4, pp.1999-2008, 2003.

[4] F. F. Li and T. J. Cox, "Speech transmission index from running speech: A neural network approach," Journal of Acoust. Soc. Am., Vol. 113, Issue 4, pp.1999-2008, 2003

[5] F. F. Li and T. J. Cox, "A Neural Network Model for Speech Intelligibility Quantification," Applied Soft Computing, Vol. 7, Issue 1, pp. 145-155, January, 2007.

[6] P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang and J. A. Chambers, "Blind Estimation of Reverberation Parameters for Non-Diffuse Rooms," Acta Acustica united with Acustica, Vol. 93, No. 5, pp. 760-770, Sept/Oct 2007 ,

[7] H. Kuttruff, Room Acoustics, Elsevier Science Publishers Ltd., 1991

[8] R. Fletcher, Practical Methods of Optimization, John Wiley, 2000

[9] International Standard, EN ISO 3382:.Acoustics - Measurement of the reverberation time of rooms with reference to other acoustical parameters, 2000.

About Author :

Francis F. Li received a B.Eng. degree from the East China University of Science and Technology, an MPhil from University of Brighton, UK and a PhD from the University of Salford, UK. Dr Li is a senior lecturer in Acoustic and Audio Signal Processing at Salford University where he teaches a variety of modules on BSc and MSc levels, supervises PhDs, and carries out research. Prior to his current appointment, he was a senior lecturer in Computer Science at the Manchester Metropolitan University. His research interests include architectural acoustics; speech, music and multimedia signals processing; artificial intelligence and soft-computing; data and voice communications; bio-medical engineering; and instrumentation. But his major and long-standing research interest centres around computational intelligence applied to concert hall acoustics, audio signal processing and machine audition.