Soft-Computing Audio Classification as a Pre-Processor for Automated Content Descriptor Generation

Francis F. Li

***Abstract*—Soundtracks of multimedia files are information rich sources, from which much content-related information and metadata can be extracted. There exist many individual algorithms for the recognition and analysis of speech, music or event sounds, allowing for information embedded in audio format files to be retrieved or represented in a semantic fashion. However, soundtracks are typically a mixture these three different types of signals, and sometimes overlapped. Segmentation and classification therefore become essential pre-processors for audio based information retrieval and metadata generation. This paper stresses the importance of a universal audio indexing and segmentation pre-processor, proposes a high-level architecture for such a system, and presents signal processing algorithms based on soft-computing and two important but neglected feature spaces to improve the accuracy of classification.**

***Index Terms*—Audio indexing, classification, metadata, information extraction, soft-computing.**

I. Introduction

With the ever increasing demand for indexing and effective search of audio-video (AV) archives, and the emerging of new media technology such as multimedia content management systems, enhanced digital audio and video broadcasting and semantic web, automated audio information extraction and content related metadata generation have received much attention in recent years. The standardization of Multimedia Contents Description Interface in MPEG-7 is an important milestone in the advancement of technology, allowing the use of XML to store metadata and tag scenes alongside the actual media signals. It is apparent that these useful content descriptors or metadata are related to the audio and video contents, and therefore might be extracted via the analysis of audio and/or video signals using machine intelligence and soft-computing.

There exist a number of MPEG-7 encoders. As the standard specifies, they are not intended to encode the actual audio or video signals, but to extract and hence encode media related metadata. Most of them treat audio, video and scenes separately. Typically, they retrieve image, generate descriptors from video or extract fundamental elements of speech with the aim to use suitable automated speech recognition (ASR) to further obtain text and semantics details. The so-called the low level stacks mostly follow the trend of extracting fundamental elements from audio or video signals.

Examples of off-the-shelf software or systems include VideoAnnEx Annotation Tool by IBM, ADS system by Eptascape and Java based audio encoder from Sourceforge.

Soundtracks are thought to be information-rich, especially in the presence of speech. Other features of audio signals carry important information correlated to scenes and contents of the program, e.g. music may give information about mood; event sounds may indicate what events are happening. For these reasons, information extracted from speech, music and event sounds from soundtracks of multimedia files can be used to assist the generation of content descriptors. Although multi-modal semantic analysis approach might be more sophisticated at the cost of significantly complicated system and much heavier computational loads, audio based content and scene analysis is arguably adequate for many applications especially for movies and TV programmes that are rich of speech contents.

This paper focuses on audio only and proposes a framework to achieve the goal of automatic generation of content descriptors and keywords. The framework adopts a cascade pattern recognition approach to the problem. Speech and music are first detected and separated out, the residual audio excerpts are assumed to be event sounds. For each of these three categories, further classification and recognition are performed as necessary to obtain information of interest and generate tags or keywords. All these classifications are done using supervised machine learning with purposely designed signal pre-processors. The framework adopts an open architecture and therefore can work with existing off the shelf commercial applications or open source software for speech recognition, music information retrieval and event sound detection. The main function of the framework is to perform a search along soundtracks to index and segment speech, music and event sounds. The separated segments are then made ready for existing tools of users own choice to obtain information of their interests depending upon applications.

II. The Framework

Fig. 1 illustrates the architecture of the proposed framework. At the first classification and segmentation layer audio signals are processed to obtain segments of 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 signals are cleaned using appropriate de-noising algorithms. Signal to noise ratios are estimated to determine whether the excerpts are suitable for further dedicated recognition. For multiple talker speech signals, source separation may be performed where possible. Three dedicated audio recognition or classification sub-systems are used: an automated speech

Manuscript received November 23, 2013; revised January 25, 2014.

Francis F. Li is with the University of Salford, M5 4WT, UK (e-mail: f.f.li@salford.ac.uk).

DOI: 10.7763/IJCCE.2014.V3.300

recognition sub-system, a music transcription sub-system, and an event sound classification system. Final stage gathers information from the 3 dedicated sub-systems and a text mining tool is used to generate metadata or support query.

Over the past few decades, there has been much research into automatic speech recognition (ASR), automatic music transcription, noise or event sound detection, and soundscapes analysis, accumulating a number of methods, commercially available products and open source software. The proposed universal framework should allow the integration and deployment of existing knowledge and tools.

It is worth noting that ASRs and music information retrieval (MIR) tools are mostly developed with generally clean speech or noise free music. In the presence of non-trivial interferences as often encountered in soundtracks, e.g. speech with loud background music or significant ambient noise, the recognition performance is significantly mitigated. The built-in discourse analysis feature can sometimes further degrade the results. In handing soundtracks of AV archives, it is useful to determine the "quality" of the audio segments before they are fed into dedicated recognition sub-systems.

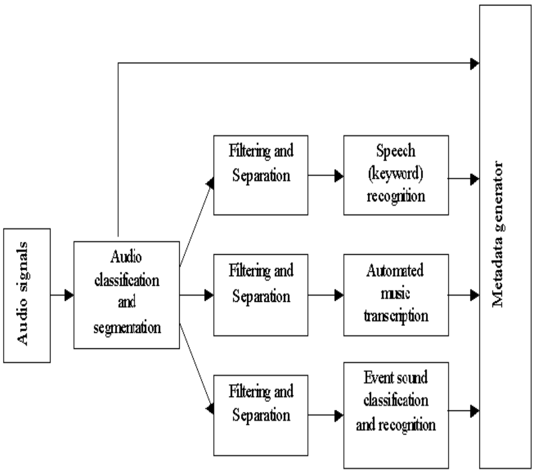
Audio signals
Audio classification and segmentation
Filtering and Separation
Speech (keyword) recognition
Filtering and Separation
Automated music transcription
Filtering and Separation
Event sound classification and recognition
Metadata generator

Fig. 1. A Block diagram of the proposed framework.

III. Audio Classification and Segmentation

The first stage in the framework aims to separate out music, speech and other event sounds. Various speech or non-speech, music or speech classification algorithms exist and their performances vary. Mel-frequency cepstral coefficients (MFCC) and zero crossing are well-known good feature spaces for speech and music. A recent study compared a number feature spaces for audio classification with neural nets as classifiers [1]. MFCC's together with zero crossing rates (ZCR) as feature spaces can offer a good performance on short frames of 30-40 ms and achieve circa 80% classification accuracy. The methods documented in [1] are used to perform the first iteration of the classification in the framework proposed.

In the case of mixed or noisy audio signals, which is common place in soundtracks of media files, the performance of classification is often mitigated. A new set of secondary classification, which is based upon long-term statistical features of speech and music signals, is further developed in this paper to perform an extra iteration of classification as confirmation.

A. Music Detection (Confirmation)

Traditional music is formed from a series of 12 equally tempered notes. MIR tools also work based on this assumption. It is therefore assumed that in the presence of music, a significant amount of signal energy will be centred on the frequencies of equal temperament notes. It is apparent this has to be calculated over a relatively long period of time (20 seconds in this study). A note-centred filter bank, with each filter tuned on a specific note in the equal temperament scale is developed to detect the signal energy. Give the fact that most melodies appear in the mid range, only four octaves from C3 (130.8Hz) to C7 (2093.5Hz) were considered, resulting in 49 outputs. Broadband energy of the signal under investigation was also calculated. The energy signals related to music notes and the signal representing total energy were fed in to a simple feed forward neural network with two hidden layers of sigmoid functions as depicted in Fig. 2.

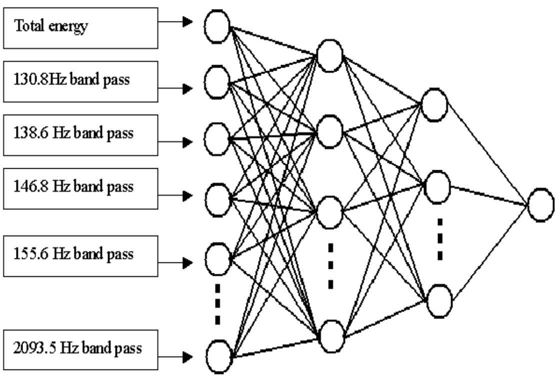

Fig 2. Music detector.

The training follows a typical supervised regime. In order to accommodate mixed cases, i.e. music with speech, such mixed cases are included in the examples for training. The purpose is this stage is to detect music, so speech is deemed as noise. Example covered cases with signal to noise ratios from infinity (noise free) to 3 dB.

As the purpose of this stage is to confirm the detected music segments and eliminate misclassified cases, testing was done with the excerpts classified as music from the first iteration. Results showed an increased recognition accuracy to 90.5%.

The presence of music means that the spectral contents will be biased towards these notes and this is likely to be better identified by comparing the note-centred spectral contents against the 1/3 octave band spectra of the audio signals. Therefore, the total energy signal is replaced with the 1/3 octave band spectral signals (extra 29 inputs to the neural network). After training the recognition rate improved to 97.0%. Thus with a simple supervised neural network and a note centred filter band, music signals are satisfactorily detected.

B. Speech Detection (Confirmation)

Speech arguably provides much more information than other two types of sounds. Speech recognition software often employs discourse analysis techniques, i.e. trying to guess unrecognised syllables or words from semantic discourse, to achieve a good recognition rate. But in the noisy signal cases, wrong recognition can have a domino effect. A secondary confirmation stage is developed to confirm speech detected from the first iteration and to give information about how intelligible or recognisable the speech excerpt is. For this particular application, it would be ideal to discard speech signals with poor intelligibility or recognisability, since this can improve the reliability of the subsequent speech recognition. A new speech detector was therefore developed to satisfy this need.

Envelope spectra of speech signals are a unique feature of speech signals. The envelope spectrum is the power spectrum of the envelope of a speech signal. The energy of the envelope signals are concentrated in the region from immediately above DC to about 15 Hz with a peak circa 4-5 Hz. Due to rhythm or pace of utterances in running speech, envelope spectra of speech are a sable feature of speech when being normalised to total signal energy [2]. Fig. 3 shows envelope spectra of several different speech signals.

The level and shape of envelope spectra was successfully used to estimate speech intelligibility or recognisability [3], [4]. Typical envelope spectra found in the signal was used in this study as an indicator for the presence of speech, while the reliability of the detection was improved by checking envelope spectra in the octave bands containing significant energy of speech signals. A speech signal has its energy concentrated from 125 to 3.4 kHz. So speech signals in five octave bands (125 Hz, 250 Hz, 500 Hz, 1 kHz and 2 kHz) were first obtained, and the envelope spectra were estimated.

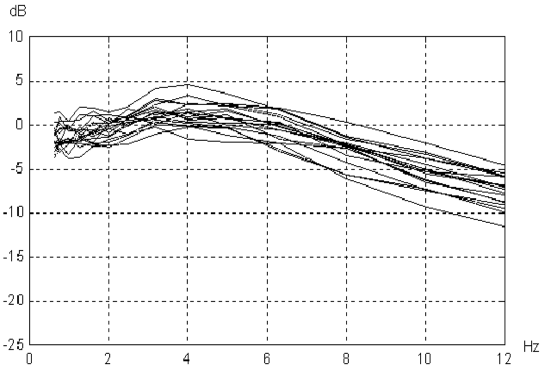

Fig. 3. Typical envelope spectra from anechoic speech signals

Hilbert transform was used to estimate signal envelopes. The envelope $e(t)$ of a sound signal $s(t)$ is obtained via

$$e(t)=\sqrt{s^2(t)+s_h{}^2(t)} \quad (1)$$

where $S_h(t)$ is the Hilbert transform of $s(t)$

$$s_h(t)=H[s(t)]\equiv\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(t-t')}{t'}dt' \quad (2)$$

Since the sound signal is digitized. The envelope signal is evaluated using discrete Hilbert transform calculated via FFT. Let s[n] denote the *nth* sample of the sound $s(t)$

$$s[n]=s(nT) \quad \text{for} \quad n\in[0,n_1,n_2,....,N-1] \quad (3)$$

where T is the sampling period and n denotes the sample number. The discrete Hilbert transform of s[n] is then calculated by:

$$H\{s[n]\}=\sum_{k=0}^{N}\{A[k]\sin\frac{2\pi kn}{N}-jB[k]\cos\frac{2\pi kn}{N}\} \quad (4)$$

where the coefficients A and B are determined by the Fourier transform

$$A[k]=Re\{\sum_{n=0}^{N}s[n]e^{-j2\pi kn/N}\} \quad (5)$$

And

$$B[k]=Imag\{\sum_{n=0}^{N}s[n]e^{-j2\pi kn/N}\} \quad (6)$$

These can be calculated efficiently using FFT. The envelope signal is obtained by

$$e[n]=\left|s[n]+jH\{s\{[n]\}\}\right| \quad (7)$$

The envelope signal e[n] is low pass filtered and power spectrum estimated to obtain the envelope spectrum from 3 Hz to 18 Hz in 1/3 octave intervals. They are normalised to total signal intensity and sent into a neural network for classification. The neural network used is similar to one in Fig. 2. The envelope spectra were obtained from 5 octave bands. One neural network is trained on envelope spectra of a specific octave band. The final decision is made, if speech envelope type of spectra appear in 4 out of 5 octave bands. Testing results show a near 100% correct classification (with the excerpts classified as speech from the first iteration).

When the envelope spectra of speech are normalised to the total signal intensity, they represent the fluctuations due to clean speech utterances above the interferences. Adverse conditions such as noise and reverberation both result in reduced levels in envelope spectra [2], [4]. This reduction in envelope spectra has empirically found to be a good indicator for recognition rate and therefore can be used to determine “quality” of speech.

In this study it is found that if envelope spectra within 2-6 Hz are below 10 dB, ASR software can hardly get any meaningful results even if speech cleaning algorithms are applied. Therefore these noisy speech excerpts are discarded.

C. Event Sound Detection

Event sound recognition is the most challenging aspect under the proposed framework. Since the ultimate goal is to recognise as many events as possible, it is assumed any residual signals above noise floor after speech and music have been detected and removed are event sounds. The event sounds can be further classified with environmental noise classification and soundscapes analysis tools.

As a pilot study, events for recognition are limited to claps, rain noise, speech babbles and car noise, the rest type of sounds are all classified in one category “others”. It is assumed that a particular event is related to the short to medium term power spectrum of its sound, i.e. the event is viewed as a function of associated spectrum of the sound. A non-linear mapping or regression model was used by assigning a scalar value to each of these events. Somewhat arbitrarily, 0, 1, 2, 3, 4 were assigned to others, claps, rain noise, babbles and car noise respectively. Thus the classification problem becomes as a regression problem.

SVM is probably best known as classifier, but it can be used to perform regression taking the advantage of high dimensional kernel induced feature space and eliminated local minima [5]-[9]. The SVM model used is the well-known ε-SVM for regression based on the epsilon-insensitive loss function proposed by Vapnik [4]. Kernel function is important in SVM development. Radial basis functions (RBF) were chosen, so the kernel function is

$$\phi(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-\mu \|\boldsymbol{x} - \boldsymbol{x}_i\|^2) \quad (8)$$

where μ is a kernel width related parameter, **x** is the input vector.

The algorithm minimises the error function

$$Min\left\{C\left(\sum(\xi_i + \xi'_i)\right) + \frac{\boldsymbol{w}^T\boldsymbol{w}}{2}\right\} \quad (9)$$

With the constraints lack variables ξ's are nonnegative;

$$y_i - \boldsymbol{w}^T\phi(\boldsymbol{x}_i) \le \varepsilon + \xi_i\,, \quad (10)$$

and

$$\boldsymbol{w}^T\phi(\boldsymbol{x}_i) - y_i \le \varepsilon + \xi'_i \quad (11)$$

where *y* is the expected output, C is a user selected scale, ***w*** is the weight matrix and ε is the parameter for epsilon SVM. Input vectors are normalised 1/3 octave band power spectra of event sounds. Since the epsilon-SVM is a regression algorithm and it yields continuous scalar outputs. They were rounded to integers to indicate the predefined types of events.

In this pilot investigation, a closed set of samples was used, and random noise with a brown spectrum was used to represent “other events”. Test result shows 82% accuracy in classification of the 5 different events.

IV. Concluding Remarks

A framework for automatic generation of keywords or metadata for AV programs from audio signals was developed. It takes relatively straightforward approach to the problem, but can potentially acquire some important information from archived AV programs.

Pilot investigation into some practical classification algorithms shows that music speech and event sounds can be accurately classified and segmented using fairly straightforward algorithms. Further development and investigation are needed to enrich the type of recognizable event sounds. It will also be interesting to develop and refine speaker independent automatic speech recognition and semantic analysis algorithms to advance the automated metadata generation techniques.

References

[1] M. Al-Maathidi and F. F. Li, “NNET based audio content classification and indexing system,” *International Journal of Digital Information and Wireless Communications*, vol. 2, no. 4, pp. 66-78, 2012.

[2] H. J. M. Steeneken and T. Houtgast, "The temporal envelope spectrum of speech and its significance in room acoustics," presented at 11th ICA conference publication, Paris, 1983.

[3] H. J. M. Steeneken and T. Houtgast, “A physical method for measuring speech transmission quality,” *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318-326, 1980,

[4] 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, 2007.

[5] V. N. Vapnik, *Statistical Learning Theory*, Wiley, 1998

[6] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[7] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*, Cambridge Press, 2000.

[8] L. Wang, *Support Vector Machines, Theory and Applications*, Springer- verlag, Berlin, 2005.

[9] J. U. Duncombe, “Infrared navigation—part I: an assessment of feasibility,” *IEEE Trans. Electron Devices*, vol. ED-11, pp. 34-39, Jan. 1959.

Francis F. Li was born in Shanghai, China, in 1963. He 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. Francis’ 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.