World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:8, 2014

Spatial Audio Player Using Musical Genre Classification

Jun-Yong Lee, Hyoung-Gook Kim

***Abstract*—**In this paper, we propose a smart music player that combines the musical genre classification and the spatial audio processing. The musical genre is classified based on content analysis of the musical segment detected from the audio stream. In parallel with the classification, the spatial audio quality is achieved by adding an artificial reverberation in a virtual acoustic space to the input mono sound. Thereafter, the spatial sound is boosted with the given frequency gains based on the musical genre when played back. Experiments measured the accuracy of detecting the musical segment from the audio stream and its musical genre classification. A listening test was performed based on the virtual acoustic space based spatial audio processing.

***Keywords*—**Automatic equalization, genre classification, music segment detection, spatial audio processing.

I. Introduction

NUMEROUS applications are currently envisioned for immersive audio systems. Sound heard over headphones is ideally suited for mobile applications. The use of stereo headphones or stereo speakers on mobile devices enables to take advantage of binaural technology which can provide an immersive sound experience for a variety of applications ranging from stereo widening of music to full 3-D positional audio. Advances in audio are going to help bring in richer multimedia, increase quality of mobile music and help create more interactive and immersive audio applications.

In this paper, we propose a smart music player that combines the musical genre classification and the spatial audio processing. In the proposed method, the spatial audio generated from the virtual acoustic space is boosted with the given frequency gains based on the musical genre in order to provide an immersive experience to the user.

A number of music genre classification methods have been proposed in [1]-[3]. In [1], three feature sets, to represent timbre texture, rhythmic content, and pitch content, are extracted. Three kinds of pattern recognition classifiers, Gaussian classifier, Gaussian mixture model (GMM), and k-nearest neighbor, are investigated. Li and Tzanetakis [2] compare the relative importance of feature sets proposed in [3]. They introduce the support vector machine and linear discriminate analysis as additional comparative classifiers. McKinney [3] evaluates the performance of four audio feature sets, including low-level signal parameters, Mel-frequency cepstral coefficients (MFCC), psychoacoustic features, and auditory filter bank temporal envelopes, and performs quadratic discriminate analysis as a classifier.

J.Y. Lee and H.G. Kim are with the Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Republic of Korea (e-mail: jasonlee88@nate.com)

The virtual acoustics concept [4] has in recent years expanded to cover even more areas, ranging from physical modeling of sound sources, via room acoustics rendering, to modeling of spatial hearing cues. Space acoustic modeling techniques [5], [6] can be divided into wave-based, ray-based, and statistical methods.

This paper is organized as follows. Section II describes the proposed method. Section III discusses the experimental results. Finally, Section IV presents the conclusion.

II. Proposed System

The proposed system is illustrated in Fig. 1. Fig. 1 shows a simplified block diagram of the investigated spatial audio player using musical genre classification.

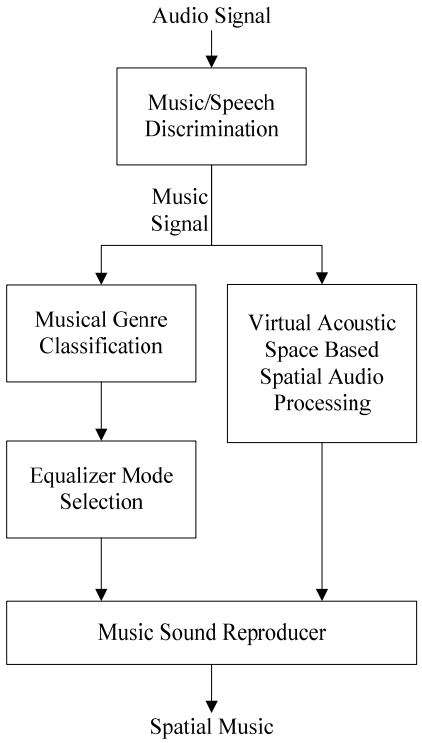


Fig. 1 Block diagram of the proposed system

The system is comprised of five modules: music/speech discrimination, musical genre classification, equalizer mode selection, virtual acoustic space based spatial audio processing, and music sound reproducer.

First, the music part is detected in the input audio stream from the music CD or the radio broadcast.

In the musical genre classification, the detected music segment is classified into one of musical genres. The equalizer mode is automatically selected based on the result of the

musical genre classification.

In the virtual acoustic space based spatial audio processing, the spatial audio is achieved by adding an effect such as artificial reverberation to the sound source. The spatial audio processing emulates a virtual acoustic space by distributing the room reflections over the stereo sound filed and decorrelating the stereo signals to add a sense of the virtual space.

Once this has been done, the musical sound is reproduced with the given frequency gains based on the selected equalizer mode by both the musical genre and the spatial impression when played back.

A. Musical Genre Classification and Equalized Mode Selection

The consecutive audio streams are first divided into audio clips by a one-second window without overlap. Each audio clip is transformed into feature vector sequences. These features consist of spectral centroid, spectral roll-off, spectral flux, zero crossing rate, and advanced MFCC (A-MFCC).

Of these five features, the A-MFCC feature plays an important role to improve the performance of classification for the musical genre classification. For extracting A-MFCC features, the audio signal is divided into overlapping frames by a hamming window function and analyzed using the short-time Fourier transform. The spectral coefficients are grouped in critical frequency bands using a series of triangular filters whose central frequencies are spaced based on the Mel-scale. The output $Mel(f,l)$ of the Mel-scale frequency range is the sum of the spectrum in each critical frequency band and is converted to the decibel scale.

$$D(f,l) = 10\log_{10}(Mel(f,l)) \tag{1}$$

where f, l denotes the Mel-scale frequency bin index and frame index.

Each decibel-scale Mel-spectral vector is normalized with the Root Mean square (RMS) energy envelope, thus yielding a normalized Mel-spectral version of $Mel(f,l)$. The full-rank features for each frame l consist of both the RMS-norm gain value R_l and the normalized Mel-spectral vector $X(f,l)$:

$$R_l = \sqrt{\sum_{f=1}^{F}(10\log_{10}\{Mel(f,l)\})^2}\ ,\ 1 \le f \le F \tag{2}$$

and

$$X(f,l) = \frac{10\log_{10}\{Mel(f,l)\}}{R_l},\ 1 \le l \le L \tag{3}$$

where F is the number of the critical bands and L is the total number of frames.

A discrete cosine transform applied to the normalized Mel-spectral vectors of the filter bank outputs results in vectors of decorrelated features. They are defined as:

$$M(d,l) = \sqrt{\frac{2}{F}}\sum_{f=1}^{F}\left(X(f,l)\times\cos\left[d\left(k-\frac{1}{2}\right)\frac{\pi}{F}\right]\right) \tag{4}$$

where d is the number of cepstrum coefficients. Usually the desired length of the cepstrum $D<<F$ is used to reduce the dimension.

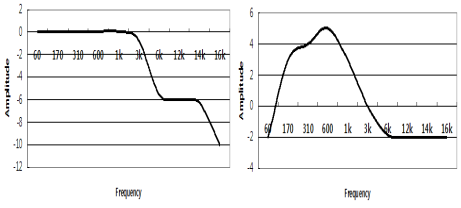
The A-MFCC is extracted based on the multiplication of $M(d,l)$ and an exponential component that depends on the outcome of $M(d,l)$. $M(d,l)$ is multiplied with an exponential component based on the mean m and variance σ of $M(d,l)$, resulting in a form as follows:

$$M_A(d,l) = M(d,l)\cdot\exp^{-\frac{1}{2}\left|\frac{M(d,l)-m}{\sigma}\right|} \tag{5}$$

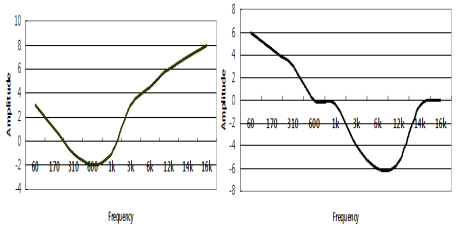
The music clip may include of a variety of musical types, including classical music, pop music, jazz music, dance music, and rock music. Any of the foregoing musical types may benefit from equalizer control, because there is a response difference based on frequency.

The musical genre classifier classifies each music clips as one of five musical genres (i.e., classic, pop, rock, dance, jazz) using low-complexity GMM, and provides a sound mode to the equalizer.

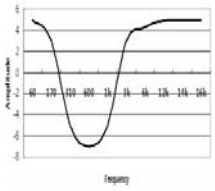
Fig. 2 represents the frequency response based on the classified music genre.



(a) Classic (b) Pop



(c) Jazz (d) Dance



(e) Rock

Fig. 2 Frequency response based on musical genre

Fig. 2 (a) shows that when the input musical data represent classical music, the frequency response is fixed from 60 Hz to 3 kHz and from 6 kHz to 14 kHz, decreases from 3 kHz to 6 kHz and from 14 kHz to 16 kHz. When the input musical data represent pop music, the response of the frequency increases from 60 Hz to 600 Hz, decreases from 600Hz to 6 kHz, and is

fixed from 6 kHz to 16 kHz (Fig. 2 (b)). When the input musical data represent jazz music, the response of the frequency decreases from 60 Hz to 600 Hz, and increases from 600 Hz to 16 kHz (Fig. 2 (c)). In the case of dance music, the frequency response decreases from 60 Hz to 600 Hz and from 1 kHz to 6 kHz, increases from 12 kHz to 14 kHz, and is fixed from 600 Hz to 1 kHz and from 6 kHz to 12 kHz (Fig. 2 (d)). In the case of rock music, the frequency response decreases from 60 Hz to 600 Hz, and increases from 600 Hz to 16 kHz (Fig. 2 (e)).

B. Virtual Acoustic Space Based Spatial Audio Processing

The virtual acoustic room based sound processing is comprised of five stages as shown in Fig. 3.

First, the dimensions of the virtual room, absorption coefficients, and source/receiver positions are selected by the user. The absorption coefficients could be also selected for each surface of the virtual room. The source position is set to be in the centre of the stage and at a height of 1.4m.

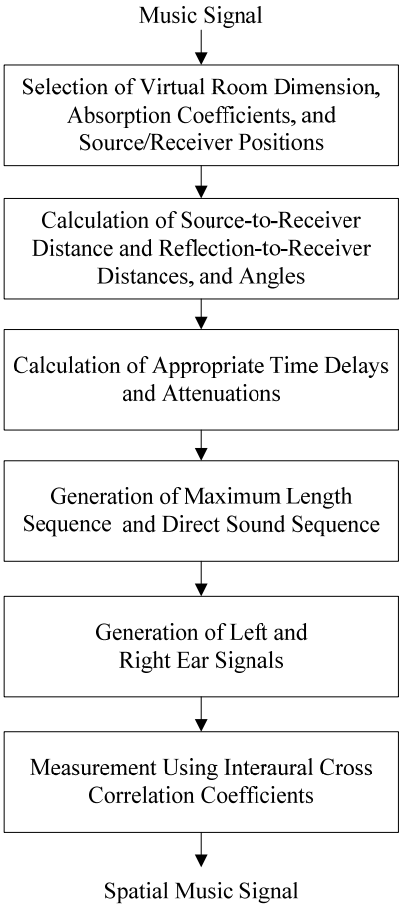


Fig. 3 Block diagram of virtual acoustic room based sound processing

The source is assumed to be omnidirectional. As the virtual room is symmetrical about its centre line, receiver positions are only required for one half of the virtual acoustic hall. For each of the 24 receiver positions, the source-to-receiver and reflection-to-receiver distances and angles are calculated using geometric methods.

For each receiver position, attenuations due to path differences and absorption, and time delays due to path differences are calculated for the six, first-order reflections. Six, single period identical copies of a 16383-pont maximum length sequence (MLS) signal are generated then delayed and attenuated accordingly.

In order to obtain interaural time and level differences, each of the seven MLS signals (the direct sound and the six reflections), particular to a virtual seat position, are convolved with a head-related impulse response (HRIR) [7] that corresponded to the source-to-receiver or reflection-to-receiver angle. The binaural ear signals are derived from a widely spaced pair of omnidirectional microphones. The HRIRs used in the virtual room based processing are taken from Gardner and Martin's set of anechoic KEMAR head measurements [8].

Whilst covering a large number of possible source positions, this set of HRIRs has a limited angular resolution ranging from 5° to 30° in azimuth and 10° in elevation. For each source and reflection-to-receiver angle calculated, the HRIR that is closest to the intended angle is selected for convolution. To obtain binaural recordings, the left and right ear signals created in Stage 4 are summed for the virtual room based sound reproductions.

The impulse response is extracted from these binaural recordings of the MLS signals and interaural cross-correlation coefficient (IACC) measurements are calculated for the spatial impression effect.

$$IACC(\tau) = \frac{\left[\int_{t_1}^{t_2} x(t)y(t+\tau)dt\right]}{\left[\int_{t_1}^{t_2} x^2(t)dt \int_{t_1}^{t_2} y^2(t)dt\right]^{1/2}} \quad (6)$$

The IACC is a measure of similarity between the signals reaching the left and right ears in a sound filed. The less similar the signals are the greater the perception of spatial impression.

If the spatial impression effect is decided by the IACC measurement, the music sound is output through the speaker.

III. Experimental Results

We investigated the proposed algorithm with an audio database collected from six different Internet radio stations. The audio database is approximately 10 hours long. It contains a wide range of speakers and five musical genres (classical, jazz, pop, rock, jazz).

Table I compares the performance of different feature extraction methods in music/speech discrimination using 9s-length musical segments and presents recall and precision values for music and non-music discrimination. Recall is the proportion of data with a true classification label correctly classified in that class. Precision is the proportion of data classified as class *i* whose true class label is indeed *i*.

In the Table I, FS A denotes a feature set including spectral centroid, spectral flux, spectral roll-off, zero crossing-rate, and the well-known MFCC. FS B represents a feature set including spectral centroid, spectral flux, spectral roll-off, zero

crossing-rate, and improved MFCC, while FS C is composed of spectral centroid, spectral flux, spectral roll-off, zero crossing-rate, and the proposed 'A-MFCC'.

TABLE I
MUSIC GENRE CLASSIFICATION CONFUSION MATRIX

	Music		Speech	
	Precision	Recall	Precision	Recall
FS A	92.25%	84.37%	94.04%	83.75%
FS B	94.55%	88.56%	96.25%	87.33%
FS C	96.39%	93.54%	97.74%	92.31%

From Table I, it can be seen that FS C achieves better results than FS A or FA B.

The performance of the musical genre classification using three feature sets is compared in Table II. The best results are obtained using FS C.

TABLE II
RESULTS OF MUSIC GENRE CLASSIFICATION

	FS A	FS B	FS C
Accuracy	84.30%	85.58%	90.24%

For measuring spatial sound quality using virtual acoustic room based sound processing, we used a mean opinion score (MOS) experiment and IACC measurements. The height, length, and width of the virtual acoustic room are 10m, 10m, and 10m, respectively.

In the MOS experiment, we asked subjects to rate the listening quality of twenty musical songs, four songs for each musical genre, using a 5-point scale: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. A total of 12 listeners participated in the test. The MOS scores for music playback are in the range of "excellent" to "good". The range of the IACC measurements in the virtual acoustic room has a minimum of 0.081 and a maximum of 0.857

IV. CONCLUSION

In this paper, we presented a spatial audio player using musical genre classification. As seen from the experimental results of the music/non-music discrimination and musical genre classification, the proposed feature, namely 'advanced MFCC', improves classification accuracy compared to MFCC or improved MFCC. The spatial sound generated by virtual acoustic space based audio processing was boosted with the given frequency gains based on the musical genre when played back and provide an immersive experience to the user.

We will apply the proposed spatial audio player using musical genre classification to various multimedia players as part of further research.

ACKNOWLEDGMENT

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1019) supervised by the NIPA (National IT Industry Promotion Agency)

REFERENCES

[1] G. Tzanetakis, P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, Jul. 2002.

[2] T. Li, and G. Tzanetakis, "Factors in automatic musical genre classification of audio signals," *in Proc. WASPAA 2003*, pp. 143-146, Oct. 2003.

[3] M. McKinney, and J. Breebaart, "Features for audio and music classification," *in Proc. ISMR2003*, pp. 151-158, Oct. 2003.

[4] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head-tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source", *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904-916, Oct. 2001.

[5] R. A. Reale, J. Chen, J. E. Hind, J. F. Brugge, "An implementation of virtual acoustic space for neurophysiological studies of directional hearing," *Virtual Auditory Space: Generation and Applications*, pp. 153-170, 1996.

[6] U. P. Svensson and U. Kristiansen, "Computational modeling and simulation of acoustic spaces," *Proc. AES 22nd Conf. on Virtual, Synthetic and Entertainment Audio*, pp. 11-30, Jun. 2002.

[7] F. Freeland, L. Biscainho, and P. Diniz, "Efficient HRTF interpolation in 3D moving sound," *Proc. AES 22nd Conf. on Virtual, Synthetic and Entertainment Audio*, pp. 106-114, Jun. 2002.

[8] B. Gardener, and K. Martin, "HRTF measurements of a Kemar dummy-head microphone," *MIT Media Lab Perceptual Computing*, May. 1994.