A Combination of MCLT Peak-Pair Based Audio Fingerprinting and Spatial Audio Reproduction

Jun-Yong Lee
Department of Electronics Convergence Engineering
Kwangwoon University
Seoul, Republic of Korea
Jasonlee88@kw.ac.kr

Hyoung-Gook Kim
Department of Electronics Convergence Engineering
Kwangwoon University
Seoul, Republic of Korea
hkim@kw.ac.kr

***Abstract*—This paper proposes a spatial audio reproduction system connected with the audio fingerprinting for providing an immersive experience to the music user. The proposed system combines the audio fingerprinting and the spatial audio processing. In the proposed system, a salient audio peak-pair fingerprint based on modulation complex lapped transform improves the accuracy of the audio fingerprinting in real noisy environments and spatial audio using diffusion metadata gives a listener a sensation of being listening to the sound in the space, where the sound is actually recorded.**

Keywords-audio fingerprinting; spatial audio reproduction; diffusion metadata; modulation complex lapped transform

I. Introduction

Recent advances in computation, displays, and networking technology have brought about many new interactive multimedia applications. Most of these applications strive to provide an immersive experience to the user, e.g., improve image quality by providing high resolution displays, improve audio quality by providing three-dimensional audio spatialization system, improve responsiveness by adopting powerful CPU/GPUs, enlarging network bandwidth and shortening network delay, improve system robustness by having quality monitoring and management, content-based multimedia information retrieval, security solutions, etc.

Thanks to these technology advances and trends, digital music has evolved in the form of new services. Recently, technologies of music identification and spatial audio reproduction have received wide attention independently of each other.

Audio fingerprinting techniques [1] are meant for successfully performing content-based audio identification even when audio signals are distorted. Common uses include query-by-example music or advertisement identification [2] [3], broadcast monitoring [4], copyright detection, and automatic audio content library organization [4][5]. A good fingerprint should capture and characterize the essence of the audio content. More, specially, the quality of a fingerprint can be measured in four dimensions: discriminability, robustness, compactness and efficiency.

Various methods [6] have been proposed to satisfy several practical requirements for a successful audio fingerprinting system. Among various algorithms, the system developed by Wang [7] has been considered as a commercially successful and widespread work. Besides, the robust hash algorithm proposed by Haitsma et al. [5] is also a well studied content-based music identification or retrieval technique. In practice, it still needs further improvement to be used in a real environment.

Spatial audio-related techniques [8] in general attempt to deliver the impression of an auditory scene where the listener can perceive the spatial distribution of the sound sources as if he or she were in the actual scene. The audio spatialization system renders a virtual sound image in order for the listener to feel as if the signals were emitted by a source located at a certain position in 3D space [9][10]. Either headphones or a small number of loudspeakers (two in our system) can synthesize such spatialized audio effects, though the latter is often more appealing in immersive applications since it does not require the user to wear headphones.

Spatial audio has been developed for many years by reproduction techniques such as ambisonics [11], wave field synthesis [12], amplitude panning [13], and binaural synthesis [14]. The wave field synthesis renders a whole sound field to the room through a large number of loudspeakers. Nevertheless, such a solution is expensive and non-scalable. Ambisonics and amplitude panning are widely used panning techniques. In both methods, the virtual sound source is rendered at various locations by controlling the output amplitude of the loudspeakers. When two loudspeakers are available, however, they can only reproduce virtual sources in the line segment between loudspeakers. In addition, results degrade significantly if the user gets closer to one of the two loudspeakers. Binaural synthesis is capable of placing the virtual sound beyond the loudspeakers' boundaries due to the use of the head related transfer functions (HRTF) that faithfully represents the transfer function between the sound sources and human ears.

Until now, audio fingerprinting and spatial audio coding techniques have been developed independently of each other.

In this paper, we propose a spatial audio reproduction system connected with the audio fingerprinting for providing an immersive experience to the music user. The proposed approach used in this paper has two advantages: (1) The proposed algorithm improves robustness of the audio fingerprinting in various noisy environments; (2) The spatial audio encoding and reproduction of diffuse sound delivers high spatial impression in multichannel surround sound systems.

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

II. THE PROPOSED SYSTEM

The proposed system using a combination of audio fingerprinting and spatial audio reproduction is illustrated in Figure 1.

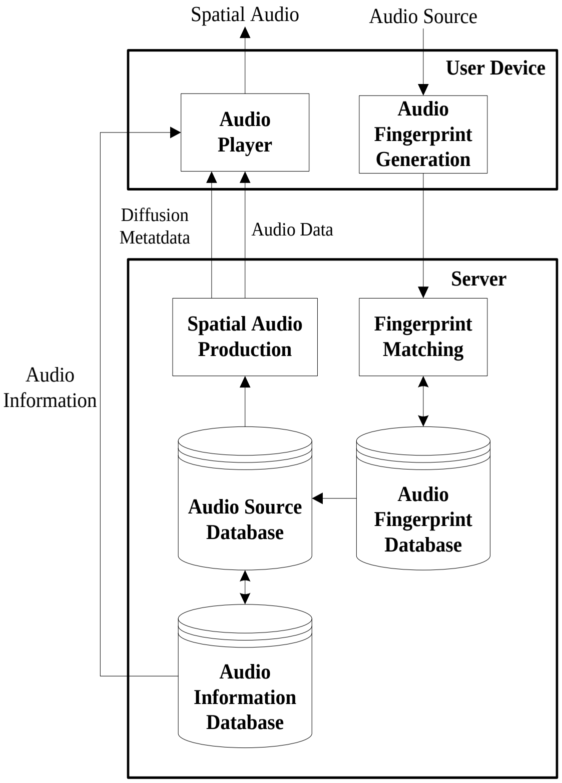


Figure 1. Block diagram of the proposed system.

The system is comprised of several modules: audio fingerprint generation, fingerprint matching, spatial audio generation, and spatial audio playback.

First, a fingerprint client, such as portable consumer device captures an audio clip that is a few seconds-long, and then extracts a robust fingerprint and submits it to the server. The extracted fingerprint is then used to query the audio fingerprint database and is compared with the stored fingerprints. If a match is found, the resulting track identifier is retrieved from the audio source database and used again as query for searching user reviews in the audio information database. If the user wants to hear the retrieved music, audio tracks are compressed and transmitted from the server to the user device in connection with synchronized spatial metadata representing diffusion and preferably mix and delay parameters. The separation of audio stems from diffusion metadata facilitates the customization of playback at the receiver, taking into account the characteristics of the local playback environment.

A. *MCLT peak-pair based audio fingerprinting*

For the robust fingerprint extraction against noise and distortion, we propose to use modulation complex lapped transform (MCLT) [16] based peak-pairs. MCLT is a cosine-modulated filter bank that maps overlapping blocks of a real-valued signal into complex-valued blocks of transform coefficients. Thus, MCLT basis functions are found in pairs to produce real and complex parts separately. These basis functions are phase–shifted versions of each other. Since MCLT has approximate shift invariance properties

As shown in Figure 2, the robust MCLT peak-pair-based fingerprint extraction method is composed of six main blocks.

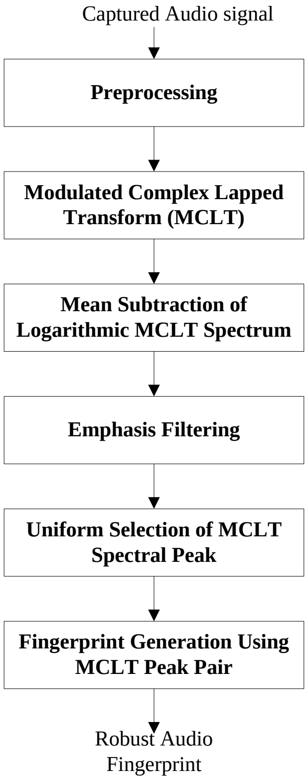

Captured Audio signal
Preprocessing
Modulated Complex Lapped Transform (MCLT)
Mean Subtraction of Logarithmic MCLT Spectrum
Emphasis Filtering
Uniform Selection of MCLT Spectral Peak
Fingerprint Generation Using MCLT Peak Pair
Robust Audio Fingerprint

Figure 2. Block diagram of the robust audio fingerprint extraction.

First, a stereo audio signal captured by a user's mobile phone is converted into mono and then downsampled to 16 kHz. The converted signal is divided into overlapping frames by the application of a Hanning window function (each of which contains 512 overlapped samples). In order to find the spectral peaks, an MCLT is then applied to each frame (1024 samples), given by

$$S_{MCLT}(k,l)=\sqrt{V(k,l)\cdot V(k,l)+V(k+1,l)\cdot V(k+1,l)} \qquad (1)$$

using

$$V(k,l)=b(k,l)\cdot U(k,l), \qquad (2)$$

$$U(k,l)=\sqrt{\frac{1}{2N}}\sum_{n=0}^{2N-1}x(n+lM)h(n)\exp\left(\frac{-j2\pi kn}{N}\right), \qquad (3)$$

$$b(k,l)=W_8(2k+1,l)\cdot W_{4K}(k,l), \qquad (4)$$

$$W_T(r,l)=\exp\left(\frac{-j2\pi r}{T}\right) \qquad (5)$$

where k is the frequency bin index, l is the time frame index, h is an analysis window of size N, M is the framing step, and $U(k,l)$ is the normalized $2N$-point FFT of input audio signal.

A log spectrum is generated by taking the log modulus of each MCLT coefficient. From the logarithmic MCLT spectrum, a frequency-time averaged MCLT spectrum is calculated and subtracted, thus yielding a normalized logarithmic MCLT spectrum.

To increase the local spectral peaks of high frequencies against attenuation distortion, an emphasis filter is applied to each normalized logarithmic MCLT spectrum.

The emphasis-filtered MCLT spectral peaks are fed into a uniform selection step shown in Figure 3, where the salient peaks are selected by applying appreciative forward and backward filtering using a dynamic peak-picking threshold.

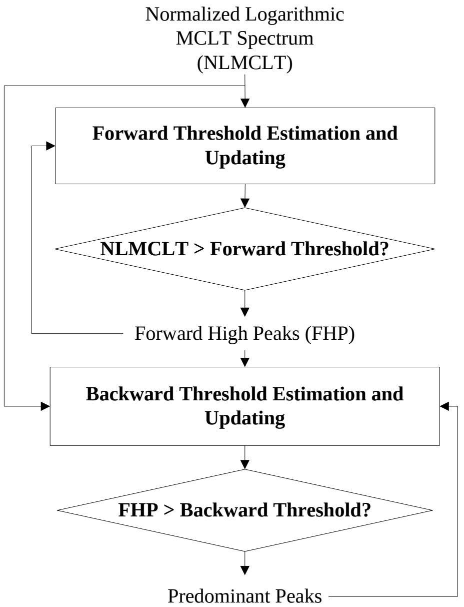

Normalized Logarithmic
MCLT Spectrum
(NLMCLT)
Forward Threshold Estimation and Updating
NLMCLT > Forward Threshold?
Forward High Peaks (FHP)
Backward Threshold Estimation and Updating
FHP > Backward Threshold?
Predominant Peaks

Figure 3. Block diagram of forward and backward filtering.

In a local target area of the frequency-time plane, nearby salient MCLT peaks are combined into a pair or landmark. Landmarks are 4-tuples of the start time, start frequency, end frequency, and time difference, and are converted into hashes with a start time.

Assuming that $P_b(k_a,l_a)$ is the anchor point and paired with another landmark point $P_b(k_p,l_p)$, its landmark L is obtain by:

$$L=(l_a,k_a,k_p,l_p-l_a)=(l_a,\Delta k,\Delta l) \qquad (6)$$

All k (in frequency bins) and l (in frames) are integers with a fixed higher bound, so each landmark point generates a fixed number of pairs. Due to the horizons, Δl is limited to a lower value than l_p.

And then, k_p, l_p and Δl are combined into a 32-bit hash

$$hash=k_a\cdot 2^{(m+n)}+\Delta k\cdot 2^n+\Delta l \qquad (7)$$

using

$$m=\log_2\Delta k \text{ and } n=\log_2\Delta l \qquad (8)$$

The robust fingerprint generated in consumer devices is submitted to the server for content-based identification.

When building the audio fingerprint database at the server, a database index is created by a fingerprint hash, and a track ID and time offset of the hash are stored according to the hash value to facilitate fast processing.

In retrieval or identification processing, the similarity searches of audio are performed in the fingerprinting domain. A query signal is fingerprinted in the user's mobile phone, and the resulting hashes are compared against the hashes stored in the database hash table. After the entire matching hashes are found, a candidate set of match segments can be obtained by combining the track ID stored in the database and the time offset of the hash in the query audio. If the files match, matching hashes should occur at similar relative offsets from the beginning of the matching file.

B. Spatial audio generation and playback

One of the primary objectives of spatial audio reproduction technique is realistic perception of the delivered contents by the consumer. The structure of our system is based on Jot's approach [16], which processes multi-channel audio by encoding, transmitting or recording audio tracks in synchronous relationship with time-variable metadata controlled by a content producer and representing a desired degree and quality of diffusion.

The spatial audio reproduction system is mainly divided into two blocks: spatial audio generation in the server side and spatial audio playback in the receiver side.

1) Spatial audio generation

Figure 4 is a system level schematic diagram of the spatial audio generation and encoding aspect.

Audio source data are converted into digital audio signals by multi-channel microphone apparatus. A metadata production engine processes the audio signal data under control of mixing, acoustic reflections, perceived direction and distance of signals via input device using multi-channel microphone apparatus, monitoring decoders, and monitoring speakers.

The metadata generated by the metadata production engine includes a representation of reverberation parameters, mixing coefficients, and inter-channel delay parameters. The metadata will be time varying in increments with the frame metadata pertaining to specific time intervals of the corresponding audio data. The time-varying audio data is encoded by a multichannel spatial audio encoder [17], to produce the encoded audio data in a synchronous relationship with the corresponding metadata pertaining to the same times.

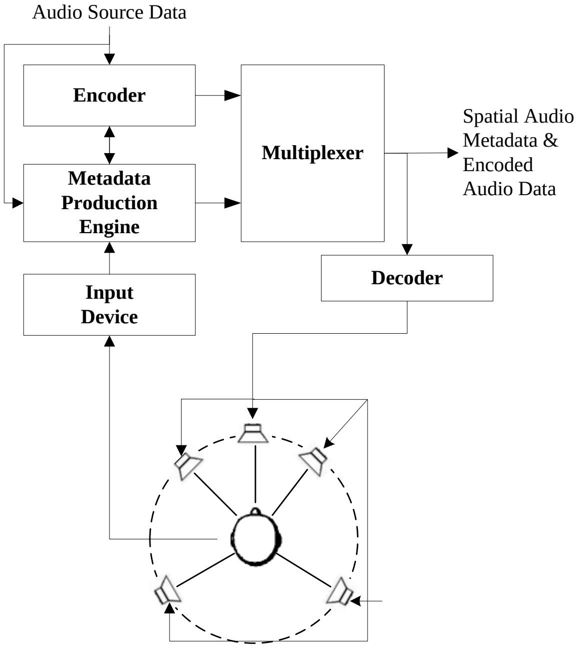

Audio Source Data
Encoder
Metadata Production Engine
Input Device
Multiplexer
Spatial Audio Metadata & Encoded Audio Data
Decoder

Figure 4. Block diagram of spatial audio generation.

Both the metadata and the encoded audio signal data are multiplexed into a combined data format by multi-channel multiplexer and transmitted to the spatial audio playback module of the user device. In order to permit monitoring during the production process of the spatial audio metadata, the monitoring decoder demultiplexes and decodes the combined audio stream and metadata to reproduce a monitoring signal at speakers. The monitoring speakers are arranged in a standardized arrangement such as ITU-R BS 775. The monitoring system allows a listener at the user device to perceive the effect of the metadata and the encoded audio.

2) *Spatial audio playback*

Figure 5 is a system level schematic diagram of the spatial audio decoding and playback aspect.

The metadata decoder receives and separates the encoded, transmitted data in a multiplexed format into metadata and audio signals data. The spatial audio decoder [17] receives the encoded audio signal data and decodes it by a method and apparatus complementary to that used to encode the data. The decoded audio is organized into the appropriate channels and output to the environment engine. The environment engine includes a diffusion engine in series with a mixing engine and operates in a multi-dimensional manner, mapping N inputs to M outputs.

The diffusion engine conditions N channel digital audio from the decoder in a manner controlled by and responsive to the metadata to add reverberation and delays, thereby producing direct and diffuse the audio data in multiple processed channels. The multiple processed channels are then mixed in a mixing engine to produce mixed digital outputs. The mixing engine mixes the N audio input channels by multiplexing and summing under control of a set of mixing control coefficients to produce a set of M output channels for playback in a user device.

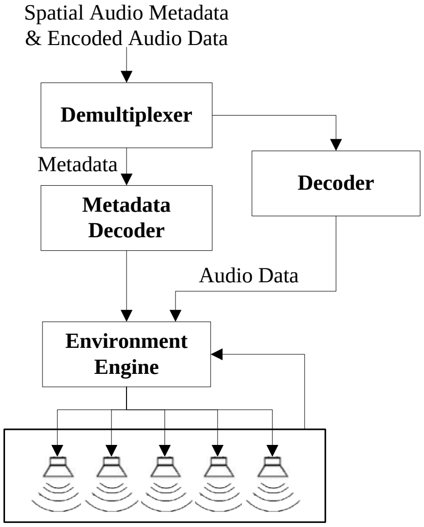


Figure 5. Block diagram of spatial audio playback.

A dedicated diffuse output from the mixing engine is differentiated for reproduction through the dedicated diffuse radiator speaker.

The multiple audio channels are then converted to analog signals, amplified by amplifiers. The amplified signals drive an array of speakers in a listening environment.

III. EXPERIMENATAL RESULTS

In this subsection, the performance of the proposed MCLT peak-pair fingerprint extraction algorithm is evaluated. Additionally, the performance of the algorithm is compared with the modified implementations of three previous methods. Method 1 is an STFT-based peak-pair fingerprint extraction method proposed by Wang [7], while Method 2 is an audio fingerprint extraction based on the masked audio spectral keypoints [18]. Method 3 is a local feature extraction from adaptively scaled patches of the time-chroma representation of the audio signal [19].

For experiments, two test database types were selected: (1) Set I consists of a database of 7,000 songs from different genres such as pop, hip-hop, jazz, and classical. (2) Set II is a database containing 4,000 TV advertisements with total time amounting to 740 hours, and each advertisement ranging from 10 to 15 minutes in length. All of the audio data are stored in PCM format with mono, 16-bit depth, and 16 kHz sampling rate converted from real audio data in consideration of portable devices such as mobile phones. Audio query clips with lengths of two, three, four, and five seconds were captured using mobile phone, which was placed 5 meter from a 2.1-channel loudspeaker connected to a TV. With the randomly created 3,000 queries, query sets are created by adding various types of noise of different levels. Five different types of noise (babble noise, moving car noise, white noise, street noise, and computer fan noise) have been artificially added to different portions of the database at signal-to-noise (SNR) ratios ranging from clean to 12 dB, and 6 dB.

Table I depicts the experimental results of the four methods when a 5-second-long query from Set I was used. MW, MC, MX, and MCLT denote Method 1, Method 2, Method 3, and the proposed method, respectively. The recognition results under the five different noisy environments are averaged for the evaluation.

TABLE I. COMPARATIVE PERFORMANCE FOUR SCHEMES WITH SET I.

SNR	Averaged Recognition Rate (%)			
	MCLT	MW [6]	MC [7]	MX [8]
clean	97.3	95.5	94.8	93.5
12 dB	96.8	93.7	89.6	78.9
6 dB	93.6	88.4	77.5	63.8
0 dB	80.7	73.6	61.7	57.6
Total	92.1	87.8	80.9	73.5

As shown in Table I, the best recognition accuracy was 97.3% for query-by-example music identification, which was obtained with the proposed MCLT. The recognition rate of MW was slightly lower than those of MCLT. MX yields the lowest identification rate, and provides worse results at SNR 0 dB.

Table II presents the results of the advertisement identification performed on a Set II database.

TABLE II. COMPARATIVE PERFORMANCE FOUR SCHEMES WITH SET II.

SNR	Averaged Recognition Rate (%)			
	MCLT	MW [6]	MC [7]	MX [8]
clean	95.5	93.6	93.5	92.6
12 dB	94.3	90.5	86.2	75.4
6 dB	91.6	85.4	74.6	60.2
0 dB	77.5	70.8	58.5	53.4
Total	89.7	85.1	78.2	70.4

As shown in Table II, the recognition accuracies for advertisement identification are not better than those of Table I for music identification, because some advertisements in Set II contain silent segments. The query was captured frequently from the silent segments and used for the matching. Also, the proposed MCLT yields better performance than MW, MC, and MX.

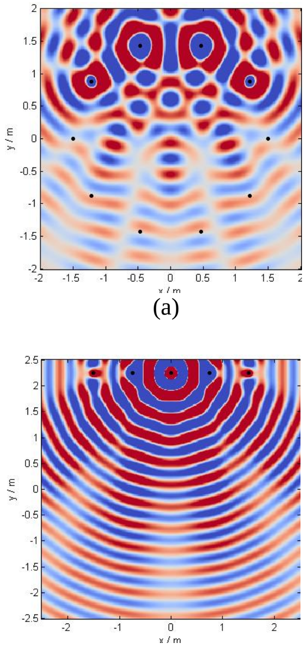
Table III shows the recognition performance of the MCLT scheme for when the query length was changed.

TABLE III. PERFORMANCE EVALUATION ACCORDING TO QUERY LENGTH.

SNR	Averaged Recognition Rate (%) By Query Length			
	2 sec	3 sec	4 sec	5 sec
clean	76.8	91.5	95.1	97.3
12 dB	71.5	90.7	94.3	96.8
6 dB	63.3	84.7	91.8	93.6
0 dB	55.6	76.5	81.7	80.7
Total	66.8	85.9	90.7	92.1

This result shows that the performance increases as the length of the query increases. Also, the proposed scheme shows satisfactory performance with 4 and 5-second-long queries, showing a recognition rate above 90%.

Figure 6 presents the simulation results of the spatial audio generation and playback.



(a)

(b)

Figure 6. Simulation results of the spatial audio generation and playback.

Figure 6 (a) shows the simulation result of the spatial audio effects, when 10 speakers are arranged in a circle. Figure 6 (b) depicts the simulation result of the spatial audio effects, when 5 speakers are arranged in a line.

We performed a Mean Opinion Score (MOS) test on music quality. The listening tests were arranged in an acoustically isolated listening room. A total of 12 listeners participated in the test. The MOS scores for music playback are in the range of “excellent” to “good”.

IV. CONCLUSIONS

In the new combination of the audio fingerprinting and spatial audio reproduction, the MCLT peak-pair based audio fingerprint improves the accuracy of the audio fingerprinting system in a real noisy environment. And spatial audio reproduction on the multi-channel loudspeaker setup improves the realism of the spatial sound experience.

ACKNOWLEDGEMENT

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)

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A2007601)

REFERENCES

[1] P. Cano, E. Batlle, T. Kalker, and J.Haitsma, "A review of algorithms for audio fingerprinting," International Workshop on Multimedia Signal Processing, December, 2002, pp. 169-173.

[2] W. Li, C. Xiao, and Y. Liu, "Low-order auditory Zernike moment: a novel approach for robust music identification in the compressed domain," EURASIP Journal on Advances in Signal Processing, August 2013.

[3] A. Sinitsyn, "Duplicate song detection using audio fingerprinting for consumer electronics devices," IEEE International Symposium on Consumer Electronics, June, 2006, pp. 1-6.

[4] J. Cerquides, "A real time audio fingerprinting system for advertisement tracking and reporting in FM Radio duplicate song detection using audio fingerprinting for consumer electronics devices," Radioelektronika, 2007. 17th International Conference, April, 2007, pp. 1-4.

[5] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," 3rd International Conference Music Information Retrieval, October, 2002, pp. 107–115.

[6] V. Chandrasekhar, M. Sharifi, and D. A. Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications," 12th International Conference Music Information Retrieval, October, 2011, pp. 801-806.

[7] A. Wang, "An industrial strength audio search algorithm," 4th International Conference Music Information Retrieval, October, 2003, pp. 7–13.

[8] C. Falch, L. Terentiev, and J. Herre, "Spatial audio object coding with enhanced audio object separation," 13th International Conference on Digital Audio Effects, September 2010.

[9] D. Cooper and J. Bauck, "Prospects for transaural recording," Journal of the Audio Engineering Society, February, 1989, pp. 3-19.

[10] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," Proceedings of the IEEE, May, 1998, pp.941-951.

[11] D. Malham and A. Myatt, "3-D sound spatialization using ambisonic techniques," Journal of the Computer Music, vol. 19, no. 4, 1998, pp. 58-70.

[12] A. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," Journal of the Acoustical Society of America, vol. 93, 1993.

[13] V. Pullki, "Virtual sound source positioning using vector base amplitude panning," Journal of the Audio Engineering Society, June, 1997, pp. 456-466.

[14] A. Mouchtaris, J. Lim, T. Holman, and C. Kyriakakis, "Head-related transfer function synthesis for immersive audio," IEEE Second Workshop on Multimedia Signal Processing, December, 1998, pp. 155-160.

[15] H. Malvar, "Fast algorithm for the modulated complex lapped transform," IEEE Signal Processing Letters, January, 2003, pp. 8-10.

[16] J.-M. Jot, S. R. Hastlings, and J. D. Johnston, "Spatial audio encoding and reproduction of diffuse sound," Patent WO2012033950, March, 2012.

[17] ISO/IEC JTC/SC29/WG11 (MPEG), Document N6455, "Call for proposal on spatial audio coding,", 2004.

[18] X. Anguera, A. Garzon, and T. Adamek, "MASK: robust local feature for audio fingerprinting," International Conference on Multimedia and Expo, July, 2012, pp. 455-460.

[19] M. Malekesmaeili and R. K. Ward, "A novel local audio fingerprinting algorithm," 14th International Workshop on Multimedia Signal Processing, September, 2012, pp. 136-140.