2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing

Stereo Goes Mobile: Spatial Enhancement for Short-distance Loudspeaker Setups

Konstantinos Drossos*, Stylianos Ioannis Mimilakis†, Andreas Floros*, Nikolaos G. Kanellopoulos*
*Dept. of Audiovisual Arts
Ionian University, Corfu, Greece
Email: kdrosos@ionio.gr, floros@ionio.gr, kane@ionio.gr
†Dept. of Sound & Musical Instruments Technology
Technological Educational Institute of Ionian Islands, Lixouri, Greece
Email: smimilak@teiion.gr

***Abstract*—Modern mobile, hand-held devices offer enhanced capabilities for video and sound reproduction. Nevertheless, major restrictions imposed by their limited size render them inconvenient for headset-free stereo sound reproduction, since the corresponding short-distant loudspeakers placement physically narrows the perceived stereo sound localization potential. In this work, we aim at evaluating a spatial enhancement technique for small-size mobile devices. This technique extracts the original panning information from an original stereo recording and spatially extends it using appropriate binaural rendering. A sequence of subjective tests performed shows that the derived spatial perceptual impression is significantly improved in all test cases considered, rendering the proposed technique an attractive approach towards headset-free mobile audio reproduction.**

***Keywords*-Stereo image enhancement, Mobile devices, Source separation, Binaural mixing**

I. Introduction

Legacy techniques for sound recording and reproduction usually employ stereophonic mixing. Hence, a large volume of stereo-encoded music content exists, despite the recent advances on multichannel audio formats and their widespread adoption by many digital audio distribution means. Based on a number of panning laws alternatives, a limited spatial audible effect can be achieved, which is acceptable for a wide range of consumer applications in both home and professional environments.

The recent advent of multiple types of hand-held devices equipped with enhanced capabilities for video and sound reproduction represent a new application field of multimedia content playback. In addition, a variety of peripheral devices exists, such as the modern smart-phones/mp3-players docking stations, serving as functional extensions to these capabilities. A common realization fact related to the sound reproduction performance of this kind of devices is the short distance between their stereo loudspeaker units imposed by their limited physical dimensions. This obviously affects the overall stereo spatial perception, rendering it nearly monophonic in nature.

In order to overcome the above drawback, a variety of stereo enhancement techniques has been already proposed in the literature, including time and frequency-domain processing of the original stereo audio signal [1]. These techniques can significantly benefit from the developments and the efficiency of blind source separation strategies for extracting the original tracks of the stereo mix [2], for example using the spatial positioning of the independent sources [3] or the frequency characteristics of the signal [4]. The extracted audio tracks may then be binaurally processed, in order to assign the separated channel recordings to virtual sound sources, with spatial properties defined through binaural rendering [2].

This work extends the above approach by employing two different blind source separation algorithms belonging to the aforementioned categories. The implementation of the overall strategy is intended for non-real time processing. Hence, the stereo audio data can be processed for example during the audio files transfer to the local music library. Within the scope of this work, we also evaluate the perceived spatial impression that is obtained when applying the proposed strategy in a typical short-distant stereo loudspeaker setup.

The rest of the paper is organized as following: an overview of the state-of-the-art for stereo image enhancement is presented in Section 2. Section 3 includes a short description of the proposed technique implementation, while the results obtained during the subjective tests performed are presented in Section 4. Finally, Section 5 concludes the work and defines some issues that can be thoroughly addressed in the future.

II. Technology overview

As stated in [1], stereo image enhancement techniques may range from recording conditions manipulation (i.e. under appropriate microphones' placement), to methods that incorporate level or frequency-based transformations of the original stereo signal. An overview of some commercial products is provided in [5]. Generally, aim of these techniques is the improvement of the spatial immersion under stereo playback (through headphones or loudspeakers), at the expense of a possible reduction in the sound quality, mainly in terms of the perceived spectral balance and distortion.

Towards this aim, various approaches exists, such as the addition of reverberation that corresponds to a listening space [6]. A recent work [7] proposed a real-time

978-0-7695-4712-1/12 $26.00 © 2012 IEEE
DOI 10.1109/IIH-MSP.2012.111

enhancement algorithm that maps stereo audio to virtual sound sources distributed around the listener, using binaural technology. The same approach was followed for realizing the "Music Widening" algorithm presented in [8], targeted to sound externalization using simple mirror-image virtual sources within an enclosure.

Lately, [2] proposed a combination of blind source separation and binaural rendering for creating the enhanced stereo image. Some blind source separation algorithms make use of sources' spatial positioning [3], but experience the appearance of artifacts when neighboring sources exist in the original audio signal, while others employ harmonic and percussive sources' frequency characteristics [9] but exhibit reduced performance in western polyphonic music [10].

III. Implementation details

Briefly, in this work, blind source separation is achieved by retrieving the original stereo panning information using a time-frequency domain metric. Subsequently, the original stereo input is separated into three layers, which correspond to sound sources located at the left, center and right part of the stereo field. This information is processed by a Harmonic - Percussive Separation algorithm, which results in harmonic and percussive components for each layer. The harmonic - percussive component separation results into six sources, two for each of the three above layers. At the final stage, the six separated sound source signals are spatially processed, producing the binaural mixing outcome.

Panning information retrieval is performed using the methodology proposed in [3]. Briefly, a frequency domain metric is employed, obtained by the comparison of each stereo channel time-frequency representation as:

$$\psi(m,k) = 2\frac{|S_l(m,k)S_r^*(m,k)|}{|S_l(m,k)|^2 + |S_r(m,k)|^2} \tag{1}$$

where k and m indicate the frequency and time index of consecutive STFT blocks of the original stereo data, $*$ denotes complex conjugation and $S_l(m,k)$ and $S_r(m,k)$ are their time-frequency representations.

Eq. 1 results into values proportional to the applied panning coefficients a, provided that a sinusoidal energy-preserving panning law is applied [2], [3]. The original panning coefficients can be then derived using the equation:

$$\psi(m,k) = 2a\sqrt{1-a^2} \tag{2}$$

For resolving the ambiguity caused by Eq. 2, expressions with partial similarity are defined as:

$$\begin{aligned} \psi_l(m,k) &= \frac{|S_l(m,k)S_r^*(m,k)|}{|S_l(m,k)|^2} \\ \psi_r(m,k) &= \frac{|S_r(m,k)S_l^*(m,k)|}{|S_r(m,k)|^2} \end{aligned} \tag{3}$$

The panning information is finally obtained from:

$$\Psi(m,k) = [1-\psi(m,k)]\hat{\Delta}(m,k) \tag{4}$$

where $\hat{\Delta}(m,k)$ is defined as $\psi_l(m,k) - \psi_r(m,k)$ and equals to a) 0, if $\Delta(m,k) = 0$, b) 1, if $\Delta(m,k) > 0$, and c) -1, if $\Delta(m,k) < 0$.

Following the definition provided in [2], the above panning indices values can be further transformed to angle values as:

$$\theta(i) = 45^o(panning\_index(i) + 1) \tag{5}$$

The original sound source signals can be separated using the aforementioned time-frequency zones (m,k), where the time-frequency panning index presentation $\Psi(m,k)$ has values equal to $panning\_index(i)$. These zones can be used for synthesizing the separated source signals through inverse STFT operations and windowing functions, defined as:

$$W_i(\Psi(m,k)) = \nu(1-\nu)e^{\Xi} \tag{6}$$

where ξ is the window width that represents a parameter of accuracy in source separation and distortion accused by neighboring sources and Ξ equals to:

$$\Xi = \frac{\Psi(m,k) - panning\_index(i)}{2\xi} \tag{7}$$

The separated sound source signals are then obtained from inverse STFT operations applied on the signal:

$$S_i = W_i[\Psi(m,k)](S_l(m,k)S_r(m,k)) \tag{8}$$

The harmonic-percussive separation takes place by appropriately modeling the above separated sound source signals. More specifically, the sum of the harmonic (H) and the sum of the percussive (P) components can be factorized into:

$$\begin{aligned} \Sigma_H(m,k) &= \imath_H(m,k)R_H(m,k) \\ \Sigma_P(m,k) &= \imath_P(m,k)R_P(m,k) \end{aligned} \tag{9}$$

where $\imath_H(m,k)$ and $\imath_P(m,k)$ are scalar time varying spectral variances, which encode the spectral temporal power of harmonic and percussive instruments respectively. $R_H(m,k)$ and $R_P(m,k)$ are $Ix1$ full-rank, time-varying spatial covariance matrices, since we focus on non-multichannel separated audio signals.

As mentioned in [9], the spectrum of the percussive instruments is usually smooth over the frequency axis, while the spectrum of the harmonic ones is smooth over the time axis in the time-frequency domain. Thus, following the aforementioned characteristics of percussive and harmonic instruments and using $\imath_H(m,k)$, for $m > 1$, and $\imath_R(m,k)$, for $k > 1$, the following Markov chain priors can be introduced:

$$\begin{aligned} \rho(\imath_H(m,k)) &= \Im\Gamma(\imath_H|\alpha_H, (\alpha_H - 1)\imath_H(m-1,k)) \\ \rho(\imath_P(m,k)) &= \Im\Gamma(\imath_P|\alpha_P, (\alpha_P - 1)\imath_P(m,k-1)) \end{aligned} \tag{10}$$

where $\Im\Gamma(\imath|\alpha,\beta)$ denotes the Inverse-Gamma Density, with shape parameter $\alpha > 0$, $\beta > 0$ whose mean is $\frac{\beta}{\alpha-1}$, and

$$\Im\Gamma(\imath|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\imath^{-\alpha-1}e^{\frac{-\beta}{\imath}} \tag{11}$$

From the spectral continuity information obtained from $\Im\Gamma$ computation, a spectral parameter θ_{SPC} can be derived as $\theta_{SPC} = (\imath_H(m,k), \imath_P(m,k))_{m,k}$, using the complete data set of STFT coefficients of the harmonic and percussive components in all frequency bins denoted as $c = (H(m,k), P(m,k))_{m,k}$ [10]. Regarding the expectation step of the Expectation Maximization (EM) algorithm, defined in [10], the expected covariance matrices $\widehat{\Sigma}_H(m,k)$ and $\widehat{\Sigma}_P(m,k)$ are updated using the Wiener filters defined in Eqs. 13 and 13 by the process described in Eqs. 14 and 15 respectively.

$$W_H(m,k) = \Sigma_H(m,k)\Sigma_{x^{-1}}(m,k) \tag{12}$$

$$W_P(m,k) = \Sigma_P(m,k)\Sigma_{x^{-1}}(m,k) \tag{13}$$

$$\begin{aligned}\widehat{\Sigma}_H(m,k) &= W_H(m,k)\widehat{\Sigma}_x(m,k)W_H(m,k) \\ &+(1-W_H(m,k))(\Sigma_H(m,k))\end{aligned} \tag{14}$$

$$\begin{aligned}\widehat{\Sigma}_P(m,k) &= W_P(m,k)\widehat{\Sigma}_x(m,k)W_P(m,k) \\ &+(1-W_P(m,k))(\Sigma_P(m,k))\end{aligned} \tag{15}$$

where $\widehat{\Sigma}(m,k)$ is analytically defined in [12].

The following maximization step of EM algorithm is implemented by the usage of the auxiliary function Q defined as:

$$Q^{MAP}(\theta|\theta^{old}) = \log(c|\theta) + \gamma \log \rho(\theta_{SPC}) \tag{16}$$

where $\rho(\theta_{SPC}) = \prod_{m,k} \rho(\imath_H(m,k))\rho(\imath_P(m,k))$ and γ is a parameter determining the contribution of the spectral prior [10].

A similar computation can be approached by setting the gradient of the above auxiliary function Q to zero. Since we consider monophonic audio data obtained from the separation of stereophonic ones, the above computation will result in a second order polynomial form of the source variances with a single positive solution:

$$\imath_H(m,k) = \frac{(-\beta + \sqrt{\beta^2 - 4\alpha_H c_H})}{2\alpha_H} \tag{17}$$

$$\imath_P(m,k) = \frac{(-\beta + \sqrt{\beta^2 - 4\alpha_P c_P})}{2\alpha_P} \tag{18}$$

where:

$$\begin{aligned}\alpha_H &= \tfrac{\gamma(\alpha_H-1)}{\imath_H(m+1,k)}, \\ \alpha_P &= \tfrac{\gamma(\alpha_P-1)}{\imath_P(m,k+1)}, \\ \beta &= \gamma + 1, \\ c_H &= -tr(\widehat{\Sigma}_H(m,k)) - \gamma(\alpha_H - 1)\imath_H(m-1,k), \text{ and} \\ c_P &= -tr(\widehat{\Sigma}_P(m,k)) - \gamma(\alpha_P - 1)\imath_P(m,k-1)\end{aligned}$$

with $\imath_H(m,k)$ equals to zero for all k when $m = 1$ and $\imath_P(m,k)$ equals to zero when $k = 1$.

Finally, the separated components are obtained by the application of Wiener filtering on the original signal S, that is:

$$\begin{aligned}\mathcal{H}_{(m,k)} &= W(m,k)S(m,k) \\ \mathcal{P}_{(m,k)} &= W(m,k)S(m,k)\end{aligned} \tag{19}$$

Table I
THE SELECTED MUSICAL PIECES FOR THE SUBJECTIVE EVALUATION

Title	Group/Artist	Style
Rasputin	Boney M	Disco
Stand By Him	Ghost	Doom Metal, Heavy Metal
Reelin' & Rockin'	The Head Cat	Rock 'N' Roll, Rockabilly
Pearls For Swine	The Kilimanjaro Darkjazz Ensemble	Future Jazz

Table II
THE DIFFERENT RANGES FOR PANNING INDICES

Version Name	Panning Indices Values Range		
	Left	Center	Right
0.1	[-1, -0.11]	[-0.1, +0.1]	[+0.11, +1]
0.3	[-1, -0.31]	[-0.3, +0.3]	[+0.31, +1]
0.6	[-1, -0.61]	[-0.6, +0.6]	[+0.61, +1]

and the application of inverse STFT on $\mathcal{H}_{(m,k)}$ and $\mathcal{P}_{(m,k)}$.

The resulting separated audio components were used as sources for the binaural mixing process. A Graphical User Interface (GUI) was implemented to handle the binaural processing of the separated source signals using the location positions acquired from eq. 5, in terms of degrees.

IV. RESULTS

For the subjective evaluation of the proposed system, four musical pieces were selected from an equal number of different musical styles (see Table I), derived from the on-line service Discogs [11]. For each musical piece, three versions were created, based on different values of the panning index discrimination among left, center and right. These values are listed in Table II. The different ranges for the panning indices were applied before the harmonic-percussive separation and the source extraction. When the sources were finally extracted, they were upmixed using binaural processing. A total of 25 participants listened to

Table III
RATED AS THE WIDEST STEREO IMAGE SCORE FOR EACH VERSION

Musical Piece	Versions' Scores			
	Stereo	0.1	0.3	0.6
Rasputin	12%	28%	24%	36%
Stand By Him	24%	24%	24%	32%
Reelin' 'N' Rockin'	24%	12%	12%	52%
Pearls For Swine	44%	16%	12%	28%

the four groups of musical pieces. Each group consisted of the original stereo version of the corresponding music piece, plus the three modified versions presented in Table II. The reproduction of each musical piece was made in random order, using a stereo docking station. The dimensions of the docking station were $17.5 \times 3.5 \times 8.8\,cm$ and the distance between the loudspeaker units was equal to $10\,cm$. Each participant was asked to rate which of the reproduced

versions for each musical piece exposed the wider stereo image, with the widest stereo image rated as 1 and the narrowest with 4. Table III presents a summary of the obtained results in terms of the measured percentage of rates equal to 1. Figure 1 also shows the complete set of results for all considered rates. Clearly, the stereo versions' scores are mainly concentrated at subjects' rates equal to 1 and 4. This result indicates that there was a perceived difference, regarding the width of the stereo image, between the processed and the stereo versions, i.e. the perceived stereo versions' image width ratings are at the extremes of the narrower - wider scale. This may explain the fact that, in 80% of the rating cases, stereo versions were not rated as having an intermediate score between the wider and the narrower. In addition, as can be seen from Fig. 1 and inferred from Table III, most of the participants rated the stereo versions' having the narrower stero image. Also, from Fig. 1 it is clear that increasing the values of panning indices values for center (the values of the Center column at Table II), the width of the sound image is perceived as wider. In general, the enhanced versions have been mainly considered as having the widest stereo image from the original stereo version, since, from Table III it can be concluded that the processed versions' score that corresponds to the perceptually widest sound image reaches up to 74%.

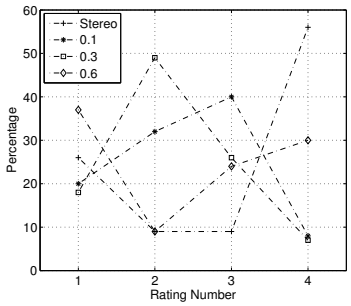

Figure 1. Rating percentage for each version

V. Conclusions

In this work, a stereo image enhancement technique was employed and evaluated, targeted to mobile / hand-held devices and environments that, due to their small size, limit the physical distance values between the stereo loudspeaker units used for audio/music playback. The above enhancement technique incorporates blind source separation and binaural mixing for deriving the spatially-enhanced stereo signal. A number of subjective tests has shown that the enhanced signal versions provide a wider stereo image perception in a total of 74% of all the test cases considered, rendering the proposed enhancement strategy an attractive alternative for headset-free music reproduction in mobile environments. The proposed technique is suitable for off-line applications, i.e. it can be applied on the desired music content during its transfer on the mobile device. Further research issues that may be considered can focus on blind source separation signal processing optimization, in order to allow for it's real-time implementation.

Acknowledgments

The authors wish to thank Kyzalas M. and Koukoudis K., of Audiovisual Arts Department of Ionian University, for their valuable contribution in organizing the subjective evaluation procedure. The research activities that led to these results, were co-financed by Hellenic Funds and by the European Regional Development Fund (ERDF) under the Hellenic National Strategic Reference Framework (ESPA) 2007-2013, according to Contract no. MIKRO2-40/E-II-A

References

[1] R. C. Maher, E. Lindemann, and J. Barish, "Old and new techniques for artificial stereophonic image enhancement," in Proc. 101st AES Convention, Los Angeles, California, U.S.A., 1996.

[2] A. Floros and N. A. Tatlas, "Spatial Enhancement For Immersive Stereo Audio Applications," in Proc. 17th International Conference on Digital Signal Processing, Corfu, Greece, 2011.

[3] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, U.S.A., 2003.

[4] D. Fitzgerald and M. Gainza, "Single Channel Vocal Separation using Median Filtering and Factorisation Techniques," in ISAST Transactions on Electronic and Signal Processing, vol. 4, pp. 62-73, 2010.

[5] S. Olive, "Evaluation of Five Commercial Stereo Enhancement 3D Audio Software Plug-ins," in Proc. 110th Convention of Audio Engineering Society, Amsterdam, Netherlands, 2001.

[6] G. S. Kendall, W. L. Martens, and M. D. Wilde, "A Spatial Sound Processor for Loudspeaker and Headphone Reproduction," in Proc. 8th International AES Conference: The Sound of Audio, Washington, D.C., U.S.A., 1990.

[7] C. Tsakostas, A. Floros, and Y. Deliyiannis, "Binaural Rendering for Enhanced 3D Audio Perception," in Proc. Audio Mostly! 2nd Conference on Interaction with Sound, Ilmenau, Germany, 2007.

[8] P. Minnaar, "Enhancing music with virtual sound sources," in Hearing Journal, vol. 63, pp. 38-40,42-43, 2010.

[9] N. Ono, et al., "A Real-Time Equalizer Of Harmonic And Percussive Components In Music Signals,"in Proc. ISMIR 2008, Philadelphia, Pennsylvania USA, 2008.

[10] N. Q. K. Duong, et al., "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in Proc. IEEE ICASSP, Prague, Czech Republic, 2011.

[11] Discogs. (2012, Mar. 07). Explore Releases on Discogs. [Online]. Available: http://www.discogs.com/

[12] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation,"in Proc. 9th international conference on Latent variable analysis and signal separation, St. Malo, France, 2010.