Extracting and Modifying the Spatial Information in Stereo Audio

Hui-yu Tseng and Chia-Ming Chang

Dept. of Computer Science and Engineering
Tatung University, Taiwan
lookcarl223@yahoo.com.tw, cmchang@ttu.edu.tw

ABSTRACT

In this paper, a method to extract the spatial information and signals in stereo, and then synthesis the new sound field is proposed. The objective is to synthesize appropriate sound field corresponding to different listening conditions. The discussed situation is limited to the recording of chorus, multi-sources playing the same melody by the same music instruments and being aligned in line. By assuming that the frequency magnitudes of spectrogram of each source are similar and the procedure of signals received by microphone is similar to image corrupt of liner distortion, we apply the concept of image restoration to extract the spatial information and single source. The simulation is performed to confirm the method and the results demonstrate that a sound field similar to the original sound field can be synthesized using the extracted single source and spatial information. Also the spatial information can be modified to synthesize a different sound field.

1: INTRODUCTION

The objective of this research is to improve the sound field reproducing in reality while the listening environment inconform with original recording conditions, *eg*, loudspeaker numbers, layout of loudspeaker,etc. To synthesize a new sound field with flexible listening suit is further application. Obviously, the information of original un-mixing sources and spatial relation is desired. We derive the synthesis solution by decomposing the received signals to get spatial information and single representing signal.

There are some related techniques of similar concepts such as source separation [1, 2], source location [3], direction of arrival (DOA), time delay of arrival (TDOA) [4, 5], MUltiple Signal Classification (MUSIC) [6], delay and sum [7], independent component analysis (ICA), up-mix system [8, 9]. These techniques generally discussed the standard conditions. The number of microphones must be greater than multiple-sources and the sources are uncorrelated each other. In multi-source situation, the condition of independent sources or large microphone number such as microphone array is indispensable. In stereo system, the number of sources is restricted with two independent sources. Most techniques of source separation are based on the given priors information such as signal model, source number, azimuth, etc.

The application of theses techniques in sound are focused on recognizing the source directions, identifying different source objects, distinguishing or separating the source objects, solving cocktail party problem, segregation of different instruments or speakers, source classification with sound database construction, etc, such as teleconferencing, interactive information systems, large conference rooms, noise control, live entertainment, etc [10], and few of them are applied to synthesis. The up-mix techniques are addressing to enhance synthesizing for stereo over multi-speaker reproduction via ambiance extraction or human perception [8, 9]. It extracts the ambiance information and the primary signal in the center to synthesize, without panning other side signal. The up-mix techniques does not desire the original sources, and does not impress on synthesizing with different spatial information.

Therefore, we purposed the method to enhancing synthesis by decomposing the received signals to spatial information and original sources. The situation we discussed in this paper is the chorus in stereo recording. That is, a group of players perform the same melody with same instrument in unison, *e.g.* a violin group of symphony orchestra.

In the section 3.1:, we take reasons of assuming that the sources have the approximate spectrogram even they are different in waveforms. In our supposition, the procedure of microphone recording is similar to the corrupt image with a motion blur functions which summing up the shifted images of continuous displacements during the time of shutter opening. In section 3.2:, the autocorrelation is evaluated with the spectrogram of the received signals to retrieve relative delays of microphone pairs. Those delays are due to the spatial information in original sound field. Then the source locations can be found by using their geometric relation and used to construct the mixing function. The single source could be obtained by applying Wiener Filter is discussed in section 3.3:. Therefore, the concept of image distoration/restoration may be applied to decompose the received signals as spatial

function and single source.

After the process of analysis and extraction, a single sound source and the spatial information can be obtained. Then the sound field can be reproduced by synthesizing the extracted single source and spatial information. Not only the original sound field but also the different sound field of modified spatial information can be synthesized in pleasure. For instance, the original source directions, distances, arrangement, and etc, can be changed.

2: BACKGROUND

2.1: Image Degradation

The procedure of an image blurred by uniform linear motion between the object and sensor during image acquisition could represented as Eq. (1) [11]

$$g(x,y) = \int_0^T f(x - x_0(t), y)dt, \tag{1}$$

where $g(x,y)$ is the blurred image, $f(x,y)$ is an object moving along the $x-$axis and $x_0(t)$ is the time varying component of motion, T is the duration of exposure. The final exposure at any point of the recording medium is obtained by integrating the instantaneous exposure over the time interval during shutter of the imaging system shutter is open.

Applying Fourier transform to Eq. (1), the following equation is obtained,

$$G(u,v) = F(u,v) \int_0^T e^{-j2\pi[ux_0(t)]}dt, \tag{2}$$

where $F(u,v)$ and $G(u,v)$ are the Fourier transform of $g(x,y)$ and $f(x,y)$, respectively. By defining

$$H(u,v) = \int_0^T e^{-j2\pi[u\ x_0(t)]}dt, \tag{3}$$

Eq. (2) can be expressed in the form

$$G(u,v) = H(u,v)F(u,v), \tag{4}$$

where $H(u,v)$ can be named as degradation function of linear motion.

2.2: Image Restoration

As the degradation function H is known, the simplest approach to restore is the direct inverse filtering. That is:

$$F(u,v) = \frac{G(u,v)}{H(u,v)}, \tag{5}$$

where $F(u,v)$ is the Fourier transform of the estimated images, and $G(u,v)$ is the degraded image. But in practice, the noise should be considered. Therefore, the Eq. (5) can be rewritten as

$$\hat{F}(u,v) = F(u,v) + \frac{N(u,v)}{H(u,v)}, \tag{6}$$

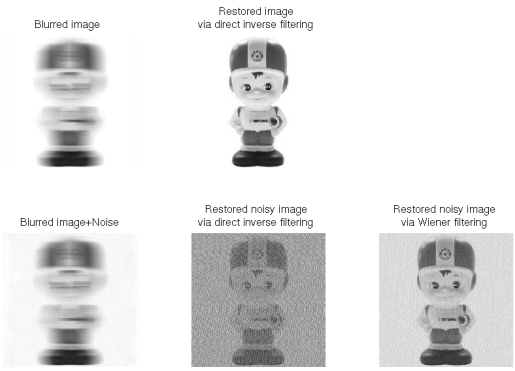

Figure 1: Image restoration

where $N(u,v)$ is the Fourier transform of the additive noise. Eq. (6) tells that even if the degradation function is known, we cannot recover the undegraded image exactly because $N(u,v)$ is a random function we never know. Furthermore, if the degradation function is zero or very small, the ratio $\frac{N(u,v)}{H(u,v)}$ would easily dominate the estimate.

Wiener Filter incorporates both the degradation function and statistical characteristics of noise into the restoration process, it is also referred to as the "minimum mean square error filter". Based on the minimum of the error function, the estimate image in frequency domain via Wiener Filter can be expressed as

$$\hat{F}(u,v) = \left[\frac{H^*(u,v)}{|H(u,v)|^2 + K}\right] G(u,v), \tag{7}$$

where

$$\begin{aligned}
H(u,v) &: \text{degrataion function} \\
H^*(u,v) &: \text{complex conjuagate of } H(u,v) \\
|H(u,v)| &= H^*(u,v)H(u,v) \\
K &= \frac{|N(u,v)|^2}{|F(u,v)|^2} \\
|N(u,v)|^2 &: \text{power spectrum of the noise} \\
|F(u,v)|^2 &: \text{power spectrum of the undergraded image.}
\end{aligned}$$

However, the power spectrum of undergraded image is usually unknown. Then the approach is specifying the constant K to approximate the estimate function as shown in Fig. 1.

3: SYSTEM DESCRIPTION

3.1: Model Description

A process to receive signals by microphones can be represented as the following function:

$$x_j(t) = \sum_{i=1}^{M} \left[h_{ij}(t) * s_i(t) + n_{ij}(t)\right], \tag{8}$$

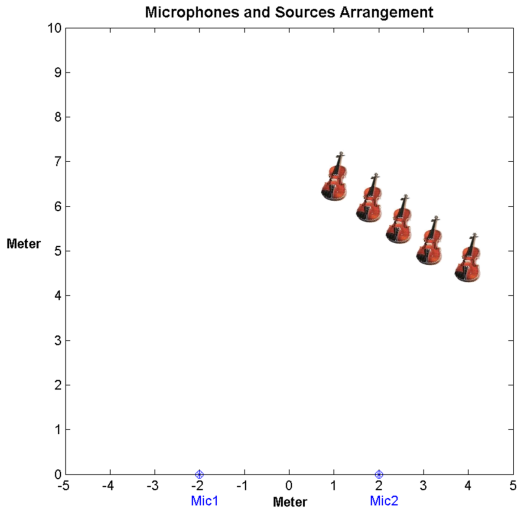

Figure 2: An example of sources arrangements with uniform linear permutation.

where $x_j(t)$ is the signal received by jth microphone and $s_i(t)$ is the signal come from ith source. The $h_{ij}(t)$ is the impulse response of channel from ith source to jth microphone. The $n(t)$ is the additive noise. In a non-reversation condition, the impulse response $h_{ij}(t)$ could be expressed as

$$h_{ij}(t) = \begin{cases} a(k_{ij}), & t = k_{ij} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

and

$$a(k_{ij}) = \alpha * d_{ij}^2, \tag{10}$$

where k_{ij} is the propagation time from ith source to jth microphone, d_{ij} is the distance from ith source to jth microphone, and a is the standard attenuation coefficient.

In this paper, the discussion is focused on the situation, that is, a subset of symphony orchestra with same music instrument. It is described more exactly that different performers in uniform linear alignment are playing the same melody of the same musical score simultaneously with the same musical instrument in a non-reverberation room as illustrated in Fig. 2.

The musical notation can be shown with a two-dimensional display to indicate frequency content that changes with time. In other words, a musical score employs a notation that corresponds to the "time-frequency" image found in the spectrogram [12]. In Fig. 3, the notations [C D E F G A B C] could be represented as the spectrogram with [262 294 330 349 440 494 523] Hz.

For our supposed conditions, each sound source is the melody of the same music score, namely it has the same musical scale at series of each short time interval and its sounds would be the same even it is played by different performers. So we can suppose that each sound source has the similar magnitude of

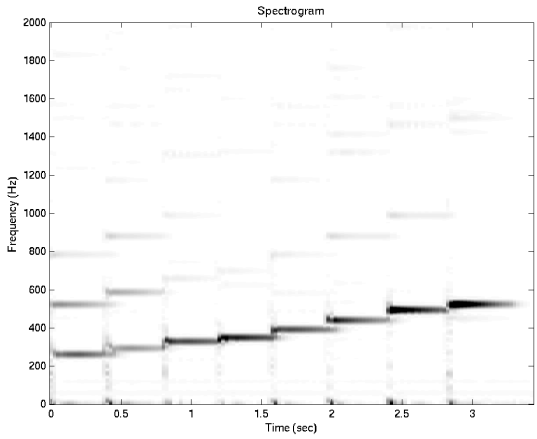

Figure 3: Spectrogram of a real sound of piano with homonymic for the C-major scale

spectrogram. The difference among sources are the phase, but human's hearing perception is insensitive to the phase of audio.

3.2: Sound Source Distortion

The degradation of image by motion blur functions is a procedure for summing up the shifted images of continuous displacements during the time when shutter is opening. And the microphone recording can be treated as a procedure for summing the sources with different time delay and energy attenuation according to the distance between microphone and sources. In our assumption, the procedure of microphone recording is similar to the image degradation while regarding the spectrogram of each source as one uncorrupted image and the signals received by microphone is a corrupted image by time delay. Therefore, the image-restoration techniques could be applied on the sound source separation.

The Eq. (8) is equal to the short-time Fourier transform in time-frequency domain :

$$X_j(f, t_m) = \sum_{i=1}^{M} h_{ij}(t_m) * S_i(f, t_m) \tag{11}$$

where f is frequency component, t_m is time frame index with length L, $X_j(f, t_m)$ is the signal received by jth microphone and $s_i(f, t_m)$ is the signal come from ith source in time-frequency domain. The $h_{ij}(t_m)$ are the impulse responses of channels from ith source to jth microphone.

For the assumption, we define the source $S(f, t_m)$ as a single source which may represent the timefrequency component property of each source $S_i(f, t_m)$. The Eq. (11) can be rewritten as

$$X_j(f, t_m) = h_{Mj}(t_m) * S(f, t_m), \tag{12}$$

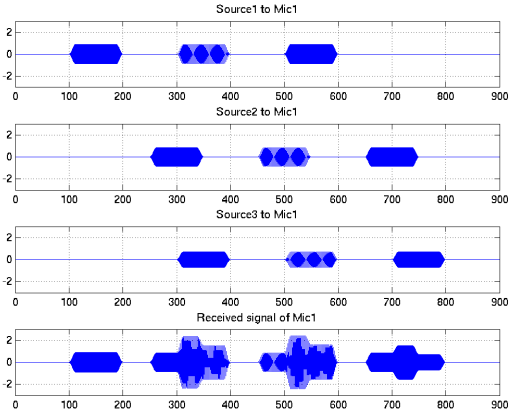

Figure 4: Waveform of received signals

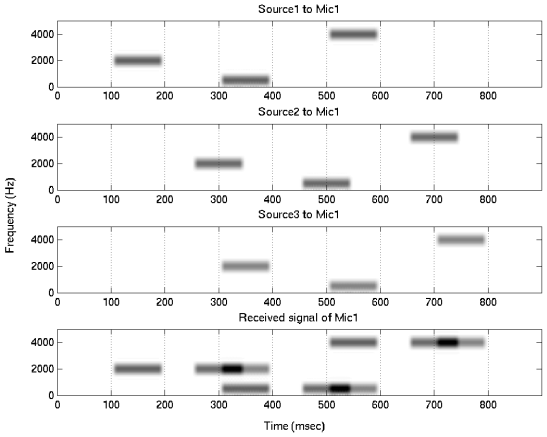

Figure 5: Spectrogram of received signals

By defining $h_{Mj}(t_m)$ to represent the summation of impulse responses form all sources to jth microphone,

$$h_{Mj}(t_m) = \sum_{i=1}^{M} h_{ij}(t_m) \tag{13}$$

where $h_{Mj}(t_m)$ is an impulse train with attenuation weight on the index of delay time for all sources to jth microphone.

As the spectrogram is treated as an image, the axes of time and frequency are noted as the $x-$ and $y-$axis in spatial domain of image, the impulse function $h_{Mj}(t_m)$ could be considered as a linear degradation function along the x-direction. Therefore, the image degradation process is similar to the procedure of microphone recording.

An example to illustrate the procedure in following diagram. Three similar sources are received by a microphone denoted as Mic1 with time delays and attenuation. The first three plots in Fig. 4 are the signals from three sources to Mic1 and shown in time domain. The received signals of Mic1 is shown in the last plot.

The Fig. 5 shows the spectrogram in time-frequency domain respective to Fig. 4. We can clearly observe that spectrogram of signal received by microphone is the sum of each source signal's spectrogram.

The corresponding impulse response from each source to microphone are shown in the first three plots in Fig. 6, and the last plot is an impulse train of sum of all source signal to microphone, the summation of each impulse response as discussed in Eq. (13). And the the received spectrogram of microphone shown in the last plot of Fig. 5 can be considered as the degraded image of the single source spectrogram.

3.3: Extraction of Spatial Information and Single Source

Based on the correlation theory, the arrival time frames of each source to jth microphone could be found

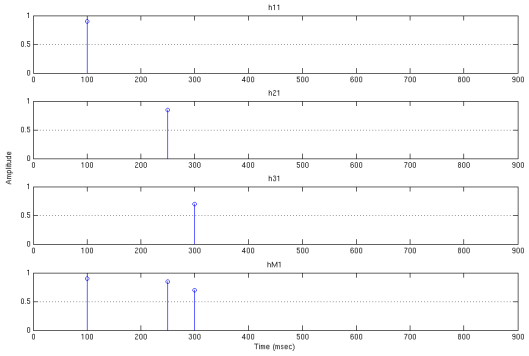

Figure 6: Impulse response at Mic1

by applying auto-correlation to received data $X_j(f, t_m)$ shown in Fig. 7. The locations of sources could be found by time delays of microphones based on their geometric relation. Then the degradation function $h_{Mj}(t_m)$ would be constructed by evaluated spatial parameters such as the number of sources, distances, etc. Therefore, we could retrieve the single representing source by applying the restoration function.

Actuality, there are some differences among the source images. The different part among the spectrogram of source may be treated as noise,

$$N(f, t_m) = \sum_{i=1}^{M} \big||S(f, t_m)| - |S_i(f, t_m)|\big|, \tag{14}$$

where the $S(f, t_m)$ and $S_i(f, t_m)$ are the spectrograms of sigle representing signal and sigle signal soure, respectively.

The Wiener Filter described in Eq. 7 is used to estimate the spectrogram of a single sound source. The corresponding parameters to the sound source degradation are explained

$\hat{F}(u, v)$: Undegraded image as the spectrogram of original single source $S(f, t_m)$ in

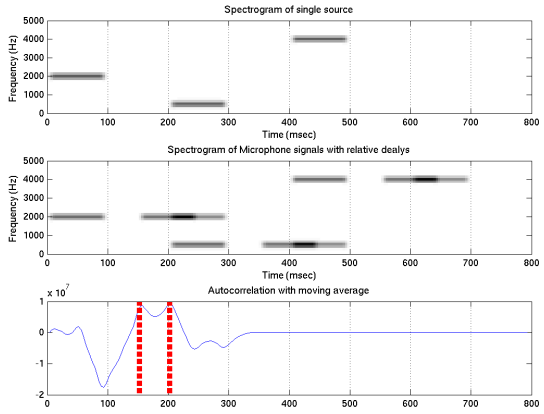

Figure 7: Auto-correlation of time delays finding

frequency domain.

$\hat{G}(u,v)$: Degraded image as the received spectrogram of microphone $X_j(f,t_m)$ in frequency domain.

$H(u,v)$: Degration function as the impulse $h_{Mj}(t_m)$ of microphone in frequency domain.

Because the the noise and original sources are unknown, we estimate the original single source by adjusting the constant K to let the evaluated error E within the error tolerance after some iterations. And the error E,

$$\begin{aligned} E &= |G(u,v) - \hat{F}(u,v)H(u,v)| \\ &= \left||X_j(f,t_m)| - |\hat{S}(f,t_m) * h_{Mj}(t_m)|\right| \quad (15) \end{aligned}$$

was evaluated by the difference with the degraded image and the re-degraded image of estimated undegraded image, the received spectrogram of microphone and remixed spectrogram of estimated single source respectively. We choose the restorted image $\hat{F}(u,v)$ with the least residue of two microphones as the extracted single source.

3.4: Sound Synthesis

After extracting the single time-frequency image and spatial information. In order to synthesize sound field with chorus effect in listening, we need to generate the similar sources from extracted standard time-frequency image $S(f,t_m)$.

In our supposition, the most difference among sources are wave phases. And the standard single time-frequency image $S(f,t_m)$ has the relation with each source $S_i(f,t_m)$.

$$\begin{aligned} S(f,t_m) &= |S_i(f,t_m)| + n_i \quad (16) \\ \Rightarrow S_i(f,t_m) &= S(f,t_m)\exp(\omega_S + \omega_i, t_m) \quad (17) \\ &\quad + n_i \end{aligned}$$

We shift the $S(f,t_m)$ phase ω_S with ω_i to generate each source $S_i(f,t_m)$, n_i is the slight difference of single image and source image. Then we can resynthesize the sound field by extracted single source $S(f,t_m)$ with spatial function h_{Mj}. And the different spatial parameters can be given to h_{Mj} to generate different sound fields.

4: SIMULATION

4.1: Simulation Structure

In order to hold the variable factors and verify the parameters. We use the simulation to implement the system. The recording environment is same to Fig. 2. Two microphones are set at the bottom side in a non-reverberation room. The sound sources aligned in line with same distance are placed in the room. The distance between two microphone is 4 meters.

We assume the wave propagation is a point-to-point propagation while considering a microphone is a pointer receiver and a source is a point source. The propagation velocity is constant under fixed temperature, and simulates in non-reverberation room and without interference. Therefore, the signal attenuation coefficient from one source to one microphone is a constant coefficient proportional to the square of its distance. To generate the similar sources of one solo source, we use the phase shifting and random samples shifting to generate the effect of difference and asynchronous among sources. Each source has a random phase shift in all time duration, and the each time duration has a little phase shift in one source.

4.2: Listening Test

Two pieces of music in violin playing was put to verify the system. The length of music are about 20 seconds along. The common sampling rate $44.1kHz$ and 16 bits rate are used. We take the listening test of 11 peoples to grade the results, the six score $0 \sim 5$ to grade the six questions of synthesized sound field and original sound field in Table. 1. The scores and their variance show in Table. 2, the higher score represent the higher efficiency of our system, the variance represent the opinioned variation of each tester.

Among the grading table, although the score of Q2 with Violin2 is medium, lower of other score, scores of Q3 and Q4 which are comparisons of direction conformity between the music and known sources layout are high, this shows that the hearing perceptions of direction and range between original and synthesized sound field are nearly if those two music are not hearing to comparing at the same time.

The results shows that the hearing perception of that similarity, range, and direction for the synthesized sound field of extracted single sources and original spatial information conform with the original recorded

Q1	The music similarity between original sound field and synthesized sound field
Q2	The direction and range conformity between original sound field and synthesized sound field
Q3	The direction and range conformity of original sound field
Q4	The direction and range conformity of synthesized sound field
Q5	The direction and range conformity of different re-synthesized sound field with modified spatial information
Q6	The different direction experience of sound field changing

Table 1: Question of listening test

		Q1	Q2	Q3	Q4	Q5	Q6
Violin1	Mean	4.19	2.64	4.0	3.63	4.27	4.27
	Variance	1.16	2.05	1.00	1.85	0.62	0.82
Violin2	Mean	4.82	4.64	3.91	3.64	4.27	4.27
	Variance	0.16	0.25	1.30	1.45	0.42	0.42
Average	Mean	4.50	3.64	3.95	3.5	4.28	4.28

Table 2: The score of listening test

sound field. And that the changing of direction and conformity for new sound field of modified spatial information can be obviously felt.

5: Conclusions and Future Work

The synthesis solution was proposed by decomposing received signals as spatial function and single sources. The simulation result shows our method with assumption is feasible. Spatial information could be evaluated under ideal situation. The single source of representing property among each source can be extracted by image process with evaluated spatial information while regarding the similar sources as an approximate spectrogram in time-frequency domain. And a new sound field can be synthesize by different spatial parameters for different playback conditions.

Estimating time delays by auto-correlation with spectrogram is difficult in practice. To find better methods to estimate time delays is an important feature work. There will be certain compression result to apply the concept of decomposing and re-synthesizing in paper with multi-channel processing in the future.

References

[1] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 10, no. 6, Sep. 2002.

[2] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, Jan. 2006.

[3] S. N. T. Nishiura, T. Yamada and K. Shikano, "Localization of multiple sound sources based on csp analysis with a microphone array," *Proceedings. 2000 IEEE Interional Conference on Acoustic, Speech, and Signal Processing*, vol. 2, no. 6, pp. II1053–II1056, 2000.

[4] P.-Y. L. Jen-Tzung Chien, Jain-Ray Lai, "Microphone array signal processing for far-talking speech recognition," *Wireless Communications, 2001. (SPAWC '01). 2001 IEEE Third Workshop on Signal Processing Advances in20-23*, pp. 322 – 325, Mar. 2001.

[5] "A design of audio-visual talker tracking system based on csp analysis and frame difference in real noisy environments," *Multimedia Signal Processing,*, vol. ,2004 IEEE 6th Workshop on 29 Sept.-1 Oct., pp. 63–66, Oct. 2004.

[6] R. . SCHMIDT, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Audio, Speech, and Language Processing on [legacy, pre - 1988]*, vol. 34, no. 3, pp. 276 – 280, Mar 1986.

[7] M. O. D. Giuliani, M. Matassoni, "Hands free continuous speech reconnition in noisy environment using a four microphone array," *Acoustics, Speech, and Signal Processing, ICASSP-95.*, vol. 1, nos. 9–12, pp. 860 – 863, May. 1995.

[8] C. Avendano, "Ambience extraction and synthesis from stereo signals for multi-channel audio upmix," *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, vol. 2, no. 2, Sep. 2002.

[9] R. Dressler, "Dolby surround pro logic ii decoder principles of operation," *http://www.dolby.com/assets/pdf/tech_library/209_Dolby_Surround_Pro_Logic_II_Decoder_Principles_of_Operation.pdf.*

[10] J. L. F. Harvey F. Silverman, William R. Patterson, "The huge microphone array," *Concurrency, IEEE*, vol. 6, no. 4, pp. 36–46, Oct.-Dec. 1998.

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing.* Prentice Hall, 1992.

[12] R. W. Schafer, *DSP FIRST: A multimedia approach.* James H. McClellan, 1994.