Visualization and manipulation of stereophonic audio signals by means of IID and IPD

Giorgio Presti, Goffredo Haus
LIM - Laboratorio di Informatica Musicale
Dipartimento di Informatica (DI)
Universitá degli Studi di Milano
Via Comelico 39/41, 20135 - Milan (Italy)
giorgio.presti@unimi.it

Davide A. Mauro
Department of Architecture and Arts
IUAV University of Venice
Dorsoduro 2196, 30123 - Venice (Italy)
dmauro@iuav.it

ABSTRACT

In this paper we will discuss a model aimed at improving the spectral data representation of stereophonic audio in a way that allows efficient stereophonic data visualization and linear manipulation of arbitrary parts of the stereo image. The stereo pair is here interpreted as a single spectrum with additional dimensions, expressing the Interaural Intensity Difference (IID) and Interaural Phase Difference (IPD) for each FFT bin. These dimensions are evaluated assuming that the stereo signal is an instantaneous mixture with a residual amount of convolutive phenomena. Even if this assumption is not generally true for the majority of music signals it is applicable to single stems or submixes used during music production or other signals that comes in pairs.

After a brief overview of the state of the art in stereo data representation, we will introduce the proposed dimensions, then we will show how they can be displayed and finally we will suggest a technique to manipulate the stereophonic data in realtime.

1. INTRODUCTION

State of the art stereo representation is generally aimed at improving audio coding, compression, and transmission and relies on perceptual cues [1] [2] [3]. In [2] it is also possible to perform manipulation tasks thanks to [4]. An interesting case of stereo data representation is [3], where a preliminary analysis on the input signal refers to IID and IPD in a way which is similar to our approach, but extracted data is then used for different purposes (mainly statistical compression).

Beside the classical Mid-Side coding, where the stereo signal is encoded as sum (Mid) and difference (Side) of the left and right channels, approaches which focuses on low level information retrieval for high quality manipulation seems to be left to patented techniques only; with the exception of [5] which provides a clever way to manipulate stereo signals, but it is not very flexible and in fact it is only oriented to source separation.

We propose here a simple rearrangement of low level information that allows to gain a new perspective on data, enabling the ability to investigate and manipulate stereophonic audio in a simple but novel and more straightforward way.

2. ADDING POSITION AND CLARITY INDEXES TO SPECTRAL DATA AS CUES OF IID AND IPD

Generally, instantaneous stereophonic mixtures are obtained mixing monophonic sources distributed in the stereo image [1] via panning and level weighting. A common implementation of the pan-pot tool can be found in equation (1) (other pan laws are differentiated by a gain factor which is here negligible). This applies in both time and frequency domain, so in case of a monophonic input the equation (1) can be considered as the weighting function of each FFT bin used to generate the stereophonic output.

$$\begin{aligned} L &= source \cdot \cos(pan + \tfrac{\pi}{4}) \\ R &= source \cdot \sin(pan + \tfrac{\pi}{4}) \end{aligned} \tag{1}$$

Where $source$ is an FFT bin (a complex number) coming from the monophonic input signal; L and R are the pan-pot output bins; and pan is the position in the stereophonic image α where the source should be placed. Negative values of pan shift the signal to the left and positive values to the right. The $\frac{\pi}{4}$ factor is introduced to have $pan = 0$ for the centre position.

The case of $pan = \frac{\pi}{4}$ corresponds to the right loudspeaker position ($L = 0; R = source$) and $pan = -\frac{\pi}{4}$ to the left loudspeaker position ($L = source; R = 0$).

Since L and R vary consistently and continuously while changing the pan parameter, we can interpret the panned signal as a continuous distribution on the azimuth plane α, and the pan-pot output L and R as observations in correspondence of $\alpha = \frac{\pi}{4}$ and $\alpha = -\frac{\pi}{4}$. In optimal circumstances the perceived source position is $\alpha = pan$ as shown in Figure 1.

The curve in Figure 1 represents the magnitude of a source in every position of the stereo image. The function we used to trace the curve is shown in equation (2) and it is derived

[1] In this context we refer to the stereophonic image as the azimuthal plane α of the spherical coordinates system centred on a virtual listener's head.

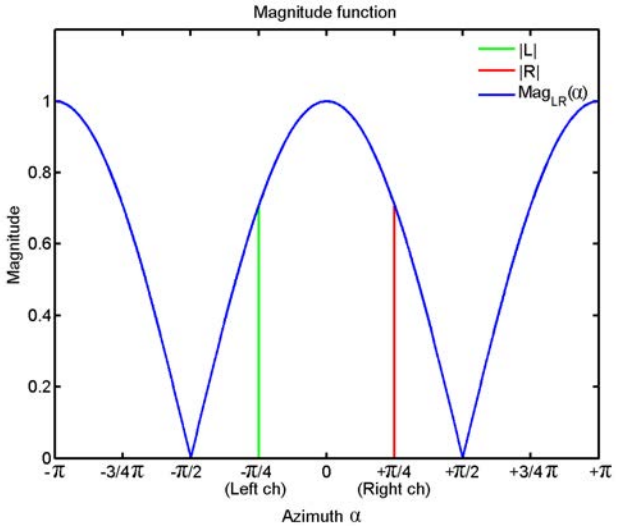

Figure 1. Magnitude of a single $source$ bin as function of the azimuth: The curve has been interpolated by the values of L and R, which are thought as two observations of a continuum in the stereo image. In this case $source = 1$, $pan = 0$ and $IPD = 0$

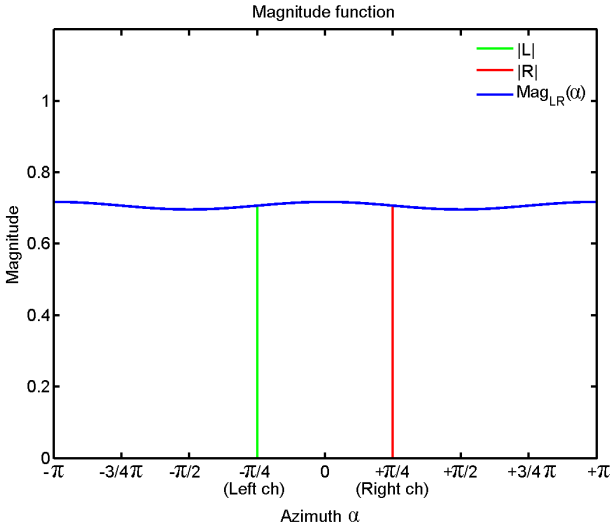

Figure 2. Here IPD is very close to $\frac{\pi}{2}$, indicating a frequency widely diffused in the stereo image. In this case inaccurate results in the computation of σ can occur.

from the interpolation of the complex values L and R obtained with equation (1).

$$Mag_{LR}(\alpha) = \tfrac{1}{\sqrt{2}}[(|S|\sin\alpha + |M|\cos\alpha)^2 + \\ + |M|\,|S|\sin(2\alpha)\cdot(\cos\Delta\varphi_{MS} - 1)]^{\frac{1}{2}} \quad (2)$$

Where $M = R + L$, $S = R - L$ and $\Delta\varphi_{MS}$ is the phase difference between M and S bins. This function has a periodicity of π and peaks in correspondence of an angle σ with a value equal to the $source$ bin magnitude. We can infer σ by finding the zeroes of the α derivative of equation (2). The resulting function is shown in equation (3):

$$\sigma = \frac{1}{2}\arctan\frac{2\,|S|\,|M|\cdot\cos(\Delta\varphi_{MS})}{|M|^2 - |S|^2} \quad (3)$$

Now, every pair of L, R bins can be described as a single point in the azimuth plane at an angle σ and magnitude $Mag_{LR}(\sigma)$. In case the assumptions are fulfilled L and R are in phase and equation (2) gives consistent results ($Mag_{LR}(\sigma) = |source|$ and $\sigma = pan$). Valid results are also provided in the case that L and R are in anti-phase, since equation (1) consider $pan = \frac{\pi}{2}$ as a legitimate pan position which generates anti-phase L and R as output.

In case of different sources, with overlapping harmonics or convolutive phenomena, a significative phase difference between L and R can occur. In this case equations (2) and (3) provides ambiguous results, as shown in Figure 2.

This happens when differences in phase are close to $\frac{\pi}{2}$. In this case L and R have minimum absolute correlation and the signal can be thought as equally distributed in the azimuth plane.

With equation (4) we introduce a clarity parameter C which describes the conformity to the assumptions, the accuracy of σ evaluation, and the spread of the signal along the azimuth plane:

$$C = \frac{1}{2}[\cos(2\Delta\varphi_{LR}) + 1] \quad (4)$$

C will approach the value 0 for phase differences of $\frac{\pi}{2}$ (which indicates a diffuse sound), and 1 for phase differences of $k\pi$ (which indicates a strongly localized sound). Please note that in equation (4) we used the phase difference between L and R, and not the difference between M and S as in the previous equations.

Since every function described above depends on the FFT, it is very important to choose an appropriate FFT frame size. From a preliminary study we found that windows smaller than 1024 samples at 44.1 kHz result in a overall fall of the C value due to the loss of frequency resolution, which causes harmonics of distinct sources to be merged in a single bin. Similarly, windows bigger than 4096 samples introduce the risk of merging sources which are separated in the time domain and reduces the ability to track sources with pan-pot automations. Ultimately we found that frame sizes of 2048 samples can handle the majority of the tested material.

3. DATA VISUALIZATION

When considering the sonogram of a stereophonic signal, a naive approach is to display the sonogram of the sum of the channels. In this way those components which are not in phase are cancelled from the plot. So, a more refined way to mix the spectra is to sum only the magnitude after the FFT process. Anyhow, in the classic $L + R$ sonogram, no information about the relations between the channels is presented. At the same time, the display of two separate sonograms for the inspected channels is not really easy to interpret and introduces a lot of redundancy.

This lead us to render a new kind of sonogram (see Figure 3), which encodes $Mag_{LR}(\sigma)$ with $brightness$, σ with hue and C with $saturation$. This method reduces the redundancy of having two separate sonograms and highlights the

differences of the channels, without discarding any binaural information. Moreover low-level visual cues such as brightness, hue and saturation are processed by our brain faster than the pattern recognition task needed to compare two separate sonograms [6].

For example, if a colourful image appears like in Figure 3 (left), it means that the observed signal fully satisfies the assumptions and sources are easily identifiable in the sonogram. If a greyscale image appears, like in Figure 3 (right), it means that the spectrum is very diffused in the stereo image and the channels have a very low absolute correlation.

This two samples were chosen to represent a well fitting dataset (in case of Coltrane's "In a sentimental mood" excerpt) and a very bad fitting dataset (in the Grieg's "Morning mood" excerpt).

This kind of sonogram helps interpreting the stereo image more as a continuum than a discrete set of channels, and packs a wide range of information in a compact area, letting the user recognize the signal properties at a glance. On the other hand, the subjective visual brightness of different colours could reduce the perception accuracy of the components loudness, but this issue is not critical, since precise loudness measurements are more affordable in the classical 2D spectrum magnitude profile plot.

Other information can be collected and plotted from the proposed space, like the dispersion of bins in the stereo image. However plotting the mere distribution of bins σ can produce misleading graphs, because bins with different magnitude have the same influence on the graph. Also the simple $Mag_{LR}(\sigma)$ weighting can be misleading, because bins with $C = 0$ are placed at an angle σ which is very likely incorrect. So, to plot the bins dispersion in the stereo image, we decided to use the product of $Mag_{LR}(\sigma) * C$ to obtain a plot which weights more the bins fitting the assumptions than the unfitting ones.

Figure 4 presents the distribution of $Mag_{LR}(\sigma) * C$ over the azimuth-frequency plane, while Figure 5 presents the distribution of the same product over the azimuth-time plane. These plots can give hints about the stereophonic balance along time and frequency. The first sample is easily interpretable, but also the second sample reveals patterns in the data distribution (like the clear bias towards the left channel).

Figure 6 shows the overall $Mag_{LR}(\sigma) * C$ accumulation over the stereo image (this plot is similar to the one introduced by [5]). Since the distribution is weighted by the factor C, the range of the plots can give hints about the general fitness of the data to the assumptions: the left sample of Figure 6 have a peak over 0.4, while the right sample peaks under 0.08. This difference reflects the fitness to the assumptions of the two samples. (input data has been normalized to keep the excerpts comparable).

4. DATA MANIPULATION

$Mag_{LR}(\alpha)$, σ and C functions can be exploited to perform two different kinds of data manipulation: Spectral masking and Azimuth resampling.

Spectral masking is a simple way to control the level of specific components of the stereo image using σ as key to mask parts of the spectrum. This can be useful in mastering, restoration, and noise suppression processes. Also C can be used to generate masks useful to split the instantaneous components of the mix from the convolutive or overlapping components.

Please note that spectral masking with σ as key is similar to the process presented by [5] where phase cancellation and gain scaling are exploited in the frequency domain to seek and separate sources across the azimuth plane. By using masking instead of resynthesis we are able to obtain a linear deconstruction of the input, which is not true in the case of the resynthesis technique used by [5].

Azimuth resampling is a resynthesis process which relays on the $Mag_{LR}(\alpha)$ function to rotate the listening space or generate virtual listening points. Rotation is made choosing any α and $\alpha + \frac{\pi}{2}$ to create new output channels. Similarly, virtual listening points can be rendered from the stereo input choosing any α as argument for the magnitude function (see equation (5), where β and γ are the new listening positions). For example it is possible to generate multichannel audio from a stereo track interrogating $Mag_{LR}(\alpha)$ in more than two points.

$$\begin{aligned} |_{new}L| &= Mag_{LR}(\beta) \\ |_{new}R| &= Mag_{LR}(\gamma) \end{aligned} \tag{5}$$

If β and γ are close to $\pm\frac{\pi}{4}$ the output can be resynthesized with the same bin phase of L and R. Otherwise, for β and γ closer to $k\frac{\pi}{2}$, the phase of M and S can be used (since those positions of the azimuth corresponds to the Mid and $Side$ components of the stereo image).

The two aforementioned techniques can be combined to perform selective phase correction along the spectrum, or can be used to create special spatialization effects.

To listen to some examples and to see other minor plots browse http://www.lim.di.unimi.it/stereocode

5. CONCLUSIONS

We presented a model that can improve realtime stereophonic data visualization and permits spectral transformations in specific stereo coordinates. In the proposed space it is possible to investigate and manipulate the frequency distribution along the stereo image and render different perspectives of the input space.

This technique is easy to implement and can be used in spatialization, mastering and restoration tasks, but also provides a general purpose data structure that can be used to improve existing algorithms, such as [5] or other stereo-dependent algorithms.

Beside the data visualization tasks, this technique can be exploited to perform phase correction, spatialization, stereo-localized equalization and dynamic processing and noise reduction.

Future works will also take in account ITD (interaural time difference) to extend the algorithm to convolutive mixtures.

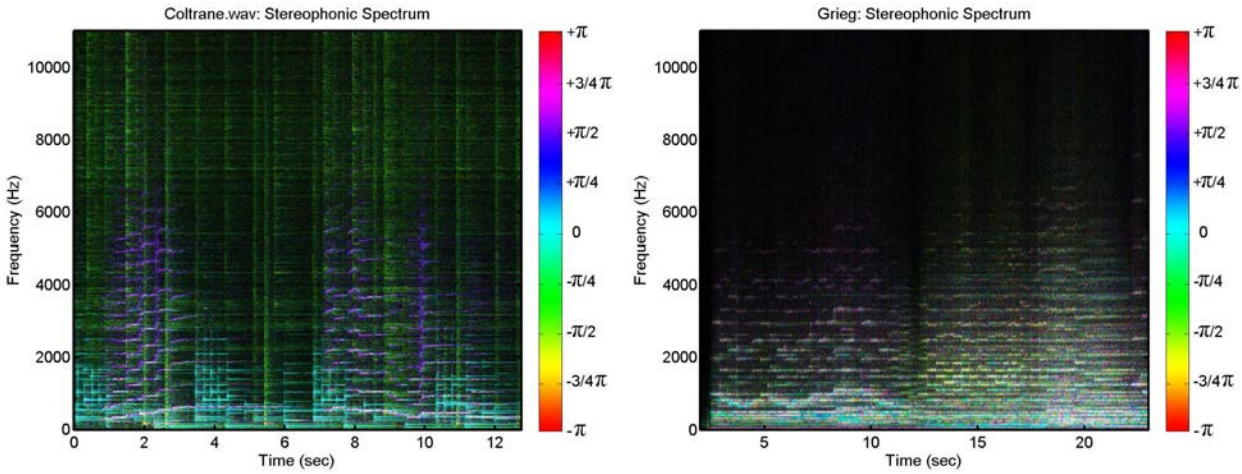

Figure 3. Sonogram with colour-coded binaural data: Brightness is related to magnitude, hue to IID and saturation is a function of IPD. (Left) Coltrane's "In a sentimental mood" excerpt: in this example it is possible to spot sax (blue), piano (cyan), drums and bass (different shades of green). (Right) Grieg's "Morning mood" excerpt: in this example sources are not clearly visible, due to reverberation, ITD and sources overlapping.

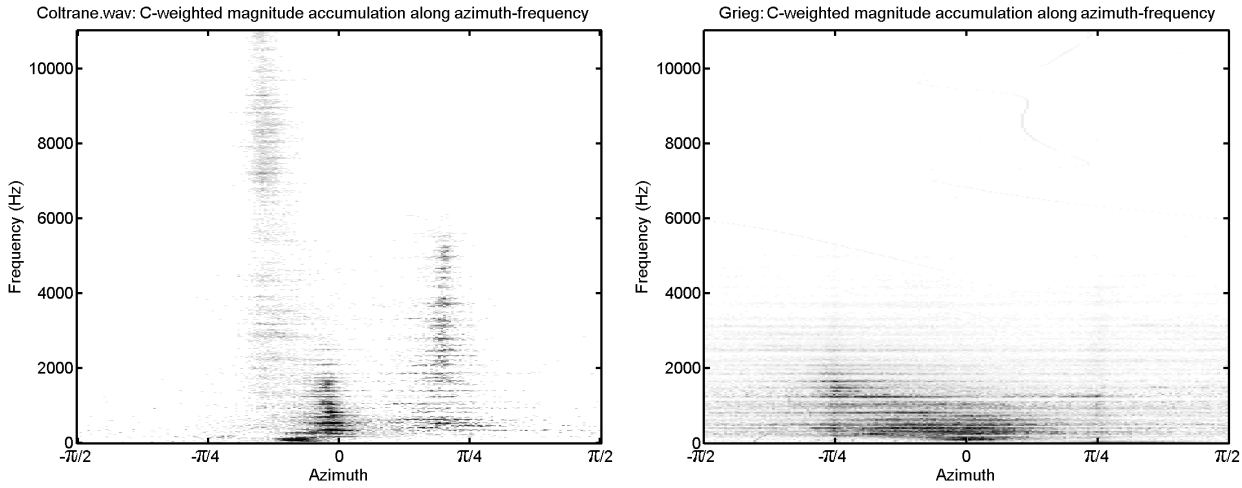

Figure 4. Projection of clarity-weighted magnitude over azimuth-frequency plane. (left) Coltrane's "In a sentimental mood" excerpt: here the spectrum occupied by each source is clearly visible. (right) Grieg's "Morning mood" excerpt: the only visible pattern here is an overall bias towards the left and a higher concentration of high frequencies always on the left

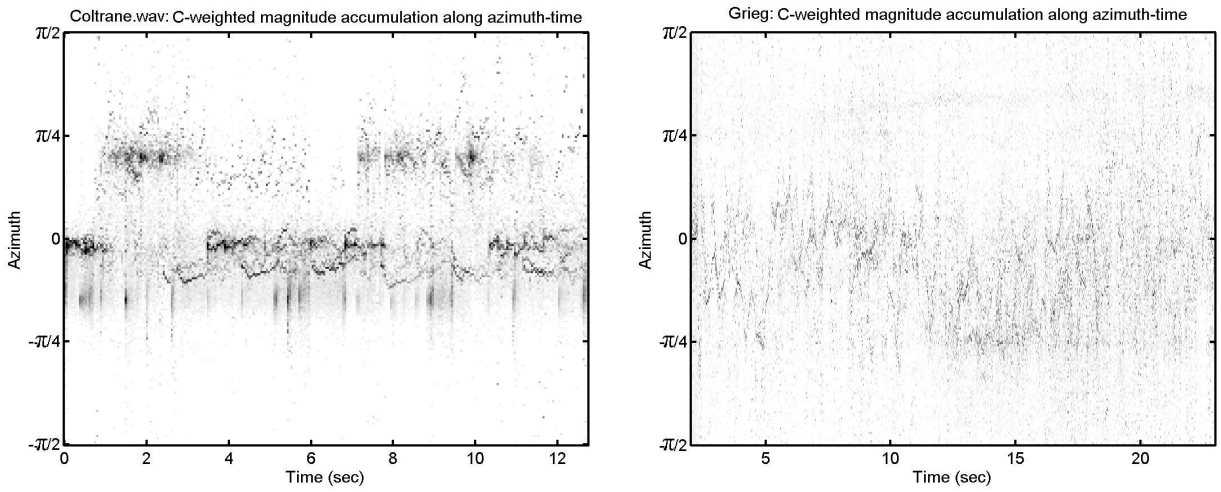

Figure 5. Projection of clarity-weighted magnitude over azimuth-time plane. (left) Coltrane's "In a sentimental mood" excerpt: in this plot it is possible to localize sources in time. (right) Grieg's "Morning mood" excerpt: from this plot it is possible to tell when the overall balance bias occurs

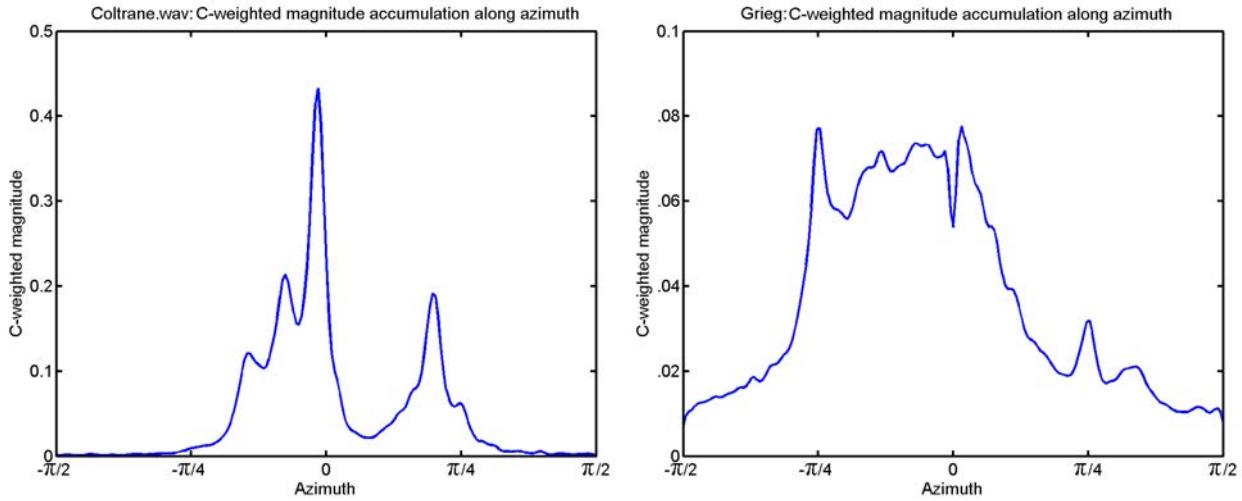

Figure 6. Distribution of clarity-weighted magnitude in the stereo image. (left) Coltrane's "In a sentimental mood" excerpt: in this plot it is possible to count the number of sources panned in the stereo image. (right) Grieg's "Morning mood" excerpt: the peak of this graph suggests an overall lack of clarity C.

6. REFERENCES

[1] C. Faller and F. Baumgarte, "Binaural cue coding: A novel and efficient representation of spatial audio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2. IEEE, 2002, pp. II–1841.

[2] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[3] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1305–1322, 2005.

[4] M. Kallinger, G. Del Galdo, F. Kuech, D. Mahne, and R. Schultz-Amling, "Spatial filtering using directional audio coding parameters," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 217–220.

[5] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discriminiation and resynthesis," 2004.

[6] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.