International Journal of Fuzzy Logic and Intelligent Systems, vol. 12, no. 2, June 2012, pp. 101−107
http://dx.doi.org/10.5391/IJFIS.2012.12.2.101 pISSN 1598−2645 eISSN 2093−744X

A Basic Study on the Conversion of Sound into Color Image using both Pitch and Energy

Sung-Ill Kim*

Department of Electronic Engineering, Kyungnam University, Changwon, Kyungnam, 731-601, Korea

Abstract

This study describes a proposed method of converting an input sound signal into a color image by emulating human synesthetic skills which make it possible to associate an sound source with a specific color image. As a first step of sound-to-image conversion, features such as fundamental frequency(F0) and energy are extracted from an input sound source. Then, a musical scale and an octave can be calculated from F0 signals, so that scale, energy and octave can be converted into three elements of HSI model such hue, saturation and intensity, respectively. Finally, a color image with the BMP file format is created as an output of the process of the HSI-to-RGB conversion. We built a basic system on the basis of the proposed method using a standard C-programming. The simulation results revealed that output color images with the BMP file format created from input sound sources have diverse hues corresponding to the change of the F0 signals, where the hue elements have different intensities depending on octaves with the minimum frequency of 20Hz. Furthermore, output images also have various levels of chroma(or saturation) which is directly converted from the energy.

Keywords : Sound-Image Conversion, Synesthesia, Pitch, Fundamental Frequency, Energy, HSI model

1. Introduction

A synesthesia literally means a joined perception that is a neurological condition in humans characterized by involuntary cross-activation of the senses[1-3]. The human synesthesia can be represented by five bodily senses by which human being can perceive information from a outside world. Multiple forms of synesthesia exist, including distinct visual, tactile or gustatory perceptions which are automatically triggered by a stimulus with different sensory properties. For example, one sense such as hearing is simultaneously perceived as if by one or more additional senses such as sight, so that it can be possible for synesthetes to see colors when hearing music.

Hitherto, many studies in the field of philosophical- and neurological-related studies[2,3] have been active. However, there have been little previous studies of synesthetic perception from a standpoint of engineering applications. The simplest method of converting sound into image is a waveform representation with both time and amplitude, and also its analysis tools such as sound spectrum and spectrogram can be examples of sound-to-image conversion. From a sound-to-image conversion viewpoint, moreover, there have been several commercial sound players such as window media player synchronized with the sound or music.

Manuscript received Feb. 16, 2012; revised Jun. 15, 2012; accepted Jun. 16, 2012

*Corresponding Author: Sung-Ill Kim(kimstar@kyungnam.ac.kr)

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 (2011-0022511).

Sight and hearing, particularly, account for a great part of bodily senses. Even though color and sound are different in frequency bands, they are identical in physical attributes because they can be explained by a wave or a vibration. However, the studies on mutual conversion between sound and color image have not been done actively both at home[4-6] and abroad so far[7-10].

The senses of both sound and vision have always coexisted in human beings. Sound is the propagation of mechanical vibrations through any material medium. The frequency of the vibrations is what we sense as the tone of the sound. Light, on the other hand, is the propagation of the oscillations of the electric and magnetic fields. It needs no material substance in which to propagate. The frequency of oscillations of visible light is what we perceive as the color of the light.

Fig. 1 shows the spectrum on both audible and visible frequency bands. Sound waves perceptible to human ear oscillate approximately between 20 Hz and 20Khz, whereas electromagnetic waves perceptible to human eye oscillate between 390THz and 750THz. On the basis of the similarity in the physical frequency information between light(or color) and sound, it is possible to mathematically map the frequency rate of audible band into the range of visible one.

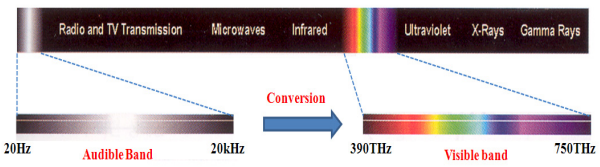


Fig. 1. Audible and visible frequency bands

In this study, we attempted to explore the visual expression of sound. The present study, particularly, focuses on both

feature extraction from an input sound source and its synesthetic conversion methods. As feature elements, pitch signals and energy were used in this study. Pitch signals as one of the extracted features from input sound sources are then converted into musical scales and octaves, where energy signals as the other feature of input sounds are converted into chroma(or saturation). Finally, a color image is synthesized to create a RGB color image with BMP file format through the HSI-to-RGB conversion.

This study can contribute to or be helpful for developing a totally new type of the applications and solutions for digital devices, advertising media, aid equipments for both blind and deaf people, educational contents, and also intelligent robot systems with an ability of synesthesia cognition, etc.

2. The Fundamental Theory on both Sound and Color Image

Modern Western instruments divide the octave into 12 equal-sized semitones. Fig. 2 shows the frequency ratios for twelve-tone musical scales[11,12] in which the frequency of each note in the chromatic scale is related to the frequency of the notes next to it by a factor of $\sqrt[12]{2}$.

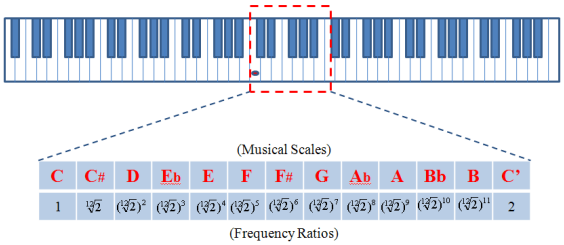


Fig. 2. The frequency ratios for twelve-tone musical scales

For some reference frequency f_R, we obtain the frequency f_k of any equal-tempered scale k(k=0,1,..,11) within the five octave by computing

$$f_k = f_R \times 2^{k/12} = f_R \times \sqrt[12]{2} \approx f_R \times 1.05946 \qquad (1)$$

in which $\sqrt[12]{2}$ is approximately 1.05946. For example, the pitch with two semitones above f_R =440Hz is $f_2 = f_R \times 2^{2/12} \cong$493.88Hz.

An octave, which is divided into twelve exactly equal intervals, is an interval whose higher note has a sound-wave frequency of vibration with twice that of its lower note. Thus the international standard pitch A above middle C vibrates at 440Hz; the octave above this A vibrates at 880Hz, while the octave below it vibrates at 220Hz. Fig. 3 shows the relationship between octave and frequency in musical scales in which a pitch played an octave higher is twice as high in pitch as the original and all 12 notes are spaced evenly inside this octave.

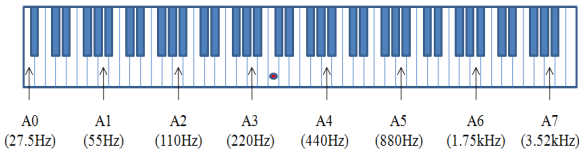


Fig. 3. The relationship between octave and frequency in musical scales

The frequency f_x of any octave x of the reference frequency f_R is

$$f_x = f_R \times 2^x \qquad x \in I \qquad (2)$$

where $x \in I$ means that x is an element of the set of all integers. If x=2, for example, then a tone with frequency $f_R \times 2^2$ is said to be two octaves higher. If x=-1, the frequency of f_{-1} is an octave below f_R because $f_{-1} = f_R \times 2^{-1} = f_R / 2$.

The HSI color model[13] is widely used for image processing applications because it represents colors similarly how human eyes sense colors. The model represents every color with three components such as H(hue), S(saturation) and I(intensity). The Hue component describes the color itself by using an angle between 0 and 360 degrees in which 0 degree means red, 120 means green, and 240 means blue. The Saturation component, which ranges from 0 to 1, describes how much the color is polluted with white color. The Intensity range is between 0 and 1 where 0 means black, 1 means white.

The RGB color space is the most widely used color model, especially used in monitors, digital cameras, etc. In this model, each color is represented by three components such as R(red), G(green) and B(blue), located along the axes of the Cartesian coordinate system. The components of RGB are available to be in the range of between 0 and 1. The black is represented as (0, 0, 0), whereas white is represented as (1, 1, 1) or (255, 255, 255). Gray scale colors are represented with identical R, G, B components. The below figure illustrates three components of both HSI and RGB color spaces, respectively, to represent colors.

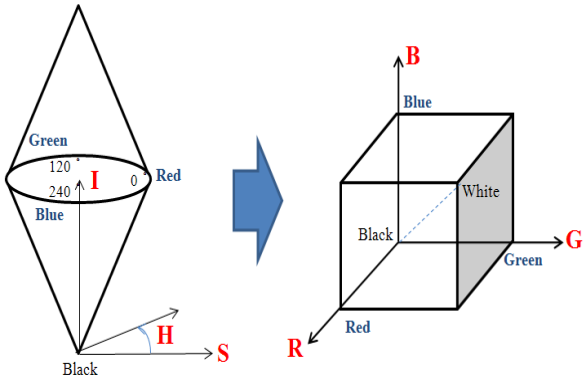


Fig. 4. The HSI and RGB Color Model

In this study, HSI to RGB color model conversion is used so that a RGB color image is finally created from an input audio signal. The conversion is made by the following equations by which HSI color space can be converted into RGB one.

$0° \le H \le 120°$, then

$$B = \frac{1}{3}(1-S)$$
$$R = \frac{1}{3}[1+\frac{S\cos(H)}{\cos(60-H)}] \quad (3)$$
$$G = 1-(R+B)$$

$120° < H \le 240°$, then

$$H = H-120$$
$$R = \frac{1}{3}[1+\frac{S\cos(H)}{\cos(60-H)}] \quad (4)$$
$$G = \frac{1}{3}(1-S)$$
$$G = 1-(R+G)$$

$240° < H < 360°$, then

$$H = H-240$$
$$B = \frac{1}{3}[1+\frac{S\cos(H)}{\cos(60-H)}] \quad (5)$$
$$G = \frac{1}{3}(1-S)$$
$$R = 1-(G+B)$$

3. The proposed method of converting sound into color image

For a conversion of a sound source into an color image, in this study, we attempted to deduce a scale and octave from the input sound signal so that each deduced element is corresponded to each one of HSI model. Fig. 5 shows the major concept of the conversion of sound elements into color ones. Both scale and octave, which are derived from F0, are converted into hue and intensity, respectively, and energy is converted into saturation as a chroma of a color.

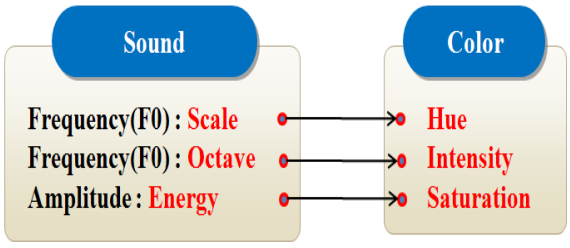


Fig. 5. A principle of a conversion of sound elements into color ones

In order to realize the conversion, a feature extraction[14,15] from an input sound signal should be first done. Fig. 6 shows a flow diagram of extracting features such as energy and F0 from an input sound with the WAVE file format. In this study, we used the normalized energies divided by a maximum value over the whole frames.

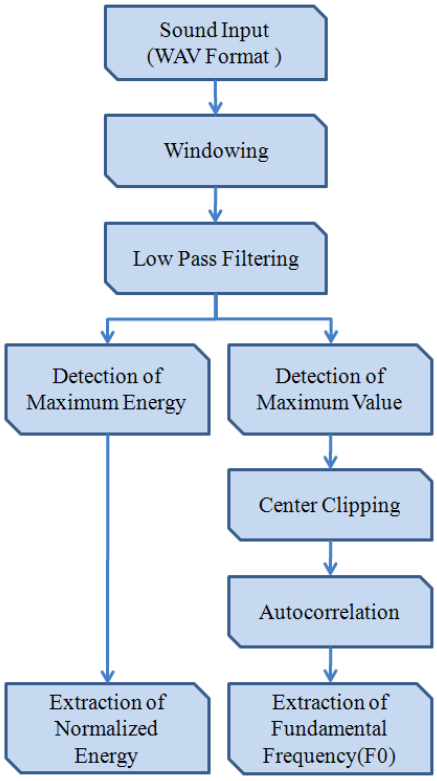


Fig. 6. A flow diagram of extracting features(energy and F0) from an input sound

The equation (6) defines the short-time energy for a sampled signal x(n) where N is the length of the rectangular window in samples.

$$Energy = \sum_{n=0}^{N-1} x^2(n) \quad (6)$$

A center clipping, which works by clipping a certain percentage of the waveform, is used for calculating an autocorrelation function. Therefore, the output from the center clipper is as follows:

if $|x(n)| > CL$

$$y(n) = |x(n)| - (MaxValue \times CL) \quad (7)$$

else

$$y(n) = 0$$

where the MaxValue is the maximum amplitude of an input signal x(n), and the CL is the clipping level with 0.64(64%) of the MaxValue in this study.

The equation (8) defines the short-time autocorrelation function which is often used as a means of detecting periodicity in signals. In this study, the autocorrelation function is used to extract a pitch from an input sound signal.

$$R_{xx}(k) = \sum_{n=1}^{N-k} x(n)x(n+k) \tag{8}$$

Fig. 7 shows the extracted features such as F0 and energy, through the process of the feature extraction as shown in Fig. 6, from an input sine wave which has nine different frequencies increasing from 320Hz to 550Hz at a same rate.

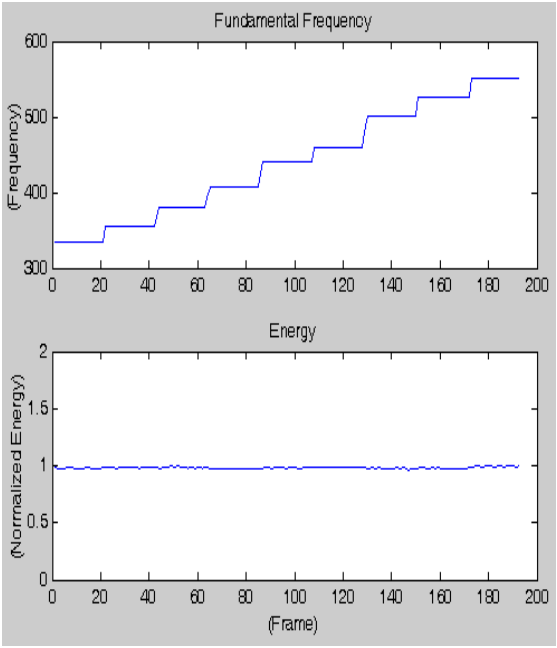


Fig. 7. The extracted features such as F0 and energy from an input sine wave with nine different frequencies

The equation (9) derives from the equation (2) where the reference frequency f_R is 20Hz as a minimum frequency of the audible frequency band. After obtaining octaves, musical scales are then calculated from the equation (10) and (11). Finally, simple equations of (12), (13) and (14) show that scale, energy and octave can be converted into three elements of HSI model such hue, saturation and intensity, respectively, as shown Fig. 5.

$$Octave = \log_2 \frac{F0}{20} \tag{9}$$

When Octave=0, 1, 2,…, 9

$$MusicalScale = \frac{F0 - 2^{Octave} \times 20}{(2^{Octave} \times 20)/12} \tag{10}$$

Otherwise

$$MusicalScale = 0 \tag{11}$$

$$Hue = MusicalScale(0,1,2,...,11) \times 23.2 \approx (0,1,2,...,255) \tag{12}$$

$$Saturation = NormEnergy(0.0,...,1.0) \times 255 \approx (0,1,2,...,255) \tag{13}$$

$$Intensity = Octave(0,1,2,...,9) \times 28.3 \approx (0,1,2,...,255) \tag{14}$$

Fig. 8 illustrates an output color image of the input sine wave with nine different frequencies, as a result of sound-to-image conversion.

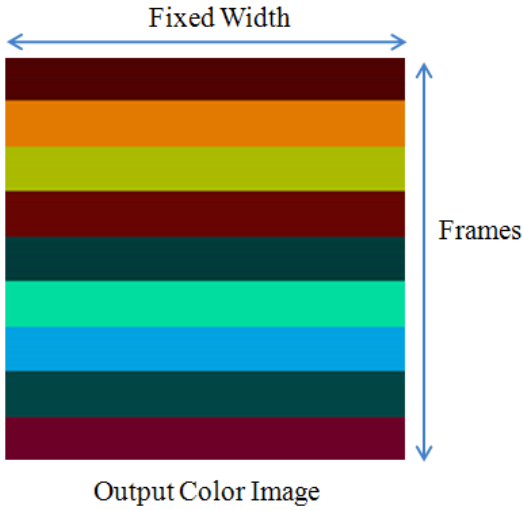


Fig. 8. The output color image of an input sine wave as a result of sound-to-image conversion

The output image has the nine different hues corresponding to the change of the F0 signals illustrated in Fig. 7(the higher part). The hues have the same intensity because they vary within the same octave range which ranges from 320Hz to 640Hz. In addition, the output image also has a nearly maximum and uniform chroma(saturation) which is directly affected by the energy illustrated in Fig. 7(the lower part). In this study, the width of the output image was fixed to 256 pixels with 24bit true colors. The hight of the output image, on the other hand, is equivalent to the number of the frames of the input sound source, so that it can be variable depending on input frame length.

Fig. 9 shows a flow diagram of converting an input sound signal into a color image as an output. An input sound file with the WAVE file format is given to the system, so that it extracts acoustic features such as F0 and energy from each frame. The energy is then normalized to be converted into a saturation which is one of three components of the HSI color model. Furthermore, the scale and octave, which are extracted from F0, are then converted into a hue and intensity, respectively. Through the process of the HSI-to-RGB conversion, a color image with the BMP file format is finally created as an output.

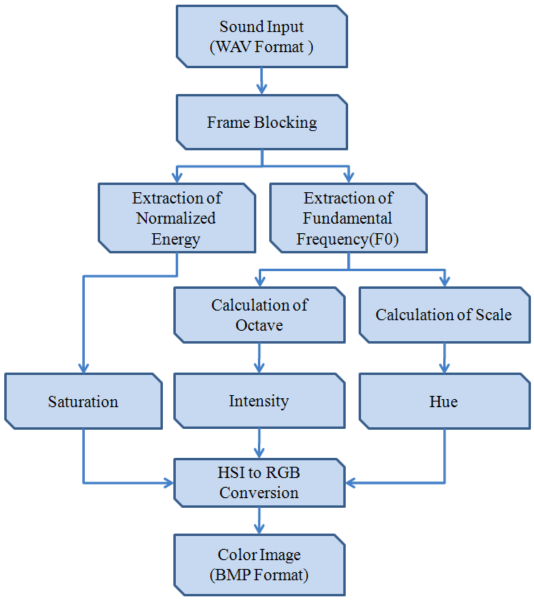


Fig. 9. A flow diagram of the conversion of sound into color image

4. Experiments and results

The input monophonic audio signal with the WAVE file format was sampled at 11kHz, quantized at 8bits. The acoustic features were then extracted from each frame using a 20ms rectangular window with a 10ms shift. Fig. 10 illustrates the features of both F0 and normalized energy, which were extracted from an input female voice.

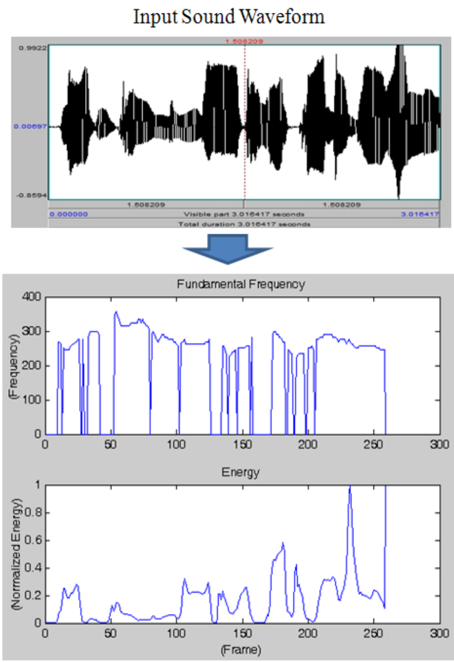


Fig. 10. The features of both F0 and normalized energy, which were extracted from a female voice

Fig. 11(a) shows examples of the features extracted from 0 to 11 frames of the female voice, and also shows the conversion of F0 into both octaves and scales.

```
##########################################################
##       Information on Input WAV Sound File            ##
##########################################################

[1] The number of channel = 1 (mono = 1, stereo = 2)
[2] Sampling rate = 11025 [Hz or samples/sec]
[3] Sample resolution = 8 [bits/sample]
[4] Play time = 3.000000 [sec]


##########################################################
##       Information on Output BMP Image File           ##
##########################################################
[1] Image Size = 256 * 259
    (Height = Input Sound length / frameshift = 33256 / 128 = 259 frames)
[2] Total Image Size(Color) = 54(Header) + 768(Width) * 259(Height) = 198966

#########################################################
##             Sound-to-Image Conversion               ##
#########################################################

Frame   F0          Energy                              ==>   Octave   Scale

0       0.000000    0.000196(=241.80 / 1234677.81)      ==>   0        0
1       0.000000    0.000203(=250.43 / 1234677.81)      ==>   0        0
2       0.000000    0.000205(=253.39 / 1234677.81)      ==>   0        0
3       0.000000    0.000322(=397.65 / 1234677.81)      ==>   0        0
4       0.000000    0.000309(=381.76 / 1234677.81)      ==>   0        0
5       0.000000    0.000211(=260.09 / 1234677.81)      ==>   0        0
6       0.000000    0.000285(=352.49 / 1234677.81)      ==>   0        0
7       0.000000    0.000675(=832.94 / 1234677.81)      ==>   0        0
8       0.000000    0.014607(=18034.56 / 1234677.81)    ==>   0        0
9       268.902439  0.049081(=60599.59 / 1234677.81)    ==>   3        8
10      268.902439  0.080112(=98912.73 / 1234677.81)    ==>   3        8
11      262.500000  0.093790(=115800.09 / 1234677.81)   ==>   3        7
```

(a) Examples of the extraction of features, such as F0 and normalized energy, and the conversion of F0 into both octaves and scales

Octave	Scale	H	S	I	==>	R	G	B
0	0	0	0	0	==>	0	0	0
0	0	0	0	0	==>	0	0	0
0	0	0	0	0	==>	0	0	0
0	0	0	0	0	==>	0	0	0
0	0	0	0	0	==>	0	0	0
0	0	0	0	0	==>	0	0	0
0	0	0	0	0	==>	0	0	0
0	0	0	0	0	==>	0	0	0
0	0	0	3	0	==>	0	0	0
3	8	185	12	106	==>	104	100	112
3	8	185	20	106	==>	104	97	116
3	7	162	23	105	==>	95	100	119

(b) Examples of the conversion of HSI into RGB model

Fig. 11. The feature extraction and the color model conversion as a result of sound-to-image conversion

Fig. 11(b), on the other hand, shows examples of the conversion of HSI into RGB model where the values of H, S and I are derived from a scale, an energy and an octave.

Fig. 12 illustrates the output color image created from the input female voice. The output image has diverse hues corresponding to the change of the F0 signals. The hue elements have different intensities because they vary from third to fourth octave which ranges from 160Hz to 640Hz. Furthermore, the output image also has various levels of chroma(or saturation) which is directly converted from the normalized energy as shown in fig. 10.

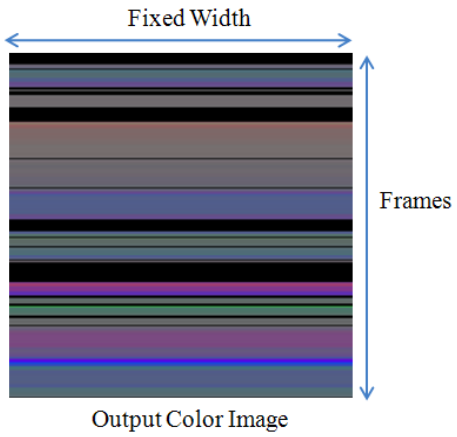
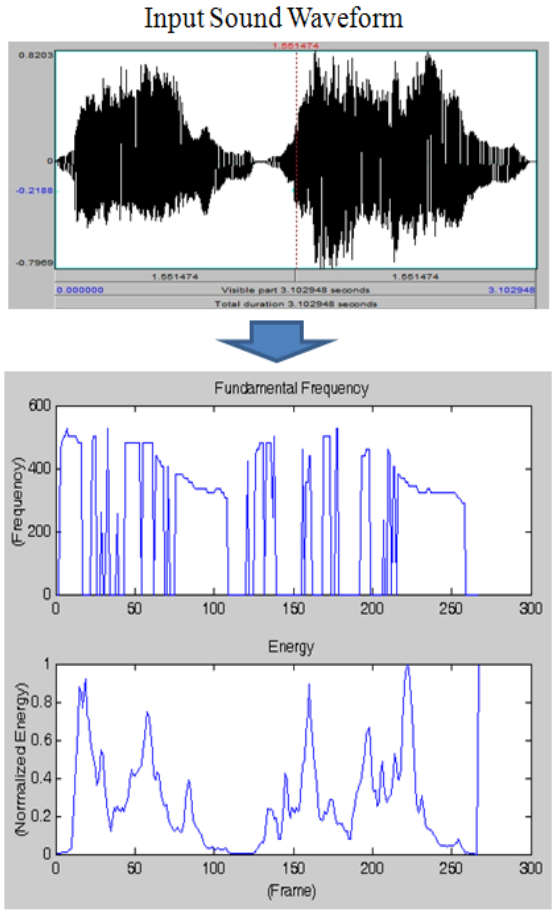
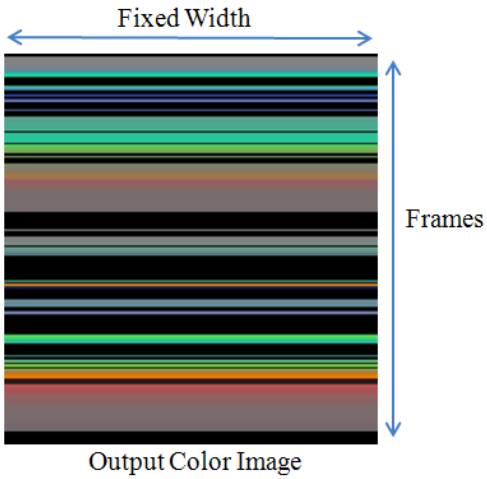


Fig. 12. The output color image created from the input female voice

Fig. 13(a) illustrates an another example of the features which were extracted from an input baby's crying sound. Fig. 13(b) illustrates its output color image with diverse hues, intensities and various levels of chroma, which are converted from the values of both F0 and normalized energy.



(a) The features of both F0 and normalized energy, which were extracted from a baby's crying sound



(b) The output color image created from the input baby's crying sound

Fig. 13. The feature extraction and the output as a result of sound-to-image conversion

5. Conclusion

As a preliminary study on mutual conversions between color images and sounds, this study presented the approach to an sound-to-image conversion emulating human synesthetic skills. The simulation results showed that output color images created from input sound sources have a wide variety of colors corresponding to the change of the F0 signals where each color has a different intensity depending on the value of its octave with the reference frequency of 20Hz. Moreover, we could see that output images also have various levels of saturation which is directly converted from the normalized energy. In the present study, unfortunately, the system dealt with only voice signals and also used a simple one-to-one temporal correspondence between sound and image in conversion. The temporal information of a sound is simply converted into the spatial information of a color image. As a result, the height of the output image is depending on the temporal order.

As future studies, the current system should be developed to explore more diverse acoustic features as well as to find more natural conversion methods. In order to do this, we should deal with music signals as input sound sources, so that we will be able to explore a totally new type of conversion method handling three elements of music such as rhythm, melody and harmony. In addition, the extracted features of a music should be converted into basic elements of an image such as color, texture and shape, so that the conversion of temporal information of sound into spatial one of image will be able to be realized in reality.

References

[1] O. Teng, "Synesthesia: Beyond the Five Senses," *Executive intelligence review*, vol. 38, no. 5, pp. 6-9, 2011.

[2] R. E. Cytowic, "Synesthesia: A Union of the Senses," The MIT Press, 2002.

[3] L. C. Robertson, N. Sagiv, “Synesthesia: Perspectives from Cognitive Neuroscience," Oxford University Press, 2004.

[4] G. H. Kim, J. G. Beak, Sound Color Harmonism(in Korean), *Impress*, 2003.

[5] G. H. Kim, Method and Apparatus for harmonizing Colors by Harmonics and converting Sound into Colors mutually(in Korean), Korean Intellectual Property, 10-99-34242, 1999.

[6] G. H. Kim, The Sound-to-Color Conversion Table using a Low of Harmony(in Korean), Korean Intellectual Property, 10-2001-0087651, 2001.

[7] J. Ward, B. Huckstep, E. Tsakanikos, "Sound-Colour Synaesthesia: to What Extent Does it Use Cross-Modal Mechanisms Common to us All," *Cortex; a journal devoted to the study of the nervous system and behavior*, vol. 42, no. 2, pp. 264-280, 2006.

[8] Thórisson, K. R., Donoghue, K., "Synthetic Synesthesia: Mixing Sound with Color," *InterChi Adjunct Proceedings*, pp. 65-66, 1993.

[9] Leonard N. Foner, "Artificial synesthesia via sonification: A wearable augmented sensory system," *Mobile networks and applications : MONET*, vol. 4, no. 1, pp. 75-81, 1999.

[10] Peter B.L. Meijer, “An Experimental System for Auditory Image Representations," *IEEE trans. on bio-medical engineering*, vol. 39, no. 2, pp. 112-121, 1992.

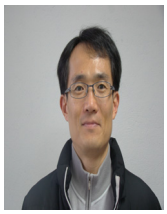
[11] G. Loy, "Musimathics: The Mathematical Foundations of Music (Volume 1)," The MIT Press, 2006.

[12] G. Loy, J. Chowning, "Musimathics: The Mathematical Foundations of Music (Volume 2)," The MIT Press, 2007.

[13] Michael Freeman, “Mastering Color Digital Photography," Ilex Press, 2004.

[14] John R. Deller Jr., John H. L. Hansen, John G. Proakis, "Discrete-Time Processing of Speech Signals," Wiley-IEEE Press, 1999.

[15] Speech Signal Processing Toolkit (SPTK), http://sp-tk.sourceforge.net/



Sung-Ill Kim

1994: B.S. from Dept. of Electronic Eng., Yeungnam Univ., Korea.

1997: M.S. from Dept. of Electronic Eng., Yeungnam Univ., Korea.

2000: Ph.D. from Dept. of Computer Science & Systems Eng., Miyazaki Univ., Japan.

2000-2001: Researcher at the National Institute for Longevity Sciences, Japan.

2001-2003: Researcher at the Center of Speech Technology, Tsinghua Univ., China.

2003-2006: Full-time lecturer at the Div. of Electrical & Electronic Eng., Kyungnam Univ., Korea.

2006-2010: Assistant professor at the Dept. of Electronic Eng., Kyungnam Univ., Korea.

2010-Current: Associate professor at the Dept. of Electronic Eng.,Kyungnam Univ., Korea.

Phone : +82-55-249-2632

E-mail : kimstar@kyungnam.ac.kr