From Shapes to Sounds: A perceptual mapping *

Vikas Chandrakant Raykar
vikas@umiacs.umd.edu

Abstract

In this report we present a perceptually inspired mapping to convert a simple two dimensional image consisting of simple geometrical shapes to a one dimensional audio waveform consisting of simple harmonic complexes. More specifically we map objects to harmonic complexes where the pitch, timbre and location of the complex corresponds to the size, shape and the position of the object respectively.

1 Motivation

On the outset audition and vision appear to be two completely different sensory modalities. While visual perception has a two dimensional input (the image on the retina), the input to the auditory system is a one dimensional pressure waveform incident on the eardrum. Each of the modalities has its own percepts. Spatial location, depth, motion, size, color, symmetry, texture contribute to a rich set of visual percepts. We make sense of the world we see in terms of the different objects we see and the percepts associated with them. In a similar way we make sense of the auditory scene in terms of the auditory percepts like source direction, range, timbre and pitch. Even though these two modalities look different a computational frame work exists which can explain both these seemingly different perception in a unified framework [1]. There exists a interesting medical condition called *synesthesia* where there exists confusion between these two senses, where people reportedly hear shapes [2]. This could be probably because of the cross-wiring between the two areas in the brain. There is an interesting theory which explains how evolution of language is related to the shapes of objects. Sounds can be metaphors for images, for example sounds can be described as bright or dull. The sounds and shapes of the objects have characteristics in common that can be abstracted, say a sharp, cutting quality of a word, and the shape it describes - also called *Bouba/kiki effect* based on the results of an experiment with two shapes and asking people to related the nonsense words bouba and kiki to them [3].

*This report was written for the course project for ENEE632: Speech and Audio Processing offered in Spring 2004 by Prof. Shihab Shamma.

[1] Shamma S. "On the role of space and time in auditory processing" in Trends Cogn Sci 2001 Aug 1;5(8):340-348

[2] See Richard Cytowic's book The Man Who Tasted Shapes for more interesting detailed account

[3] See the article Hearing Colors, Tasting Shapes: The Puzzle of Language by Vilyanur Ramchandran in Scentific American

2 Goal of the project

In this project we are concerned with the following concrete problem. Given a two dimensional visual input we would like to sonify the image into a one dimensional auditory waveform. We would like to do it in such a way that there is a convincing perceptual map between the visual and auditory percepts. Consider the image shown in Figure 1 which has a square and a circle next to each other. We recognize the image in terms of the objects. We say there are two objects of different sizes and different shapes and at different locations. We would like to map these visual percepts to a suitable auditory percept. The potential candidates are pitch, timbre and location.

The task would involve the following three steps:

- Given a 2D image extract the visual percepts in the image which we would like to map to. This involves identifying how many objects are there in the image, their position sizes and their shapes.
- Deciding which auditory percept to map to which visual percept.
- Generating a auditory waveform corresponding to these percepts.

Each of these is discussed in detail in the next three sections.

Figure 1: We recognize this image as consisting of a square and circle of different sizes placed next to each other.

3 Symmetry as a tool to extract the visual percepts

Given a image our task is to extract the following visual percepts

- Find the number of distinct objects in the image.
- Their spatial location in the images.
- The size of each of the objects.
- A convenient description which encodes the shape of the objects.

We will be discussing with Figure 1 as our example. We use the concept of symmetry to localize the objects in a scene. Symmetry is an important mechanism by which we identify the structure of objects. Most of the natural objects (animal and plants) and also man made objects show a high degree of symmetry. An object is considered symmetric if it remains invariant under some transformation. Two kinds of symmetry which we are familiar are the bilateral and radial symmetry. A object is bilaterally symmetric about an axis if it is invariant to a reflection about that axis. A object is rotationally symmetric if it is invariant under a rotation. For example a square has four axis of bilateral symmetry, while a circle has infinite axes of bilateral symmetry. Most mammals are bilaterally symmetric. Clearly these are not the only two kinds of symmetry. Consider the leaf shown in Fig 2 which is symmetric about its stalk. The stalk may not be exactly vertical. Note that when defining

symmetry we did not specify what kind of transformation. Also symmetry is exhibited at various scales. Certain kinds of fractals have symmetry at all possible scales. We need a multi-scale, multi-directional quantitative measure of symmetry. To this end we use the even and odd Gabor wavelets to define a quantitative measure of symmetry.

Figure 2: A leaf which is symmetrical about its stalk.

3.1 Gabor wavelets

Gabor wavelets are plane waves restricted by a gaussian envelope. The Gabor filter consists of a even symmetric part and a odd symmetric part which are defined as follows:

$$\Phi^{even}(x,y) = (\frac{k_1^2+k_2^2}{\sigma^2})\exp[\frac{(k_1^2+k_2^2)(x^2+y^2)}{2\sigma^2}]cos(k_1x+k_2y) \quad (1)$$

$$\Phi^{odd}(x,y) = (\frac{k_1^2+k_2^2}{\sigma^2})\exp[\frac{(k_1^2+k_2^2)(x^2+y^2)}{2\sigma^2}]sin(k_1x+k_2y) \quad (2)$$

Compactly we can write it as a complex filter.

$$\Phi^{even}(x,y) = (\frac{k_1^2+k_2^2}{\sigma^2})\exp[\frac{(k_1^2+k_2^2)(x^2+y^2)}{2\sigma^2}]\exp[i(k_1x+k_2y)] \quad (3)$$

Sometimes a DC correction is also added to the filter. This makes sure that the integral over the filter is zero. The output becomes independent of the mean gray level of the image under consideration.

$$\Phi^{even}(x,y) = (\frac{k_1^2+k_2^2}{\sigma^2})\exp[\frac{(k_1^2+k_2^2)(x^2+y^2)}{2\sigma^2}]\{\exp[i(k_1x+k_2y)]-\exp[\frac{-\sigma^2}{2}]\} \quad (4)$$

k_1 and k_2 can be written as

$$k_1 = rcos(\theta) \quad (5)$$

$$k_2 = rsin(\theta) \quad (6)$$

r controls the scale of the filter and θ controls the orientation of the filter. σ controls the number of excitatory and inhibitory lobes in the filter. There are a number of ways parameterize the Gabor wavelets and this is one of them. Figure 3 shows a example of the even and the odd gabor filters for a particular scale and orientation zero degrees.

Gabor wavelets are useful models for simple cell receptive fields in the visual cortex [4]. Gabor showed that these function achieve the theoretical limit for the joint representation of information in the 2D spatial and fourier domains. Pollen and Ronner showed that simple cells exist in quadrature-phase pairs as in the even and the odd symmetric part. We can use a series of Gabor filters corresponding to different scales and orientation to build a multi-scale multi-orientation representation of the image. Figure 4 shows the Gabor filters for different orientations and scales.

[4] J.G. Daugman, " Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by TwoDimensional Visual Cortical Filters," J. Optical Soc. Amer., vol. 2, no. 7, pp. 1,160-1,169, 1985

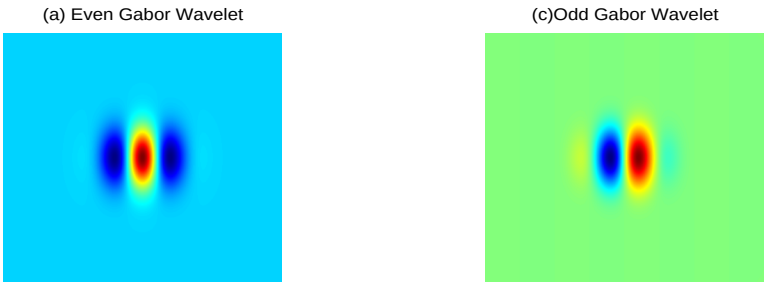


Figure 3: A sample gabor wavelet showing the even part and the odd part.

3.2 Measure of Symmetry

Figure 5(a) shows the original image. Figure 5(b) and Figure 5(c) show the even and odd Gabor filter for a particular scale and orientation zero degrees. Figure 5(e) and Figure 5(f) shows the output when the given image is filtered with these two Gabor wavelets. For the even filter the response is high at points where the image is symmetric and for the odd filter the response is high where the image is anti symmetric. So at the given point if the image is symmetric in that orientation then the even filter will give high response and the odd filter will give a low response. So we can define a measure of symmetry as the difference between the even and the odd part. For a given image $I(x, y)$ and given a gabor wavelet $\Phi(x, y, r, \theta)$ corresponding to a particular scale r and orientation θ we can define symmetry $Sym(x, y, r, \theta)$ as

$$Sym(x, y, r, \theta) = |I(x, y) \odot \Phi^{even}(x, y, r, \theta)| - |I(x, y) \odot \Phi^{odd}(x, y, r, \theta)| \quad (7)$$

where $\odot$ is the convolution operation. Figure 5(d) shows the symmetry. As can be seen the output is high at points of symmetry. Figure 6 shows the same results for a different orientation of $\theta = 45^o$.

Symmetry can occur at different orientations and scales. This can be clearly seen for a test image as shown in Figure 7 which shows the symmetry response at different scales and orientations.

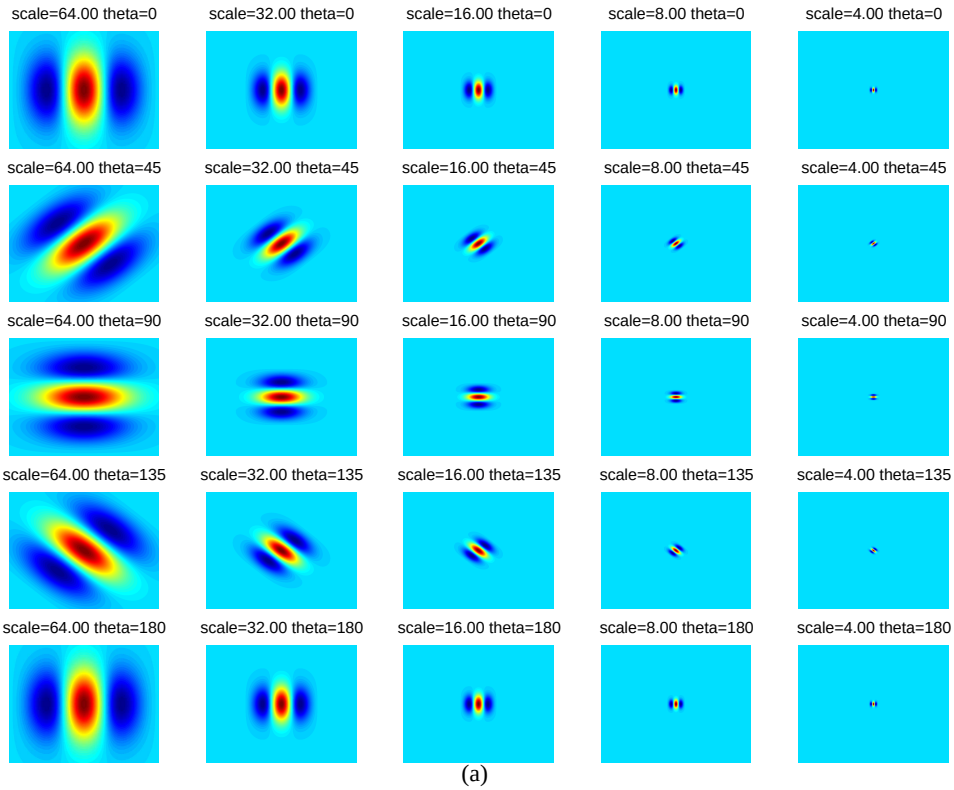
We can sum up the symmetry response at different scales and orientations we get a complete measure of symmetry. Local maximas in this representation will correspond to points of very high symmetry.

$$TotalSymmetry(x, y) = \sum_{\theta} \sum_{r} Sym(x, y, r, \theta) \quad (8)$$

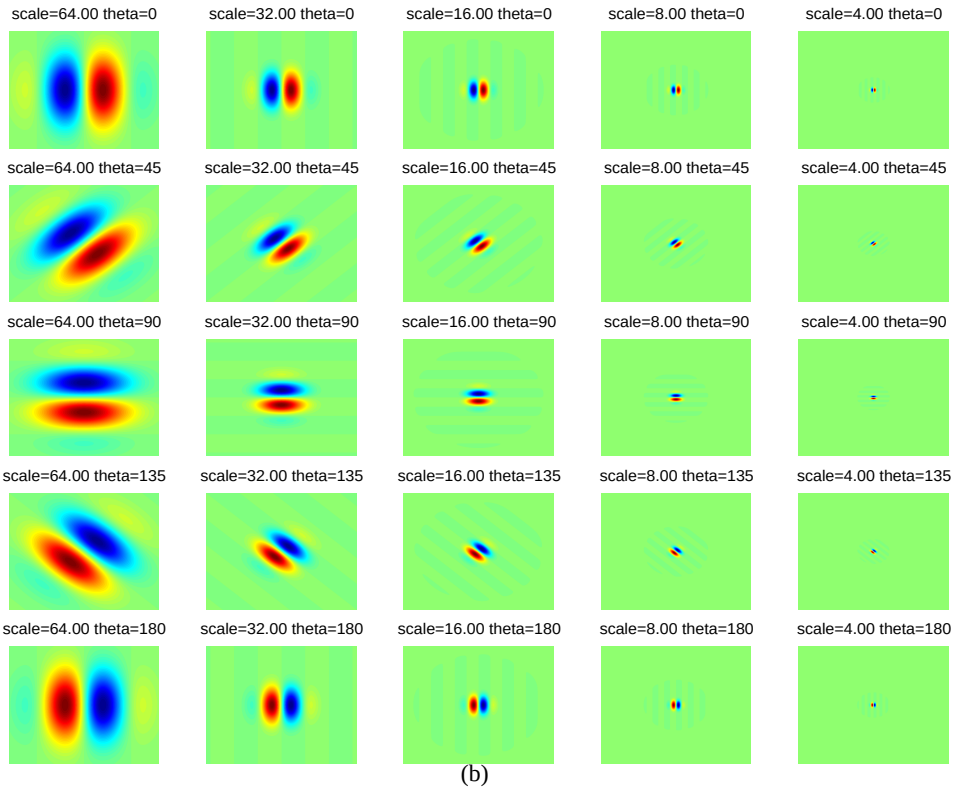
Figure 8 shows the total symmetry response for the test image. The output is high at the center of the circle and the square. Also note that there is a strong response in between because there is certain degree of symmetry overall considering the square and the circle as one object.

3.3 Extracting the visual percepts

From the total symmetry image we can easily extract the visual percepts. The location of the objects can be got by finding the local maxima in the total symmetry response. The size of the objects can be determined by finding where the intensity drops below a certain threshold from the points of local maxima. For mapping to the auditory domain we need only a measure of the relative sizes of the objects. So the threshold can be set quite heuristically. Figure 9 shows the two objects marked. Currently I am ignoring the high



(a)



(b)

Figure 4: (a) The even and (b) odd gabor wavelets for different scales and orientations.

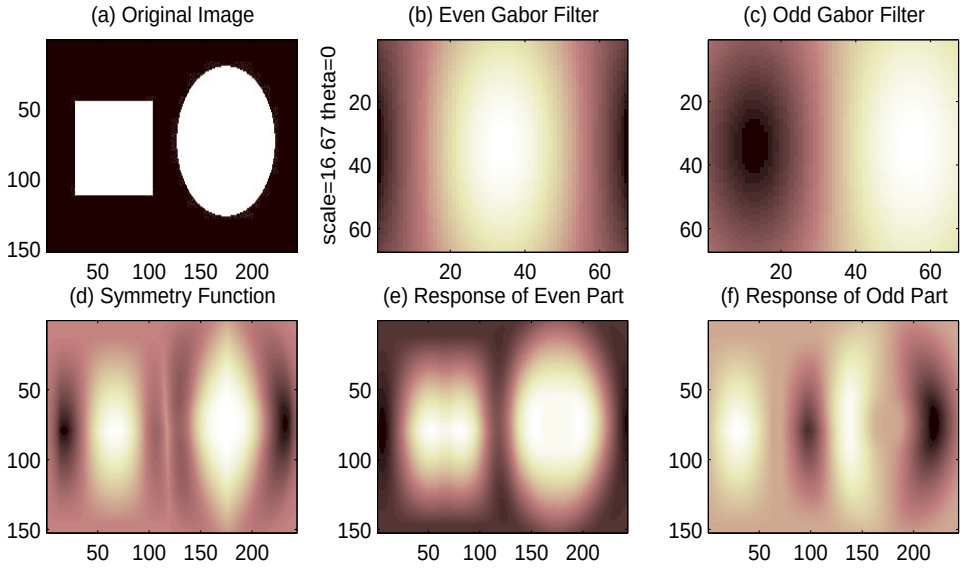


Figure 5: Output of a sample Gabor wavelet.

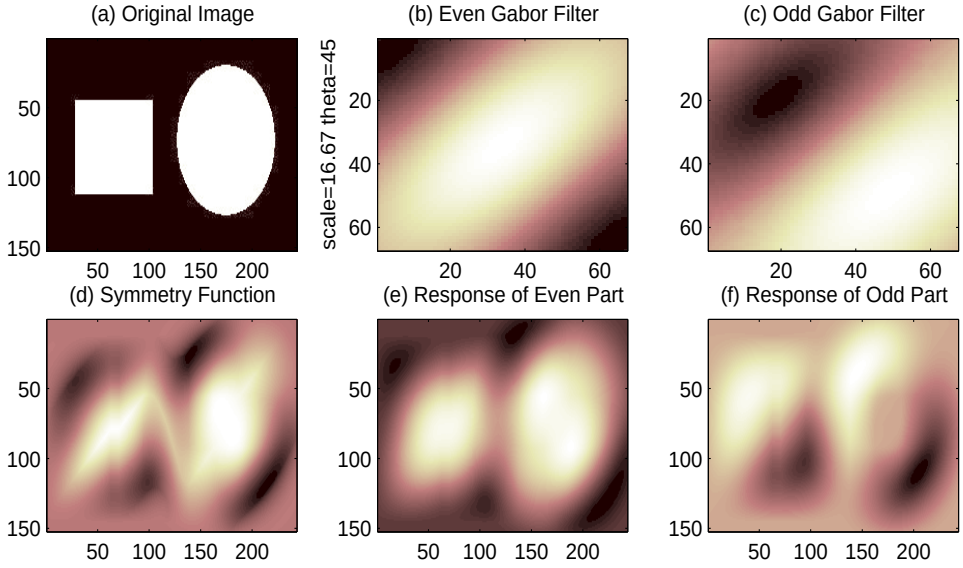


Figure 6: Output of a sample Gabor wavelet.

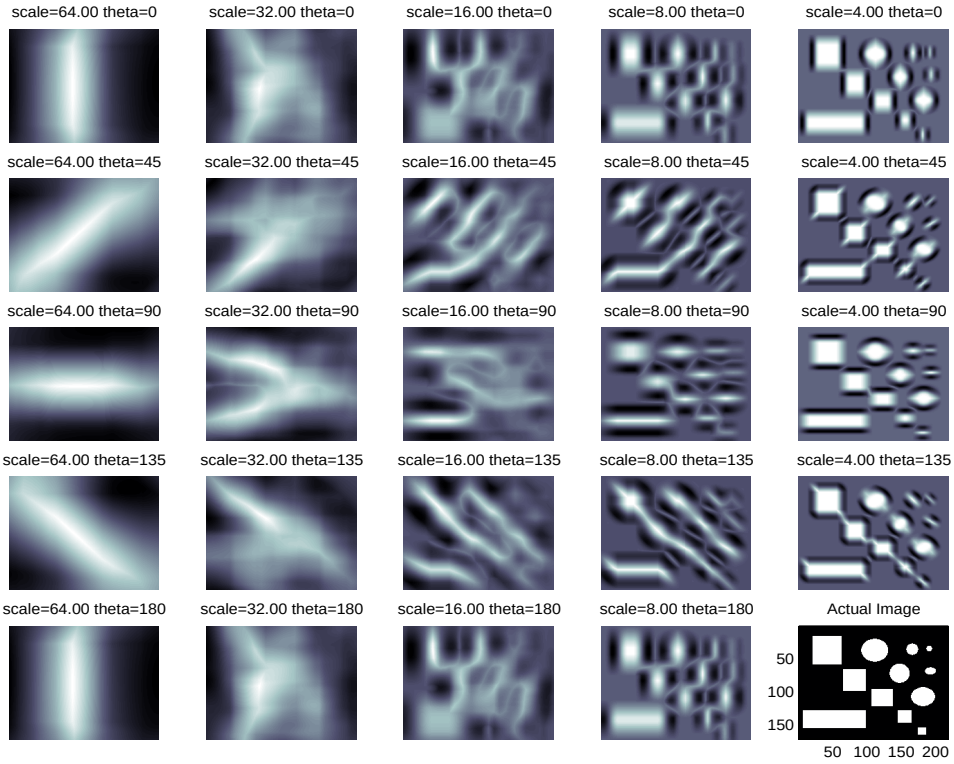


Figure 7: Symmetry response at different scales and orientations.

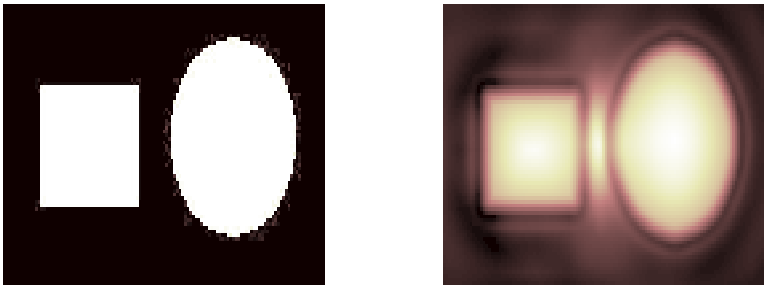


Figure 8: The test image and the total symmetry response.

response in between the images. However it can also be considered as an object. It is also possible to eliminate this by searching for the local maxima over a wider window.

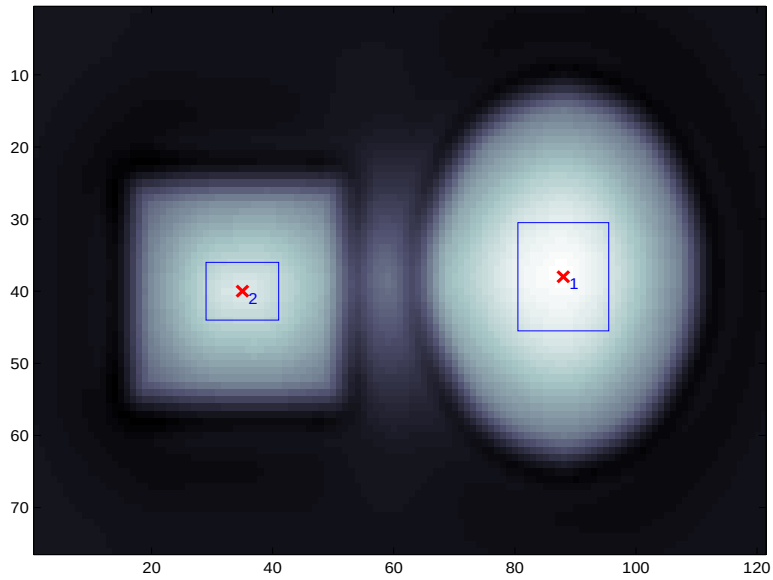


Figure 9: The total symmetry response and the two objects marked.

3.3.1 Shape Descriptor

Now we need a descriptor for the shape of the object. For this we use the sum of the response across all scales as a function of orientation, and normalizing it. Figure 10 shows this as a function of the orientation for the two objects. For the circle the variation is very less and the shape descriptor is almost a constant function of θ. However for the square the variation is quite high. Different shapes will have different shape descriptors.

4 Auditory percepts and the Perceptual mapping

4.1 Object $\equiv$ Harmonic Complex

Harmonicity is a major cue used by human auditory system for organizing complex acoustical environments. Human listeners automatically fuse partial harmonics of complex tones into unitary perceptual entities. A harmonic complex consisting of N tones is given by

$$s(t) = \sum_{i=1}^{N} Asin(2\pi i f_o t) \tag{9}$$

where f_o is the fundamental frequency. For each object extracted we assign a harmonic complex corresponding to a different fundamental frequency.

4.2 Size $\equiv$ Pitch

The perceived pitch of the harmonic complex is equal to the fundamental frequency. The larger the object we assign it a lower pitch. We can think of all the objects as resonant cavities. Larger the resonant cavity lower its fundamental. Perceptually we try to associate

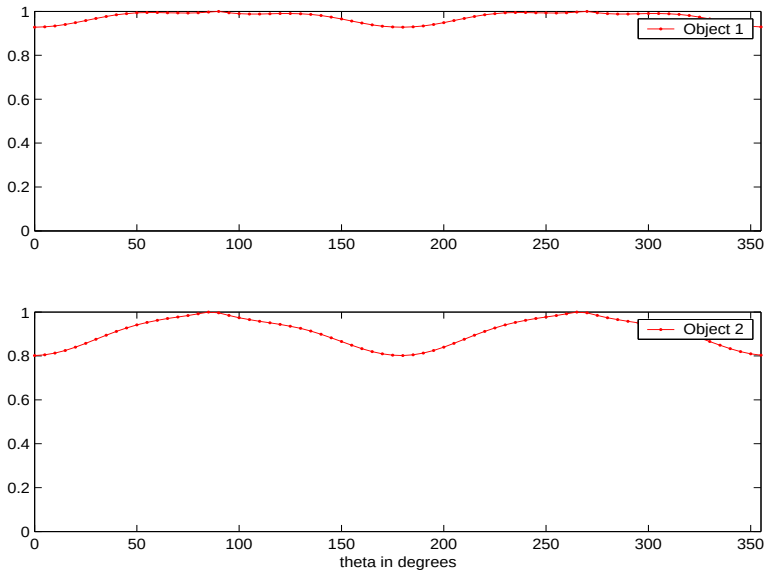


Figure 10: The shape descriptors for the two objects.

lower pitch with larger objects. Based on the ares of the extracted objects we map the relative areas to harmonic complexes with fundamental frequencies between 200 Hz to 1000 Hz [5]. For the example shown the circle would be given a lower pitch and the square a higher pitch.

4.3 x,y Position ≡ Elevation,Azimuth

As with vision hearing is also three dimensional. Humans have an amazing ability to localize a sound source, i.e., determine the range, elevation and azimuth angles of the direction of the sound source. The major mechanisms responsible for the directional capability of the human hearing system has been fairly well understood though not completely. One of the primary cues responsible for localization of the sound source is the Interaural Time Difference(ITD) and the Interaural Level Difference(ILD). However ITD and ILD cues alone do not completely explain the source localization mechanism. For example, for all points lying in the hyperboloid of revolution with vertex as the center of the head and passing through a point, the ITD and ILD cues are essentially same. Also perceptual experiments done with virtual sources rendered using just ITD and ILD cues show that while the lateral placement of the source is correct, the perceived range and elevation are not. This is because there are additional important static and dynamic acoustic cues that arise from the scattering of the sound by the head, torso and the pinnae. This can be explained in terms of the spectral filtering provided by the torso, head and the pinnae. This filtering can be described by a complex frequency response function called the Head Related Transfer Function (HRTF). The corresponding impulse response is called the Head Related Impulse Response (HRIR). By manipulating the cues responsible for the directional hearing capability a virtual audio system which can place the sound to any given location can be built by using just a pair of headphones.

The spatial position of the objects are mapped to elevation and azimuth with the center of the image corresponding to zero elevation and azimuth. Elevation increases from 0 to

[5] I am not sure of what ranges of fundamental to use. This can be modified in the program.

45 in the positive y direction, and azimuth varies from 0 to 90 in the positive x direction. Objects were rendered to the mapped spatial location be convolving them the HRTFs of the KEMAR corresponding to the particular elevation and azimuth [6]. For the example shown the square would be placed towards the left ear and the circle would be placed towards the right ear.

4.4 Strength of Symmetry ≡ Distance

We vary the range of the objects depending on the strength of the symmetry. For the example shown the circle would be placed closer than the square.

4.5 Shape ≡ Timbre

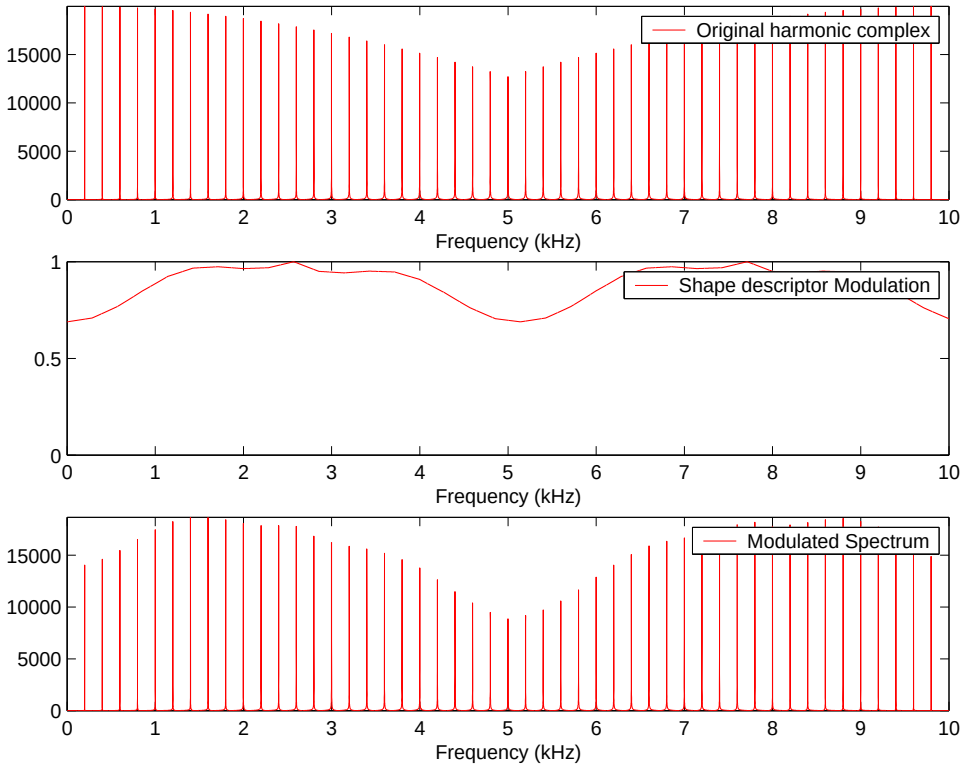
The spectro-temporal modulation decides the timbre of the sound. Using the analogy of each object as a resonating cavity the spectrum of the harmonic complex is modulated depending on the shape of the object. We use the shape descriptors shown in the previous section to modulate the harmonic complex. We stretch the range of the shape descriptors by raising the shape descriptor to a certain power. Figure 11 shows these modulations for the two objects shown in the example above.

The following shows the two objects extracted and their properties.

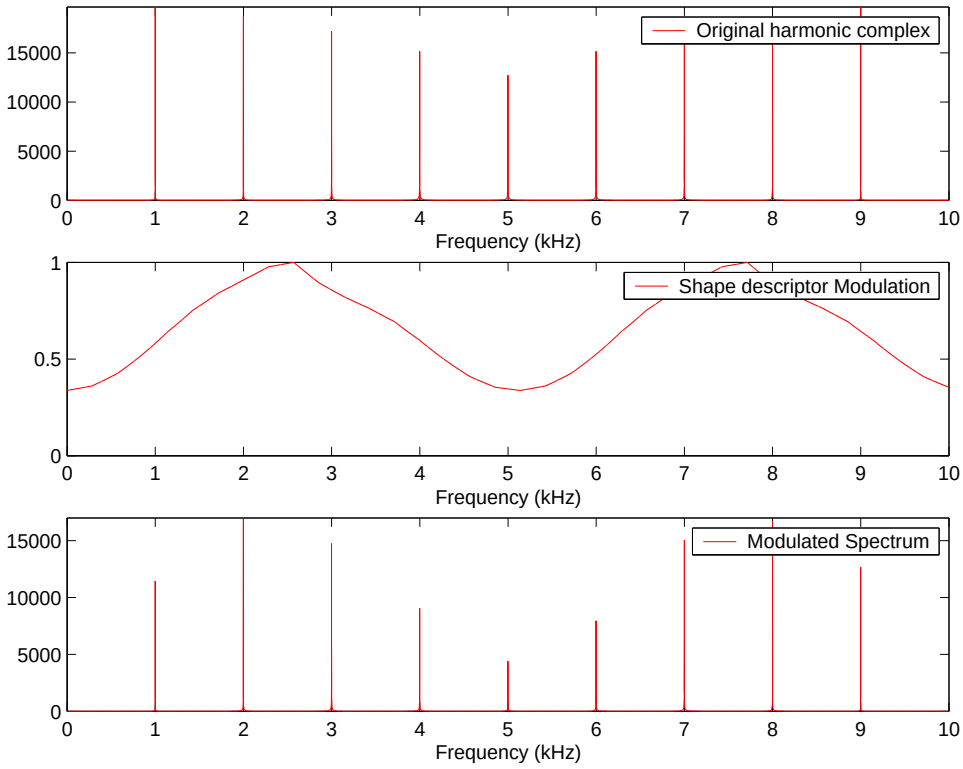
```
>> audio_data.properties(1)
ans =
            elevation: 0
              azimuth: 35.4098
    elevation_rounded: 0
      azimuth_rounded: 35
             distance: 1
                   fo: 200
          fo_rounded: 200
        num_of_tones: 50
             spectrum: [1x20002 double]
             waveform: [40002x2 double]

>> audio_data.properties(2)
ans =
            elevation: -2.3684
              azimuth: -34.0984
    elevation_rounded: 0
      azimuth_rounded: -35
             distance: 0.8823
                   fo: 1000
          fo_rounded: 1000
        num_of_tones: 10
             spectrum: [1x20002 double]
             waveform: [40002x2 double]
```

[6] We use the public-domain CIPIC HRTF database. V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, "The CIPIC HRTF Database," Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics, pp. 99-102, Mohonk Mountain House, New Paltz, NY, Oct. 21-24, 2001.



(a)



(b)

Figure 11: (Spectral modulation to change the timbre corresponding to (a) object1 and (b) object2.

5 More results

The waveforms can be heard in the accompanying power point presentation. More results for different images can be seen/heard in the presentation.

6 Some ideas for future work

- Currently I am extracting the visual percepts from the image. A more natural way would be to use all the available response in the symmetry response. We could define a degree of harmonicity and assign it 1 only where there is a local maxima and decrease it as we move away from it. In this way the percepts should stand out by themselves rather than we extracting them explicitly.
- Is there any work on how many harmonic complexes can a human listener distinguish if they are presented at once ? Does it depend on the spacing? When looking at the image we can easily count the number of objects. However in the auditory domain I just knew there are a lot of harmonic complexes. I was not sure about the number of auditory sources. Probably I should do something about the spacing of the pitch and the range allowed.
- If a harmonic complex is crowded by some harmonic complexes close to its fundamental does it have any effect on the perception of the original harmonic complex.
- Incorporate more pleasant sounds rather than just pure tones. We could use musical notes or a speech like production mechanism with the vocal tract response depending on the shape of the object.
- Probably we can try to model the evolution of language. Given a vocal tract like system and different images can the system learn sounds corresponding to different images.
- What perceptual attribute can we map color to?
- Try out in more natural images.
- Make the object extraction more robust.

7 Conclusions

We presented a perceptually inspired mapping to convert a simple two dimensional image consisting of simple geometrical shapes to a one dimensional audio waveform consisting of simple harmonic complexes.