

CAPeD: Calm Audio-controlled Personalized Display

(A Ubiquitous Computing project by Ryder Ziola and Sidharth Nabar)

The vision of calm computing includes computing devices that exist in the periphery of our attention and almost blend into the environment. When attention is directed towards these applications, they change their behavior and offer additional functionality. In this project, we follow this paradigm and to design and implement a personalized, audio-controlled data display. This display can be used by any group of individuals who use a common space.

Ordinarily, the screen is blank, or displays a pretty picture and blends into the background in an aesthetically pleasing way. When a user walks up to the screen and looks directly at it, the screen comes to life and becomes a responsive data display. Based on the user's voice commands, it can display data such as the Current News, Weather, the user's Calendar, etc. Moreover, this content is customized for each user based on his/her preferences.

Basic System Description:

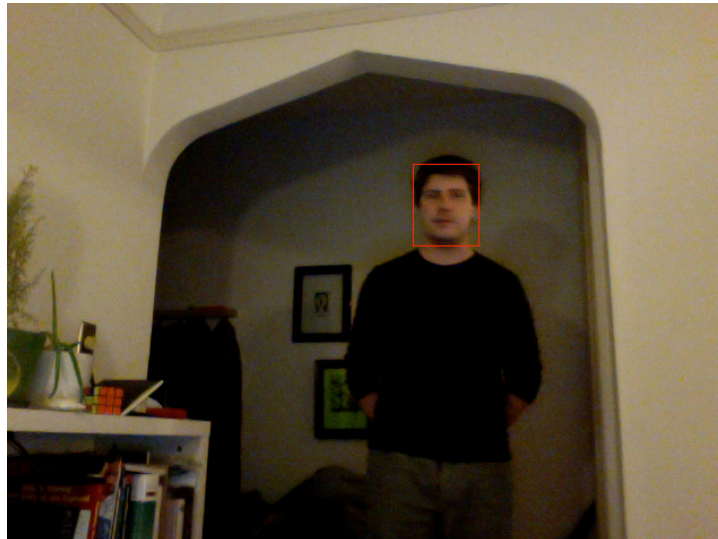
- a) **The screen:** This can be any digital frame, digital monitor, projector or screen that is connected to a computer. A webcam and a microphone are used to capture a video feed and the user's voice commands, respectively.
- b) **Computer vision:** The webcam continuously captures the video feed and our computer vision algorithm looks for faces inside the frames. If a face is found, the screen switches to the active mode. Moreover, it identifies the user based on the face and displays the welcome screen corresponding to that user.
- c) **Voice commands:** The voice commands of the user are captured using the microphone, and then passed into a signal processing block. This block converts the captured sound into a normalized spectrogram and then compares it to the previously stored training data. The final output is one out of the allowed set of commands, which is fed back to the screen. The screen then switches to the appropriate content.
- d) **User-specific content:** User preferences are pre-coded into the system. These consist of preferred sources and types of news feeds, cities of interest for weather, etc.

In the **current version**, our project uses a computer/laptop screen or a standard digital monitor as the display. Computer vision algorithms are trained to extract and identify faces corresponding to 2 users (Sid and Ryder) but can easily be extended to more users. The set of

audio commands includes “Weather”, “News” and “Calendar” which display current weather, news feeds and schedule corresponding to the user. Again, this set can easily be extended to more commands.

Implementation Details:

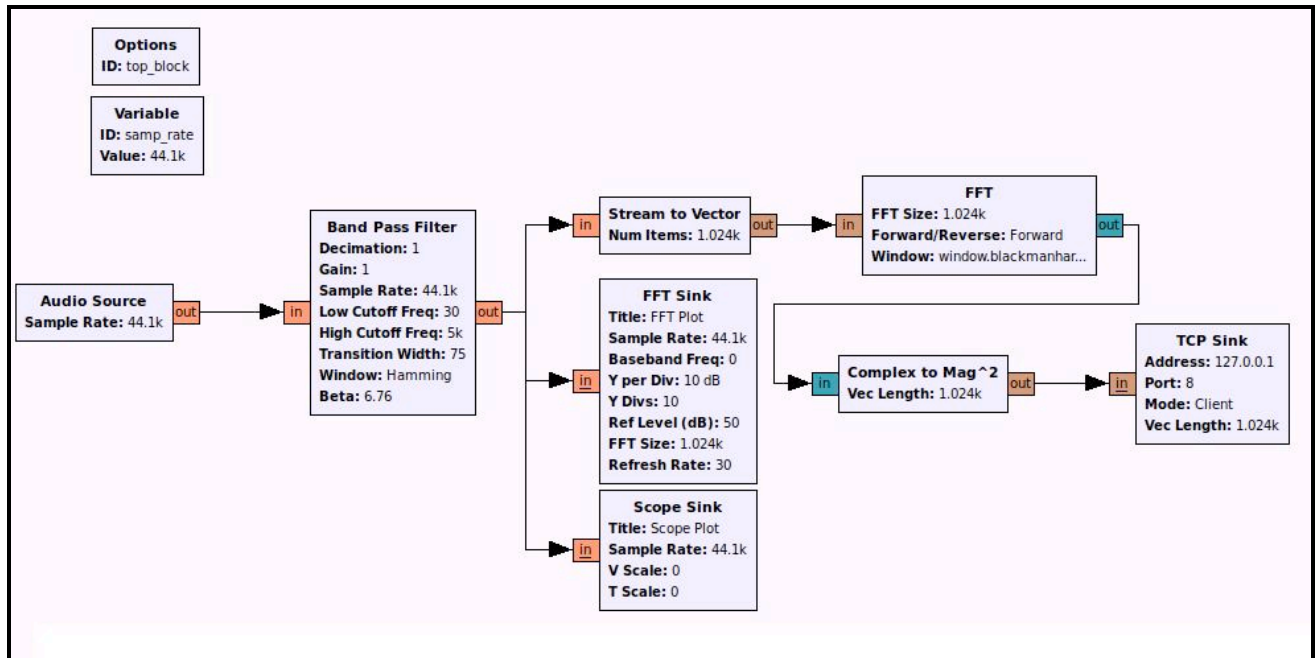
1. Computer Vision:



The vision system was implemented in C using OpenCV. First, the regions of the image containing faces are isolated using Haar-like features. The incoming 640 x 480 RGB image is converted to 320 x 240 greyscale and fed to a Haar cascade classifier, using the face data provided with OpenCV. In practice, faces could be isolated reliably up to a distance of 10 feet. The isolated faces were subsequently identified, described below under the Machine Learning section.

2. GNURadio:

The GNURadio flow graph used to capture the audio is shown below:



This block also transforms the captured audio into a 1024-point FFT and streams these floats over a TCP connection to a server. The server runs a C code that normalizes these FFTs based on the total power and concatenates them to form a spectrogram of the utterance. FFTs with total spectral power below a certain threshold are dropped in order to provide robustness to ambient noise. The drop in audio intensity is used to mark the end of the utterance.

3. Machine Learning:

Machine Learning algorithms are used in the computer vision as well as the audio commands sections.

Face Identification

As described above, regions containing faces are isolated using the Haar classifier. Presented with these regions, machine learning algorithms are used to ascertain the identity of the face with the eigenfaces technique. This consists of three phases:

1. Vector-space construction

Using a subset of the 'ORL Database of Faces' from AT&T Cambridge, eigenvectors are extracted for the classification of live instances.

2. Training

At this point, we start running the system and generating images of our faces. The images are scaled to a constant size and projected into the eigenspace. We attempt

to classify the resulting vector as either Sid or Ryder and, if the classification fails, the vector is stored in the database for future reference.

3. Classification

At runtime, the vector representing an unknown face is classified with a nearest-neighbour algorithm using a simple Euclidean distance. The identity of the closest example is assigned.

This approach is fairly sensitive to lighting and pose, requiring a training example near to every possible combination. In practice, we found that we needed about 15 examples per person in a constantly-lit environment. We were planning to increase a weighted k-nearest-neighbour approach, but found the accuracy sufficient as is. This improvement would likely be necessary if the system were extended to more than two users.

Voice Recognition

For the voice-command recognition, training utterances corresponding to all the commands are collected from each user. These are converted into normalized spectrograms and stored in the system. When the test spectrogram is computed, it is compared with all the stored spectrograms and a K-Nearest Neighbor algorithm is used to identify the command given by the user. The recognition algorithm only identifies the uttered command and not the user. This was done since identity of the user was already available through computer vision, and was more robust than that obtained through speech recognition.

Results:

The developed system was tested with a digital monitor and the performance of the video as well as the audio components were found to be very good (video uploaded on Youtube). However, some lighting conditions were found to affect the training of the face detection. The system was also demonstrated using a projector and it performed with almost 100% accuracy.