Motivating Use Case

Video conference audio may not reflect the position of the audio source relative to the camera. If one can locate the speaker in the image, 3d location information can be encoded into the audio, providing a richer experience.

Test Scenarios

Vary 3 parameters:

- moving speaker vs static speaker
- dynamic background vs static background
- 1 or 2 speakers

For each technique described in the milestones, analyze and compare the results based on the scenario.

Milestone 1: Sound Source Localization using Visual Cues (SSLVC)

Detecting speech frames using energy in spectogram

Given:

- Speech is likely to occur at certain frequency range, R
- Assume video and audio are synced
  - Time-warping could done to assure this

Then the speech corresponds to activation in the spectrogram in the range R.

Finding the corresponding frame is then a matter of mapping the audio sample to the associated frame. Because the frame rate is much slower than the audio sample rate, some simple binning will be required.

Use NMF to find locations in frame associated with sound sources

Relevant Work: Nonnegative matrix factorization for unsupervised audiovisual document structuring

Compare with simple gradient filters

- Movement likely corresponds to audio. Thus, the graident of the image likely serves as a satisfactory feature when the background and is static.

Milestone 1.5: Removing Brightness Fluctuations

Proposed Solutions include:

- Histogram Equalizing
  - Can provide probabilities that can be incorporated in Milestone 2
- Simple threshold that gradient must pass before being considered a change
- Changing color spaces and then projecting away the brightness information
  - HSV
  - YUV

Milestone 2: Offline Model biased SSLVC

- Switch to PLCA and other Probabilistic Factorizations.
- We expect the video to contain speakers looking at the camera. Thus we can develop offline models of faces and human voices.
- We can incorporate the models as priors in the factorization.

Milestone 3: SSLVC biased by previous frames

- If a sound was previously emitted from a location, then it is likely that the same or nearby location emits a sound in the next frame.

Milestone 4: Camera Positions Based on Sound Sources

- Milestones 1-3 located the position of the sound source in the frame. Given estimates for how large a human face is, one can approximate the distance and angle to the camera.
- This enables the emulation of 3-d audio relative to the camera location.
- May not be reliably done for audio since amplitudes are much more variable
  - Works well with faces and because they have roughly static dimensions

Milestone 5: (Super Stretch) Estimate Microphone Position

- Audio amplitudes may not be a reliable measure of distance.
- The microphone and camera positions are likely not overlapping. Thus given work done with The Visual Microphone: Passive Recovery of Sound from Video, one can use traditional microphone array techniques to locate the microphone.