Automatic Calibration of Microphone Arrays and Video Cameras

John McDonough

Language Technologies Institute
Carnegie Mellon University

Contributors: Kenichi Kumatani, Bhiksha Raj

September 9, 2011
Audio and video data streams are both useful for speaker tracking.

- Video can track a (potential) speaker who isn’t presently speaking.
- Audio let’s you know which face is talking.

In this project, we will perform acoustic and visual speaker tracking.

We will also simultaneously learn the relative orientations of the video cameras and microphone arrays.
Acoustic Speaker Tracking

Figure: Positions of the microphones \( \{ m_{mn} \} \) and speaker \( x \).
Central Projection

a) Perspective view  
b) Planar view

Figure: Central projection.
Figure: Central projection: image plane projection.
Coordinate Transformation

Figure: Transformation from global to camera coordinates.
Figure: Formation of the video innovation vector $\mathbf{s}_k = \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}$. 
Constrained Estimation

\[ \hat{x}_{k|k} = (I - P_k C_k) \hat{x}_{k|k-1} \]

\[ \hat{\theta}_k = C_k x = b \]

\[ \hat{x}_{k|k} = C_k (C_k C_k^T)^{-1} b_k \]

Figure: Constrained estimation as a projection operation, where

\[ P_k = K_{k|k-1} C_k^T \left( C_k K_{k|k-1} C_k^T \right)^{-1} \]