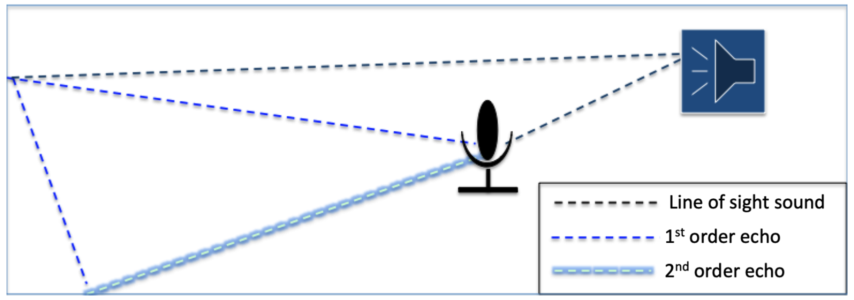
Acoustic Localization and Tracking via Machine Learning

Justin Joco and **Yanling Wu** (Advisors: Prof. Christoph Studer and Emre Gönültaş (Ph.D. Student))

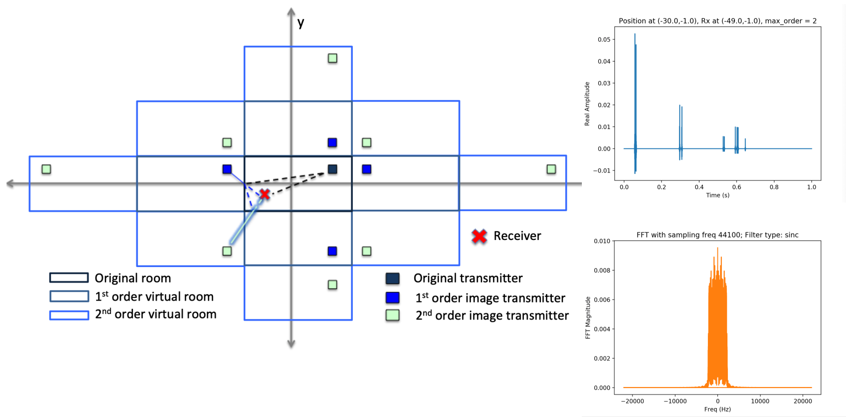
School of Electrical and Computer Engineering, Cornell University

Can we make a system that tracks users in a room using sound?



- Localizing users via sound is a novel approach that uses microphones and deep learning to track users inside a MIMO room channel
- Need to model a room with echoes to optimize deep learning models

Data generation: Modeling a room with echoes

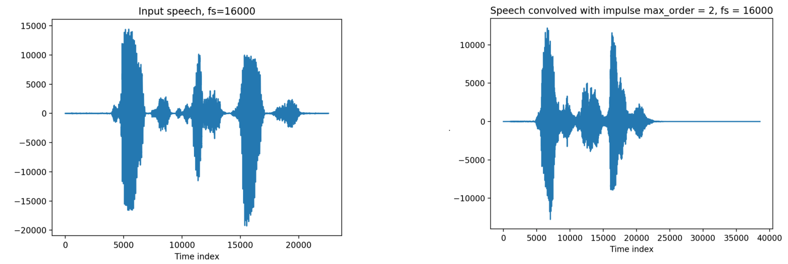


- Create impulse response (IR) from one user to one microphone
  - Specify room with adjustable sampling frequency, room size, and user and microphone locations, max echo order
  - Recursively build virtual rooms around original room (image-source)
  - Generate echoes by creating one imaginary user per virtual room
  - Pulse shape with sinc function

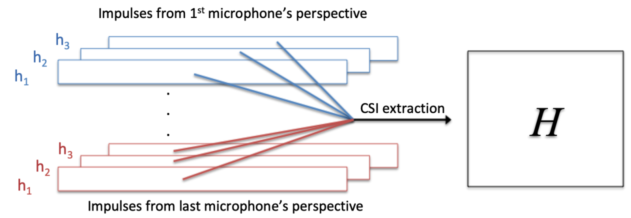
Sscaling: Populate room with multiple users and microphones

- Added a linear array of 16 microphones and increased number of users to 1600 into a 20m x 20m room for preprocessing

Verify validity of generated impulse and create channel state information (CSI) matrix H for machine learning

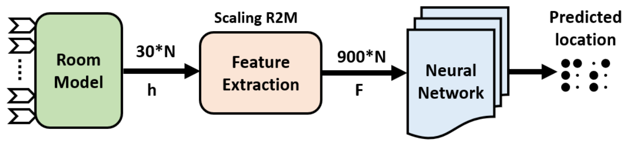


- Convolved normalized IR with speech file to successfully simulate a speaker at a set distance away from listener in room



- Filled H with the value of a specific sub-carrier of each IR
- Analyze H with deep learning models to learn data embeddings

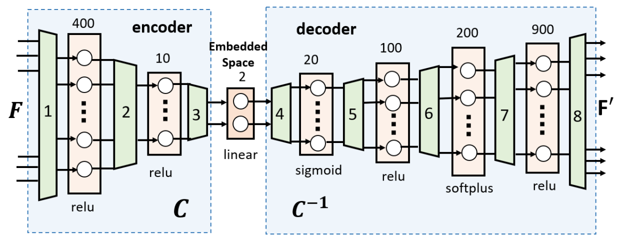
Supervised model: learn the user locations in cartesian space



- **Feature extraction**: Normalize impulse response, h $\rightarrow$ scaling the raw 2nd moment, $\tilde{H} \rightarrow F = D\tilde{H}D^H$ (D is discrete Fourier transform matrix)
- Parameters:
  - The number of dense layers: 4
  - Activation function: relu, linear
  - Loss function: Huber loss
  - Optimizer: Adam(lr = 0.001)

Supervised model needs to be retrained if room geometry changes

Unsupervised model: learn the channel embedding distribution

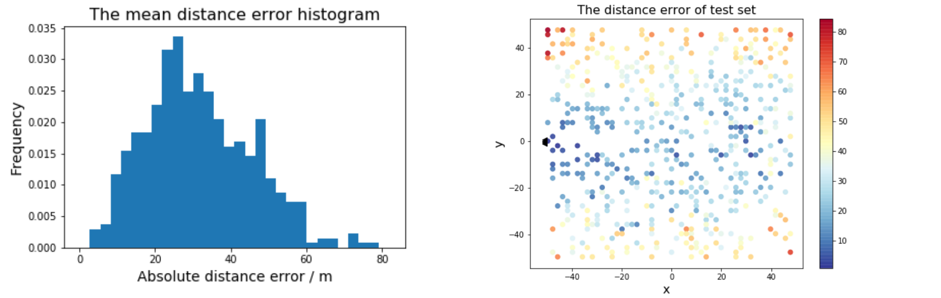


- **encoder**: output the feature embedding
- **decoder**: estimates features F'
- **Goal**: minimize the $error(F, F')$ to get better feature mapping

Unsupervised model learns the location features *only* using CSI

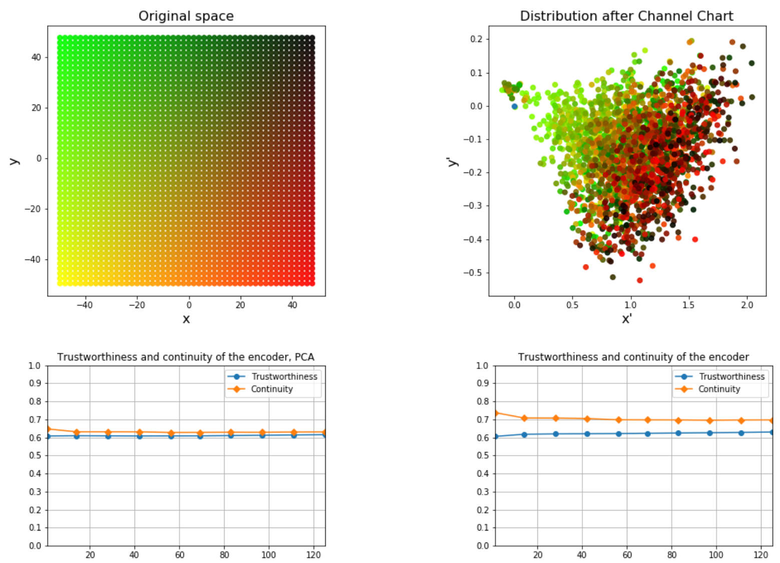
Supervised results: Localization error is least when users nearby the cartesian axis as the microphones'

- The key parameters of data (same for both deep learning models):
- Max reflection order: 2
- Channel SNR: 10dB
- # of Rx: 30
- Room size: 100x100m
- Sampling rates: 44.1kHz
- # of Users: 2500



- The mean of absolute distance error is 29m out of 100 m

Unsupervised results: the embeddings distributions



- Loss function: Huber; Optimizer: RMSprop
- The points with same color means they are neighbors in original space
- **Continuity**: Are neighbors in the original space preserved in the embedding space?
- **Trustworthiness**: How well do the features avoid introducing the false relationships in embedding space?

Autoencoder learning of modelled room has higher continuity than PCA, though both have similar trustworthiness

Combine consistent room data and a robust deep learning model to make a sound tracking system

- Generating room data required image-source techniques for scalability, using supervised or unsupervised learning on this data is application-specific, and determining optimal parameters required much trial and error.