Imperial College London

Robot audition in rooms
- The challenge of reverberant speech

Alastair H Moore, Christine Evers and Patrick A Naylor

OVERVIEW

Aim of robot audition

- Use sound to facilitate compelling human-robot interactions in real world environments

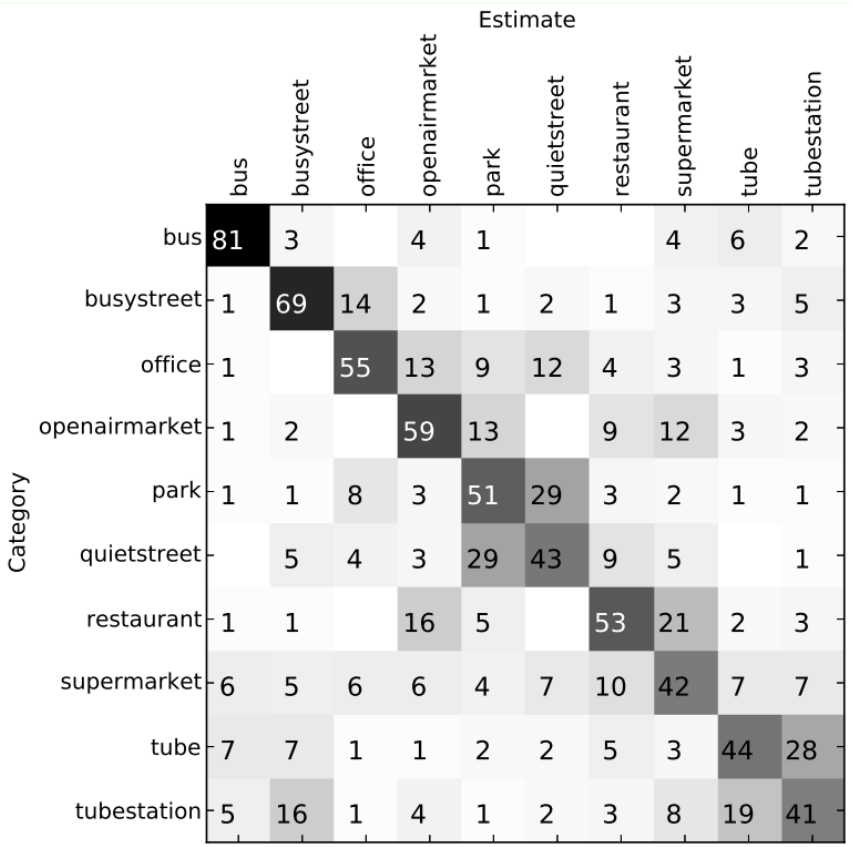
Key enabling capabilities

- environment awareness
- robust ASR
- *sophisticated dialogue system for natural HRI*

ENVIRONMENT AWARENESS

General - “Where am I?"

- D-CASE challenge [Giannoulis2013]
- Accuracy ~80%
- Confusions between semantically similar spaces
  - park vs quiet street
  - tube vs tube station
- Visual disambiguation is presumably quite simple



Specific - “What is in my vicinity?"

- sound source localization
  - real source vs image source
  - difficult because reflections are correlated

- sound source classification
  - speech vs non/speech

Attention – “Who should I listen to?”

- Can a robot attend selectively to just one source and choose which source correctly?
  - Nearest?
  - Loudest sound?
  - Keyword spotting?
  - Attention grabbing movements (vision)?

- Can a robot attend to more than one source simultaneously?(superhuman capabilities)

ROBUST ASR – "WHAT ARE THEY SAYING?"

Current state of the art

- ASR can work well
  - with close-talking microphone
  - well matched training and test conditions
- Distant talking ASR is more challenging
  - Likely need “front-end” processing prior to ASR
- Microphone array processing is now commonly applied in telecommunications to address
  - acoustic feedback path (mature)
  - background noise (mature)
  - reverberation (nascent)

Additional challenges in robot audition

- Expect sources to move
  - “chair free” vs “hands free”
- Expect robot to move
  - robot self noise
  - additional degrees of freedom/uncertainty

Dereverberation for robots

- Aim is to improve results of ASR
  - traditional signal-based and perceptual metrics not necessarily relevant
  - ASR results are affected by many factors
    - language models
    - machine learning structure (e.g. HMM, DNN)
    - choice of training and test conditions

- Reverberation damages ASR even when training data contains reverb
  - aim to remove it
  - but first 50 ms may be helpful [Maas2012]
    - extra speech energy
    - colouration dealt with using cepstral mean subtraction

Taxonomy of dereverberation methods

- Spatial filtering
  - Steering a beam in the direction of source improves direct-to-reverberation ratio (DRR)
- Spectral subtraction (SS)
  - Estimate and subtract the power spectral density (PSD) of ‘noise’ which includes diffuse reverberant sound
- Long-term linear prediction based deconvolution
  - Requires an IIR estimate of the acoustic model
- Multichannel equalisation (MCEQ)
  - Requires an FIR estimate of the acoustic channel

Baseline dereverberation tests using simulated data

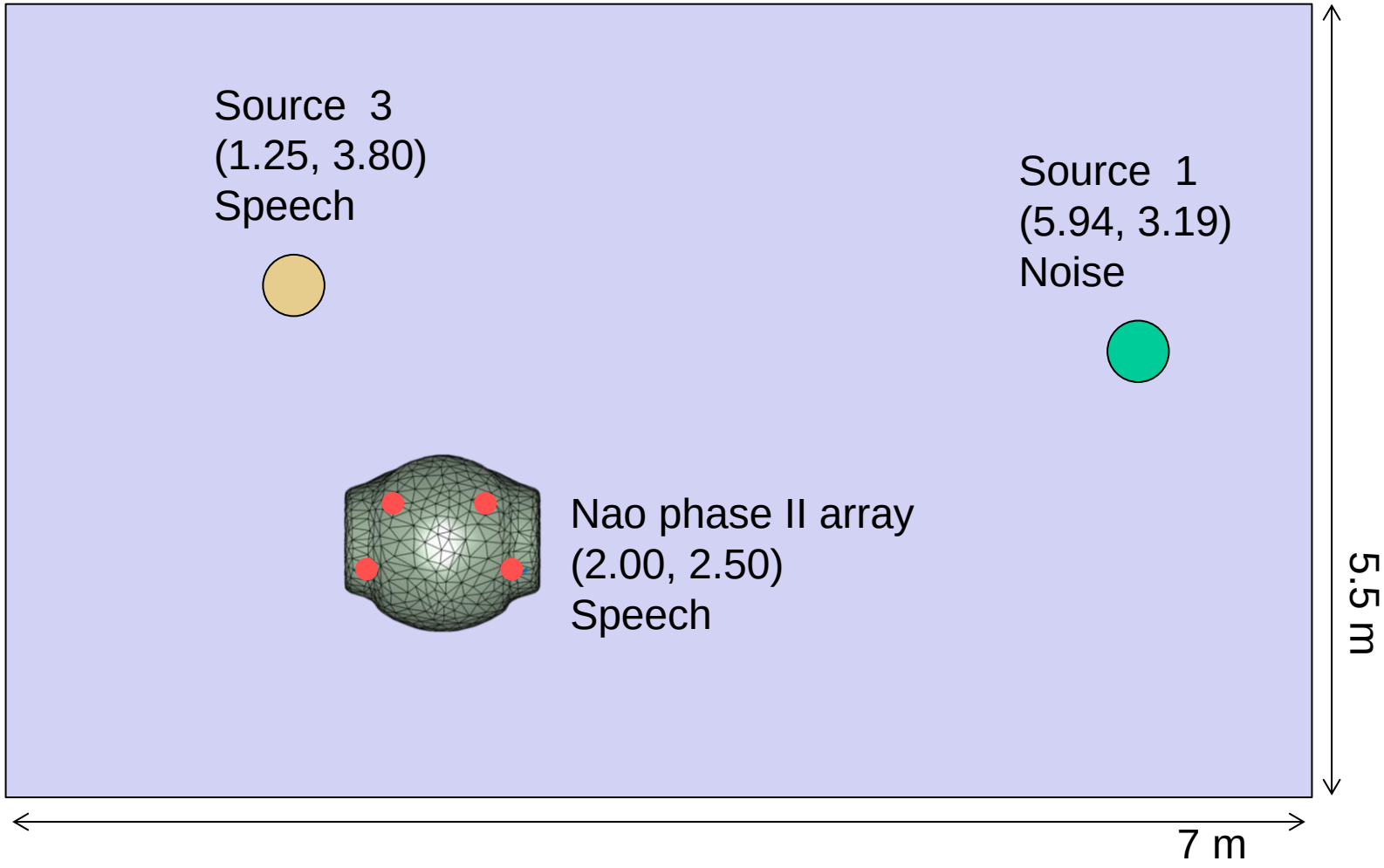
- Aim: Establish a test framework using baseline dereverberation algorithms
- Use “specific example” dataset provided by BGU as example
- Image-source room impulse responses

	125Hz	250Hz	500Hz	1kHz	2kHz	4kHz
T30 (s)	0.9040	0.7397	0.6191	0.6122	0.4703	0.3736

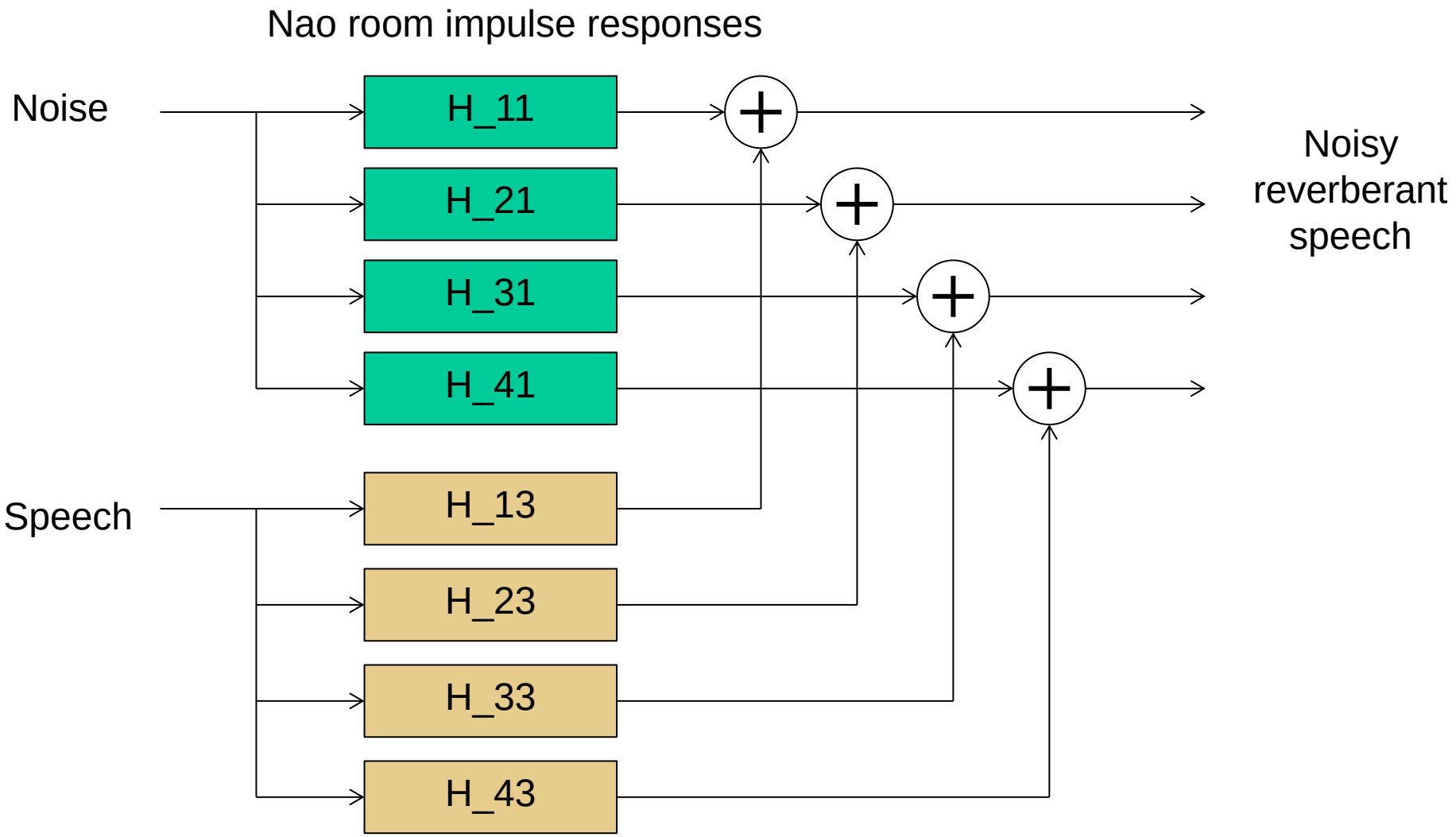
- Nao phase II head simulated HRTFs using boundary element method (BEM)
- TIMIT speech
- Specialized white Gaussian noise
  - 20 dB SNR at 1st mic

Simulated room setup

Plan view (Room height: 3.2 m. All sources/receivers: 1.5m)



Signals


Nao room impulse responses
Noise
Speech
H_11
H_21
H_31
H_41
H_13
H_23
H_33
H_43
Noisy reverberant speech

Dereverberation algorithms (blind)

- Spectral subtraction (SS) [Xiang2014]
  - single channel method
  - applied to beamformer output or channel 1 if multichannel
- Delay and sum beamformer (DSB)
  - delay estimated using GCC-PHAT (at 16 kHz before downsampling)
  - applied as phase shift in frequency domain
- MVDR beamformer
  - noise estimated using minimum statistics [Martin2001, Martin2006]
  - noise correlation matrix estimated as mean across utterance

Dereverberation algorithms (blind)

- Generalised Weighted Prediction Error (GWPE) method (linear prediction) [Yoshioka2012, Delcroix2014]
  - calculates linear prediction filter based on estimate of source power
  - iterates between estimating LP coefficients and the source power
  - operates in STFT domain
  - linear prediction ignores the first two frames to avoid over-whitening
  - implementations uses maximum of 5 iterations

Dereverberation algorithms (with known RIR)

- MINT [Miyoshi1988]
  - with perfect knowledge of RIR equalization is perfect
  - equalized impulse response (EIR) is a delayed dirac delta function

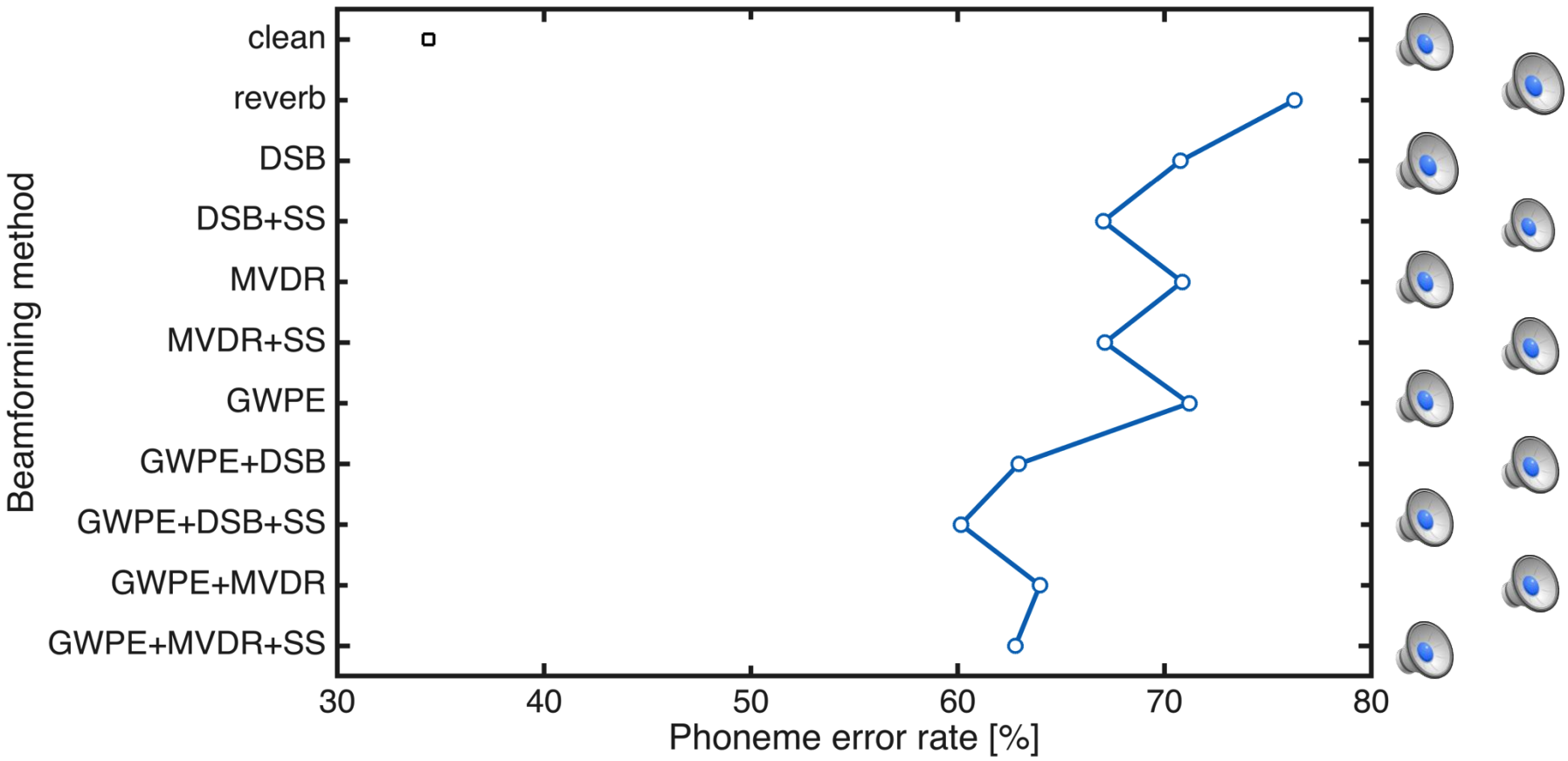
Solutions designed to be robust to system identification errors:

- Relaxed multichannel least squares (RMCLS) [Zhang2010]
  - L_w initial coefficients in EIR are unconstrained
  - these correspond to early reflections
- RMCLS with Constrained Initial Coefficients (R-CIC) [Lim2014]
  - $L_{cic} < L_w$ earliest coefficients corresponding to early reflections in EIR are constrained to be same as the RIR
- RMCLS with Envelope Constraint (R-EC) [Lim2014]
  - Coefficients after L_{cic} and before L_w are constrained to follow decaying envelope

Kaldi ASR engine

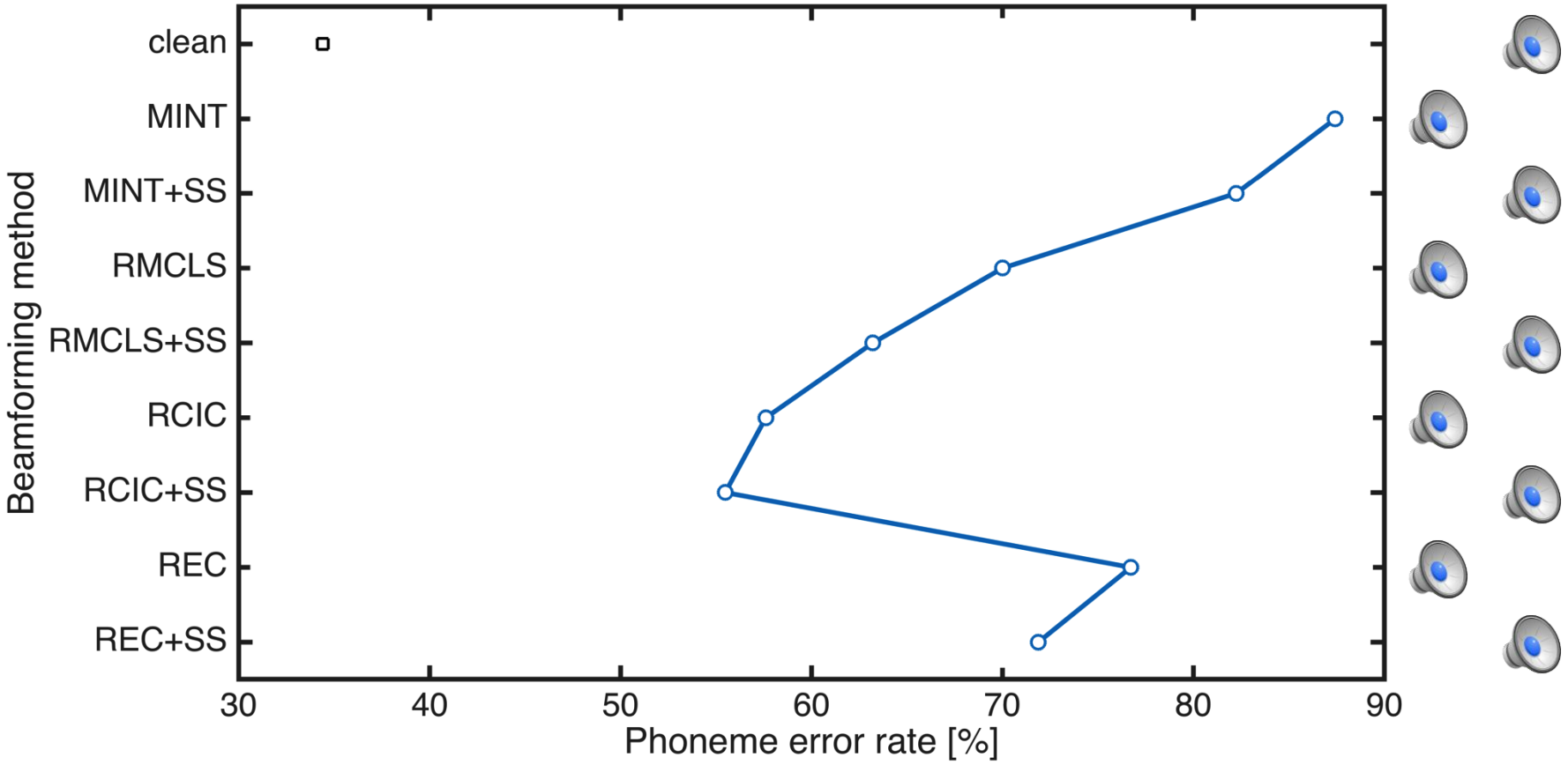
- Uses the ‘s5’ script distributed with Kaldi
  - Trained on clean TIMIT ‘train’ set
  - Mel-frequency cepstrum coefficient features
  - HMM-GMM acoustic modelling
  - WFST-based decoding
  - ‘tri1’ system type
    - recognize triphones
- Trained at 8 kHz sample rate for compatibility with limitations of acoustic signal processing
- Phoneme recognition only (no language model)

Example dereverberation ASR results



- Still room for improvement

Example dereverberation ASR results



- R-CIC looks like the most promising, but BSI is unsolved problem

Outlook

- Need to develop methods which are robust to source/robot movement

- Currently developing methods to deliver
  - Good source tracking
  - Image-source tracking

- Method switching based on estimated extent and speed of movement.
  - RIR based methods in (near-)static cases
  - Online dereverberation methods for moving cases must deal with early reflection colouration as well as late tail

Thank you for your attention