Progress report on automatic speech recognition studies

Guy Brown, Kalle Palomäki and Amy Beeston

Overview

- Summary of modelling results on AI corpus
  - Issues with May system
  - Matched training and semi-forced alignment
  - Problems with modelling confusions
  - Template-based recognition using “frozen” speech
- Speech recognition experiments using the L-shaped room
  - MFCC baseline
  - FDLP
  - Reconstructed of reverberation-corrupted regions using missing data imputation (Kalle)

Modelling listener performance in Amy’s AI corpus study

Aims

- Aim to develop a ‘perceptual constancy’ front-end for automatic speech recognition (ASR).
- Should be compatible with Tony’s findings but also validated on a ‘real world’ ASR task.
  - wider vocabulary
  - range of reverberation conditions
  - variety of speech contexts
  - naturalistic speech
  - consider phonetic confusions in reverberation in general
- Initial ASR studies using articulation index corpus
- Compare human performance (Amy experiment) and machine performance on same task

Initial work (May meeting)

- HMM-based phone recogniser
  - implemented in HTK
  - monophone models
  - adapted from scripts by Tony Robinson/Dan Ellis
- Bootstrapped by training on TIMIT then further 10-12 iterations of embedded training on AI corpus
- Good performance on 'clean' test signals, but
  - Mismatch between clean training data and the test signals, which are near (0.32m) reverberated, low-pass filtered to 4kHz and have headphone correction applied
  - High error (~40%) even in near-near condition

Matched training and semiforced alignment

- Training data for the ASR system is now matched to test conditions:
  - Low pass filtered to 4kHz
  - Reverberated with near (0.32m) impulse response
  - Headphone correction filter applied
  - Error cut by half (now ~20%)
- Semi-forced alignment is also used
  - Errors in recognition of context words had knock-on effect on recognition of test words
  - Now use semi-forced alignment in which ASR system knows the context words for each utterance and must only identify the test word

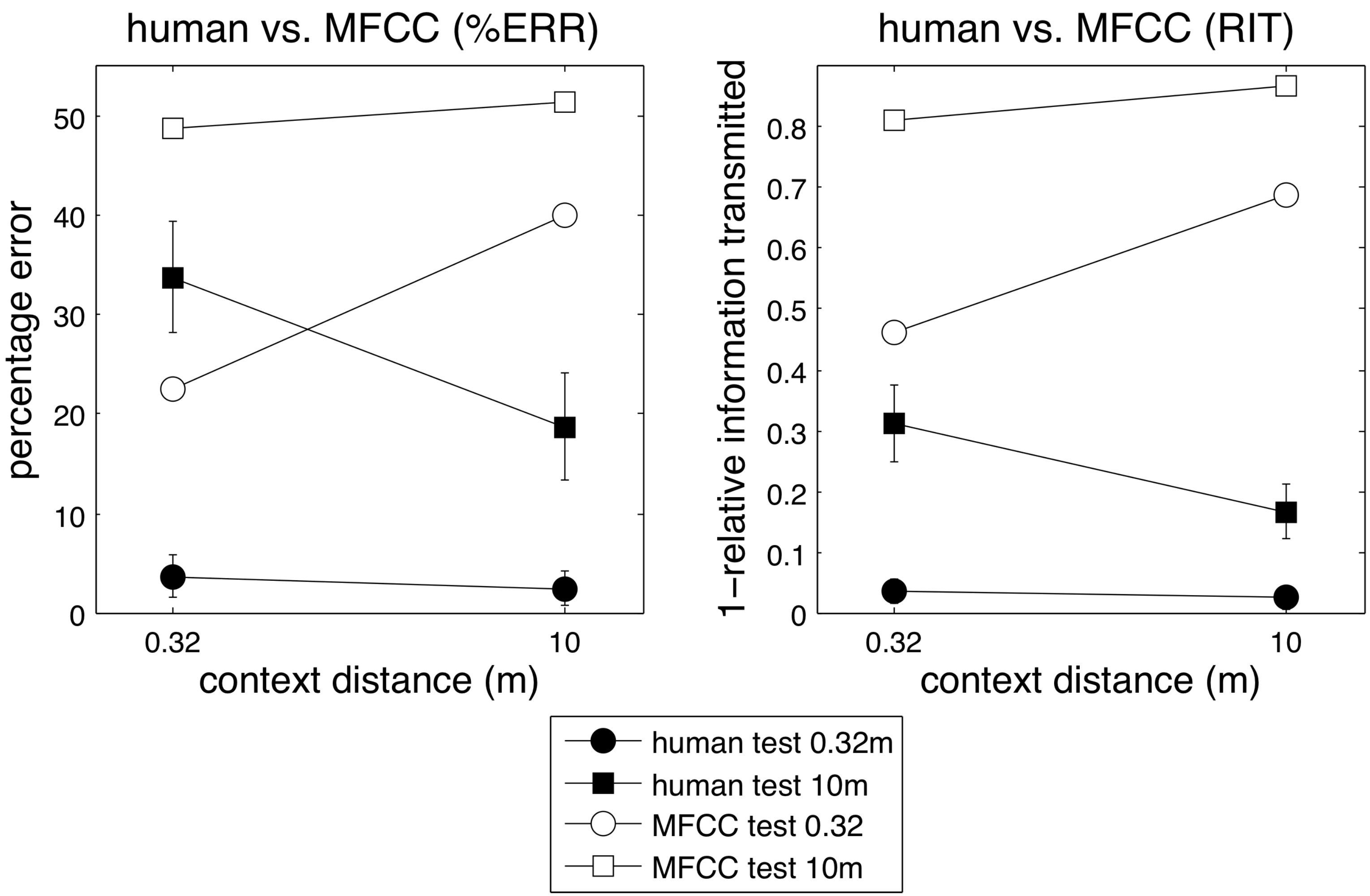
Evaluation metrics

- Model performance expressed in terms of
  - Percentage test words correct
  - 1-RIT
- Relative information transmitted (RIT) is an information-theoretic metric that reflects the distribution of errors in the confusion matrix:

$$RIT = H(X:Y)/H(X)$$

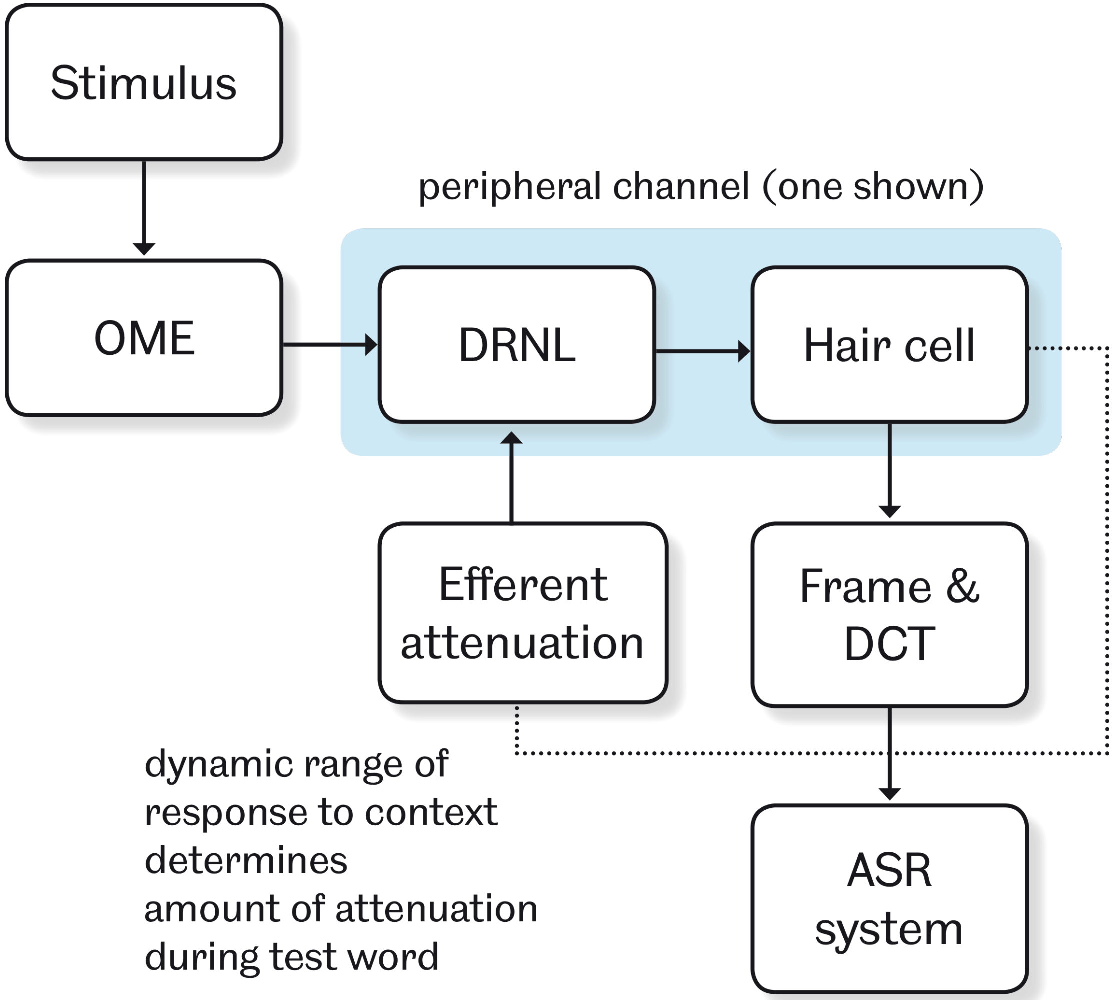
- H(X:Y) is the average mutual information of the input X and output Y, and H(X) is the average self-information (entropy) of the input

Human performance vs. baseline ASR

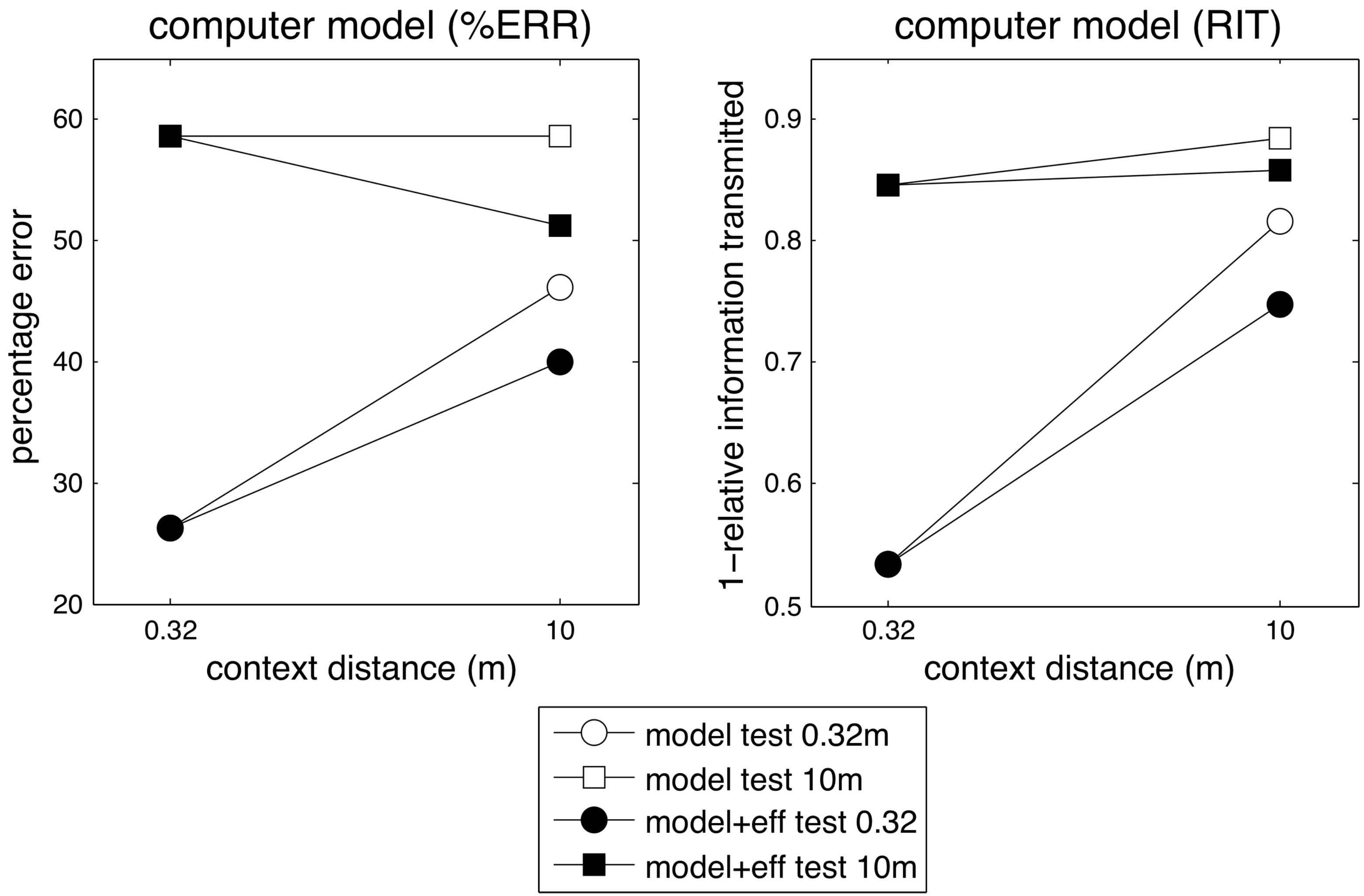


Auditory model with efferent circuit

- Simplified version of Amy's model in which efferent attenuation is manually tuned
- Full model involves a feedback loop in which efferent attenuation depends on dynamic range of AN response



Model performance in Amy's test



But ... pattern of confusions is different

	SIR	SKUR	SPUR	STIR
SIR	18	0	0	2
SKUR	3	15	0	2
SPUR	7	2	10	1
STIR	8	1	1	10

Human near-far

	SIR	SKUR	SPUR	STIR
SIR	16	1	1	2
SKUR	0	16	0	4
SPUR	2	1	14	3
STIR	1	0	0	19

Human far-far

	SIR	SKUR	SPUR	STIR
SIR	5	12	0	3
SKUR	1	12	3	4
SPUR	1	14	5	0
STIR	2	4	3	11

Model near-far

	SIR	SKUR	SPUR	STIR
SIR	11	3	2	4
SKUR	3	12	1	4
SPUR	1	10	7	2
STIR	5	5	1	9

Model far-far

Some thoughts

- For human listeners:
  - Predominant confusions are STIR->SIR, SPUR->SIR
  - a far context generally reduces confusions (particularly STIR->SIR)
- For the model:
  - Predominant confusion is SIR->SKUR
  - A far context reduces SIR->SKUR confusions but does not substantially improve identification of the consonant
- How to get a closer match to listener confusion patterns?
  - Gender-dependent or speaker-dependent models
  - Discriminative training
  - Simpler recogniser that uses “frozen speech”

Possible approach – "frozen speech"

- The Oldenburg group[1,2] have obtained a reasonable match to listener confusions by using "frozen speech" (testing on the training set) and a Euclidean distance metric.

- Quick test using our corpus:
  - Auditory spectrograms derived from 40-channel gammatone filterbank output, 10ms frame rate, cube root compression
  - Test word templates excised from all 80 of Amy's subset of the AI corpus
  - Matching using Euclidean distance

[1] Holube, I., and Kollmeier, B. (1996) J. Acoust. Soc. Am. 100, 1703–1716.

[2] T. Jürgens, T. Brand (2009) J. Acoust. Soc. Am. 126 (5), pp. 2635-2648.

Template matching with "frozen speech"

	SIR	SKUR	SPUR	STIR
SIR	20	0	0	0
SKUR	1	19	0	0
SPUR	0	0	20	0
STIR	1	0	0	19

near-near

	SIR	SKUR	SPUR	STIR
SIR	20	0	0	0
SKUR	13	4	3	0
SPUR	6	0	14	0
STIR	11	1	1	7

near-far

- Gives a better match to listener's confusions (mostly -> SIR, although confusion rate much higher than listeners)
- Need to try this template-matching approach with Amy's complete model
- Matching metric can incorporate a weight for each frequency region, can be optimised to fit confusions (using GA)

Comparison of ASR approaches on all L-shaped room conditions

Motivation

- Our eventual aim is to demonstrate a perceptual constancy front-end on a realistic ASR task
- Currently using the following task:
  - Amy's subset of the AI corpus, but scoring context words and test words (320 words in test set)
  - All distances from L-shaped room
- Implemented two baseline systems for comparison:
  - MFCC
  - FDLP
- Also work (with Kalle) on using missing data techniques to reconstruct 'unreliable' time-frequency regions from statistical models of speech

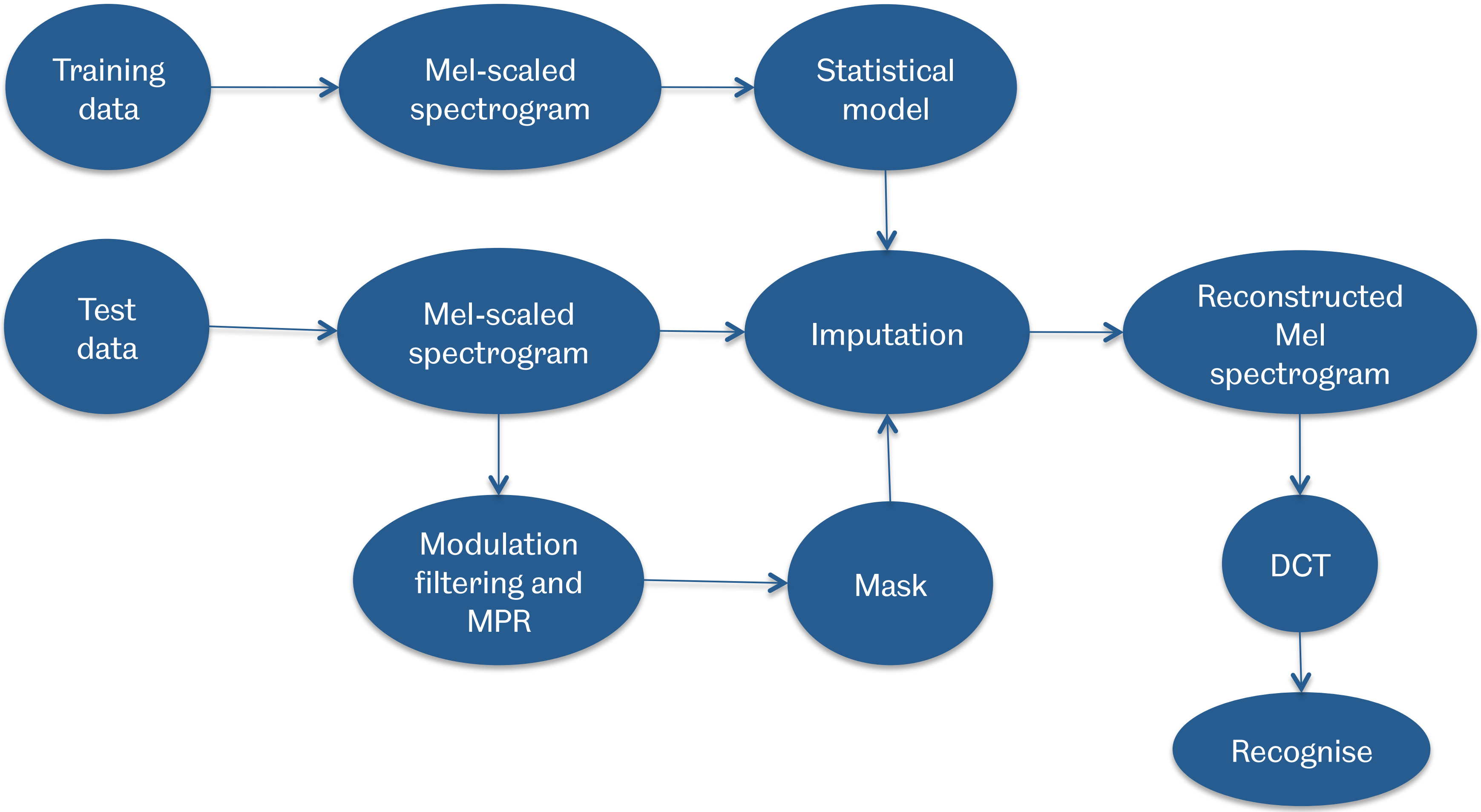
MFCC baseline

- Conventional mel-frequency cepstral coefficient front-end
  - Mel-scaled log filterbank (100Hz to 8kHz)
  - Discrete cosine transform (12 coefficients)
- First and second order temporal differences (deltas and accelerations)
- No cepstral mean subtraction (will do this shortly)

Frequency domain linear prediction (FDLP)

- Frequency domain linear prediction (FDLP) as described by Thomas, Ganapathy and Hermansky:
  - Linear prediction on a long window of DCT coefficients in order to derive an all-pole model of 96 sub-band temporal envelopes
  - Gain normalisation of the sub-band FDLP envelopes
  - Conversion to short-term cepstral features with 10ms frame rate
  - Deltas and accelerations
- Got similar results from my own code and from code kindly supplied by Sriram Ganapathy (results shown for latter).

Imputation using statistical models (Kalle)

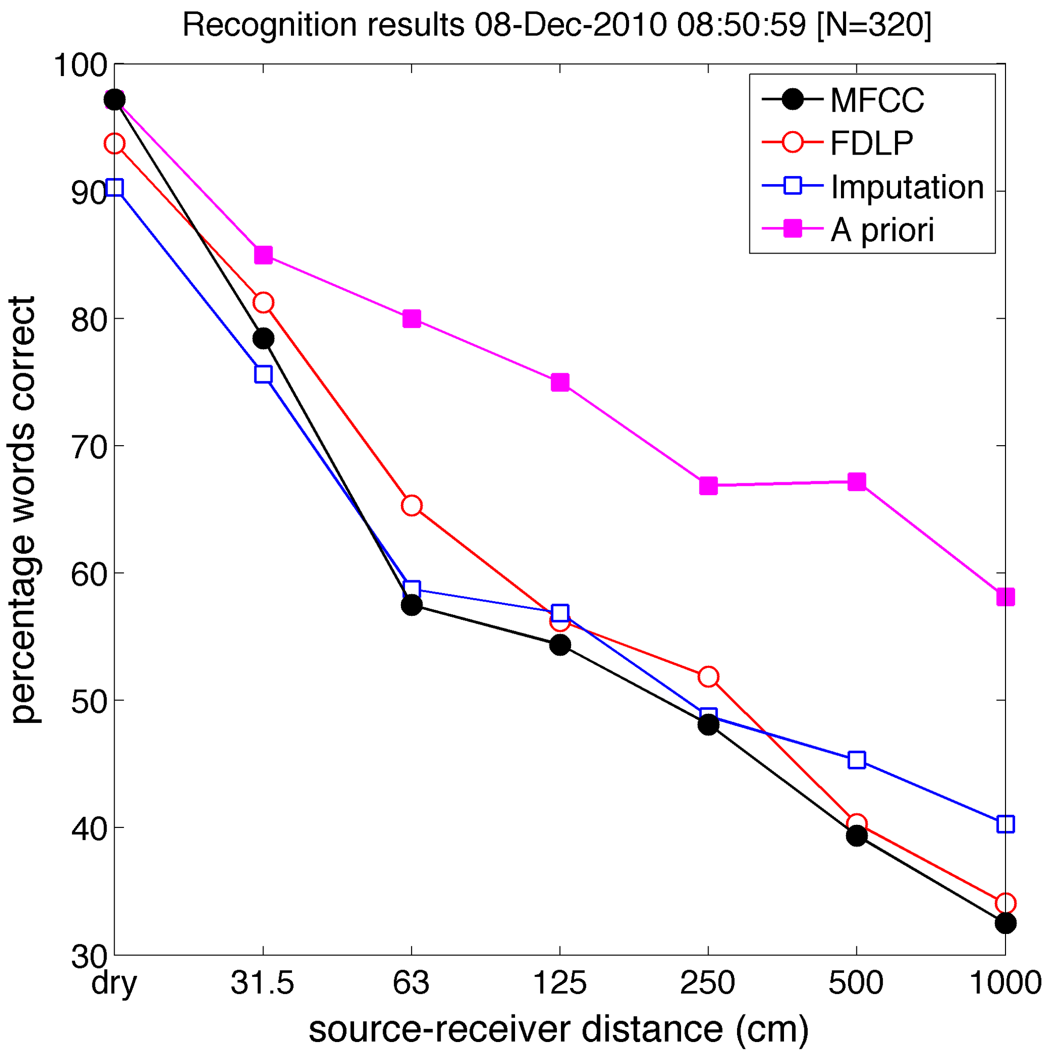


Imputation – details

- Imputation via clustering method proposed by Raj, Seltzer and Stern (Speech Communication 43, 275-296)
  - 10-component Gaussian mixture model (speech prior) trained using 2000 utterances from training set of AI corpus
  - Missing features are estimated from the statistics of the speech prior and the reliable features for each analysis frame
  - If the estimated values exceed the observed bounds, then the value is forced to the bounded value

ASR results

- FDLP better than MFCC in every condition except dry (not by much)
- Imputation should be turned off in dry
- Imputation gives largest benefit at large source-receiver distance
- Imputation with *a priori* mask shows performance limit



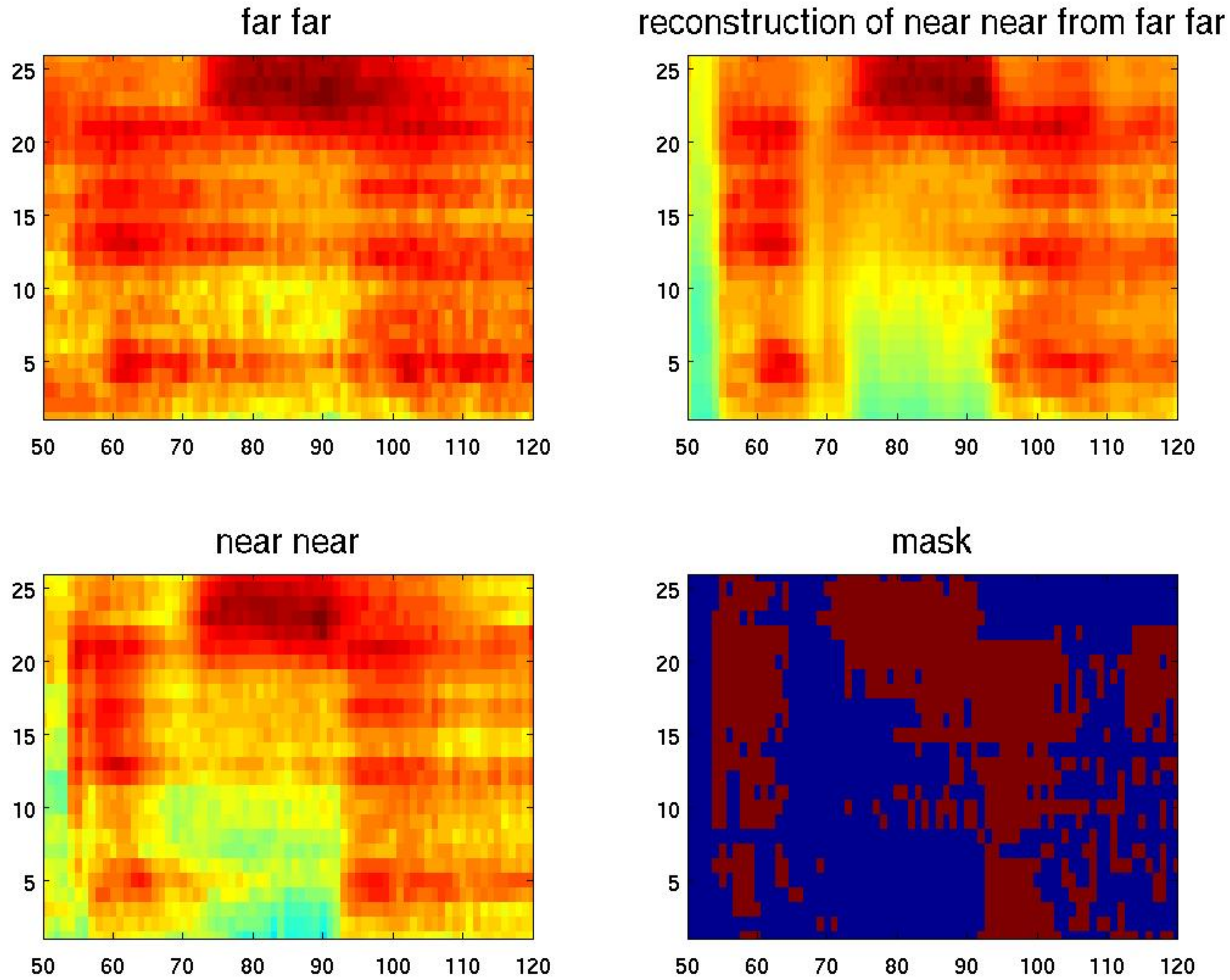
Planned work

- Focus on improving the match between listener confusions and model confusions
  - Improvements needed to recogniser architecture and matching metric
  - Complete study on modelling Amy's experimental data
- Compare within-band vs. across-band approaches to mask estimation for the imputation approach
- Incorporate more temporal context in imputation approach (currently using single frames)
- Could the imputation approach be applied to modelling Amy's experimental data?

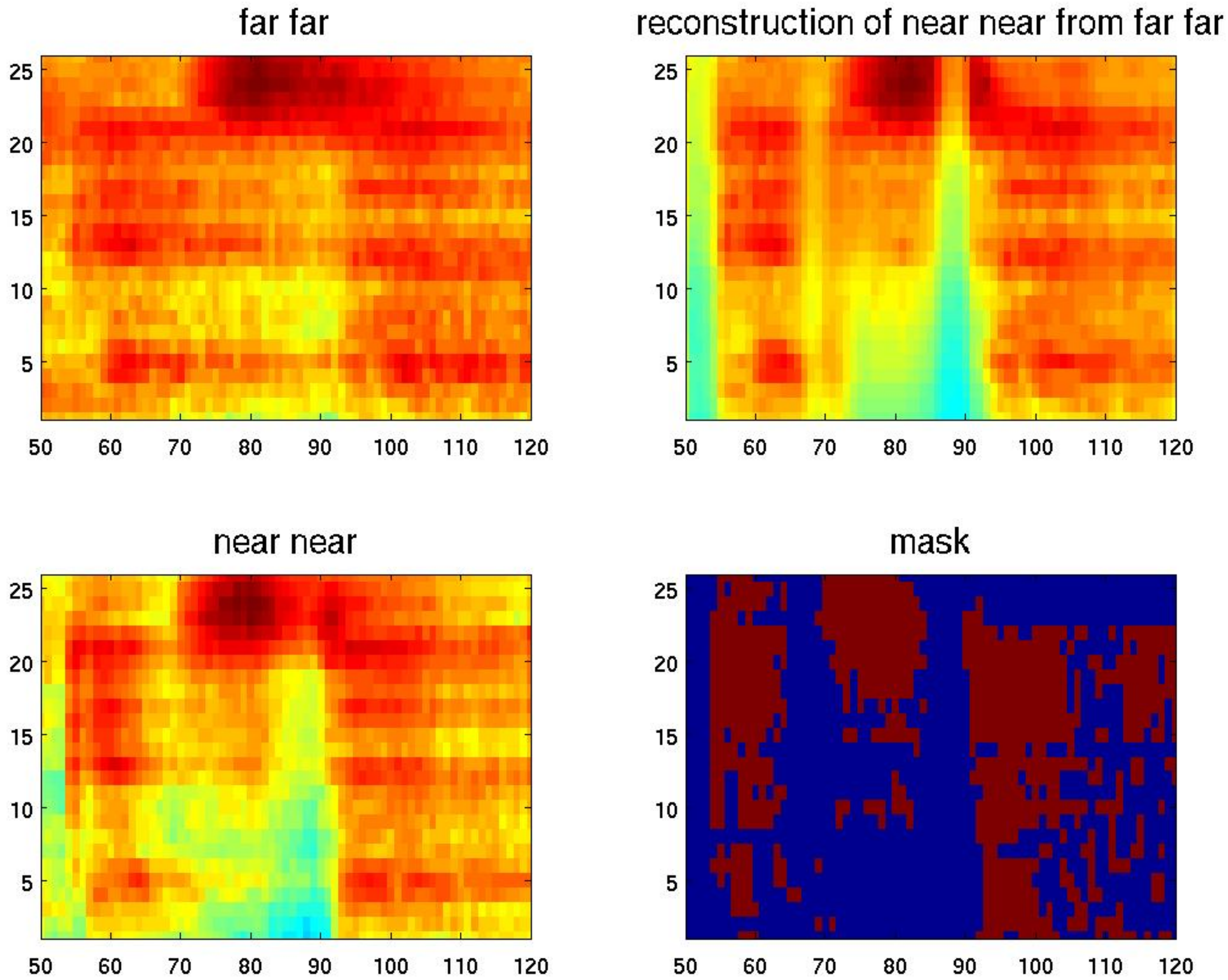
Imputation as a model of perceptual compensation (Kalle)

- Could missing data imputation be used to model Tony's sir/stir data and also the data from Amy's experiment?
- Proposed scheme:
  - Use measure of context reverberation to determine threshold for missing data mask
  - 'near' context, little evidence for reverberation tails that need to be reconstructed in the test word region -> SIR
  - 'far' context, reverberation tails in test word region are marked as unreliable in the mask and reconstructed -> STIR
- Does imputation reconstruct a 'stir' from the speech model?

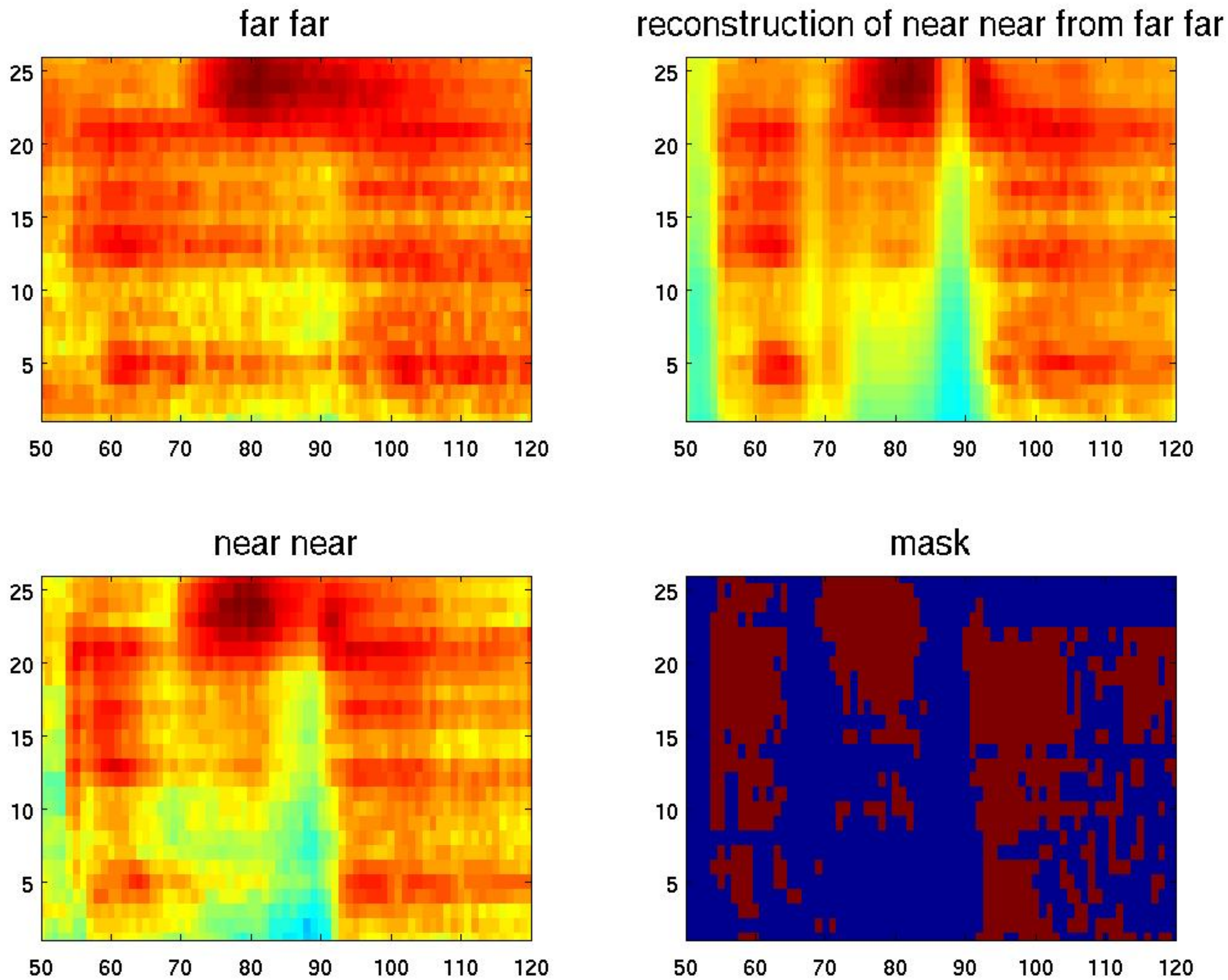
Reconstructed sir/stir step 1 (Kalle)



Reconstructed sir/stir step 8 (Kalle)



Reconstructed sir/stir step 11 (Kalle)



Comments?