A model of perceptual constancy based on acoustic feature selection

Guy Brown and Kalle Palomäki

5th July 2011

Overview

- Eventual aim is to develop a 'perceptual constancy' front-end for automatic speech recognition (ASR).
- Should be compatible with Watkins et al. findings but also validated on a 'real world' ASR task.
  - wider vocabulary
  - variety of speech contexts
  - naturalistic speech
  - consider phonetic confusions in general
- New scheme based on selection of acoustic models
  - WP4: Constancy based on statistical structure of sounds
  - WP5: Direct comparisons between human/machine

Reminder: Amy's experiment

- Amy's first experiment used 80 utterances from the Articulation Index corpus
  - 20 instances each of "sir", "skur", "spur" and "stir" test words
  - Test word embedded in 3 context words
- Overall confusion rate was controlled by lowpass filtering at 1, 1.5, 2, 3 and 4 kHz (here we consider 4kHz condition only)
- Same reverberation conditions as in Watkins et al. experiments

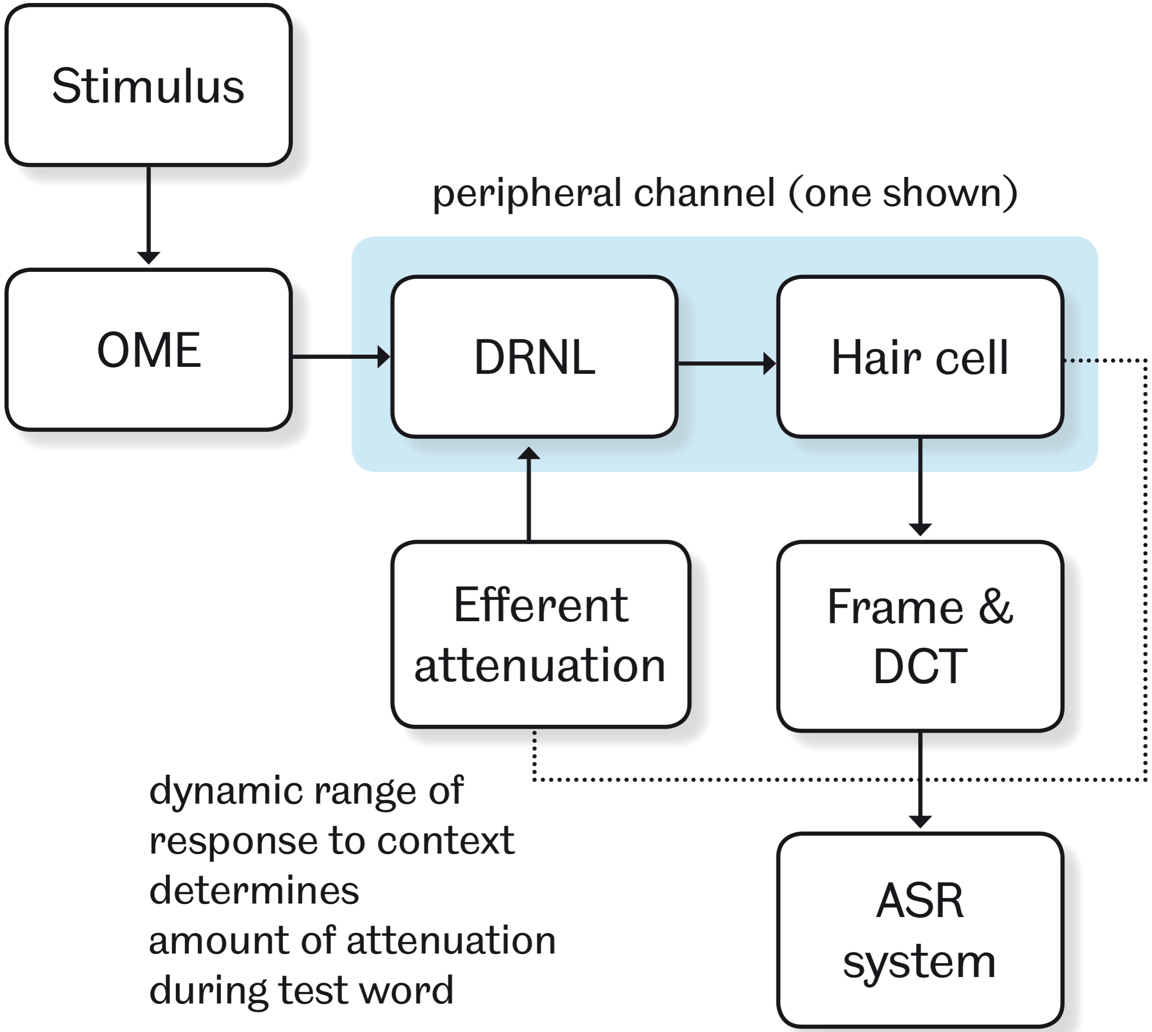
	Test 0.32m	Test 10m
Context 0.32m	near-near	near-far
Context 10m	far-near	far-far

Grammar for Amy's subset of AI corpus

```
$cw1 = YOU | I | THEY | NO-ONE | WE | ANYONE | EVERYONE | SOMEONE |
   PEOPLE;

$cw2 = SPEAK | SAY | USE | THINK | SENSE | ELICIT | WITNESS | DESCRIBE
    | SPELL | READ | STUDY | REPEAT | RECALL | REPORT | PROPOSE | EVOKE
    | UTTER | HEAR | PONDER | WATCH | SAW | REMEMBER | DETECT | SAID |
   REVIEW | PRONOUNCE | RECORD | WRITE | ATTEMPT | ECHO | CHECK |
   NOTICE | PROMPT | DETERMINE | UNDERSTAND | EXAMINE | DISTINGUISH |
   PERCEIVE | TRY | VIEW | SEE | UTILIZE | IMAGINE | NOTE | SUGGEST |
   RECOGNIZE | OBSERVE | SHOW | MONITOR | PRODUCE;

$cw3 = ONLY | STEADILY | EVENLY | ALWAYS | NINTH | FLUENTLY | PROPERLY
    | EASILY | ANYWAY | NIGHTLY | NOW | SOMETIME | DAILY | CLEARLY |
   WISELY | SURELY | FIFTH | PRECISELY | USUALLY | TODAY | MONTHLY |
   WEEKLY | MORE | TYPICALLY | NEATLY | TENTH | EIGHTH | FIRST | AGAIN
    | SIXTH | THIRD | SEVENTH | OFTEN | SECOND | HAPPILY | TWICE | WELL
    | GLADLY | YEARLY | NICELY | FOURTH | ENTIRELY | HOURLY;

$test = SIR | STIR | SPUR | SKUR;

( !ENTER $cw1 $cw2 $test $cw3 !EXIT )
```

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



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

✗ Fisher 2x4 exact test p<0.01

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?

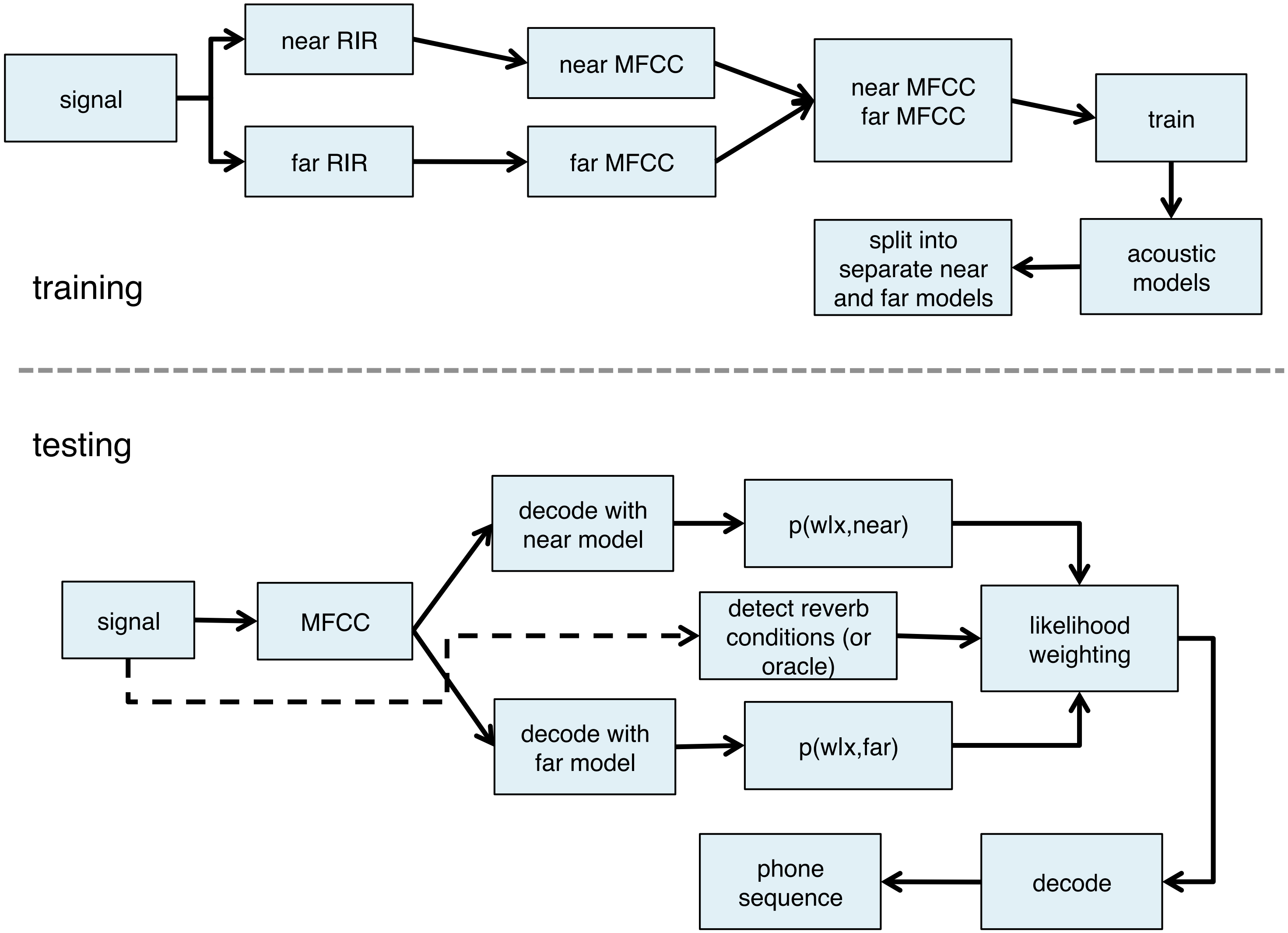
WP4: statistics of sounds in natural environments

We will develop machine hearing systems based on the idea that constancy in hearing is underlain by processes that instantiate the statistical structure of sounds encountered in natural environments.

A new approach

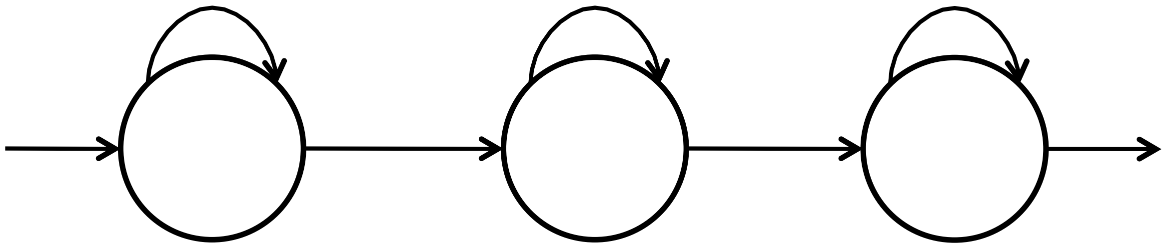
- Constancy can be modelled in terms of acoustic model selection
  - Train statistical models for speech under different reverberation conditions
  - During recognition, engage the acoustic model that is appropriate for the environment
  - Switching models cannot be done instantaneously
  - Distance swapping (e.g., near-far) leads to model mismatch
- Links with Tony's notion of a Bayesian process; can have a prior on a particular acoustic model

Schematic of the ASR system



Training

- HMM recogniser uses 40 monophone models plus a silence model
- 3 emitting states per model, no skip, straight-through

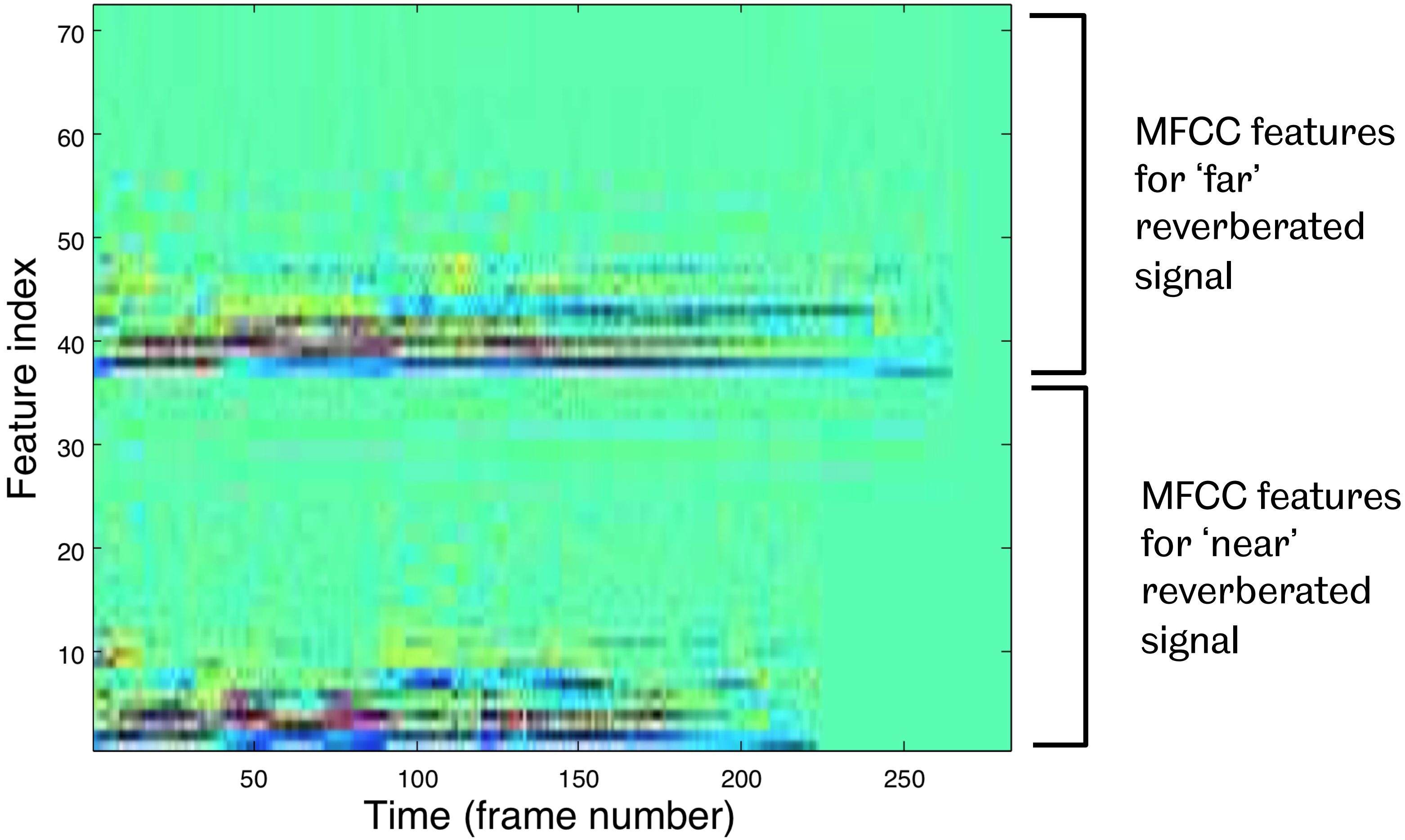


- Initial training (bootstrapping) on TIMIT corpus which has detailed phonetic transcription
- Adaptation on Amy's subset of AI corpus
  - Note: we are effectively testing on the training set
  - Necessary for near-human performance

Acoustic features and training

- 12 MFCC features + deltas + accelerations
- To avoid mismatch with Amy's test stimuli, all training utterances were:
  - lowpass filtered to 4 kHz cutoff
  - had headphone correction filter applied
- Training done by concatenating 2 x blocks of 36 features
  - one filtered with 'near' RIR
  - one filtered with 'far' RIR
- Models split after training (done this way so that both models have the same segmentation during training)

MFCC features for one training utterance



Testing

- Amy's stimuli presented to the system during testing
- MFCC features computed for the input signal and duplicated to form two feature streams
  - one set used as input to 'near' model
  - one set used as input to 'far' model
- Effectively running two recognisers in parallel, and combining the observation state likelihoods
- Used semi-forced alignment: ASR systems knows the context words and is only required to identify the test word

Combining feature streams in decoding

- During recognition, for each feature frame $x(t)$ at time t, the observation state likelihoods are computed from the HMMs for both feature streams
  - likelihood of a HMM state having generated the corresponding input feature frame $x(t)$
  - $p(x(t)|\lambda_n)$ for the 'near' acoustic model
  - $p(x(t)|\lambda_f)$ for the 'far' acoustic model
- Combined near-far observation state likelihood is a weighted sum of likelihoods in the log domain

$$\log[p(x(t)|\lambda_{n,f})] = \alpha(t) \log[p(x(t)|\lambda_n)] + (1-\alpha(t)) \log[p(x(t)|\lambda_f)]$$

Determining the weighting factor $\alpha(t)$

- The weighting factor is adjusted dynamically according to the prevailing acoustic conditions
  - Low value of $\alpha(t)$ if reverberant environment
  - High value of $\alpha(t)$ if dry environment
- Three schemes investigated here:
  - Use an ‘oracle’ value of $\alpha(t)$, assuming that context reverberation condition is know
  - Adjust $\alpha(t)$ according to the mean-to-peak ratio of the context speech envelope
  - Adjust $\alpha(t)$ according to maximum likelihood estimates from the near and far acoustic models

Evaluation metrics

- Model performance expressed in terms of
  - Percentage error in identifying test words
  - 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
- Also compare human/machine confusions

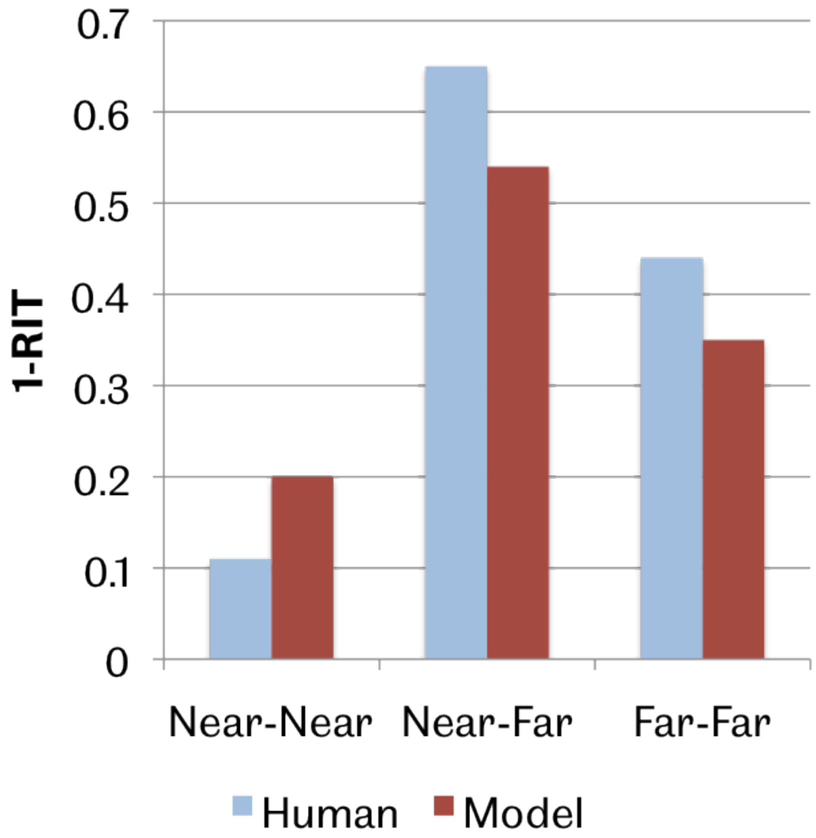
Analysis of confusions

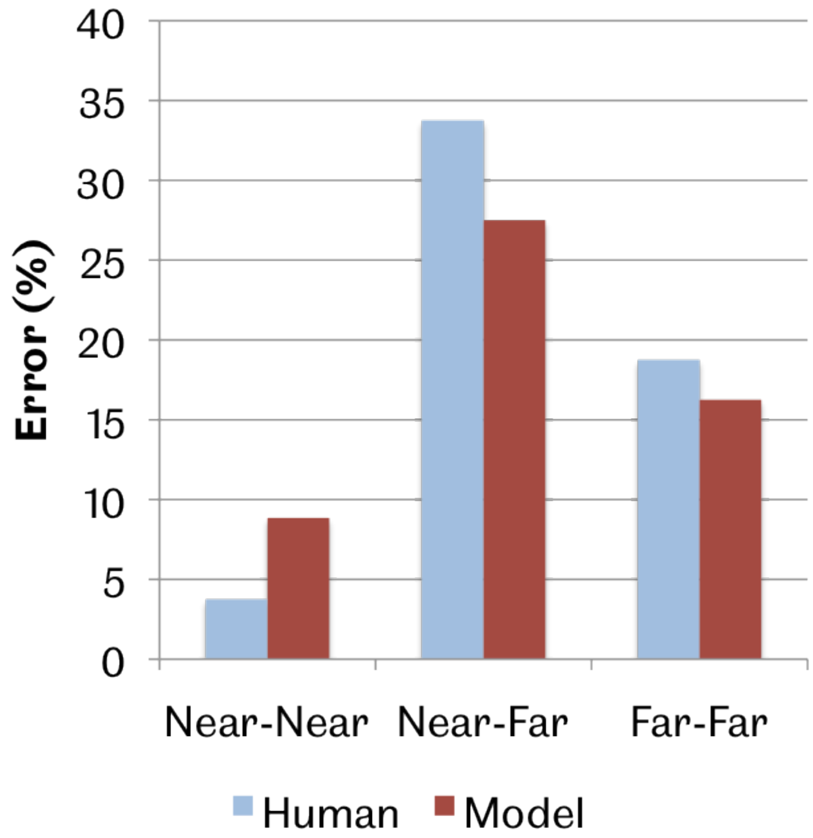
- Used two tests to determine similarity of human and model confusion matrices (applies to each row)
- Pearson's phi-squared test (normalised form of chi-squared test)
  - For identical distributions $\Phi^2=0$
  - For non-overlapping distributions, $\Phi^2=1$
  - Concerned about validity of this since sample is small
- Fisher's exact test for 2x4 contingency tables
  - Null hypothesis is that there is no difference between the human and model confusions
  - No evidence for rejecting N.H. in any condition (good!)

Oracle feature stream selection

- In this condition we adjust the weighting $\alpha(t)$ based on a priori ('oracle') knowledge of the context reverberation condition
  - 'near' set $\alpha(t) = 1$
  - 'far' set $\alpha(t) = 0$
- Gives an upper limit on model performance
  - No error in classification of the reverberation environment
- Simple idea
  - if the reverberation condition of the context and target word are different, the acoustic model is mismatched and performance will fall

Oracle feature stream selection





Confusions: oracle feature selection

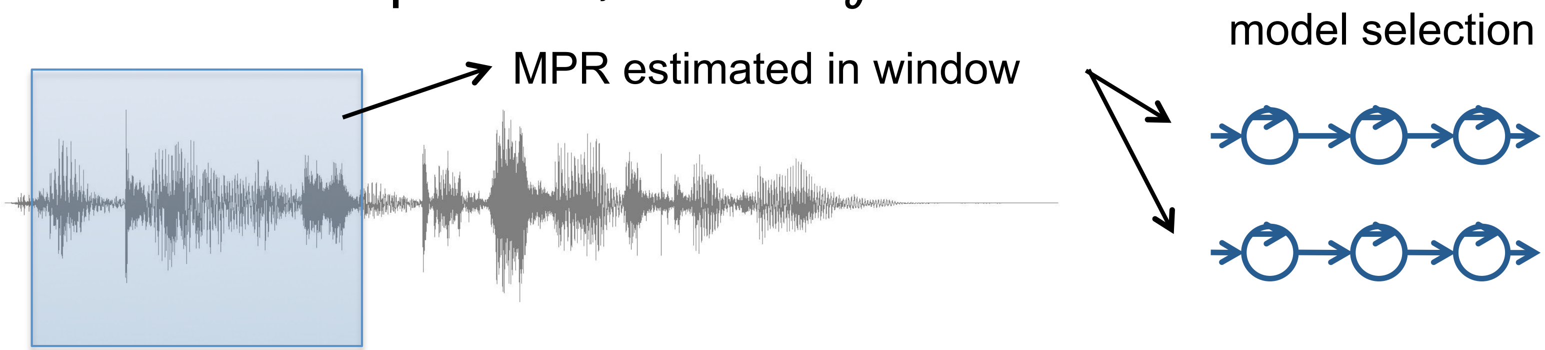
Human Near-Near					Model Near-Near						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	19	0	0	1	SIR	16	0	0	4	0.0514	0.3416
SKIR	0	20	0	0	SKIR	0	19	0	1	0.0256	1.000
SPIR	0	1	18	1	SPIR	1	0	19	0	0.0757	1.000
STIR	0	0	0	20	STIR	0	1	0	19	0.0256	1.000
Human Near-Far					Model Near-Far						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	18	0	0	2	SIR	18	1	1	0	0.1000	0.4872
SKIR	3	15	0	2	SKIR	3	17	0	0	0.0531	0.5793
SPIR	7	2	10	1	SPIR	3	1	15	1	0.0733	0.4211
STIR	8	1	1	10	STIR	9	3	0	8	0.0570	0.6001
Human Far-Far					Model Far-Far						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	16	1	1	2	SIR	11	2	2	5	0.0720	0.4773
SKIR	0	16	0	4	SKIR	1	18	0	1	0.0729	0.2617
SPIR	2	1	14	3	SPIR	2	0	18	0	0.1125	0.2623
STIR	1	0	0	19	STIR	0	0	0	20	0.0256	1.000

Interim discussion

- Overall model performance is similar to human listeners
  - Model error rate is higher than humans in the near-near condition, but lower in the other conditions
  - Similar results in terms of 1-RIT and percent error
- Pattern of confusions made by the model is plausible
  - In near-far condition, predominant confusion is STIR ➔ SIR but also SPIR ➔ SIR and SKIR ➔ SIR
  - These confusions are resolved in the far-far condition
  - Fisher test indicates no difference between the distributions of model and listener responses for all test words

Feature selection by MPR

- The 'oracle' model requires prior information about the reverberation condition of the context
- In general, must estimate the reverberation condition from the signal
- Use the mean-to-peak ratio of context envelope as a measure of reverberation present, as in Amy's model



- Gaussian classifier used to detect near/far condition
- Currently working across all frequency bands (see later)

Feature selection by MPR

- The mean-to-peak ratio of the context speech envelope is computed from the Hilbert envelope

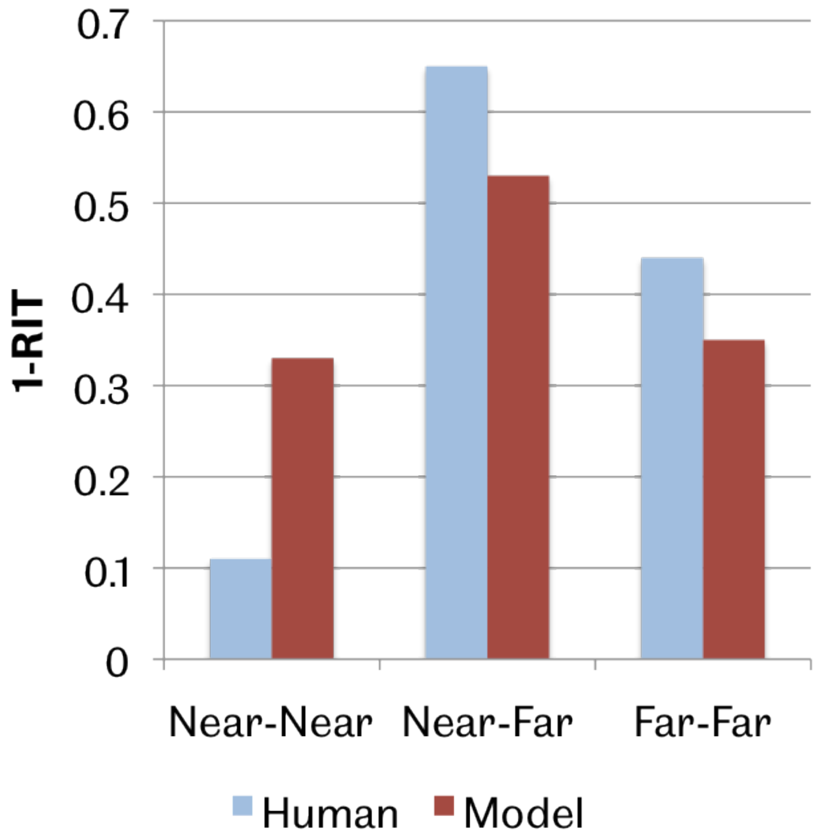
$$MPR = \frac{1}{T}\sum_{1}^{T} e(t) \Big/ \max_{t}[e(t)]$$

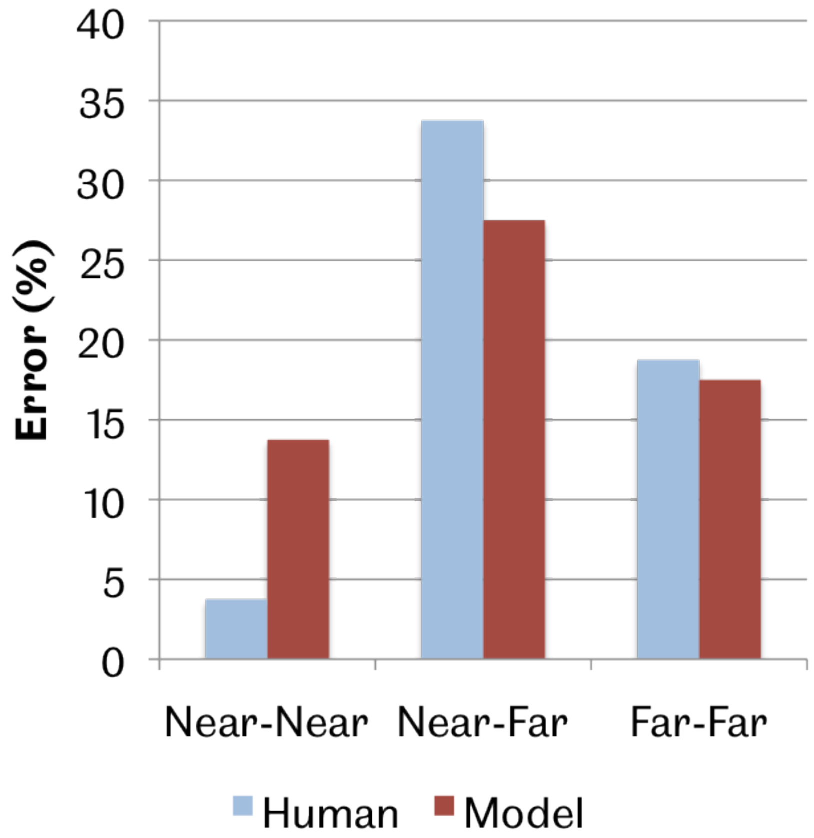
- Here T is 500 ms
- Gaussian classifier trained on MPR to distinguish between 'near' and 'far' conditions. Compute the log odds:

$$d = -\frac{1}{2} * \left[\frac{(MPR-\mu_n)^2}{\sigma_n^2} - \frac{(MPR-\mu_f)^2}{\sigma_f^2} + \log \sigma_n^2 - \log \sigma_f^2 \right]$$

- If $d \geq 0$ the context speech classified as 'near', otherwise 'far'
- 83% correct classification on test set

Feature stream selection by MPR





Confusions: feature selection by MPR

Human Near-Near					Model Near-Near						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	19	0	0	1	SIR	16	0	0	4	0.0514	0.3416
SKIR	0	20	0	0	SKIR	0	19	0	1	0.0256	1.000
SPIR	0	1	18	1	SPIR	1	0	17	2	0.0590	1.000
STIR	0	0	0	20	STIR	1	1	1	17	0.0811	0.2308
Human Near-Far					Model Near-Far						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	18	0	0	2	SIR	18	0	1	1	0.0333	1.000
SKIR	3	15	0	2	SKIR	3	17	0	0	0.0531	0.5793
SPIR	7	2	10	1	SPIR	5	1	14	0	0.0583	0.5255
STIR	8	1	1	10	STIR	8	3	0	9	0.0513	0.6947
Human Far-Far					Model Far-Far						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	16	1	1	2	SIR	14	1	2	3	0.0167	0.8650
SKIR	0	16	0	4	SKIR	2	16	0	2	0.0667	0.4152
SPIR	2	1	14	3	SPIR	3	0	16	1	0.0583	0.6483
STIR	1	0	0	19	STIR	0	0	0	20	0.0256	1.000

Interim discussion

- Fully autonomous system still shows the right overall pattern
  - Constancy effect
  - Plausible pattern of confusions
- However, note that overall error rate is higher (due to occasionaly misclassification of the context)

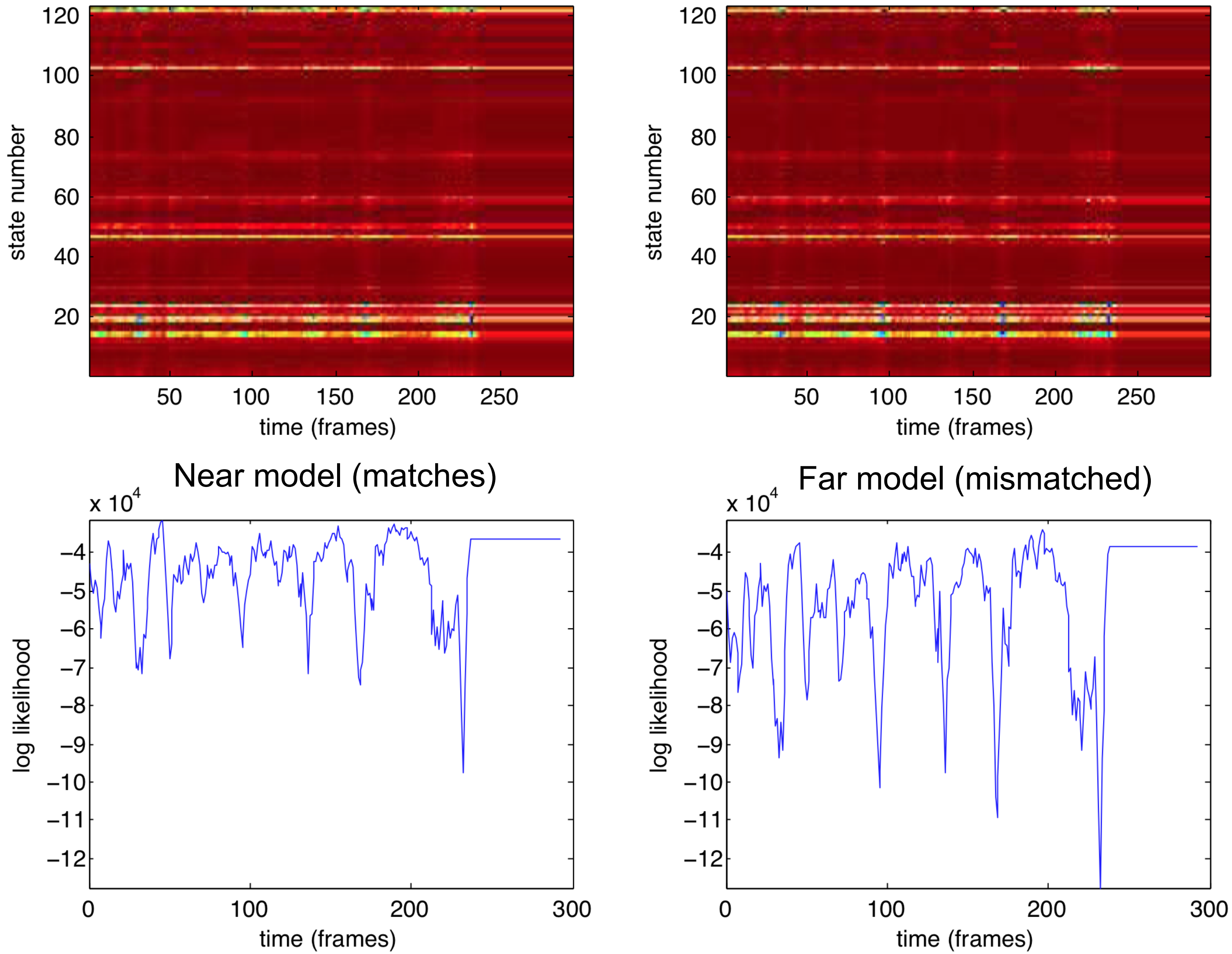
Feature selection by maximum OSL

- Can also use the acoustic models *themselves* to direct the model selection
- Observation state likelihoods for the first 100 frames of the speech are examined
- Classify as 'near' if

$$\sum_{t=1}^{100} \max_q \log[p(x(t) \mid \lambda_n, q)] > \sum_{t=1}^{100} \max_q \log[p(x(t) \mid \lambda_f, q)]$$

- Correct classification of near/far on test set was 88% using this approach (better than MPR)

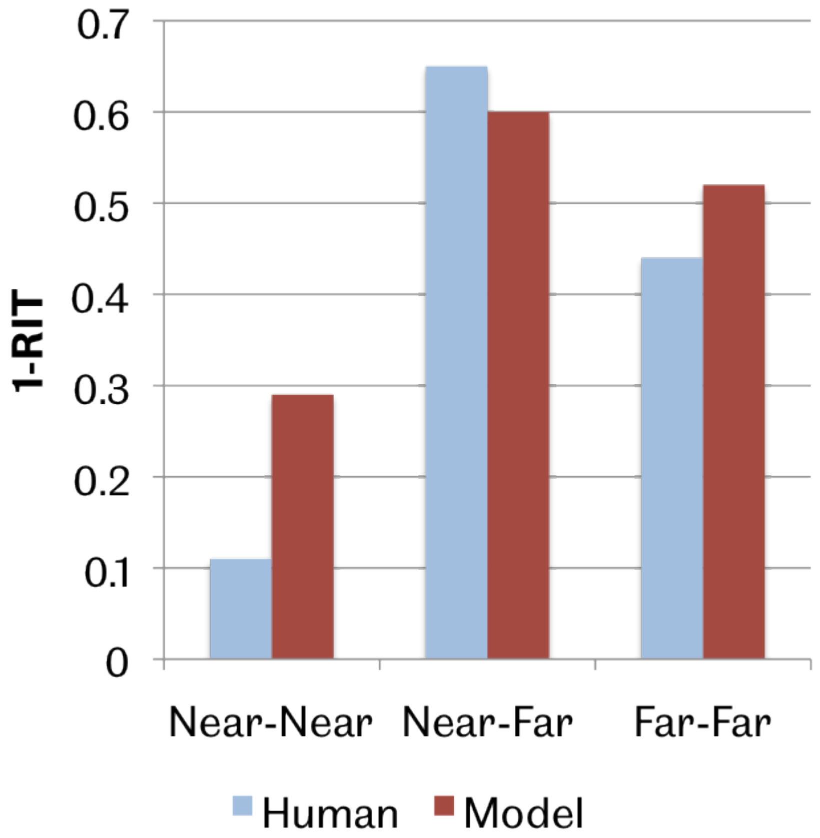
Matching ‘near’ and non-matching ‘far’

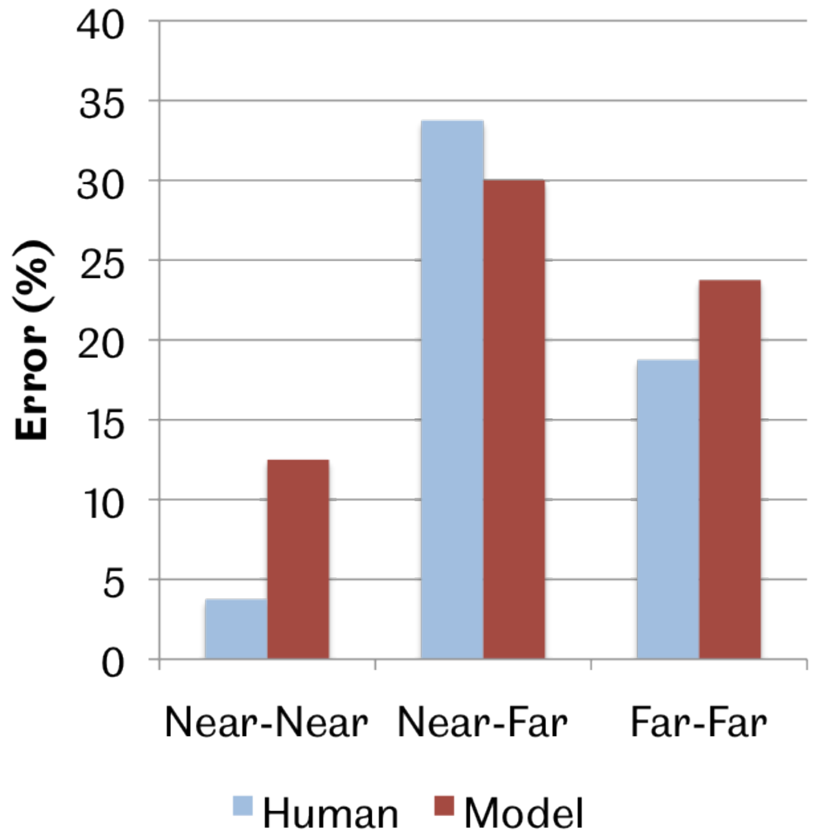


Observation state likelihoods

Maximum log likelihood across all states

Feature stream selection by MOSL





Confusions: feature selection by MOSL

Human Near-Near					Model Near-Near						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	19	0	0	1	SIR	16	0	0	4	0.0514	0.3416
SKIR	0	20	0	0	SKIR	1	18	0	1	0.0526	0.4872
SPIR	0	1	18	1	SPIR	2	0	18	0	0.1000	0.4872
STIR	0	0	0	20	STIR	0	1	1	18	0.0526	0.4872
Human Near-Far					Model Near-Far						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	18	0	0	2	SIR	18	1	1	0	0.1000	0.4872
SKIR	3	15	0	2	SKIR	2	16	1	1	0.0391	0.9346
SPIR	7	2	10	1	SPIR	5	1	13	1	0.0264	0.7890
STIR	8	1	1	10	STIR	9	2	0	9	0.0361	0.8534
Human Far-Far					Model Far-Far						
	SIR	SKIR	SPIR	STIR		SIR	SKIR	SPIR	STIR	ϕ^2	Fisher
SIR	16	1	1	2	SIR	12	1	2	5	0.0548	0.5200
SKIR	0	16	0	4	SKIR	3	15	0	2	0.0925	0.2228
SPIR	2	1	14	3	SPIR	2	1	17	0	0.0823	0.3936
STIR	1	0	0	19	STIR	1	1	1	17	0.0528	0.7367

Interim discussion

- Similar performance to the MPR version of the model
  - But error is higher in far-far condition, which somewhat reduces the magnitude of the compensation effect
  - Less impressive match to confusions in far-far condition (but still acceptable, and no statistically significant different from human confusion pattern)

Conclusions

- All versions of the model
  - Exhibit a constancy effect in the same manner as the listeners in Amy's experiment
  - Provide a good match to the pattern of consonant confusions made by listeners

Planned work for next period

- A further extension of the model is to perform feature selection and combination on a band-by-band basis
  - Divide the features into, say, 8 bands
  - Train 'near' and 'far' HMMs for each band
  - During decoding, have a dynamic weight $\alpha(t,b)$ which is determined by reverberation estimate in band b
- Could allow modelling of Tony's experiments using noise-vocoded speech
- Probably necessary to do this with spectral, rather than cepstral, features

Comments?