Generalising from sir-stir

The University Of Sheffield.

EPSRC 36-month meeting · 29 Mar 2012

Amy Beeston and Guy Brown

Overview

1. Modelling sir-stir (i) across-band
2. (ii) within-band
3. Generalising from sir-stir
4. Constancy front-end for ASR

Generalising from sir-stir

naturalistic speech stimuli

- do Watkins' findings hold for naturalistic speech?
- Articulation Index (AI) Corpus
  - includes sir and stir
  - more context words
  - more talkers
- each AI corpus utterance uses different talker, vocabulary, speech rate, pitch contour, stress pattern etc.
  - cancel excess variability?
  - analyze results with regard to this variability?

Wright (2005). Articulation Index. Linguistic Data Consortium, Philadelphia.

<more>

Generalising from sir-stir

ideals

- naturalistic speech
  - real world listening
  - ASR compatible
- increase data per participant
  - increase subset of Articulation Index Corpus
  - with {s, sk, sp, st} can have {e, i, E, I, @, R, (a, o)}
  - further consonant/vowel sets?
- minimize manual handling
  - word boundaries located via (HTK) forced-alignment

Generalising from sir-stir

extending sir-stir

- subset of corpus

  sir · skur · spur · stir

🔈 🔈 🔈 🔈

- unvoiced stop consonants
- place of articulation

  /p/ front · /k/ back· /t/ middle

Generalising from sir-stir

relative information transferred (RIT)

- no category boundary
- misclassifications

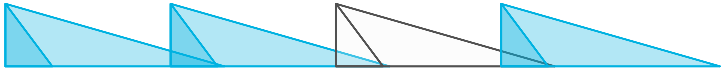
@ nf	sir	skur	spur	stir
sir	37	0	0	3
skur	6	29	2	3
spur	16	3	19	2
stir	16	2	1	21

- RIT
  - regards participants as channels
    - accept input stimuli
    - produce output responses
  - measures their information transfer characteristics

Miller and Nicely (1955). *J Acoust Soc Am*, 27, 338-352.

<more>

Generalising from sir-stir



‘cutoff’

- Is it possible to replicate compensation for reverb?
- Probably necessary to increase overall error rate
  => low pass filtered to avoid ceiling effects
- same and mixed distance sentences

  {near, far} context + {near, far} test

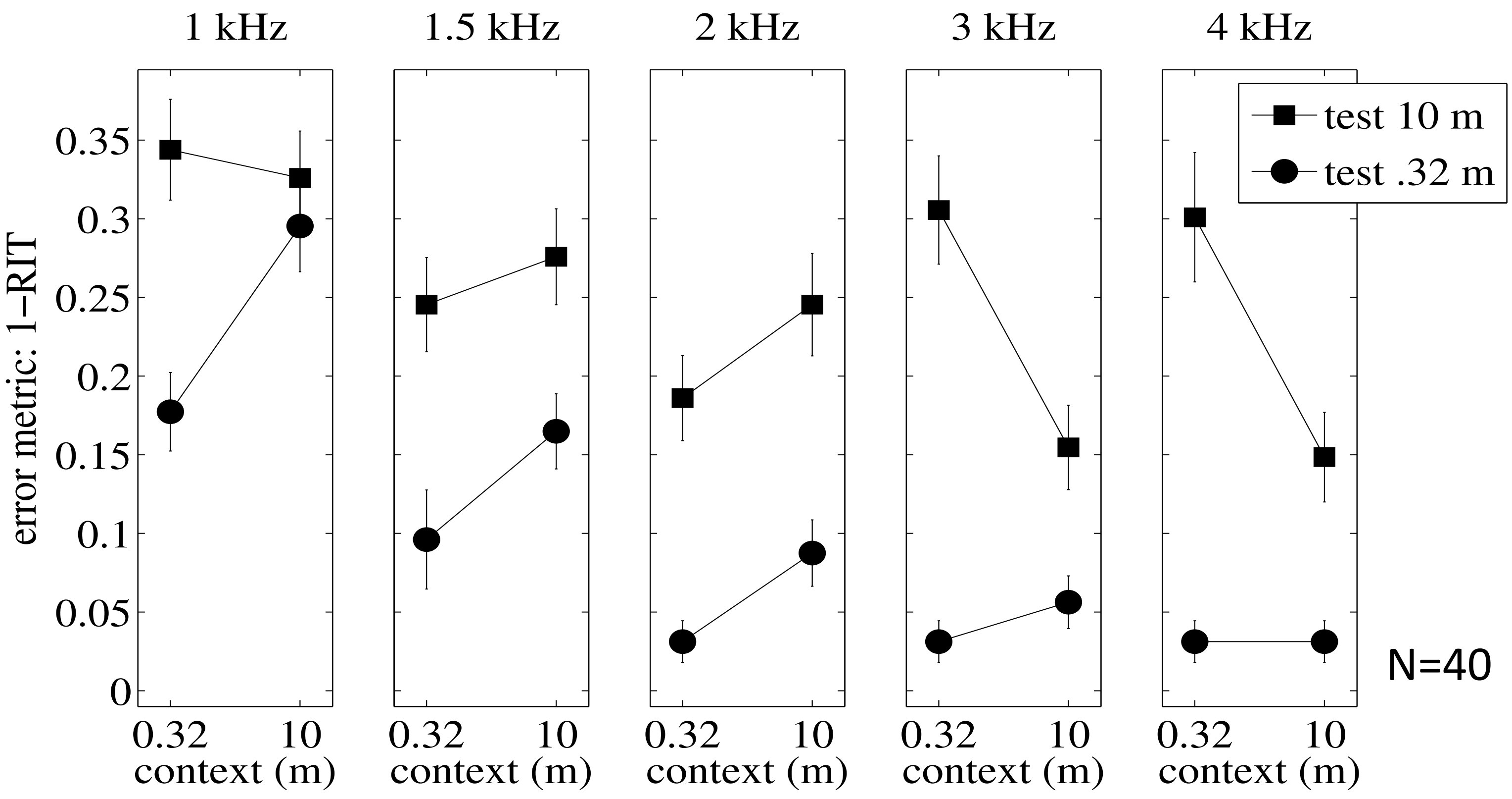
  {1, 1.5, 2, 3, 4} kHz low-pass filter cutoff
- 1600 stimuli partitioned across 20 listeners (N=40)

  4 targets X 20 talkers X 4 distances X 5 filters

Generalising from sir-stir

‘cutoff’

- errors incr. as low-pass filter cutoff frequency decr.
- compensation apparent when high freqs are present



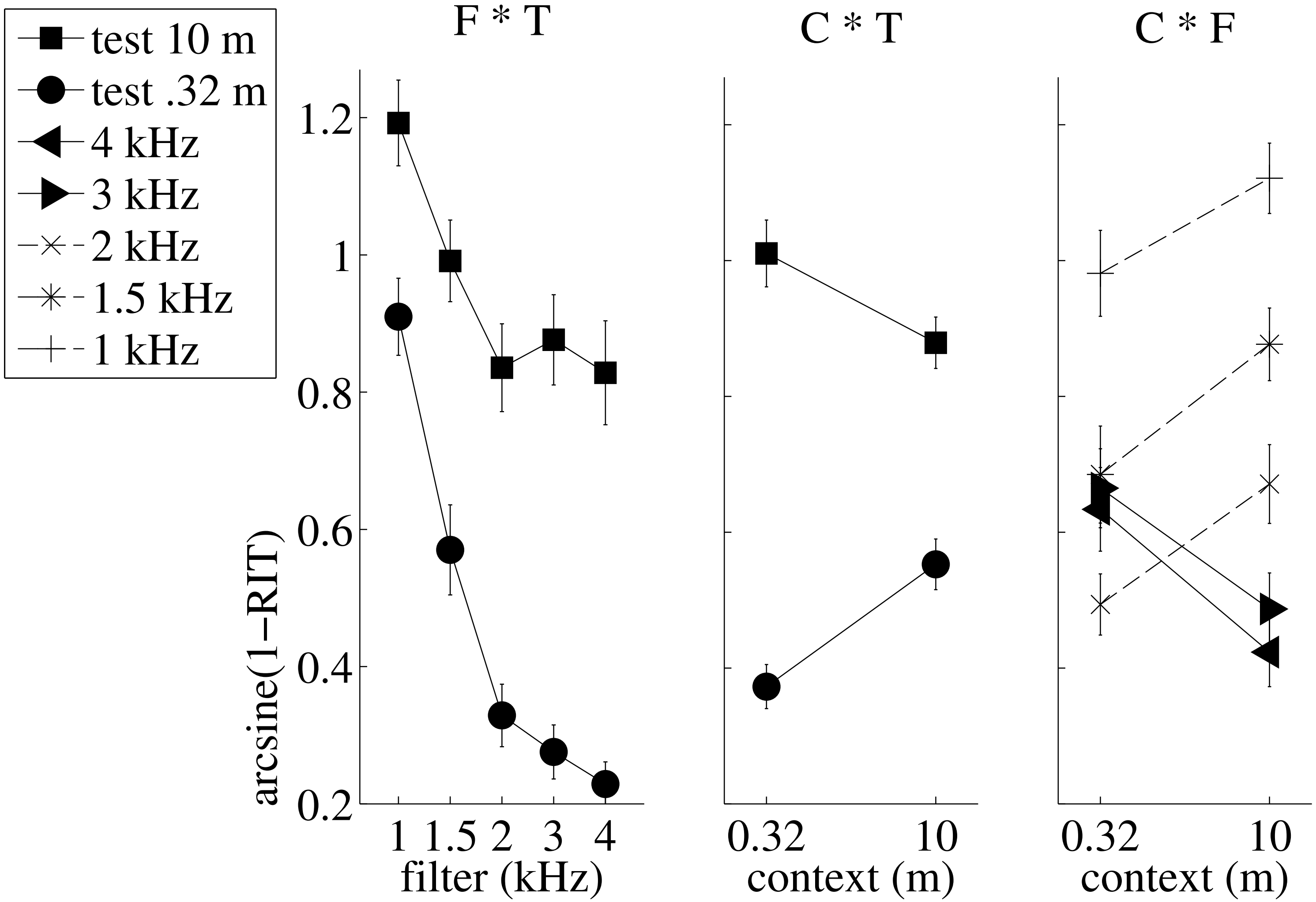
Generalising from sir-stir

ANOVA ‘cutoff’

- 3-way repeated measures, all within-subject factors
- independent variables
  - test word distance (2 levels)
  - context distance (2 levels)
  - low pass filter cutoff (5 levels)
- dependent variable: arcsine-RIT
- significant main effects
  - test, filter
- significant interactions (no 3-way, all 2-way)
  - test X filter, context X test, context X filter

Generalising from sir-stir

ANOVA ‘cutoff’



Generalising from sir-stir

word-level analysis

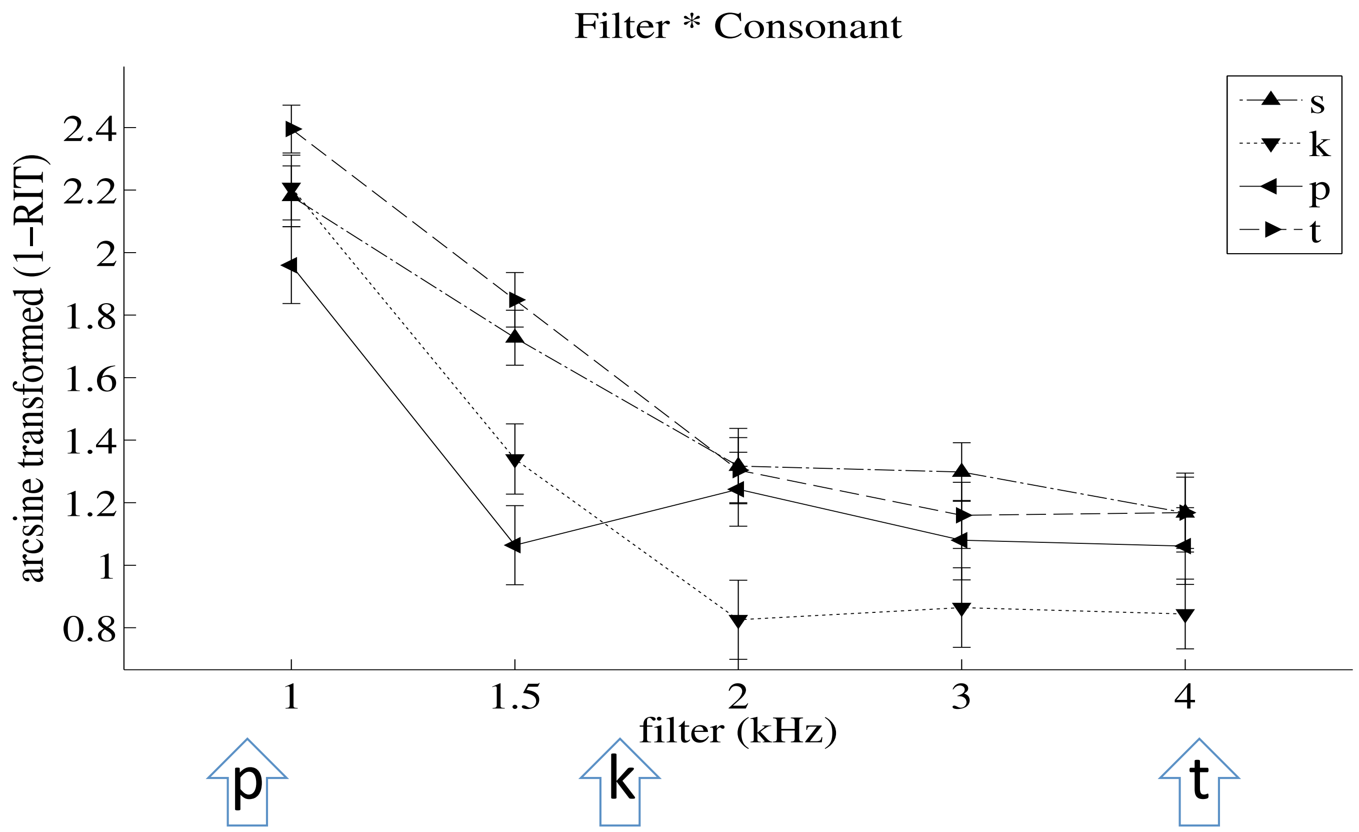
- 2-way repeated measures ANOVA aggregating across context and test distances
- Independent variables: filter condition, consonant
- Dependent variable: arcsine-RIT (per consonant presented)
- Allen and Li: {/t/, /k/, /p/} identified by burst frequency
  /t/ at 4 kHz; /k/ at 1.4 – 2 kHz; /p/ at 0.7 – 1kHz

Allen and Li (2009). IEEE Signal Process. Magazine 73-77.

Generalising from sir-stir

word-level analysis

- /k/ had generally fewer errors (but advantage was lost at low freqs)
- /p/ holds identity better at 1.5 kHz



Generalising from sir-stir

‘inAndExtrinsic’

- does compensation occur…
  - without following contexts?
  - without preceding contexts?
  - with reduced intrinsic (test word) information?
- H: intrinsic info not required if extrinsic info is reliable
- 5760 stimuli partitioned across 12 listeners (N=48)

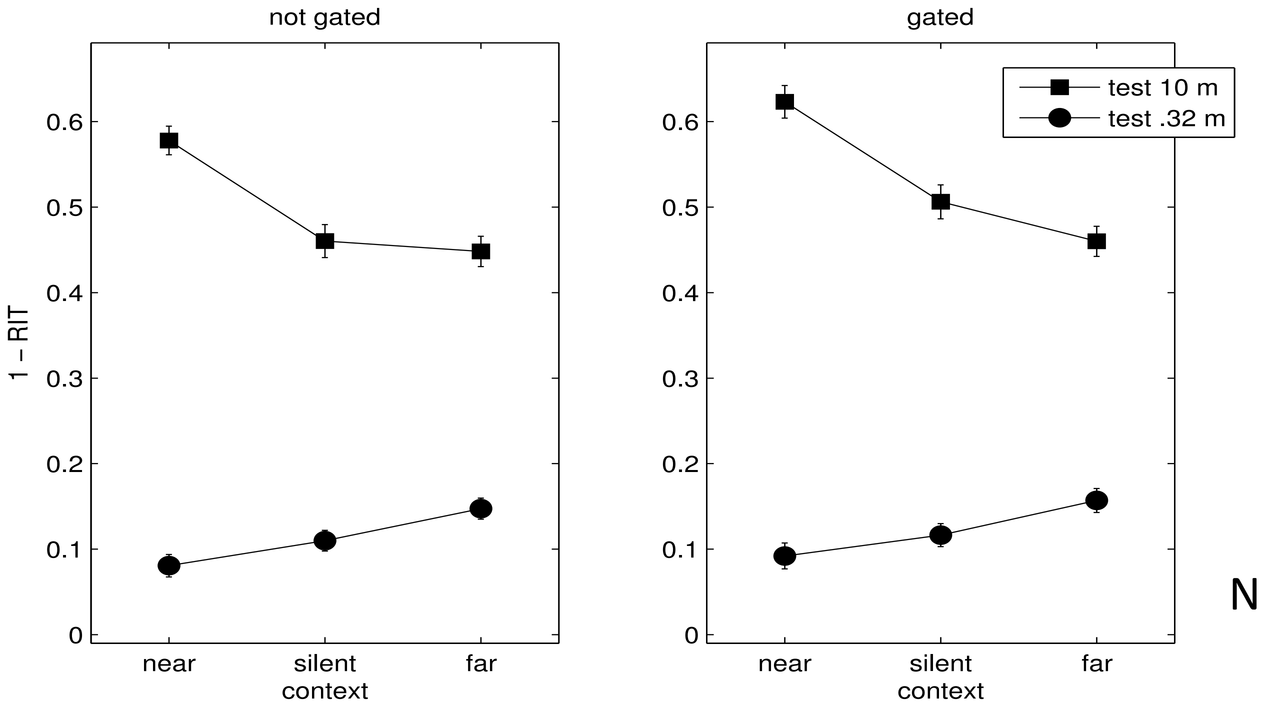
  {near, far, silent} context X {near, far} test

  4 consonants X 6 vowels X 20 talkers X 3 context conditions X 2 test distances

Generalising from sir-stir

‘inAndExtrinsic’

- Following CWs not required for compensation
- Preceding CWs not required: ‘silent’ acts like ‘far’
- Intrinsic TW information: significant but small effect



N=48

Generalising from sir-stir

ANOVA ‘inAndExtrinsic’

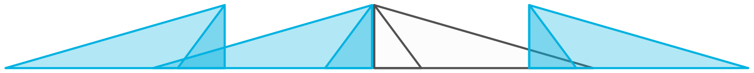
- 3-way repeated measures, all within-subject factors
- independent variables
  - context condition (3 levels)
  - test word distance (2 levels)
  - test word gate condition (2 levels)
- dependent variable: arcsine-RIT
- significant main effects
  - test, context, gate
- significant interactions
  - test X context

Generalising from sir-stir

ANOVA ‘inAndExtrinsic’

- no 3-way interaction but
- planned comparisons based on hypothesis examined effect of gate on far-distance test words
  - far context: no effect
  - silent and near contexts: small incr. in errors
- suggests intrinsic info is used when context is ambiguous (e.g. missing or inappropriate)

Generalising from sir-stir



‘reverse’

- do time-reversal procedures disrupt compensation if applied to preceding context? 🔈 🔈 🔈 🔈

- time reversed speech and/or reverberation

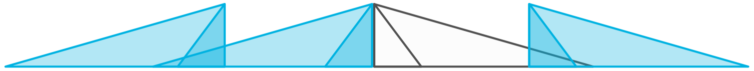
  fwd reverb: context reverb overlaps test

  rev reverb: context reverb does not overlap test

- 1280 stimuli partitioned across 16 listeners (N=64)

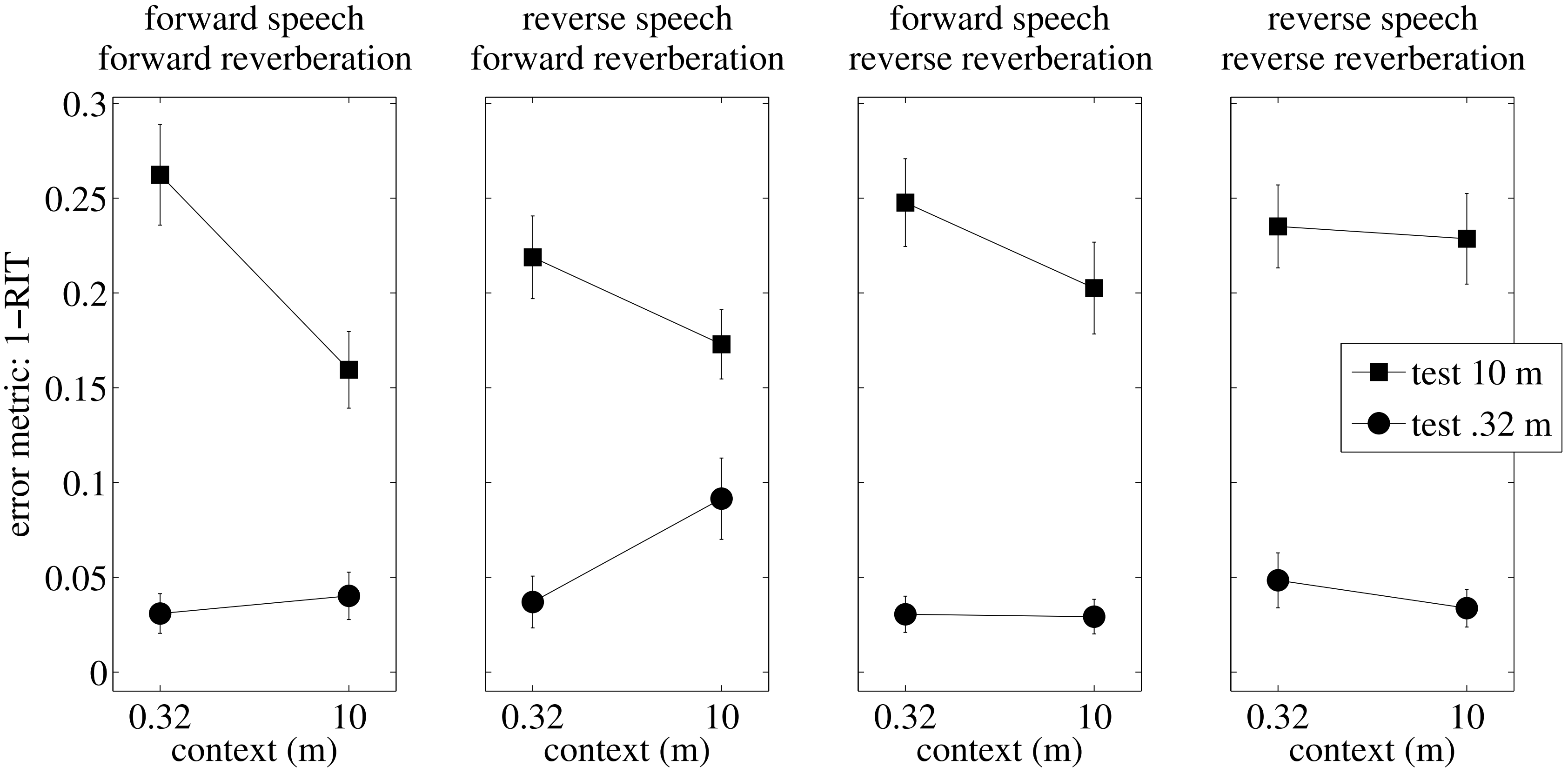
  4 targets X 20 talkers X 4 distances X 4 reversals

Generalising from sir-stir



‘reverse’

- compensation is present for forward reverberation, but abolished with reverse reverb?



N=64

Generalising from sir-stir

ANOVA i. 'reverse'

- 4-way repeated measures, all within-subject factors
- independent variables
  - test word distance (2 levels)
  - context distance (2 levels)
  - speech direction (2 levels)
  - reverberation direction (2 levels)
- significant main effects
  - test, context
- significant interactions
  - context X test, context X speech
- not reverb direction!

Generalising from sir-stir

ANOVA ii. ‘reverse’

- ? 3-way repeated measures, all within
- independent variables
  - test word distance (2 levels)
  - context distance (2 levels)
  - speech direction (2 levels)
  - ~~reverberation direction (2 levels)~~
- but results of ANOVA [C, T, C*T, C*S] then depends on averaged-arcsine-transformed-RIT scores
- If categories are combined in the confusion matrices before the RIT calculation: different results [T, C*T] i.e. no interaction with speech direction

Generalising from sir-stir

interim conclusions

- analysis methods require still more thought!
- compensation for reverberation exists for naturalistic speech despite -
- high degree of variability (cf. Watkins)
  - more talkers
  - more context words
  - more test words
- different things going on for different test words…

the end

thank you for listening

references

Allen, J.B. and Li, F. (2009). Speech perception and cochlear signal processing. IEEE Signal Process. Magazine 73-77.

Miller, G.A. and Nicely, P.E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants. *J Acoust Soc Am*, 27, 338-1265.

Wright J. (2005). Articulation Index. Linguistic Data Consortium, Philadelphia.

extra slides

Articulation Index Corpus (AIC)

```
$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;

$test = SIR | STIR | SPUR | SKUR;

$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;

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

Wright (2005). Articulation Index. Linguistic Data Consortium, Philadelphia.

<back>

relative information transmitted (RIT)

- considers consonant confusions
- regards participants as channels
  - receiving input stimuli (X)
  - producing output responses (Y)
- measures their information transfer characteristics
- RIT = H(X:Y) / H(X)

  where H(X:Y) is the mutual-information of X and Y,

  and H(X) is the self-information (entropy) of X.

Miller and Nicely (1955). *J Acoust Soc Am*, 27, 338-352.

<back>