Perceptual experiments

EPSRC 24-month meeting · 13 Dec 2010

Amy Beeston

The University Of Sheffield.

Overview

1. generalising from sir-stir
2. listening experiments so far
3. modelling approach
4. future

sir-stir

1. generalising from sir-stir
2. listening experiments so far
3. modelling approach
4. future

Watkins' paradigm

- Watkins' stimuli
  - one context sentence
  - one talker
  - artificially created /t/ in 'stir'

- sir-stir identification depends on rapidly changing amplitude modulation (envelope)
- reverberation
  - prolongs peaks and masks dips
  - overcomes processing used to create continuum

naturalistic speech stimuli

- do Watkins' findings hold for naturalistic speech?

- Articulation Index Corpus
  - includes sir and stir
  - more context words
  - more talkers

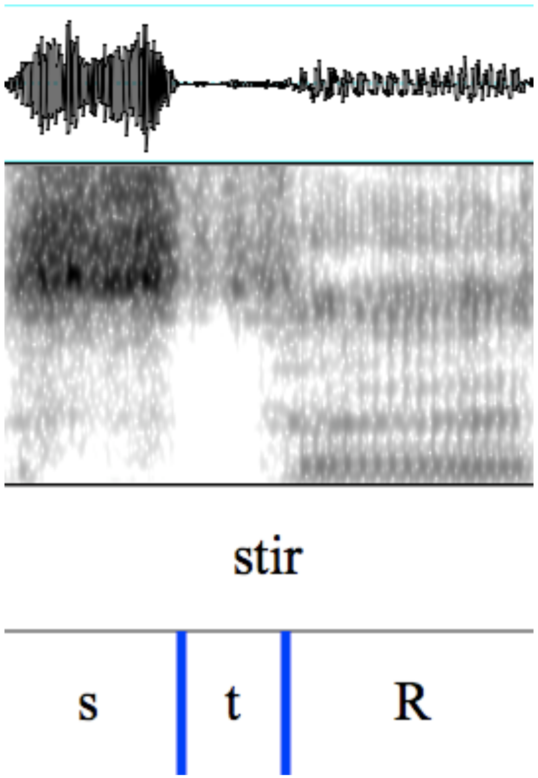
Wright (2005). Articulation Index. Linguistic Data Consortium, Philadelphia.

<more>

sir-stir

stop consonants

- esp. sensitive to reverberation
- gaps are filled
  - self masking, /s/
  - overlap masking, pre. context



Drullman et al. (1994). J Acoust Soc Am, 95(2), 1053-1064.
Nábělek et al. (1989). J Acoust Soc Am, 86(4), 1259-1265.

extending sir-stir

- subset of corpus
  - sir · skur · spur · stir
- unvoiced stop consonants
- place of articulation
  - /p/ front · /k/ back· /t/ middle

consonant confusions

- no category boundary
- misclassifications

@nf	sir	skur	spur	stir
sir	53	2	1	4
skur	11	47	2	0
spur	11	6	41	1
stir	13	2	0	45

- percentage correct
- relative information transferred (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>

listening

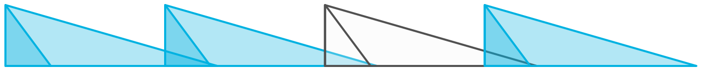
1. generalising from sir-stir
2. listening experiments so far
3. modelling approach
4. future

findings so far

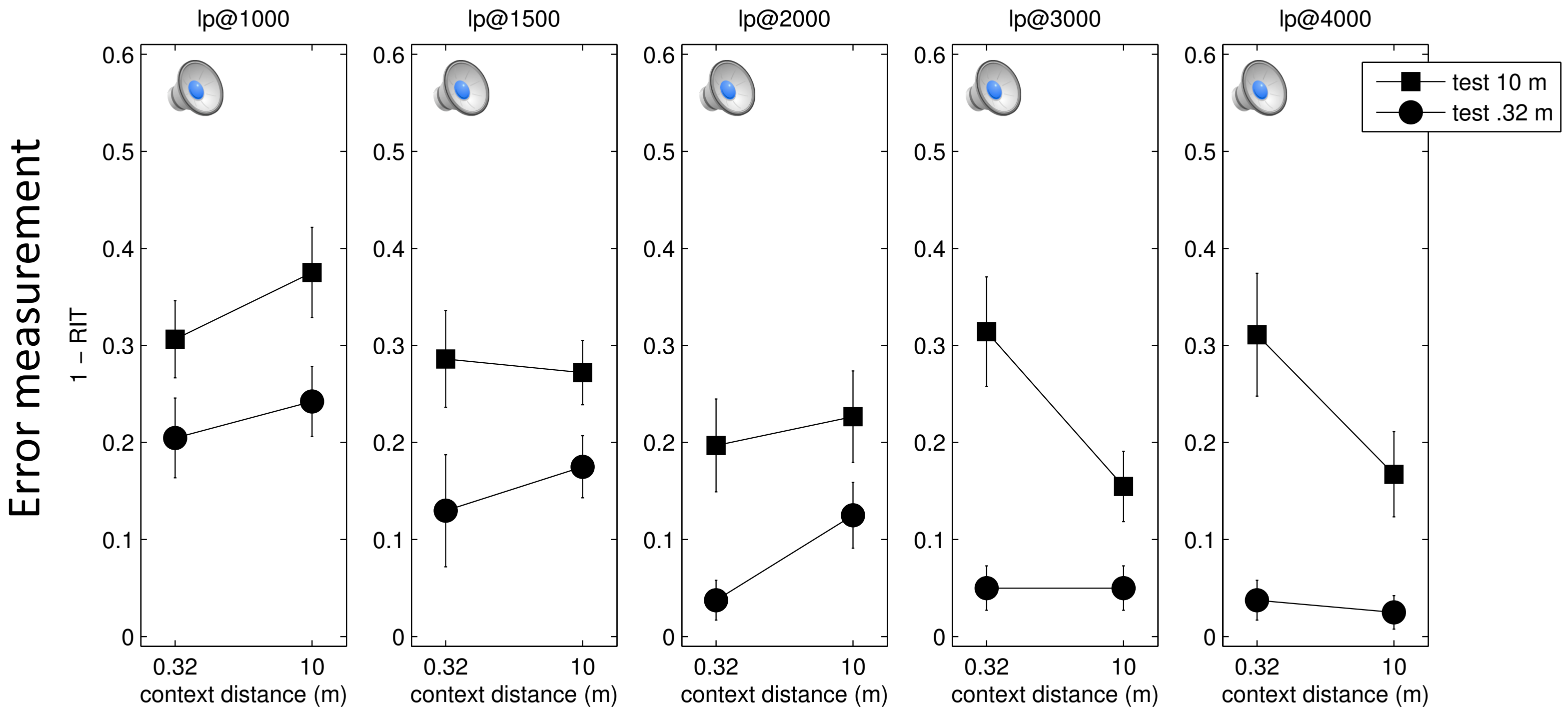
- … despite high degree of temporal uncertainty
  - more talkers
  - more context words
  - few more test words
- compensation for reverberation reported for naturalistic speech
  - freq effect
  - time reversal

Beeston et al. (2010). *British Society of Audiology.* Abstract #118

experiment 1



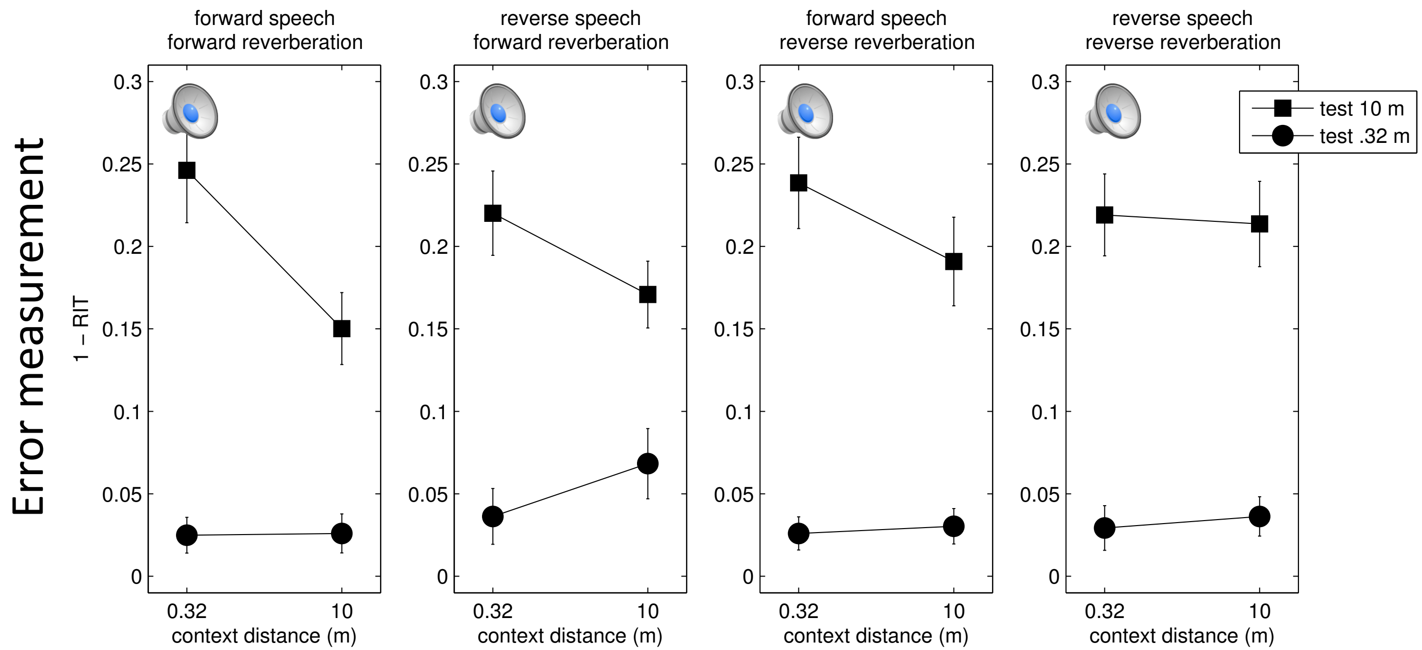
- freq. effect: low-pass versions
- compensation significant at 4kHz



<more>

experiment 2

- time reversal: preceding context
- no compensation with rev. reverb

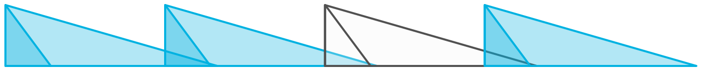


<more>

word-level analysis

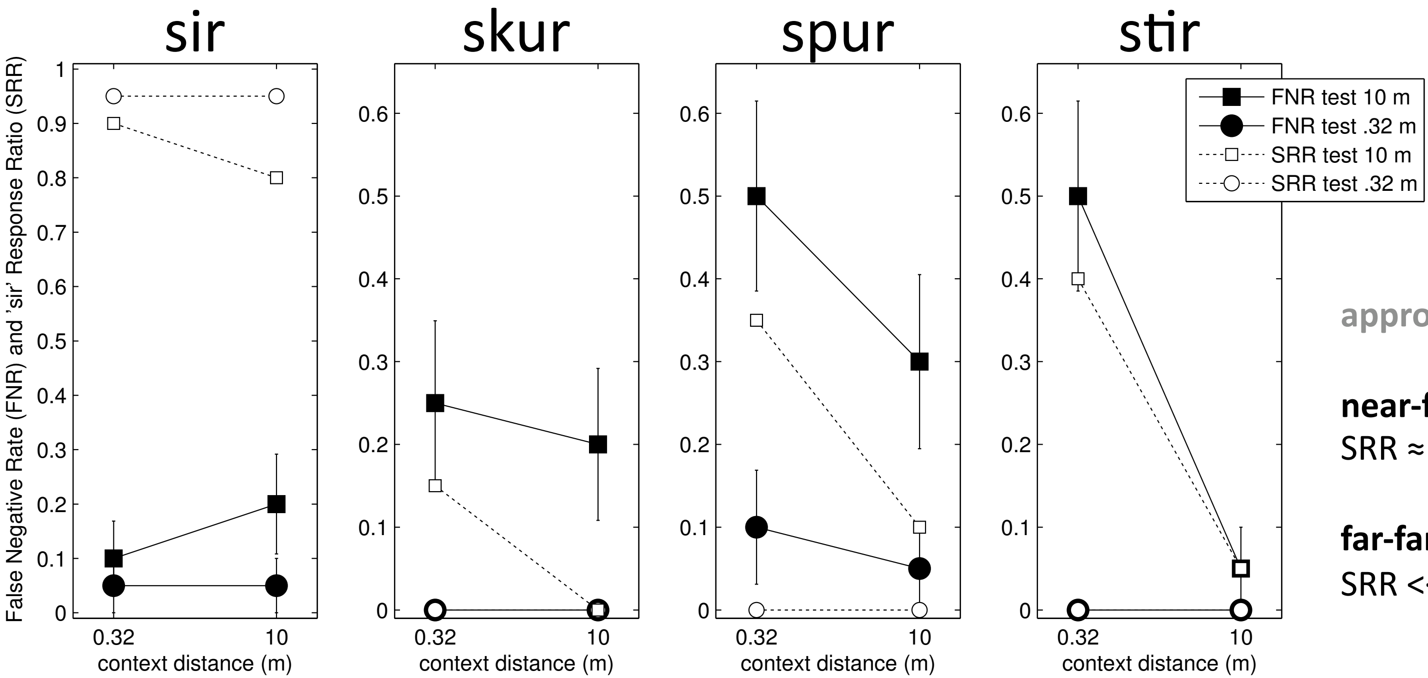
- compensation reduces mistaken ‘sir’ responses at far-far
- but confusions persist between ‘skur’, ‘spur’ and ‘stir

Beeston et al. (2010). *British Society of Audiology.* Abstract #118

4kHz 

expt 1 'normal'

- compensation reduces mistaken 'sir' responses at far-far
- but confusions persist between 'skur', 'spur' and 'stir



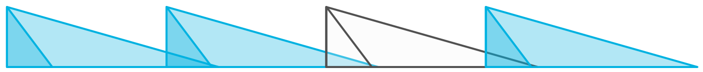
approximation:

near-far:
SRR ≈ FNR

far-far:
SRR << FNR

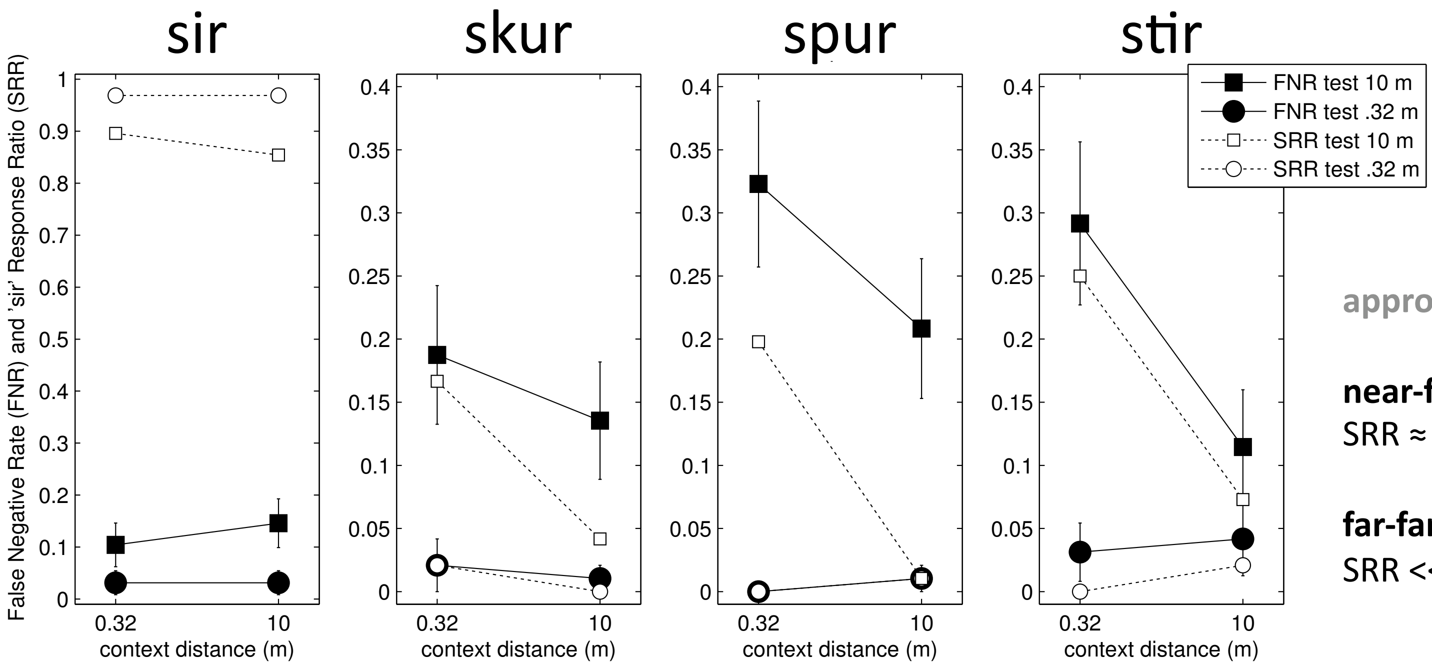
FNR = false negative rate = FN / (TP + FN)
SRR = 'sir' response ratio = proportion of 'sir' responses

fwd



expt 2 ‘normal’

- compensation reduces mistaken ‘sir’ responses at far-far
- but confusions persist between ‘skur’, ‘spur’ and ‘stir



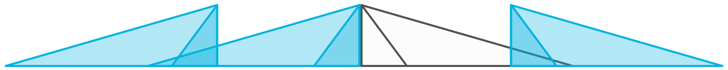
approximation:

near-far:
SRR ≈ FNR

far-far:
SRR << FNR

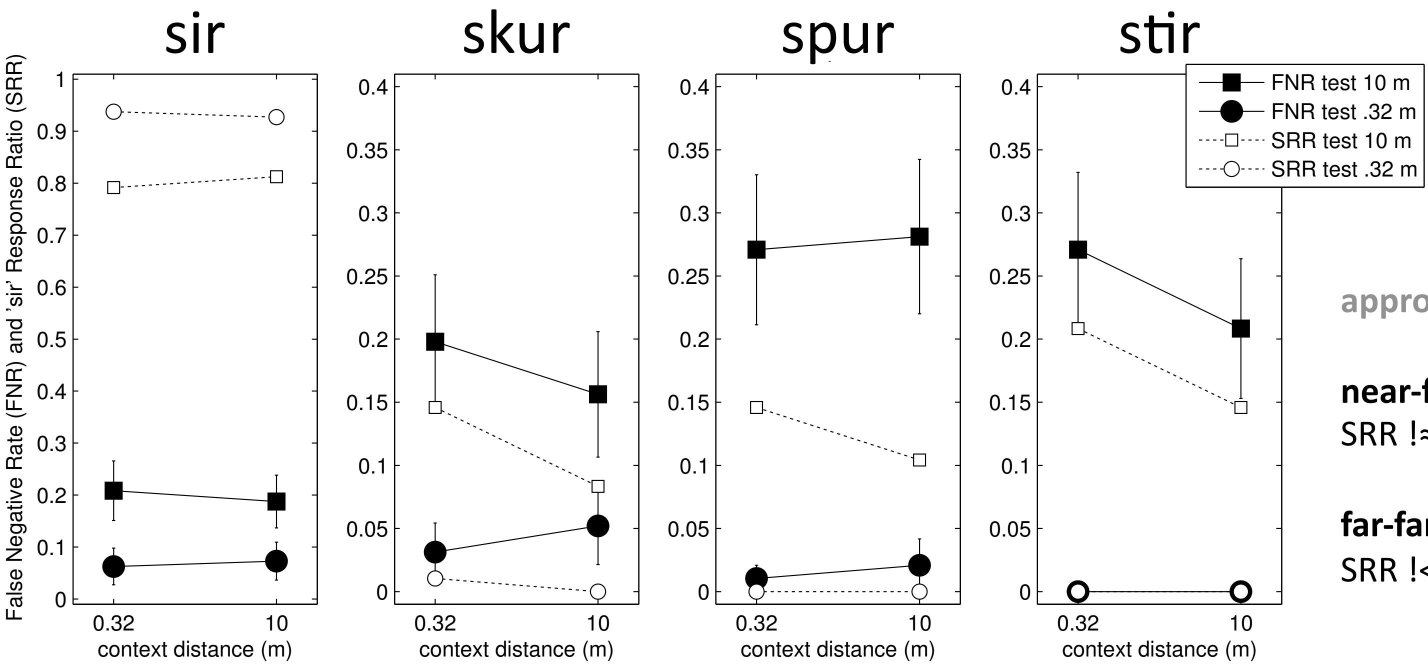
FNR = false negative rate = FN / (TP + FN)
SRR = ‘sir’ response ratio = proportion of ‘sir’ responses

rev-rev



expt 2 'reverse'

- majority of near-far errors are no longer for 'sir'
- far-far errors still include 'sir' amongst confusions



approximation:

near-far:
SRR !≈ FNR

far-far:
SRR !<< FNR

FNR = false negative rate = FN / (TP + FN)
SRR = 'sir' response ratio = proportion of 'sir' responses

modelling

1. generalising from sir-stir
2. listening experiments so far
3. modelling approach
4. future

combine approaches

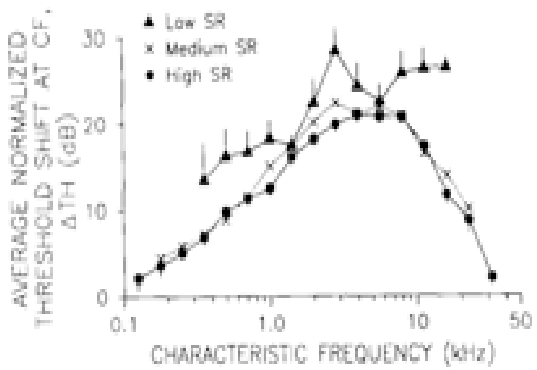
- psychoacoustic experiment design informs computational modelling questions
  - and vice versa

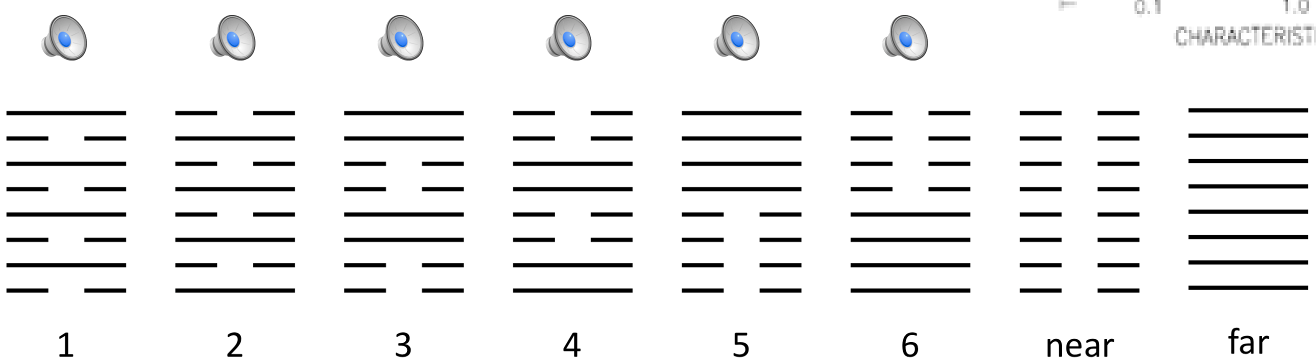
- specifically
  - awareness of preceding context (window)
- eventually
  - integrate auditory model as front-end for ASR

model

context awareness

- high freq. bands incr. important in sir-stir distinction
  - psychoacoustic data
  - physiological data





Guinan and Gifford (1988). *Hearing Res*, 31, 29-46.

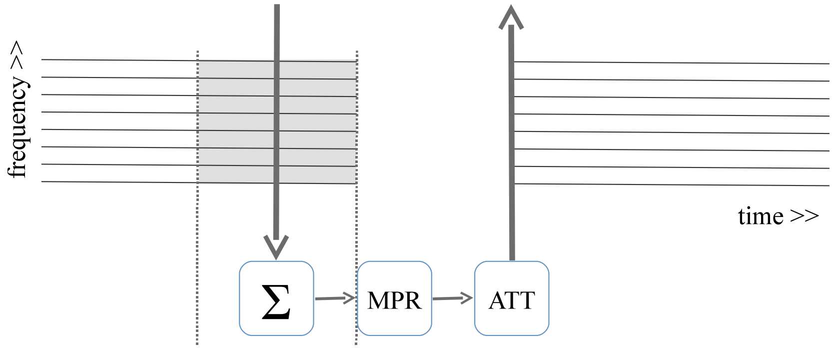
Watkins at al. (2010). *British Society of Audiology.* Abstract #116.

<more>

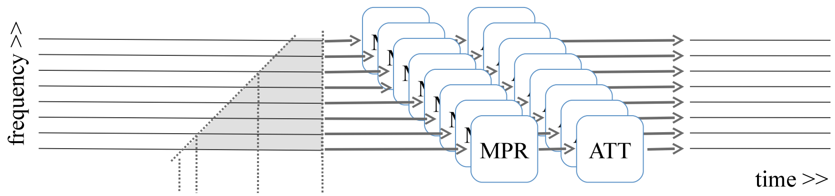
context windows

- Different contextual awareness in each channel

Across channel



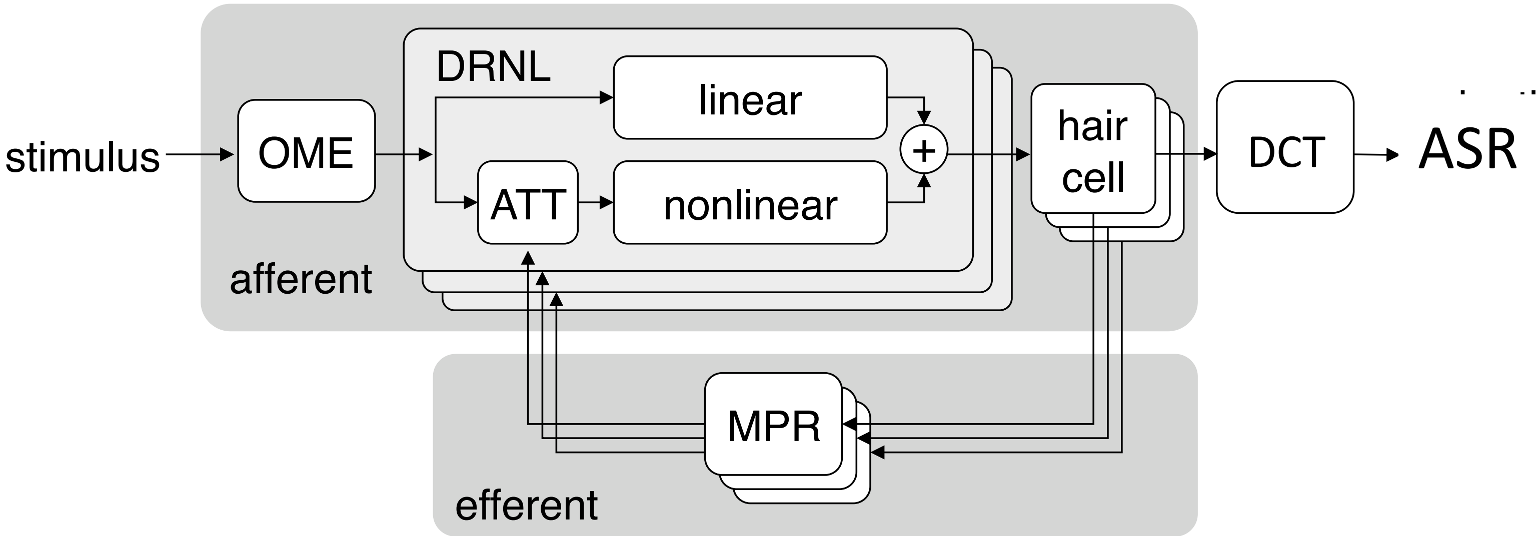
Within channel



- Freq-window shape (MPR to ATT mapping)
- Time-window length (footprint)
- Time-window shape (forgetting function)

constancy front-end for ASR

- does ASR improve for reverberant stop consonants?
  - ATT determined channel-by-channel (MPR?)
  - recognition using DCT (not STEP) features



Ferry and Meddis (2007). J Acoust Soc Am, 122(6), 3519-3526.

ASR

- compare human/machine performance
  - assess consonant confusions on same listening task

- Baseline system
  - HTK/MFCC phone recogniser
- Simplified auditory model
  - efferent circuit engaged (with more ATT at far distance)

Brown and Beeston (2010). *British Society of Audiology.* Abstract #117.

<more>

interim conclusions

- human listeners use information from preceding context to effect compensation
- conventional ASR systems do not

- simulation of compensation
  - tracking dynamic range of context
  - via efferent suppression
- much work remains
  - frequency-dependent efferent suppression
  - wider range of consonant confusions

future

1. generalising from sir-stir
2. listening experiments so far
3. modelling approach
4. future

requirements

- naturalistic speech
  - real world listening
  - ASR compatible
- minimize manual handling
  - word boundaries located via forced-alignment
  - wide bandwidth (no low-pass filter)
- increase data per participant
  - further AI corpus utterances
  - with {s, sk, sp, st} can have {a, e, i, o, xe, xi, xq, xr}
  - further consonant/vowel sets?

avoiding confounds

- 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?

- no conclusions yet!
  - much careful thought needed…

perceptual experiments

- word-by-word
- silent contexts
- longer contexts
- processed contexts

future

word-by-word

- [cw1][cw2][test][cw3]
- 16 distance conditions
- following context for naturalistic speech? (FFFF·FFFN)
- length of reverberated preceding context? (FFFF·NFFF)
- does 'near' signal reset constancy? (FFFF·FNFF)

```
NNNN, NNNF – near-near
NNFN, NNFF – near-far
FFNN, FFNF – far-near
FFFN, FFFF – far-far
NFNN, NFNF, NFFN, NFFF – mixed NF preceding
FNNN, FNNF, FNFN, FNFF – mixed FN preceding
```

silent contexts

- [~~cw1~~][~~cw2~~][test][cw3]
- ‘reference’ condition?
- silent preceding context, small dynamic range
- model predicts max ATT, large compensation
  - efferent system at higher sound levels
  - something higher up?
- equivalent to ‘far’ preceding context?
  - also small dynamic range, max ATT, large compensation

Nielsen and Dau (2010). J Acoust Soc Am, 128(5), 3088-3094.

longer contexts

- [cw1][cw2][cw3][cw4] + [cw5][cw6][test][cw7]
- more compensation with longer preceding context?
  - or is two syllables already enough?

Brandewie and Zahorik (2010). J Acoust Soc Am, 128(1), 291-299.

processed contexts

- rank or classify utterances by characteristic?
  - eg. lengths of context words, test words, stop gaps
  - eg. spectral/temporal centres-of-gravity
- treating utterances to control for some variability?
  - eg. equalise long-term average spectrum
  - eg. noise-vocoded versions
- other ideas?

the end

thanks

references

Beeston, A.V. and Brown, G.J. (2010). Perceptual compensation for effects of reverberation in speech identification: A computer model based on auditory efferent processing. *Interspeech, Japan*. Proceedings, 2462-2465.

Beeston, A.V., Brown, G.J., Watkins, A.J. and Makin, S.M. (2010). Perceptual compensation for reverberation: human identification of stop consonants in reverberated speech contexts. *British Society of Audiology Annual Conference, Manchester.* Abstract #118.

Brandewie, E. and Zahorik, P. (2010). Prior listening in rooms improves speech intelligibility. *J Acoust Soc Am*, 128(1), 291-299.

Brown, G.J. and Beeston, A.V. (2010). A computer model of perceptual compensation for reverberation: evaluation on a consonant identification task. *British Society of Audiology Annual Conference, Manchester.* Abstract #117.

Drullman, R., Festen, J.M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am*, 95(2), 1053-1064.

Ferry, R.T. and Meddis, R. (2007). A computer model of medial efferent suppression in the mammalian auditory system. *J Acoust Soc Am*, 122(6), 3519-3526.

Guinan, J.J. Jr (2006). Olivocochlear efferents: anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear Hear*, 27(6), 589-607.

Guinan, J.J. and Gifford, M.L. (1988). Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory-nerve bers. III. Tuning curves and thresholds at CF. *Hearing Res*, 31, 29-46.

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

Nábělek, A.K., Letowski, T.R., and Tucker, F.M. (1989). Reverberant overlap- and self-masking in consonant identification. *J Acoust Soc Am*, 86(4), 1259-1265.

Nielsen, J.B. and Dau, T. (2010). Revisiting perceptual compensation for effects of reverberation in speech identification. *J Acoust Soc Am*, 128(5), 3088-3094.

Watkins, A.J. (2005). Perceptual compensation for reverberation in speech identification. *J Acoust Soc Am*, 118(1), 249-262.

Watkins, A.J., Raimond, A.P. and Makin, S.M. (2010). Effects of room reflections on speech identification and the relative importance of different frequency-bands. *British Society of Audiology Annual Conference, Manchester.* Abstract #116.

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>

experiment 1

- is it possible to replicate compensation for reverb?

- same and mixed distance sentences
  {near, far} context + {near, far} test
- low-pass filtered to avoid ceiling effect
  {1, 1.5, 2, 3, 4} kHz cutoff
- 1600 stimuli partitioned amongst 20 listeners
  4 targets X 20 talkers X 4 distances X 5 cutoffs

Beeston et al. (2010). *British Society of Audiology.* Abstract #118

<next>

listening expt 1: ANOVA

- 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)
- significant main effects
  - test $F\,(1,19) = 59.27$, $p < 0.001$
  - cutoff $F\,(4,76) = 9.19$, $\varepsilon_{HF} = 0.96$, $p < 0.001$
- significant interactions
  - context X cutoff $F\,(4,76) = 2.593$, $\varepsilon_{HF} = 1.0$, $p < 0.05$

<next>

listening expt 1: chi-squared

lp@4000: only significant result

X^2 = 8.006926407
p = 0.023299381 (Bonferroni corrected)

lp@4000	#sirs	#not
near-far	36	44
far-far	19	61

lp@3000	#sirs	#not
near-far	37	43
far-far	23	57

lp@2000	#sirs	#not
near-far	28	52
far-far	21	59

lp@1500	#sirs	#not
near-far	27	53
far-far	14	66

lp@1000	#sirs	#not
near-far	19	61
far-far	22	58

<back>

experiment 2

- 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 amongst 16 listeners
    - 4 targets X 20 talkers X 4 distances X 4 reversals
- 48 participants

Beeston et al. (2010). *British Society of Audiology.* Abstract #118

<next>

listening expt 2: ANOVA

- 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 $F\ (1,47) = 189.5$, $p < 0.001$
  - context $F\ (1,47) = 5.7$, $p < 0.05$
- significant interactions
  - context X test $F\ (1,47) = 7.9$, $p < 0.01$

<next>

listening expt 2: chi-squared

	# sirs	# not sirs
near-far	168	312
far-far	121	359

forward reverb: significant

X^2 = 10.9345699957
p = 0.001886577 (Bonferroni corrected)

	# sirs	# not sirs
near-far	165	315
far-far	140	340

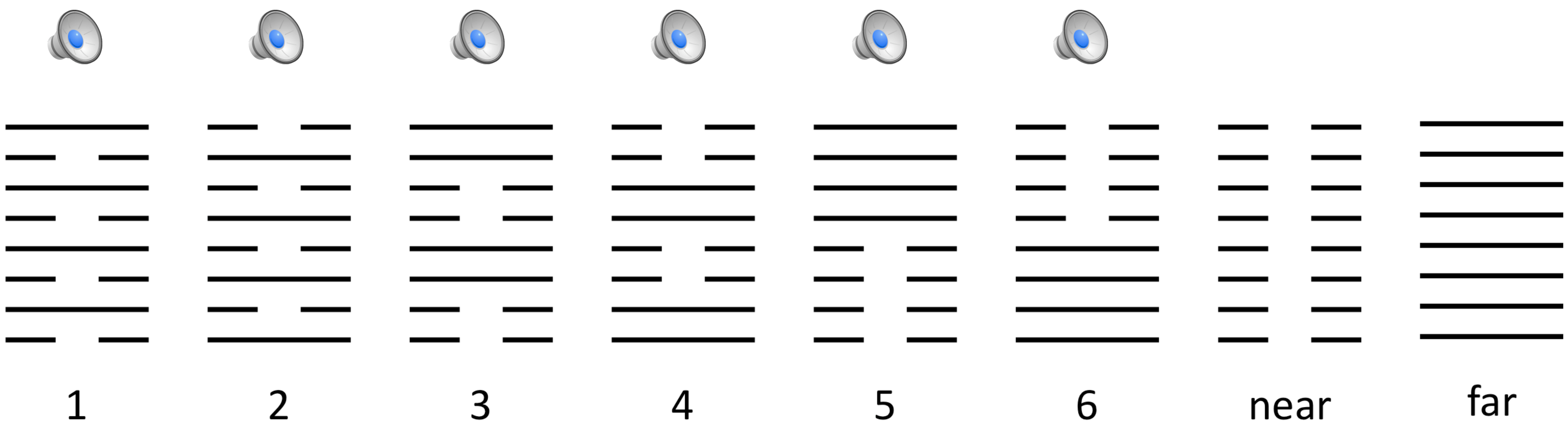
reverse reverb: not significant

X^2 = 3.003378801
p = 0.166182128 (Bonferroni corrected)

<back>

frequency importance

- new data from Watkins' lab using 8-band stimuli
- reverberation applied only on certain bands
- high freq. bands incr. important in sir-stir distinction

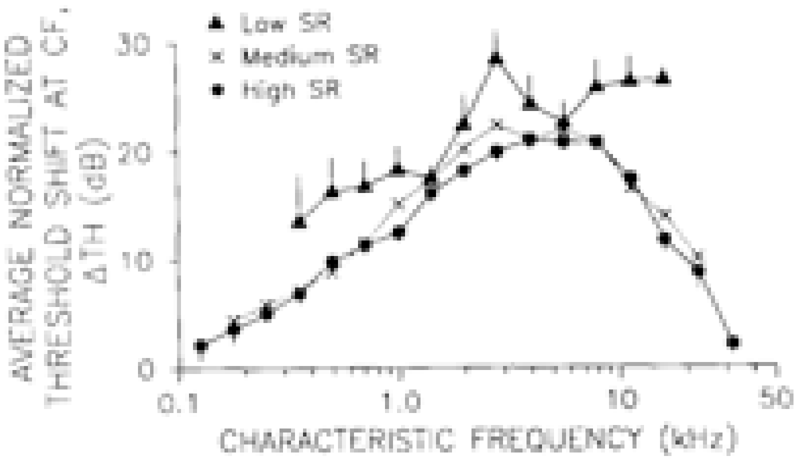


Watkins at al. (2010). *British Society of Audiology.* Abstract #116.

<next>

frequency importance

- frequency sensitivity of efferent system
- physiological data (for a cat, not a human)
  - approx. linear increase in region 100-8000 Hz



Guinan and Gifford (1988). *Hearing Res*, 31, 29-46.

<back>

ASR specifications

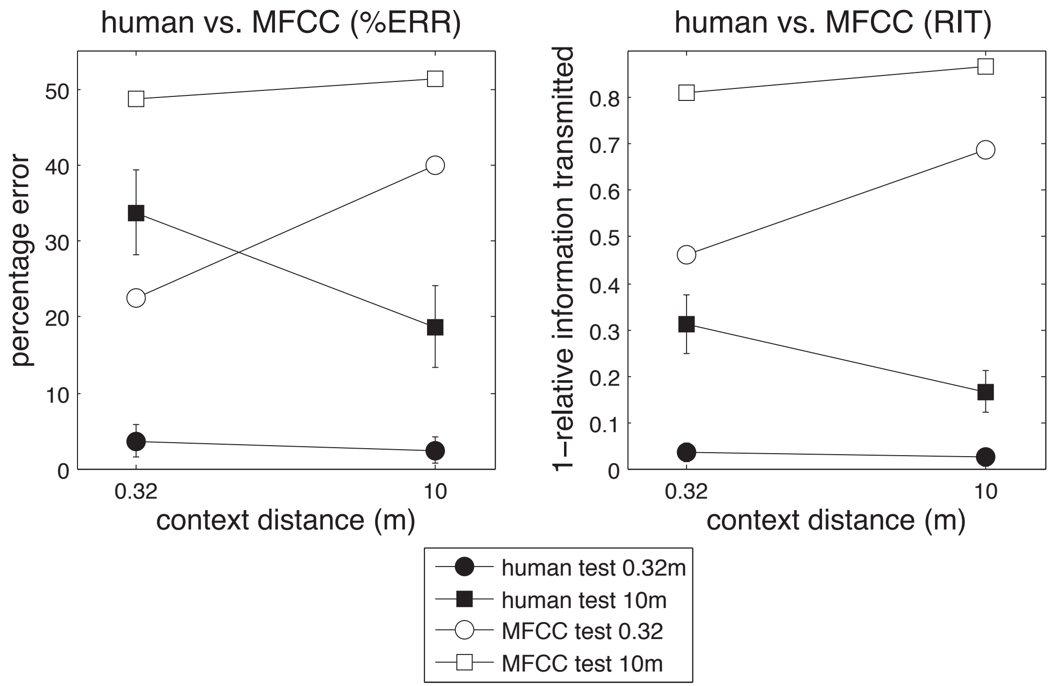
baseline system

- HTK phone recogniser
- 39 monophone models, 20 Gaussian mixtures per state
- CMU pronunciation dictionary: transcripts to phone seq.
- trained on TIMIT corpus, then adapted to AIC
- 12 MFCC features or 13 DCT-transformed auditory features
- semi-forced alignment: context words are known

Brown and Beeston (2010). *British Society of Audiology.* Abstract #117.

<next>

ASR baseline system



- ● human: compensation
- ○ machine: does not

ASR error ∝ amount of reverb on test word

near-near < far-near < near-far < far-far

<next>

ASR specifications

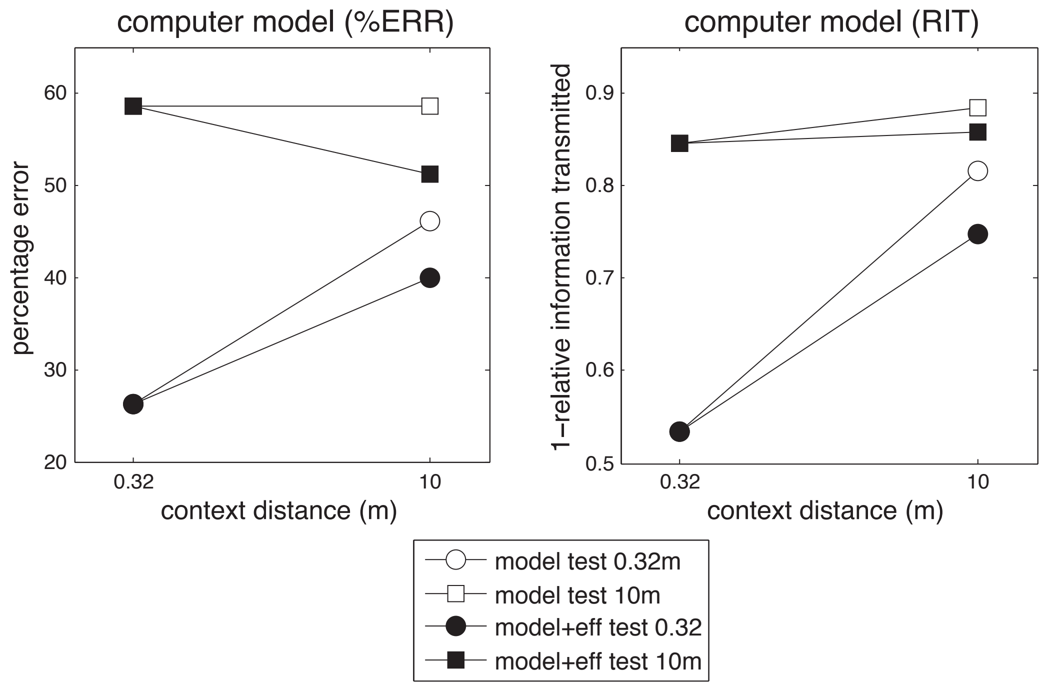
simplified auditory model

- no attenuation is applied at near context distances
- attenuation applied is fixed at 4 dB for far context distances
- model is ‘open’ loop rather than continually updating
- no frequency-dependency features are included

Brown and Beeston (2010). *British Society of Audiology.* Abstract #117.

<next>

ASR simplified auditory model



○ no efferent system: auditory features are similar to MFCC result, no compensation

● with efferent system, 4 dB ATT at far context: a little compensation if viewed with %ERR (but not with RIT)

<back>