Predicting intelligibility of connected speech and singing in adverse listening conditions

Sarah Hawkins

Centre for Music & Science
University of Cambridge

sh110@cam.ac.uk

SPIN 2015, Copenhagen, January 2015

Outline

- Data from experiments on intelligibility of words
  - in ensemble speaking and singing
  - in silence and against competing vocalisation

- Theoretical implications

Antje Heinrich

Antje.Heinrich@ihr.mrc.ac.uk

Future Leaders flare of Ageing Research

Antje Heinrich

MRC Institute of Hearing Research

Sarah Knight

Supported by wellcome trust

The Clerks (dir: Edward Wickham)
http://www.talesfrombabel.co.uk/

Some properties of vocal music

- Sung text is interesting because, especially in ensembles
  - the 'background noise' is part of the signal
  - but not necessarily part of the message
  - messages may compete, or be equally important
- Sung text can vary as much as natural speech does
  - e.g. genre, rhythmic and melodic style within genre
- Singing imposes extra constraints on intelligibility:
  - many contrasts found in speech are neutralised
    e.g. vowel amplitude
    vowel length if it would compromise the musical rhythm
  - styles that aim for constant Timbre and Loudness hinder e.g.
    - they sacrifice vowel quality contrasts by reducing vowel space
    - they may reduce formant definition by avoiding placing VT resonances at multiples of f0
  - f0 may be greater than F1

1. Ensemble speaking

Roger Go to Yellow Three

Musical analogue of CRM (Brungart) test

High predictability because small vocabulary

Number of target voices vs competing background voices

Mix of f0 (gender)

Strong rhythms

Roger Go to Yellow Three

- Spoken ensemble concerts by professional singers
- Very rhythmic CRM 'Brungart test'
  - 8 colours (black, white, yellow, red, purple….)
  - 10 numbers (1-10)
  - 1 target name (Frances)
- unpredictable colour and number combinations
- 1-6 speakers in one or two streams
- Systematically varied
  - # target talkers
  - # competing talkers
  - combinations of gender/pitch

The Clerks, dir. Edward Wickham. Composer: Christopher Fox

http://www.talesfrombabel.co.uk/tales-from-babel/audio/

http://www.bbc.co.uk/music/records/nz52rh

controlled: # singers, in one or two streams

who's singing?						gender combinations
f1	f2	m1	m2	m3	m4	
					m4	m
f1						f
		m1				m
f1					m4	f/m
f1	f2					f/f
		m1			m4	m/m
f1					m4	f/m
f1					m4	f/m
f1	f2					f/f
		m1			m4	m/m
f1		m1			m4	f/m/m
f1		m1			m4	f/m/m
f1	f2				m4	f/f/m
f1	f2				m4	f/f/m
		m1		m3	m4	m/m/m
		m1		m3	m4	m/m/m
f1		m1			m4	f/m/m
f1		m1			m4	f/m/m

orange ground = singing target name, Frances

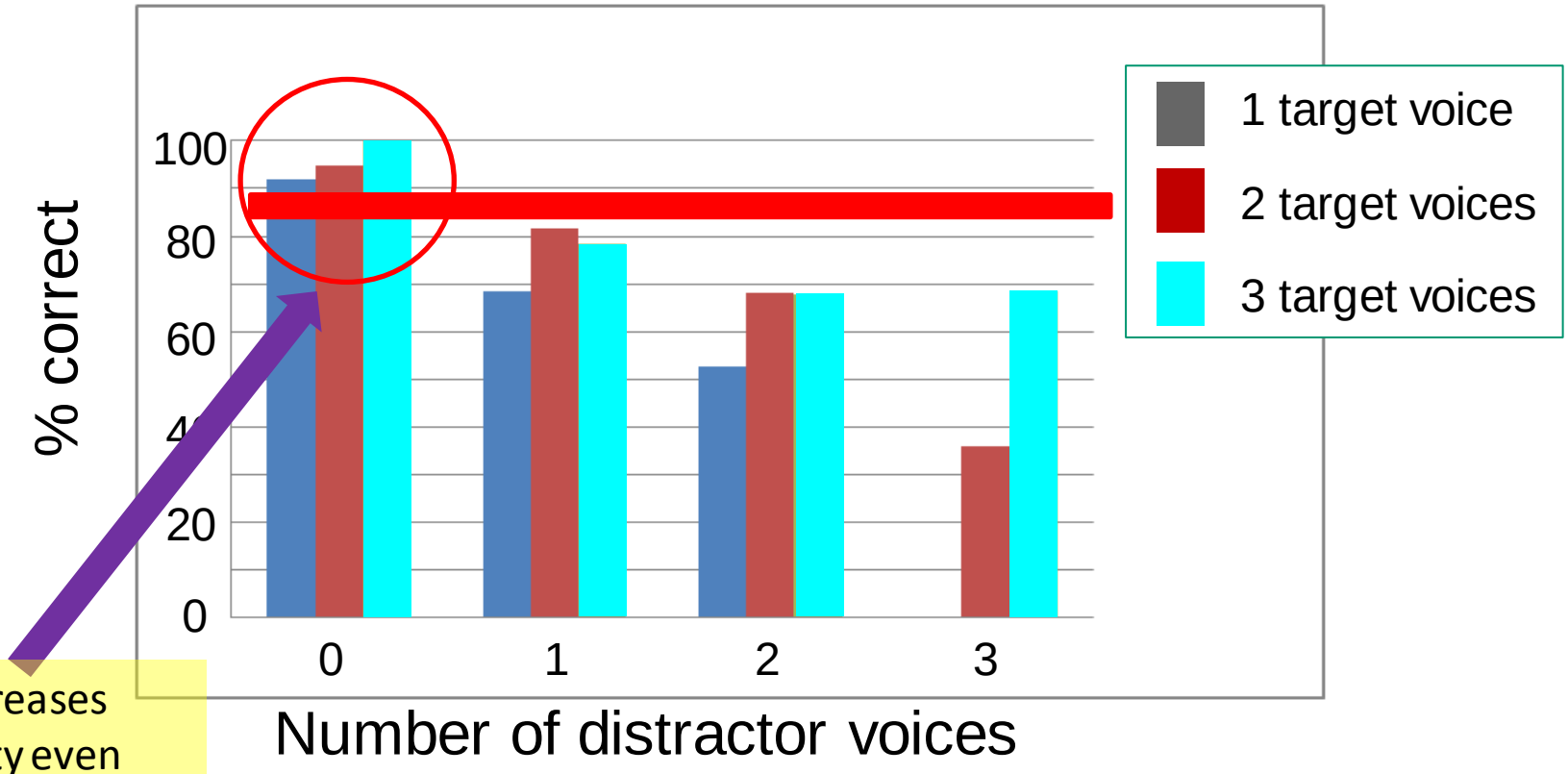
red = a target singer
black = a background singer

f/f/m/m/m/m
f/f/m/m/m/m
f/f/m/m/m/m

and so on….through to everyone singing, in 2 streams of 3 singers each

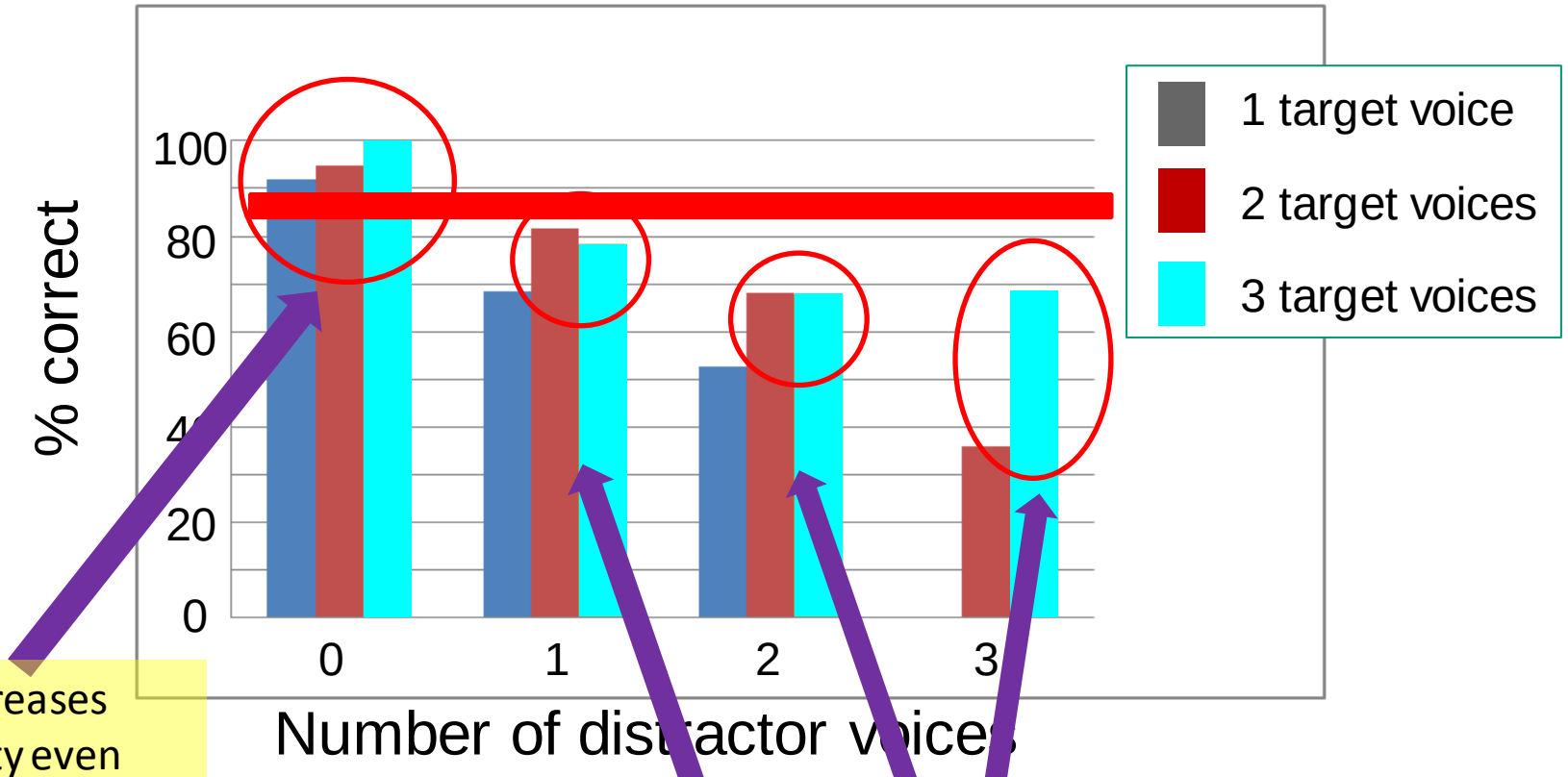
target name: FRANCES because:
- most acoustically distinctive in noise
- most discriminable at higher pitches
- and against the others:
  Matthew, Roger, Harold, Patrick, Chloe

The power of unison



Unison increases intelligibility even in the absence of distractor voices

The power of unison – and of more target than distractor voices



Unison increases intelligibility even in the absence of distractor voices

To ensure or maintain intelligibility, have AT LEAST as many target voices singing in unison as there are distractors

2. Target word predictability

Spoken and Musical analogues of the SPIN-R test

with added variables

Speech: accent of background babble, number of talkers

Music: type of background noise, level (SNR)

2a: Speech intelligibility in babble

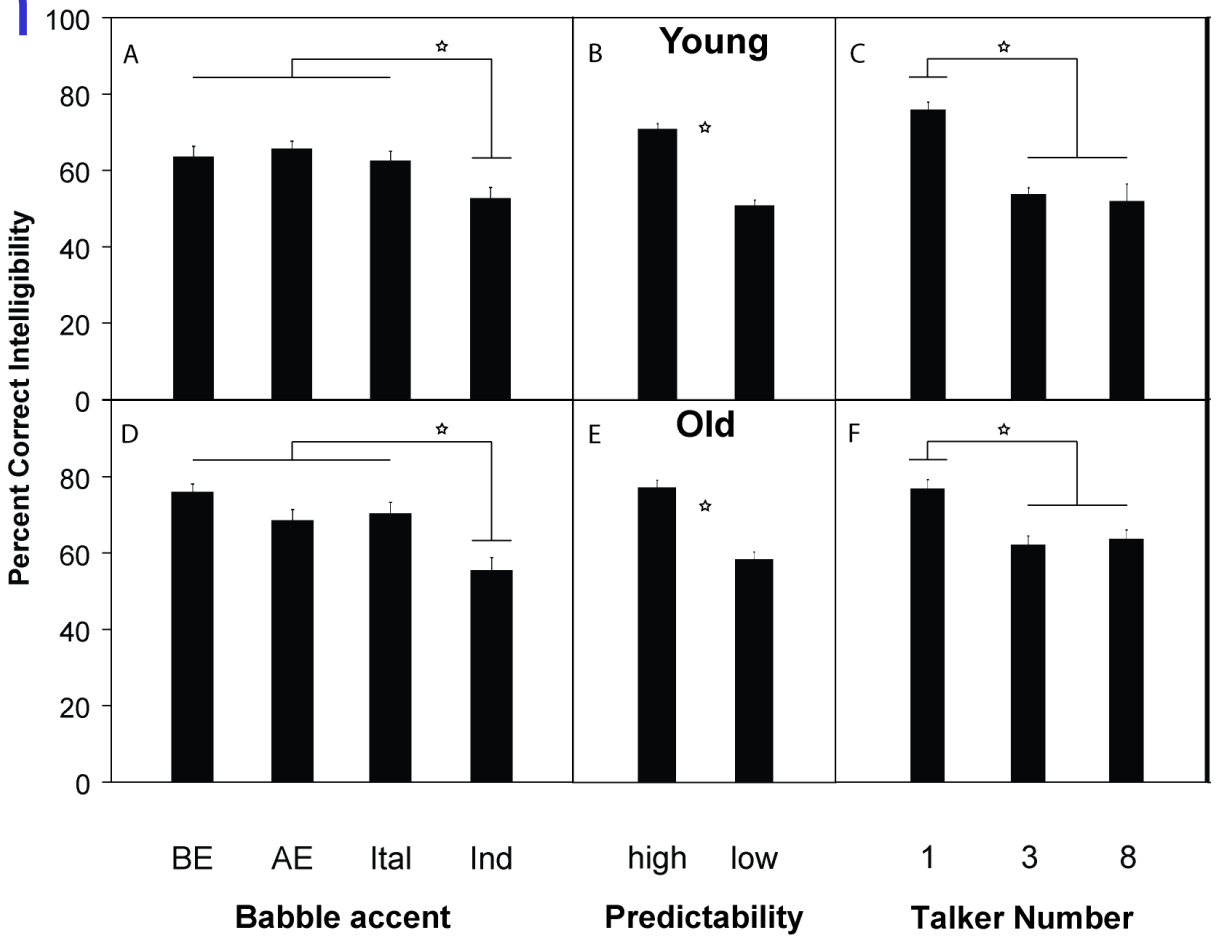
- 36 sentence pairs contrasted predictability of last (key)word
  - as in SPIN test except phonetically controlled
    - *the birds flew overhead in a huge flock*
    - *the boys knew where to look for a huge flock*
  - keywords monosyllabic + least 2 minimal pairs
  - immediate phonetic context of the keyword identical
  - bases unique, natural and meaningful
  - identical numbers of syllables and prosodic structure
  - one male speaker

Korlin Bruhn

Heinrich, Bruhn & Hawkins (2011)

In Algom, Zakay, Chajut, Shaki, Mama, & Shakuf (eds.), *Fechner Day 2011: International Society for Psychophysics 27th Annual Meeting,* 113-118.

2a. Speech in babble



- babble: 1, 3 and 8 male talkers, reading English in 4 accents (British, American, Neapolitan Italian, South Indian)
- normal-hearing listeners, young (21 years) and old (67 years) native speakers of Southern British English

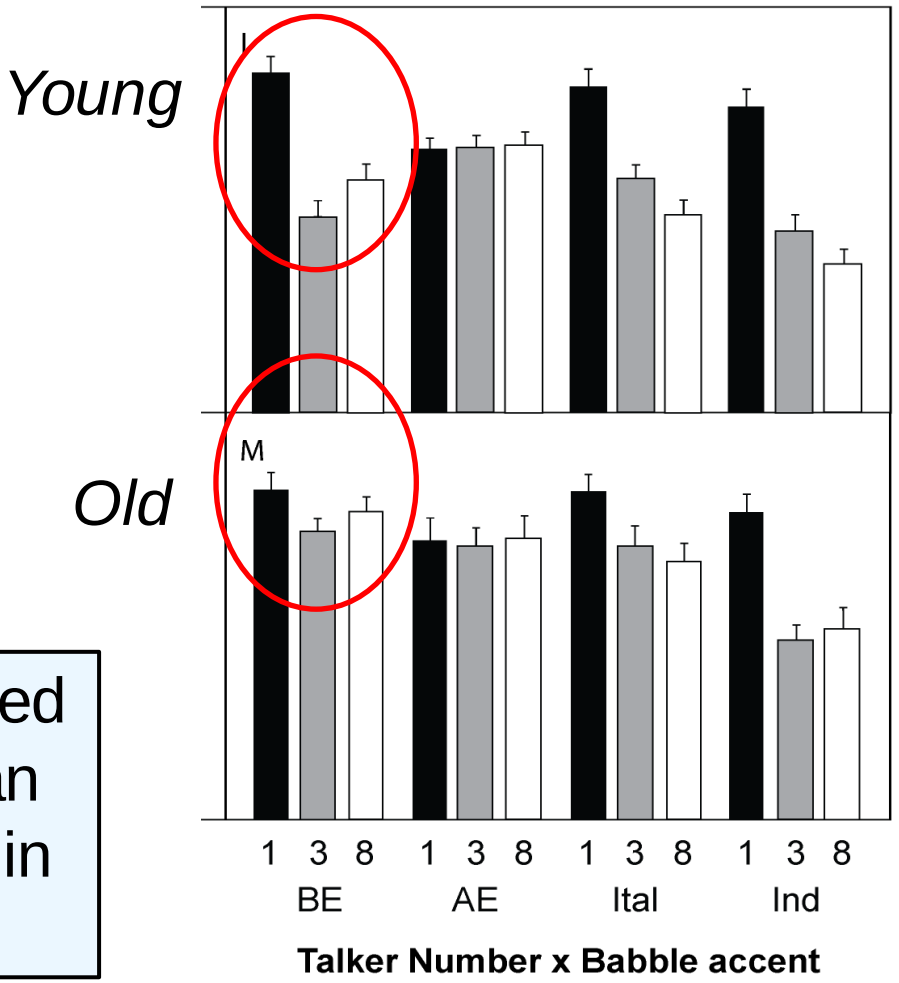
Heinrich, Bruhn & Hawkins (2011)

2a. Speech in babble

talker number x babble accent x age

F[6, 208] = 2.31, p = 0.03

older listeners less affected by number of talkers than young adults, especially in native accent



- babble: 1, 3 and 8 male talkers, reading English in 4 accents (British, American, Neapolitan Italian, South Indian)
- normal-hearing listeners, young (21 years) and old (67 years) native speakers of Southern British English

Heinrich, Bruhn & Hawkins (2011)

2b: Adaptation of test to ensemble singing Tales from Babel, "Test 2"

- a musical analogue of the SPIN-R test: final word predictability
- with added variables:
  - type of background
    - silence
    - spoken babble
    - sung competing vowels (close, dissonant harmony)
    - [ʃ] ('sh')
  - level of background noise (SNR)
  - background noises rotated across 6 live public concerts

- 10 sentence pairs from Heinrich, Bruhn and Hawkins (2011)
- one tenor target singer, 5 background voices (S A T B B)
- no visual cues

www.talesfrombabel.co.uk/tales-from-babel/

The Clerks, dir. Edward Wickham.
Composer: Christopher Fox

2b: Adaptation of test to ensemble singing Tales from Babel, "Test 2"

Procedure

- each sentence sung in silence or against competing 5-voice background (fully crossed over 6 concerts)
- 4 words were projected onto a screen after each sentence

predict-ability	Sentence precursor	Target word	phonetically similar	semantic^{y} plausible	moderate phonetic + semantic
High P	The poor bird's broken its	wing	ring	leg	limb
Low P	I'm sure Lynn spoke of its				

www.talesfrombabel.co.uk/tales-from-babel/

The Clerks, dir. Edward Wickham. Composer: Christopher Fox

2b: Adaptation of test to ensemble singing Tales from Babel, "Test 2"

Procedure

- each sentence sung in silence or against competing 5-voice background (fully crossed over 6 concerts)
- 4 words were projected onto a screen after each sentence
- e.g.

1. limb
2. wing
3. leg
4. ring

numbered

order of choices randomised between trials and across concerts

- audience members indicated which word they thought they had heard: N = 354
- using an electronic voting device (entered its number)

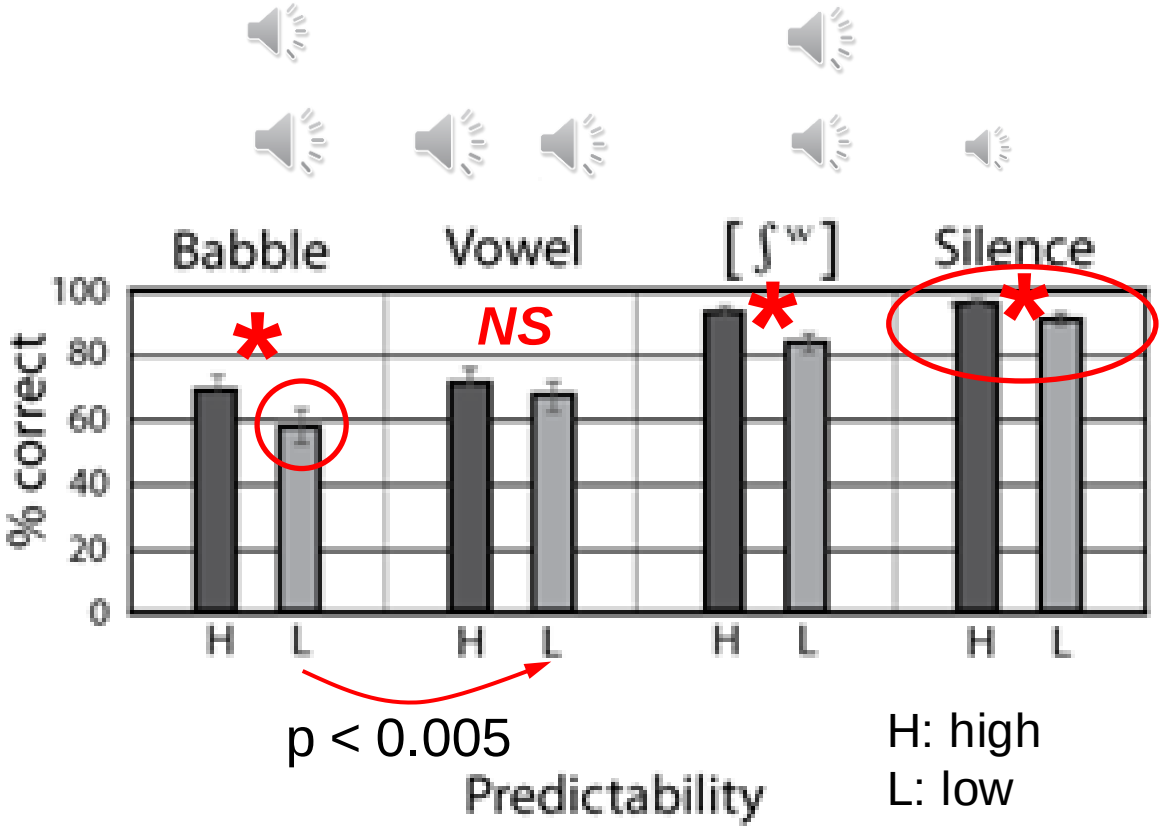
www.talesfrombabel.co.uk/tales-from-babel/

The Clerks, dir. Edward Wickham.
Composer: Christopher Fox

2b. Tales from Babel, “Test 2”

Type of background x Predictability

- good intelligibility in quiet: 96%, 91%
- predictable keyword more intelligible than unpredictable overall
- and for each condition except sung vowels
- low predictability keywords were least intelligible in babble

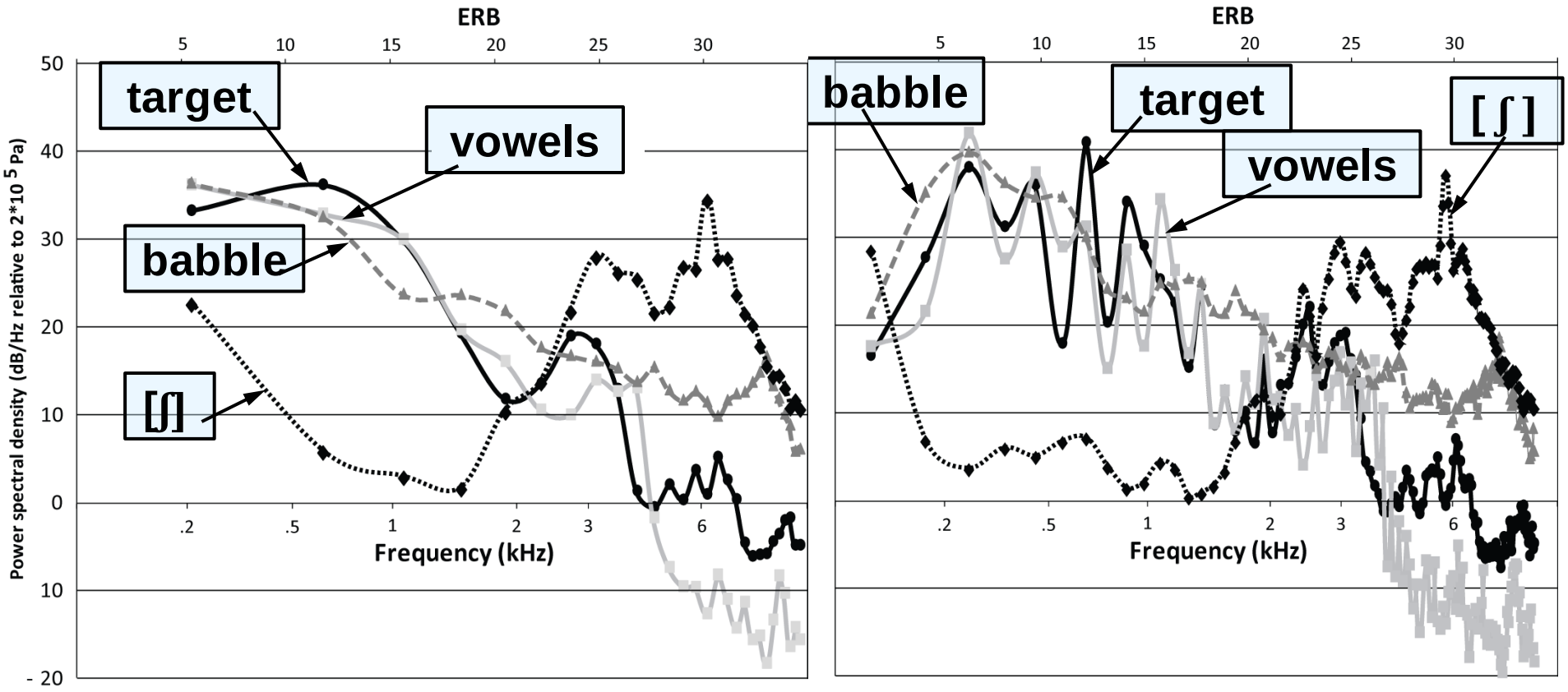


vowel background
ff *pp*

LTAS: target singing and 3 background noises

400-Hz bins

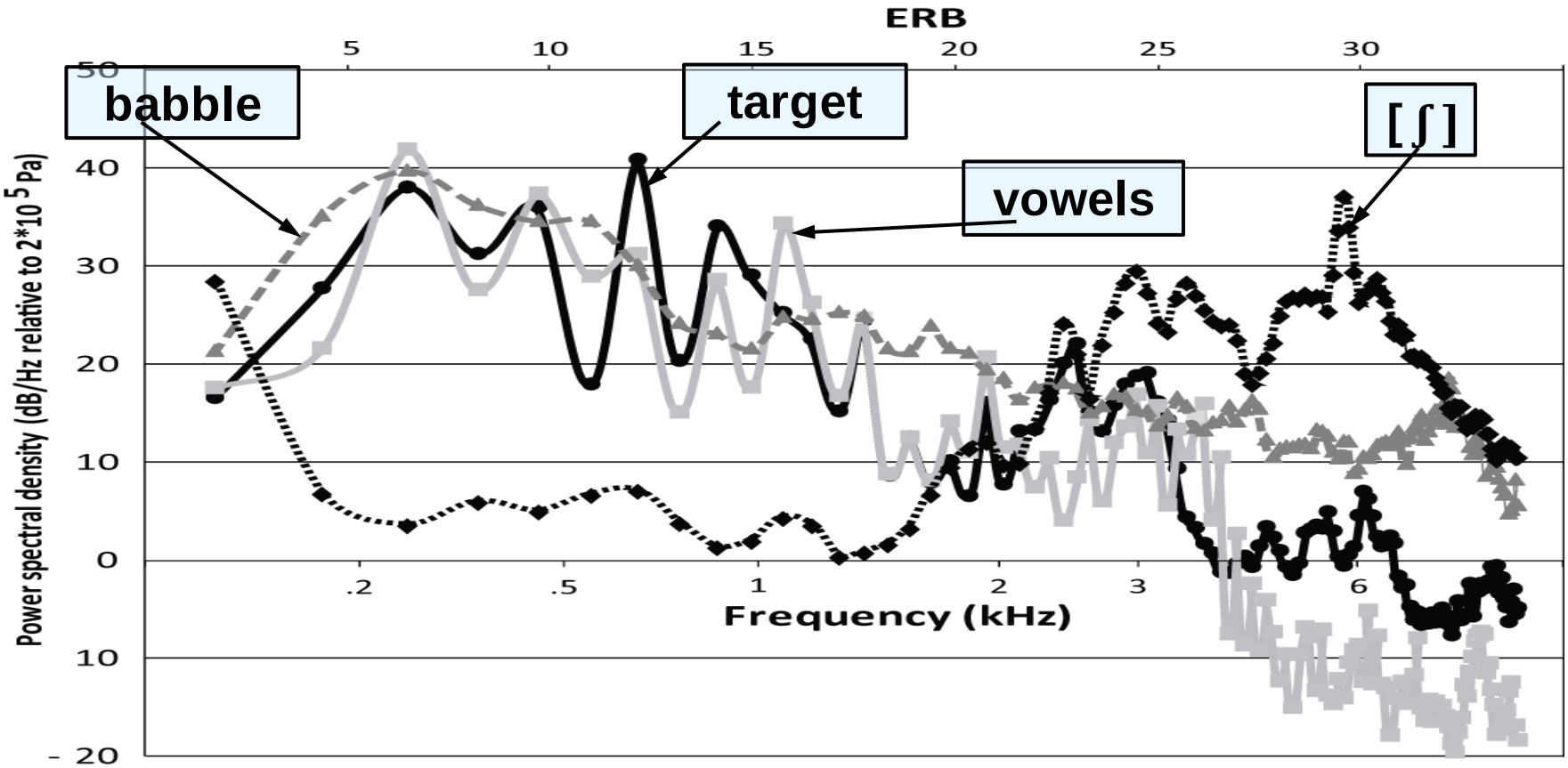
100-Hz bins



babble masks more than sung vowels because babble lacks the vowels' spectral troughs?

LTAS: target singing and 3 background noises

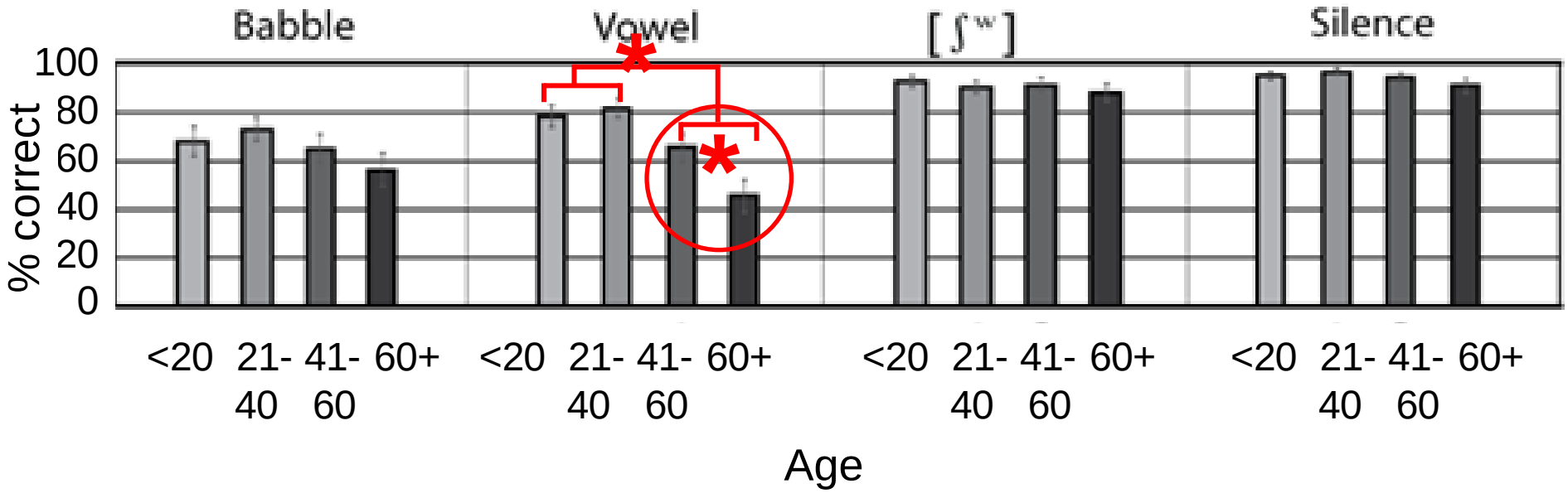
100-Hz bins



babble masks more than sung vowels because babble lacks the vowels' spectral troughs?

2b. Tales from Babel, “Test 2”

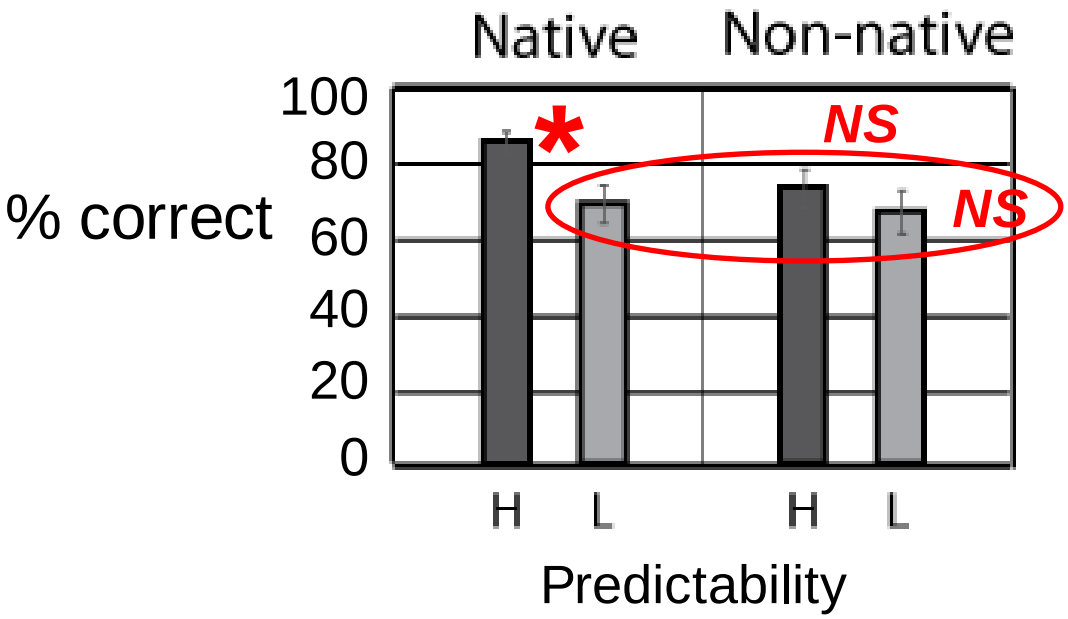
Type of background x Age



- > age 40 disproportionately bad in vowels
  – not better than in babble ($p > 0.05$)

2b. Tales from Babel, “Test 2”

Predictability x native language status



- native speakers benefitted from predictable words
- non-native speakers did not

Summary: Tales from Babel Test 2

- Well-established non-phonetic influences on spoken word intelligibility had much the same effects on sung word intelligibility, with a single professional target singer and music intended to allow intelligibility

- Predictability strongly influential for native speakers
  (not for non-native speakers)

- Good intelligibility in Silence & the poor noise masker [ʃ] (sh)

- 5-talker babble is a good masker of sung target words
- at least compared with vowels sung on similar pitches as the target singing

3. What (else) makes sung words intelligible?

Phonetic properties of the words:

- in polytextual settings

Musical properties:

- genre
- consonance vs dissonance

Funded by University of Cambridge Newton Trust

3 polytextual word-monitoring experiments

Genres

- Polyphonic (3 voice parts A T B) - 3 expts
- Lively jingle e.g. like in some advertising (S A B)

Phonetic contrasts

- "acoustic contrast" — spectral change, mainly f0 (dis)continuity
- phonological vowel length — long, short (tense, lax)
- vowel quality

high front	low/central	(high) back
[i eɪ ɪ ɛ]	[ɑ ɜ a]	[u ɔ ɒ ʌ]
heed hay hid head	hard heard had	who hoar hot hut

Polytextual polyphonic vocal music

- Polyphonic singing, esp. polytextual, is like listening to speech against a background of other talkers
  - but when each talker can be equally important ('mutual informational masking' ?) -- or not!
- All factors that influence speech intelligibility apply to singing
  - Spatial location & Rhythmic patterns can help or hinder
  - Pitch and Timbre differences are likely to be helpful
  - Seeing the singers' mouths helps
  - Knowing the words helps (enormously)
  - can probably only stream one source against one other, at best (but the other may be composite of many voices)

3. Texts and general method

- Word monitoring: 24 monosyllabic animal names, well-known, familiarised, and available during the test
- In nonsense sentences, all content words monosyllabic

with a red my lack toe buys **chick** on to peach aisle
tack peel up the bleak to his tan hill to ram and piece

heat mould to wolf by tea stock fig through years true
whale tack for loam with mire

- One keyword per critical sentence (other animal words too)
- Listen to top voice, type all animal words you hear
- Young normally-hearing, native-English Ps

3. Texts and general method: structure

- Top voice (Alto): critical sentence
- Middle voice (Tenor): competing sentence
- Lowest voice (Bass): hums

- 3 competing sentence conditions:
  - Odiff different onset
  - Vdiff different vowel
  - Cdiff different coda

whale

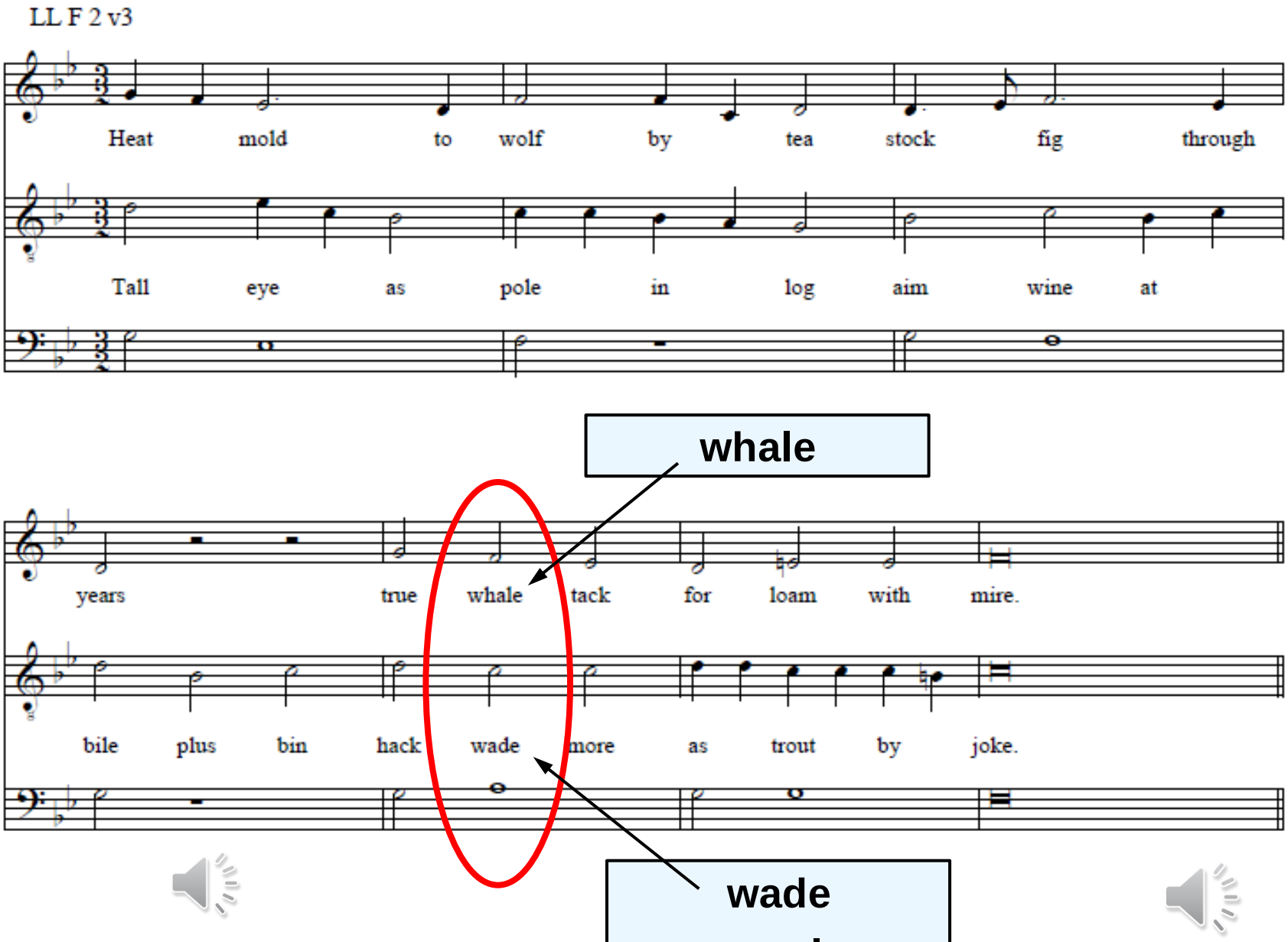
pale Onset

wh**ee**l Vowel

wa**de** Coda

Alto: heat mould to wolf by tea stock fig through years true
whale tack for loam with mire

Tenor: tall eye as pole in log aim wine at bile plus bin hack
pale / wheel / wade more as trout by joke



Expt 3a: normal blend

Expt 3b: mic 1 only

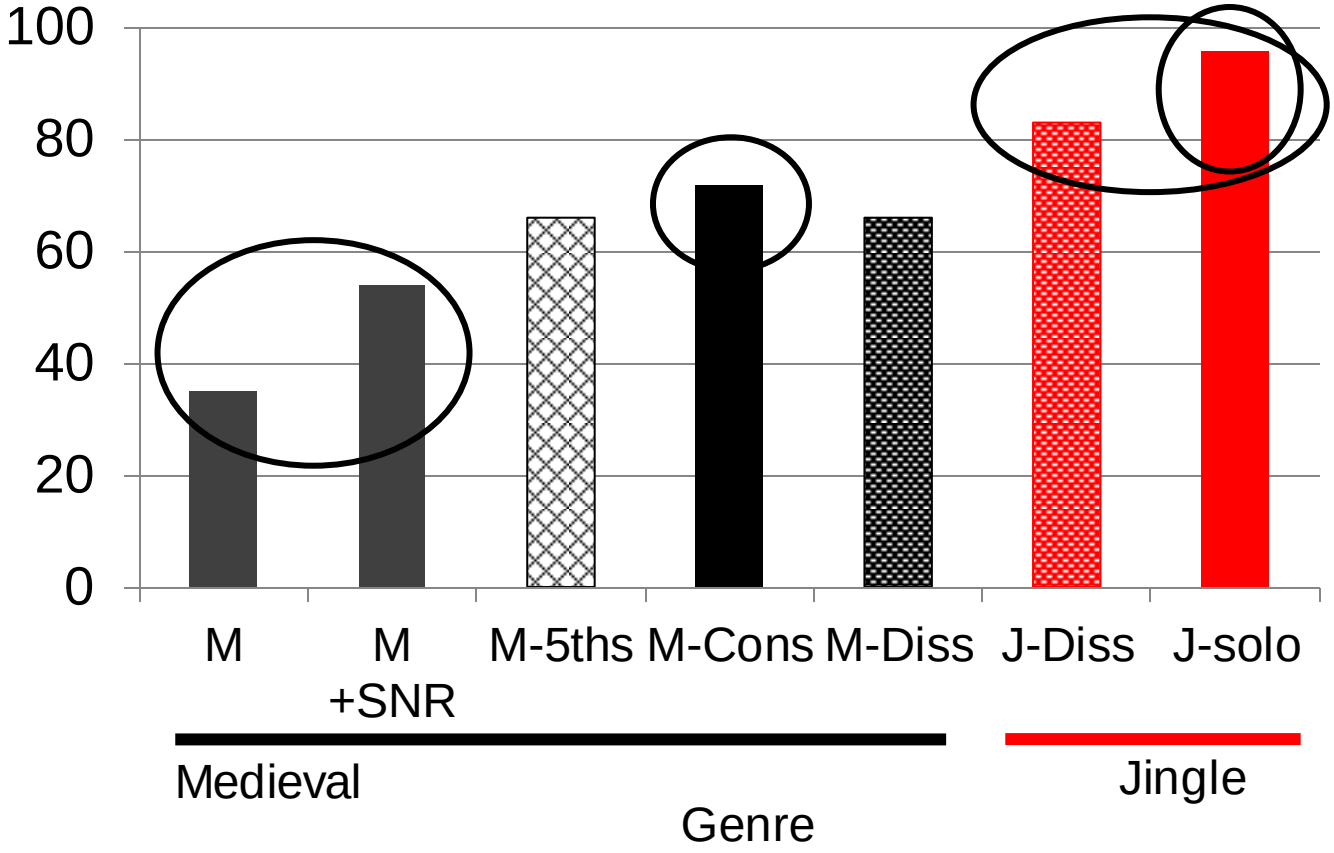
Polytextual expts

Kate Honey

Genre	E#	Timbre	Voices	Harmony condition	% √
Medieval	3a	blend	A T B	standard simple	**35**
	3b	high SNR			**54**
motet (polyphony)	3c	blend	A A B A =	‘5ths’ B-A 8; B-T 5	**66**
				Cons B-T 3; T-A 4	**72**
				Diss B-T 4; T-A tritone	**66**
Jingle		clear ----	same singer	Diss B-T 4; T-A tritone	**83**
				Solo	**96**

- no part crossing, no melisma on critical word, control where the critical word falls in the phrase and bar; etc
- each P only heard each word once, in one of the 3 conditions
- Expt 3c had only one competitor word condition: Vdiff.

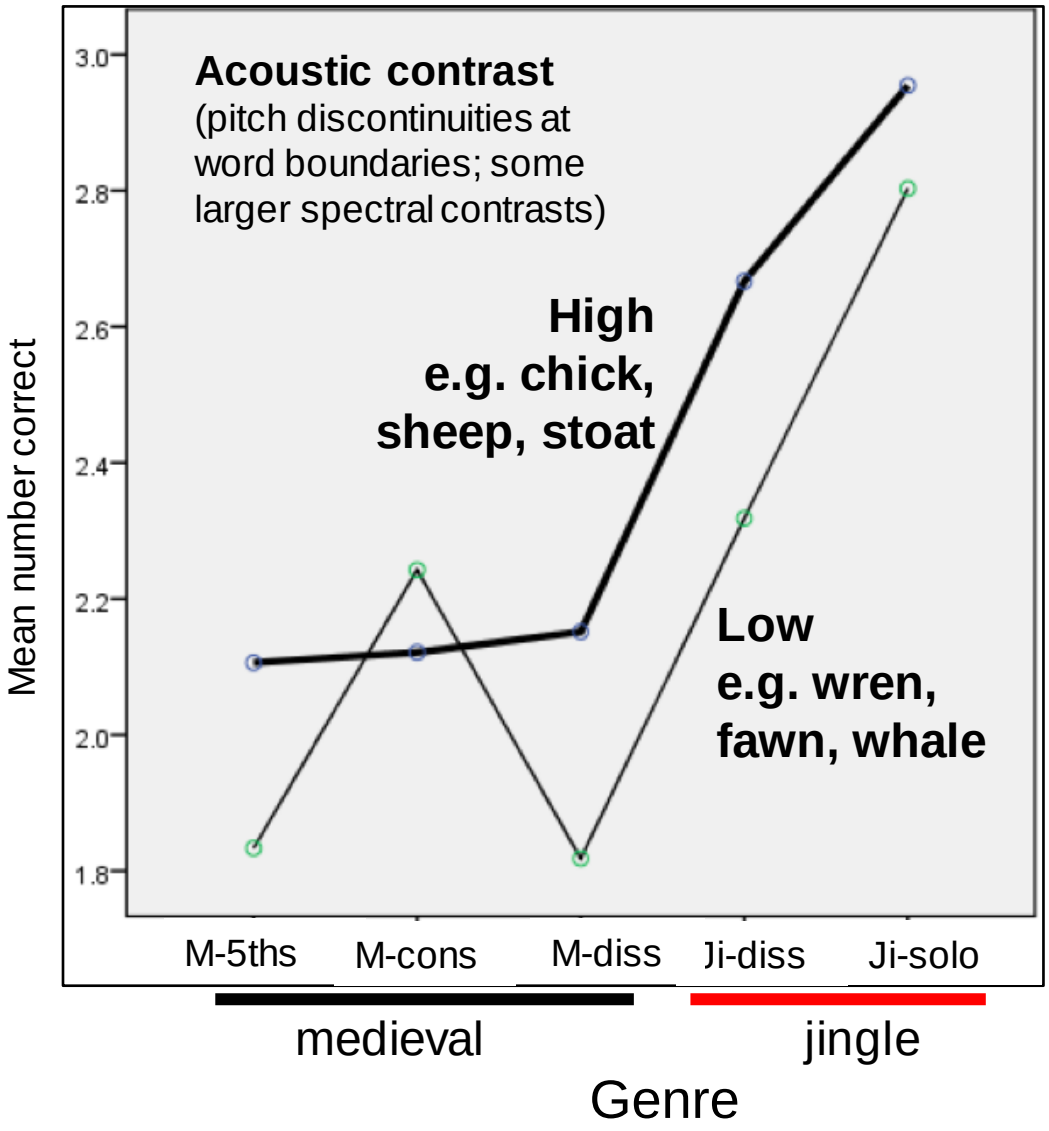
Mean % correct: each condition, each expt



- rests between words help (rhythm and pitch discontinuities?)
- anomalous text can be very intelligible in the right conditions (no competing text, jingle style, know what words you're listening for)
- medieval motets are pretty unintelligible even at good SNR! Even when you aim for very clear diction? 'Vocal blend'
- harmonic consonance may help

Phonetic parameter: Acoustic contrast

- Consistently strongly significant in Expts 3a, 3b, 3c, and a 4th expt on Lieder style (2 male voices, + piano)
- Most benefit of acoustic contrast in the 2 dissonant conditions
- Consonance effect may or may not be reliable



Other phonetic parameters

- No other phonetic parameter has a robust effect
- Effects may be significant but not consistent e.g.
  - Expt. 3a & 3c: long vowels more intelligible than short vowels
  - but Expt 3b (better SNR): long vowels = short vowels
  - a 4th expt (Lieder style, 2 voices + piano): short vowels > long

- Vowel quality differences occur, but there are many interactions: no interpretable patterns that are reliable across experiments

Conclusions: from lab experiments: genre, harmony, and phonetics

- Phonetics contributes
  - acoustic discontinuities (high acoustic contrast) easier: f0 and amplitude envelope affecting pitch continuity and rhythm – when at word boundaries
  - long vowels may be easier (more suitable to singing?)
- Musical genre is most important
- Harmonisation needs more work:
  unison > consonance > dissonance?

weak converging evidence

- Overall, musical properties seem more powerful than phonetic ones, though well-chosen phonetic properties seem likely to enhance the power of musical choices

Theory

and speech perception

What comes out of these findings?

- Speech rhythm (pitch-duration-intensity relationships) and amplitude envelopes are really important
  - word segmentation errors in connected speech-in-noise suggest the same thing
  - allophonic differences between syllable Onset and Coda consonants too
    - **we'd read the reviews** **we dread the reviews**

- Much other evidence that segmental timing is central to connected speech intelligibility, even not in noise, e.g.:

"Standard" segment durations - for stressed monosyllables spoken in isolation

Context-sensitive segment durations

fairly unintelligible

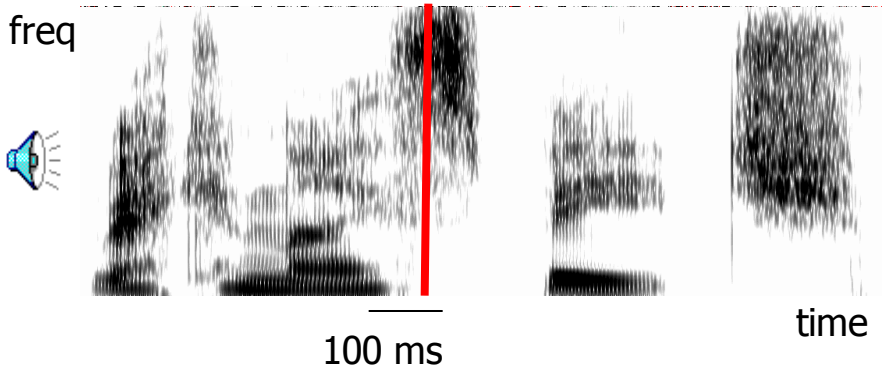
Klatt synthesis c. 1979

much more intelligible

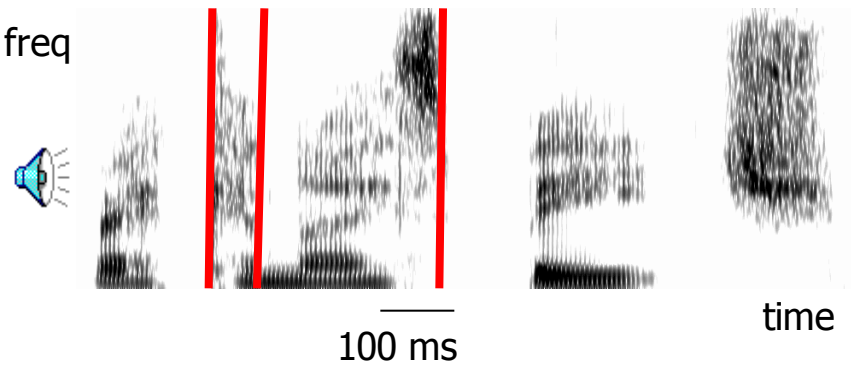
Local phonetic detail indicates (1) word boundaries

≈ perceiving discrete words from a 'continuous' signal

recognise speech



wreck a nice beach



/rɛkənaɪzspitʃ/

/rɛkənaɪsbitʃ/

What comes out of these findings?

- To model this, we have to include time explicitly: rhythm and timing: the crux of speech perception
- Neither phonemes nor allophones can do this
  speech rhythm is hierarchically organised, just as beats/pulses group into higher-order metrical rhythm in music

* Hawkins (2011). Phonetic perspectives on modelling information in the speech signal. *Sādhanā, 36*(5), 555-586.

What comes out of these findings?

- Rhythm may be the most basic requirement for accurate, effortless speech perception:
  - pattern perception structured around short-domain beats organised into long-domain metrical structure
  - distinctions of spectral detail vary systematically with these rhythmic distinctions e.g. internal acoustic structure
    - discolour (prefix) – much periodicity
    - discover (non-prefix) – much aperiodicity
  - + regions of high spectral certainty
    ‘anchor’ the matching process
    the patterns relate to all aspects of meaning and function of the communication

Hawkins (2010) *J. Phonetics 38, 60-89*
Hawkins (2010). Fougeron et al. (eds) *Laboratory Phonology 10.* de Gruyter.
Hawkins,(2011) Does phonetic detail guide situation-specific speech recognition? *ICPhS.*
Hawkins, S. (2014) in Smith et al. (eds) *Communicative Rhythms in Brain and Behaviour*. Phil. Trans. Royal Soc. B 20130398.

Systematising rhythm

- Hierarchical structure, details created via enculturation (native language vs foreign language…)

- Related to (at least)
  - pragmatic (and probably interactional) function
  - linguistic structural function (morphemes, pronouns, conjunctions etc – each operates its own system for dealing with the 'same' sounds in different contexts – the systems differ because their functions differ)

- One approach to doing this: Firthian Prosodic Analysis (FPA)

Details of references available on request: FPA authors include: Sarah Hawkins, Rachel Smith, Katharine Barden; Richard Ogden; John Local; Gareth Walker. See also Mirjam Ernestus; Harald Baayen

Rhythm perception is Not Passive

- We readily interpolate elements, even when onsets are missing in the music

- and even when there is “rhythmic interference”

- Regular rhythms are constructed, in both speech and music. Active creation of metrical components that may not be in signal; due to nonlinear coupling; learned

Examples courtesy Justin London
Muddy Water: Hootchie Cootchie man:

Rhythm perception & entrainment

- People entrain endogenous rhythms to external rhythms

- Entrainment enhances temporal sensitivity & **prediction**
  - produces **phase synchronisation** of (e.g.) theta waves between interacting dyads in playing music
  - produces **enhanced, shared periodicity** between accented syllables at ends of Questions and first accented syllable of their Answers
  - seems to transfer seamlessly between conversational speech and the pulse of improvised music – even amongst non-musicians

Sänger, Müller & Lindenberger (2012) *Frontiers in Human Neuroscience, 6*.
Ogden & Hawkins, 2014
Hawkins, Cross & Ogden, 2013

Speculation: This view is consistent with

- Poorer intelligibility for foreign listeners, and people unfamiliar with a regional accent, in any form of adverse listening situation
- Current views that adults given cochlear implants use the input signal to stimulate auditorily richer memories
- Comodulation findings?
- Informational masking influences intelligibility more when the physical signal is fairly clear – not in really adverse conditions, because the details of patterns will be uncertain or unavailable

Active construction in speech perception: a natural "pop-out" matching physical signal with known patterns

What does a pilot do?

Child, 2½ years

this slide is a reminder of a demo – the answer was given in the talk, but future demonstrations will be spoiled if it is given here. See Hervais-Adelman's paper for more on pop-out.

when you know what to listen for, meaning pops out, words sound 'real'

e.g. Hawkins (2010) J. Phonetics 38, 60-89