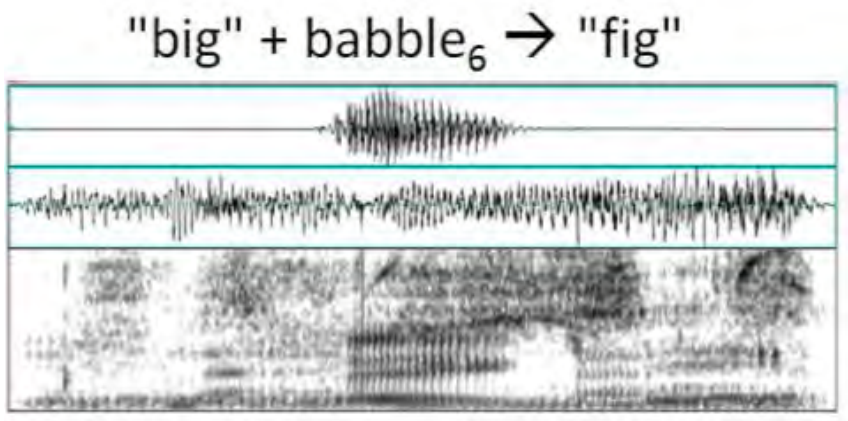
Noise-induced slips of the ear

Martin Cooke, María Luisa García Lecumberri
& Attila Máté Tóth

Ikerbasque (Basque Science Foundation)
Language & Speech Lab, University of the Basque Country, Spain

ERRARE 2013, Ermenonville

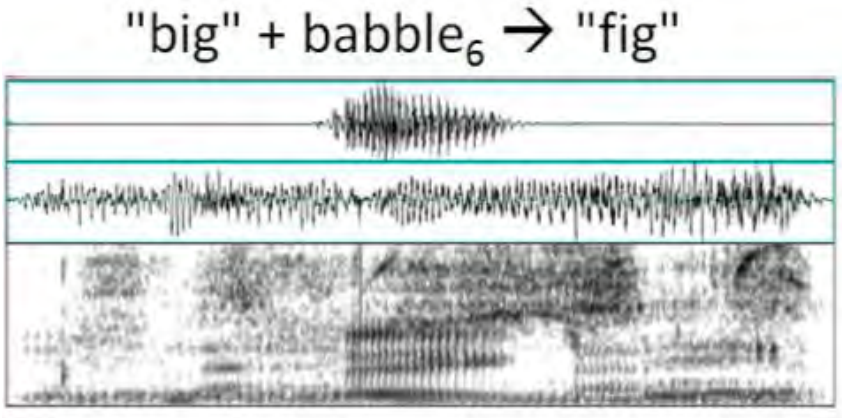
Errare Humanum Est . . .



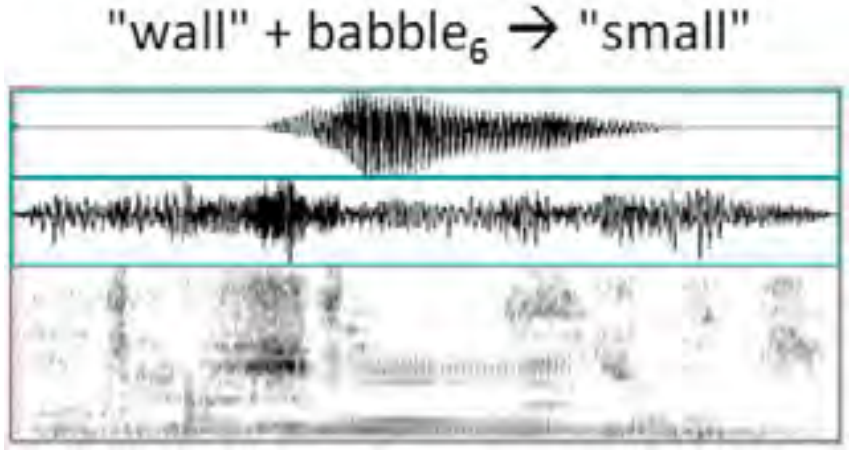
big $\mapsto$ fig

Cooke (2009), Interspeech

Errare Humanum Est . . .



big ↦ fig



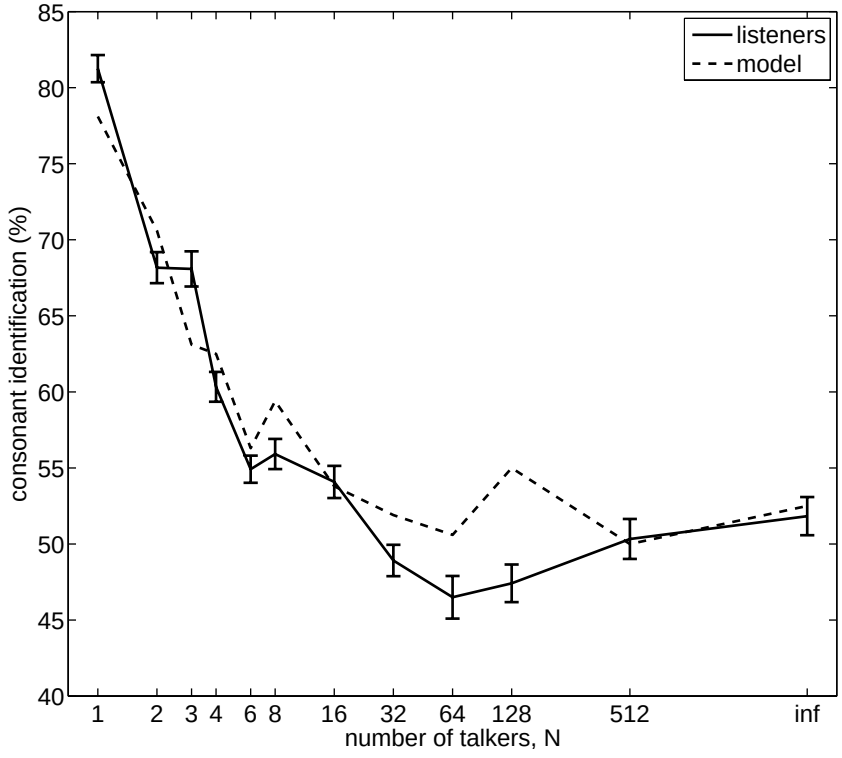
wall ↦ small

Cooke (2009), Interspeech

Key idea: Robust listener ‘errors’ are hugely informative in designing and evaluating computational models of human speech perception in realistic conditions

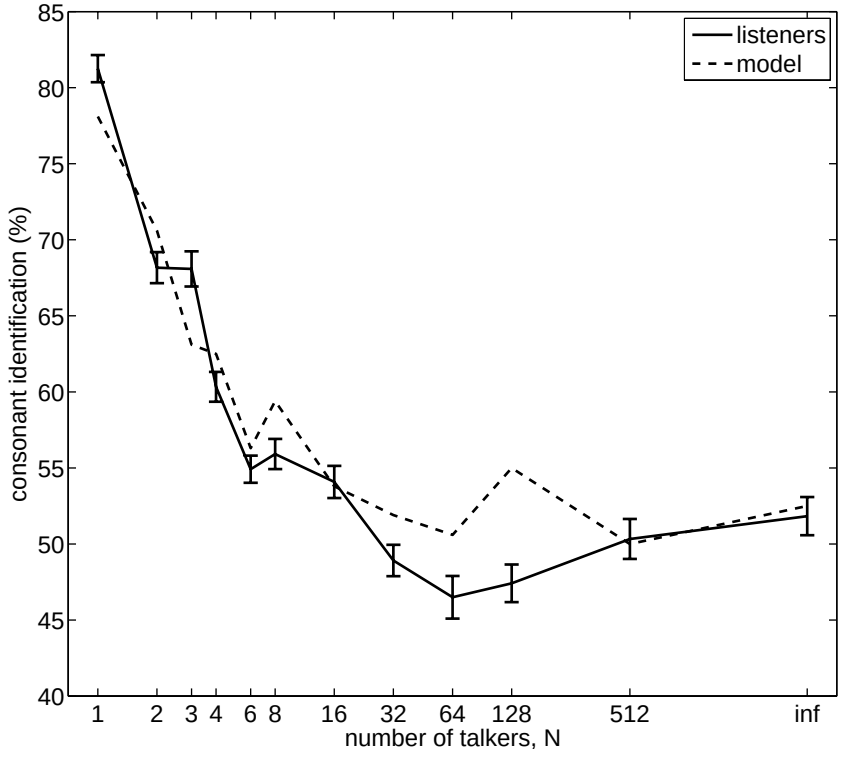
Why we need better models

Macroscopically good . . .



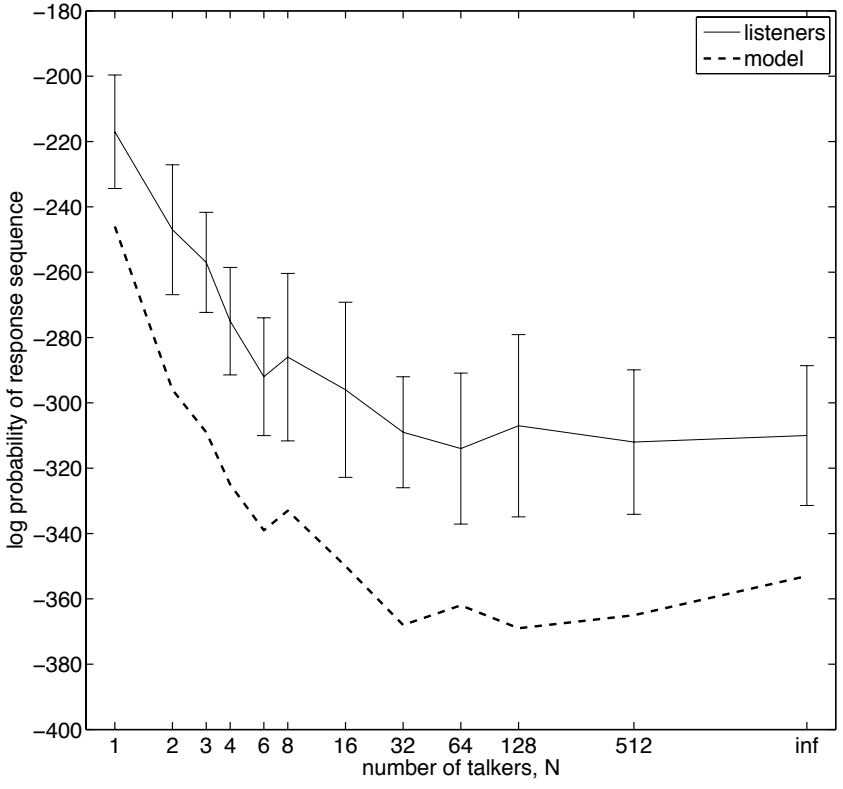
Why we need better models

Macroscopically good . . .

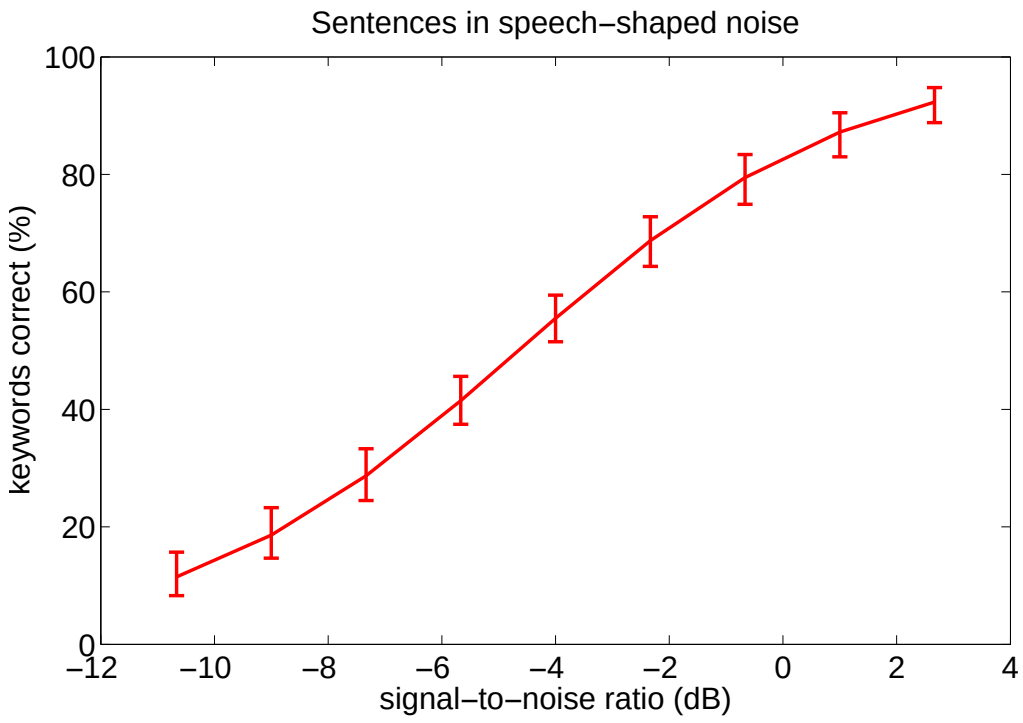


Glimpsing model (Cooke, 2006)

. . . Microscopically bad!



Mean intelligibility: Easy to fit with few parameters



Microscopic reality is harder to model

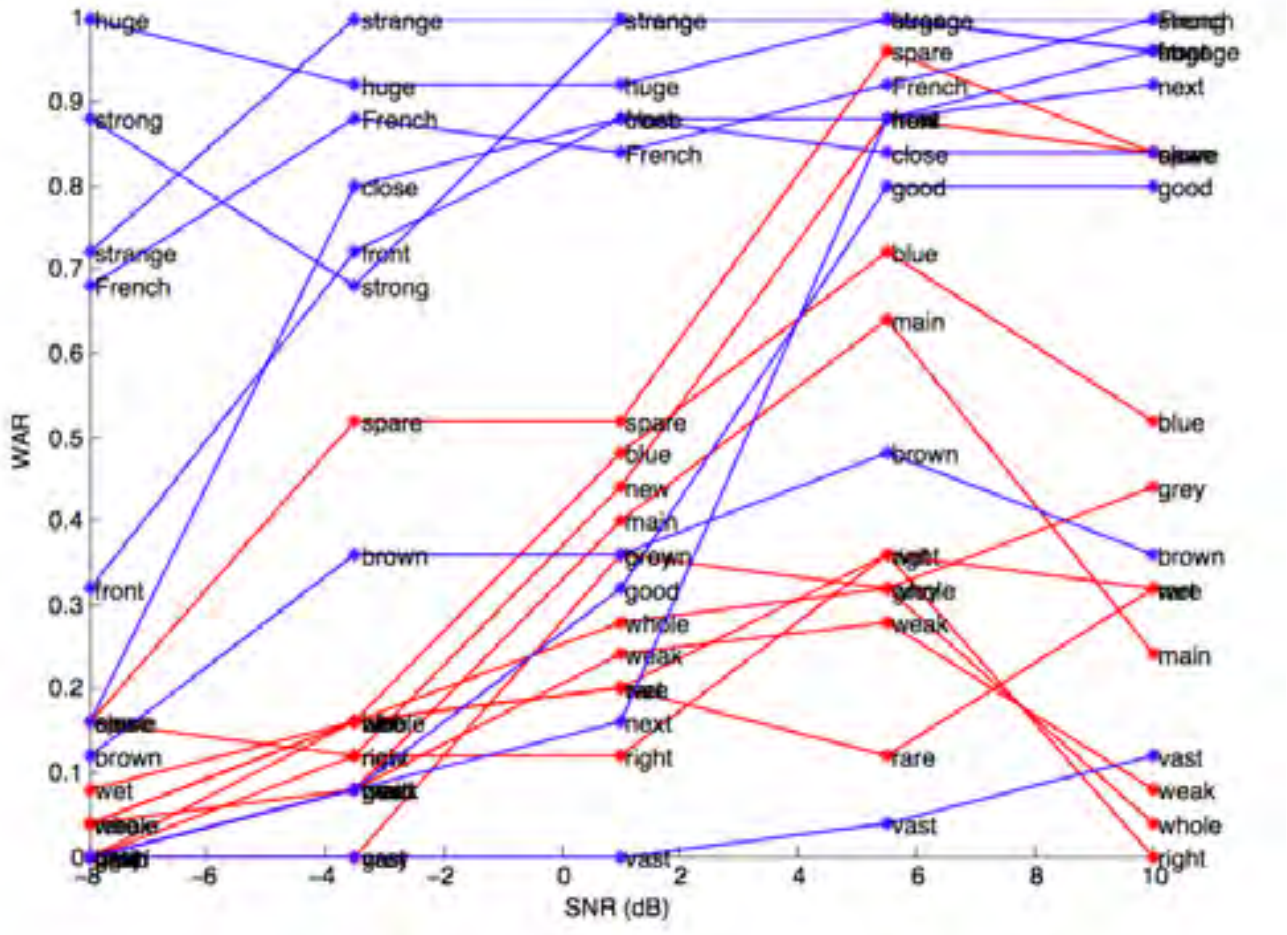


fig from Cassia Valentini-Botinhao

Outline

“Slips of the ear”

- First person reports of spontaneously-occurring speech misperceptions in the ‘wild’
  e.g., Garnes & Bond, 1975; Cutler, 1981; Cutler & Henton, 2004; Bond, 2005; Tang & Nevins, 2012

“Slips of the ear”

- First person reports of spontaneously-occurring speech misperceptions in the ‘wild’
  e.g., Garnes & Bond, 1975; Cutler, 1981; Cutler & Henton, 2004; Bond, 2005; Tang & Nevins, 2012
  - “Jean-Jacques [Rousseau/Cousteau]”

“Slips of the ear”

- First person reports of spontaneously-occurring speech misperceptions in the ‘wild’
  e.g., Garnes & Bond, 1975; Cutler, 1981; Cutler & Henton, 2004; Bond, 2005; Tang & Nevins, 2012
  - “Jean-Jacques [Rousseau/Cousteau]”
  - “how the [ducks/dutch] avoid going down the waterfall”

"Slips of the ear"

- First person reports of spontaneously-occurring speech misperceptions in the 'wild'
  e.g., Garnes & Bond, 1975; Cutler, 1981; Cutler & Henton, 2004; Bond, 2005; Tang & Nevins, 2012
  - "Jean-Jacques [Rousseau/Cousteau]"
  - "how the [ducks/dutch] avoid going down the waterfall"
- Problems: (i) N = 1 (usually); (ii) neither acoustic recordings nor context typically available for replication and further study

"Slips of the ear"

- First person reports of spontaneously-occurring speech misperceptions in the 'wild'
  e.g., Garnes & Bond, 1975; Cutler, 1981; Cutler & Henton, 2004; Bond, 2005; Tang & Nevins, 2012
  - "Jean-Jacques [Rousseau/Cousteau]"
  - "how the [ducks/dutch] avoid going down the waterfall"
- Problems: (i) N = 1 (usually); (ii) neither acoustic recordings nor context typically available for replication and further study
- Solution: elicit in the lab using e.g. faint speech (Cutler & Butterfield, 1992) or time-compressed speech (Vitevich, 2002); Elicitation in noise has been seen as a confounding factor.

"Slips of the ear"

- First person reports of spontaneously-occurring speech misperceptions in the 'wild'
  e.g., Garnes & Bond, 1975; Cutler, 1981; Cutler & Henton, 2004; Bond, 2005; Tang & Nevins, 2012
  - "Jean-Jacques [Rousseau/Cousteau]"
  - "how the [ducks/dutch] avoid going down the waterfall"
- Problems: (i) N = 1 (usually); (ii) neither acoustic recordings nor context typically available for replication and further study
- Solution: elicit in the lab using e.g. faint speech (Cutler & Butterfield, 1992) or time-compressed speech (Vitevich, 2002); Elicitation in noise has been seen as a confounding factor.

Here: elicit errors arising from the speech-masker interaction.

Speech materials

3962 Spanish words

- spoken in isolation
- high-frequency
- 1-3 syllables

Recorded by 2 male and 2 female talkers

Maskers

masker	SNRs (dB)	informational	fluctuating
Speech-shaped noise (SSN)	-4 to -7		
Speech modulated noise (BMN1)	-7 to -13		✓
3-talker babble modulated noise (BMN3)	-3 to -8		✓
4-talker babble (BAB4)	+1 to -3	✓	✓
8-talker babble (BAB8)	+1 to -4	✓	✓

All constructed from the Spanish speech material

Listeners & procedure

- 143 normal-hearing young adults studying at the Universidad del País Vasco
- monolingual in Spanish or bilingual in Spanish/Basque
- tested in sound-attenuated recording studio
- screened up to 20 blocks of 100 stimuli in two non-contiguous 1 hour sessions
- responded by typing response in custom Java applet

Online token pruning

Issue: robust confusions are actually quite rare

Online token pruning

Issue: robust confusions are actually quite rare

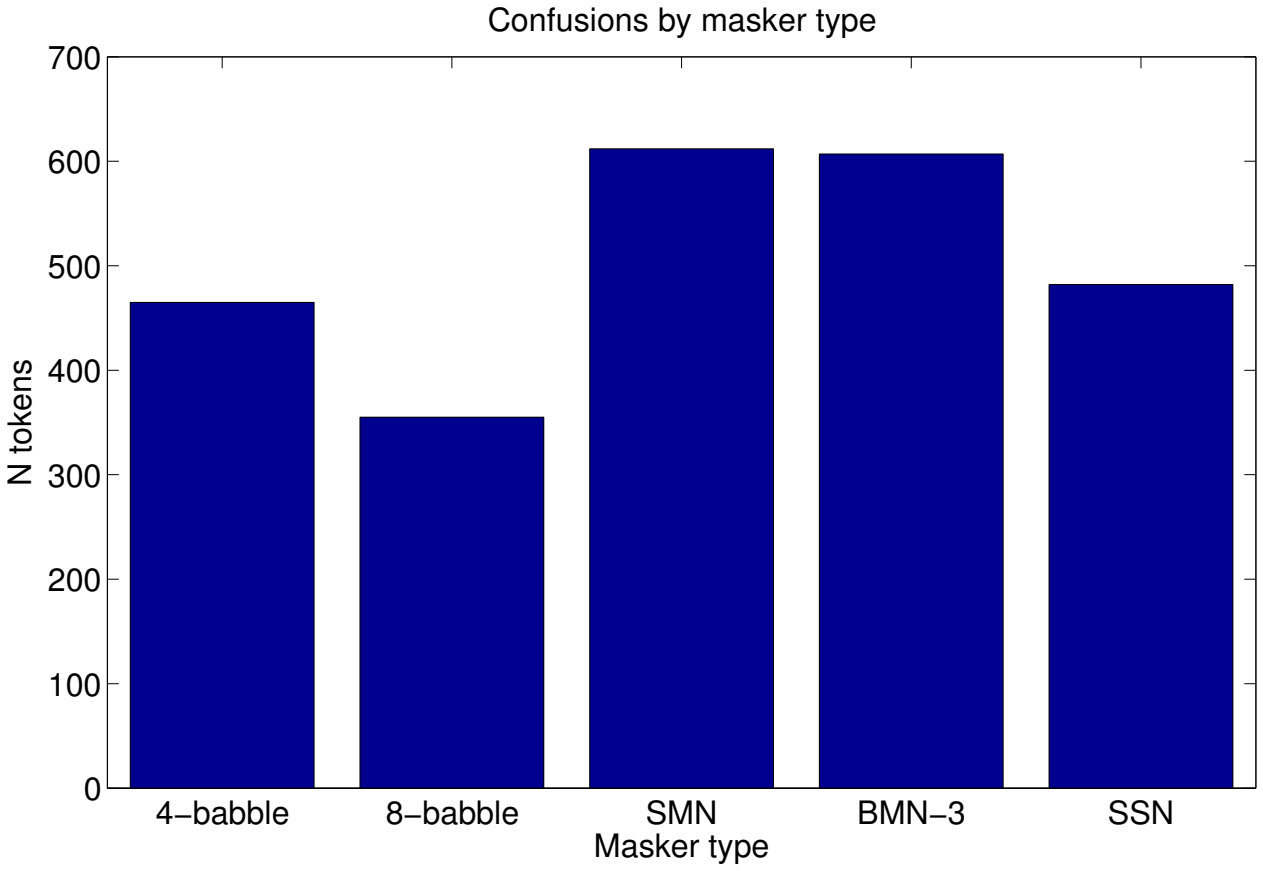
- tokens in one of 3 states:
  - **active**: token still under test
  - **discarded**: token unlikely to be a robust confusion; discard and generate new token online
  - **exhausted**: survived maximum number N_{max} 'listens' and hence potentially 'interesting'
- heuristics to remove tokens:
  - identified correctly by L_1 listeners in first N presentations; or by L_2 listeners subsequently
  - responses of first L_3 listeners all different

Outcomes to date

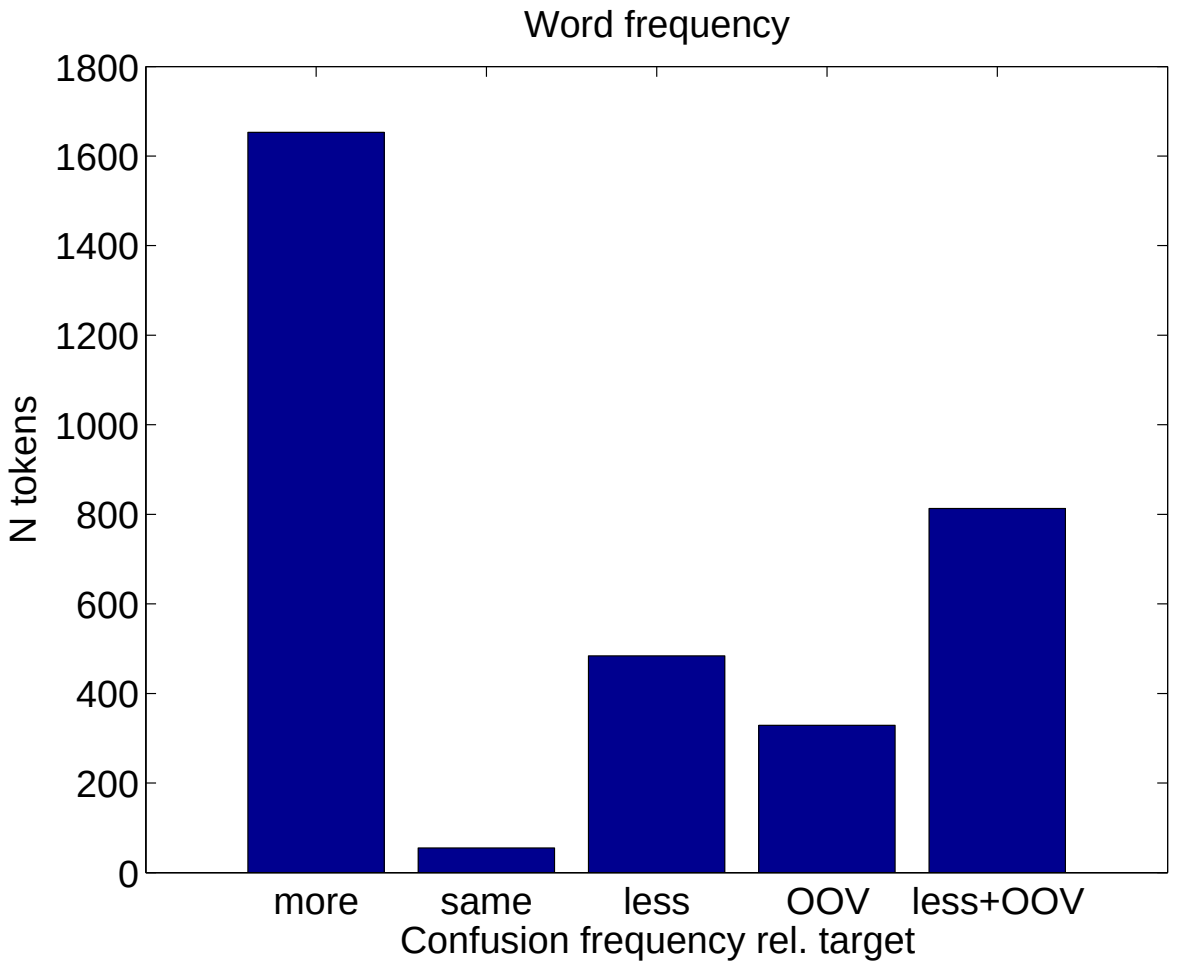
- 143 listeners
- 251 143 responses
- 43 520 different tokens screened
- 5.77 responses per token = 2.6 fold increase re. non-adaptive
- 2615 'interesting' tokens (minimum agreement of 6; max. 15)
- discovery rate: 9.3 per listener-hour = approx. 1 euro per 'interesting' case

Analysis in part based on earlier subset (N=69, 1248 tokens) presented at Interspeech 2013

Surprisingly little effect of masker type



Confusions are more frequent words than target about two-thirds of the time



Masker-independent analysis

Taxonomy is an extension of Garnes & Bond (1980)

1. **Single phoneme**: insertion, deletion, substitution
2. **Dual phoneme**: two phonemes changed
3. **Syllabic**: insertion or deletion
4. **Compounds**: more complex but still understandable reconstructions
5. **Eccentric**: defy simple explanation

NB: not really mutually-exclusive categories but treated as exclusive here

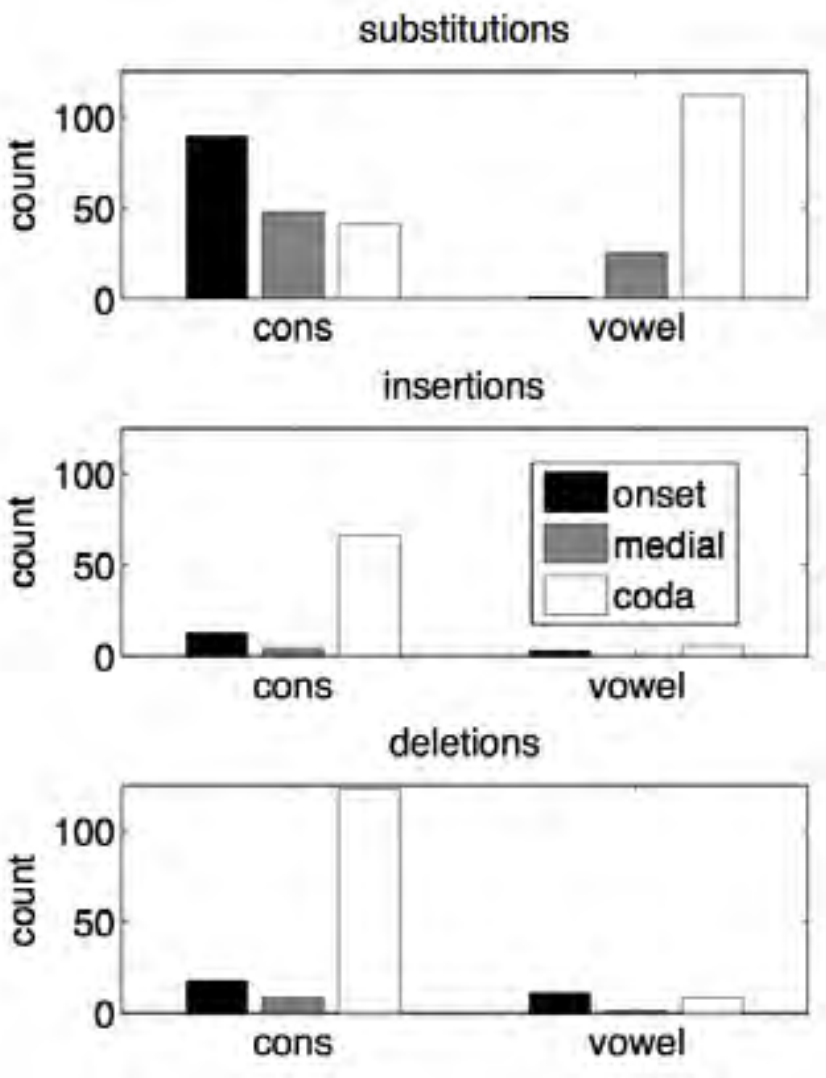
Single phoneme cases

- insertions:
  /falsa/ $\mapsto$ /falsas/
- deletions:
  /estabas/ $\mapsto$ /estaba/
  /tropa/ $\mapsto$ /ropa/
- substitutions:
  /eskutʃe/ $\mapsto$ /eskutʃa/
  /kiso/ $\mapsto$ /piso/

Mainly inflectional in the codas

	phoneme heard																							
	del	a	e	i	o	u	p	b	t	d	k	g	s	f	θ	j	ʎ	l	r	ɾ	m	n	ɲ	sum
ins	.	3	.	.	6	.	4	2	2	2	3	.	45	.	2	.	.	1	18	.	1	3	.	92
a	11	.	9	1	69	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	92
e	4	13	.	1	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	28
i	1	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	6
o	3	22	4	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	33
u	2	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3
p	5	.	.	.	.	.	.	4	1	.	8	.	.	1	.	1	.	1	.	.	1	1	.	23
b	.	.	.	.	.	.	8	.	5	3	.	1	.	3	3	.	.	3	1	.	3	1	.	31
t	9	.	.	.	.	.	1	2	.	3	3	.	.	.	.	.	.	2	.	.	.	.	.	20
d	2	.	.	.	.	.	.	.	4	.	.	.	.	1	1	.	1	1	1	.	.	.	.	11
k	.	.	.	.	.	.	5	.	.	.	.	.	.	1	1	.	.	.	.	1	.	1	.	9
g	2	.	.	.	.	.	.	1	1	1	2	.	.	.	1	.	.	.	.	.	2	.	.	10
s	49	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	50
f	1	.	.	.	.	.	3	.	.	.	1	1	.	.	.	2	.	.	.	.	.	1	.	9
θ	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	3	.	.	21	.	.	.	.	26
j	.	.	.	.	.	.	3	.	.	.	4	3	.	.	.	.	.	.	.	.	.	.	.	10
ʎ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1
l	1	.	.	.	.	.	.	.	.	1	.	2	.	.	.	.	.	.	2	.	1	.	1	8
r	5	.	.	.	.	.	.	.	.	2	.	.	1	.	.	.	.	2	.	.	.	2	.	12
ɾ	2	.	.	.	.	.	.	1	.	.	1	1	.	.	.	.	.	.	.	.	.	1	.	6
m	1	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	1	.	.	.	3	.	7
n	73	.	.	.	.	.	.	.	1	.	.	.	11	.	.	.	.	2	2	.	1	.	2	92
ɲ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0
sum	171	42	13	3	85	4	24	10	16	12	22	8	59	6	8	6	1	13	47	1	9	16	3	579

Single phoneme cases



As in previous studies:

- consonant errors more frequent
- vowel insertion/deletion rare

Unlike Cutler & Henton (2004):

- segments in final position most affected
  - due to inflectional variation; and
  - mainly in unstressed position in Spanish

Dual phoneme cases

- double insertion: /lios/ $\mapsto$ /libros/
- double substitution: /poste/ $\mapsto$ /boske/
- insertion and substitution: /lesion/ $\mapsto$ /presion/
- two-vowel errors (very rare): /beber/ $\mapsto$/bibir/ ("to drink is to live")

Syllabic

- insertions:
  - initial: /se/ $\mapsto$ /pense/
  - medial: /daras/ $\mapsto$ /dexaras/
  - final: /tan/ $\mapsto$ /taŋgo/
- deletions:
  - initial: /deten/ $\mapsto$ /ten/
  - final: /aθerka/ $\mapsto$ /aθer/

Compounds

more complex reconstructions

- metatheses:
  - /kreemos/ $\mapsto$ /keremos/
  - /medio/ $\mapsto$ /miedo/
- 3-consonant substitutions:
  - /pagado/ $\mapsto$ /θapato/
- others:
  - /suenan/ $\mapsto$ /sueɲo/
  - /komun/ $\mapsto$ /mundo/
  - /armario/ $\mapsto$ /manos/

Eccentric cases

- /fila/ $\mapsto$ /entramos/
- /nuebe/ $\mapsto$ /dutʃas/
- /dato/ $\mapsto$ /kaktus/
- /komiendo/ $\mapsto$ /fumar/

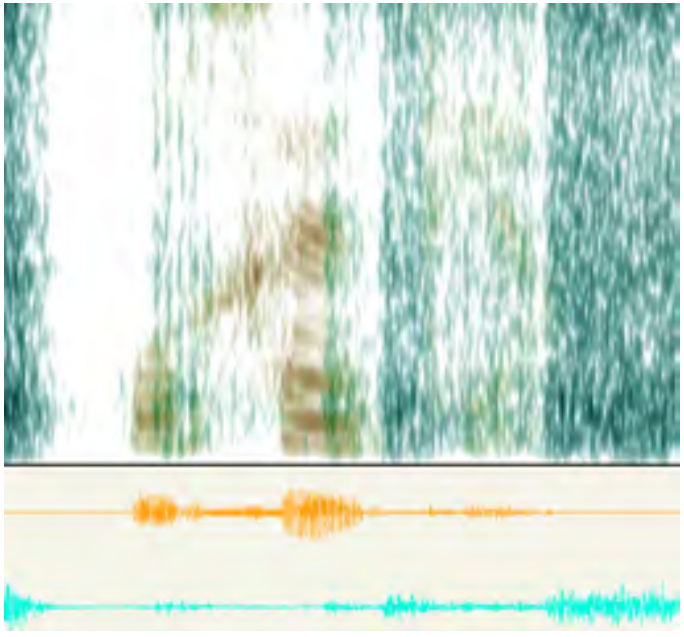
Masker-dependent analysis

- Goal is to explain the robust misperception in terms of the way speech and masker interact
- 'Traditional' energetic versus informational masking distinction too crude
- New 3-way categorisation based on the **degree to which information from the masker appears in the misperceived word**; Largely orthogonal to preceding segmental taxonomy

  1. Minimally $\mapsto$ **Reinterpretation**
  2. Partially $\mapsto$ **Blend**
  3. Totally $\mapsto$ **Override**

Reinterpretation

The reported word is based solely on reinterpreting those audible components of the target word which escaped energetic masking, using none of the masker components.



/urɤenθia/ + SMN @ -9.6 dB SNR ↦ /muxer/

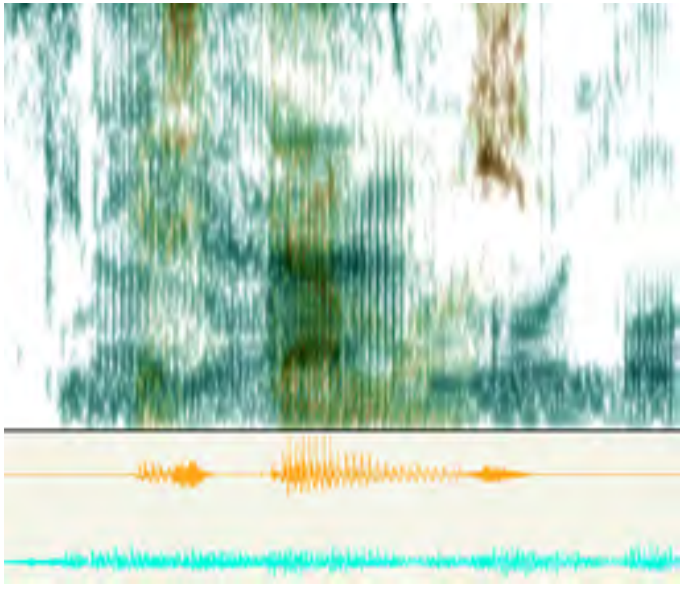
$N_{agree} = 11$

Others: s, empujar, empuje, mujre

Word suffers little masking initially, but both the weak fricative /θ/ and unstressed final syllable are masked; 'mujer' chosen as most likely lexical candidate in spite of no evidence of an initial nasal.

Blend

The reported word is composed of parts of both the target word and the masker. Blends can make use of elements of various types, ranging from subphonemic cues to segments or entire syllables. They can also incorporate prosodic information from the masker.



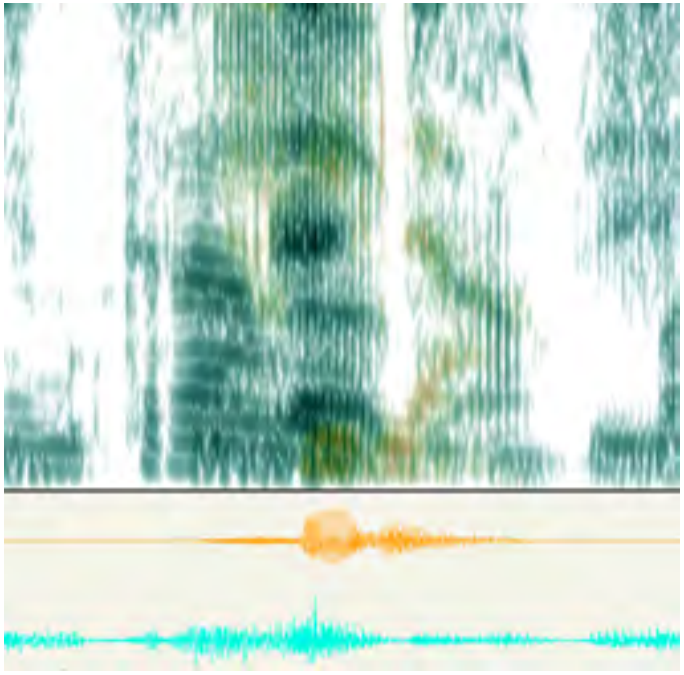
/estamos/ +
4-babble @ -0.2 dB SNR
↦ /kristal/

$N_{agree} = 10$
Others: cristales (2), quien esta, crital, estamos

Transformation appears to require the blending of the target word sequence /sta/ with elements from the masker.

Override

A word contained within the masker is reported in its entirety.



/fila/ +

4-babble @ -2.9 dB SNR

$\mapsto$ /entramos/

$N_{agree} = 14$

Other: entrar

Although the SNR is not particularly adverse, the critical stressed vowel /i/ of the target was masked and the start of the reported word happened to coincide with that of the masker.

Automated procedures to identify the origins of misperceptions

- for **Reinterpretations**, use energetic masking (EM) model and missing data recogniser (bounded marginalisation, Cooke et al., 2001)

Automated procedures to identify the origins of misperceptions

- for **Reinterpretations**, use energetic masking (EM) model and missing data recogniser (bounded marginalisation, Cooke et al., 2001)
- for **Overrides**, analyse ASR word recognition likelihoods directly in the mixture (or simpler: just look at the babble constituents)

Automated procedures to identify the origins of misperceptions

- for **Reinterpretations**, use energetic masking (EM) model and missing data recogniser (bounded marginalisation, Cooke et al., 2001)
- for **Overrides**, analyse ASR word recognition likelihoods directly in the mixture (or simpler: just look at the babble constituents)

ongoing for **Blends**, use EM and glimpse decoder (Barker, Cooke & Ellis, 2005) to determine which assignment of evidence best matches reported confusion

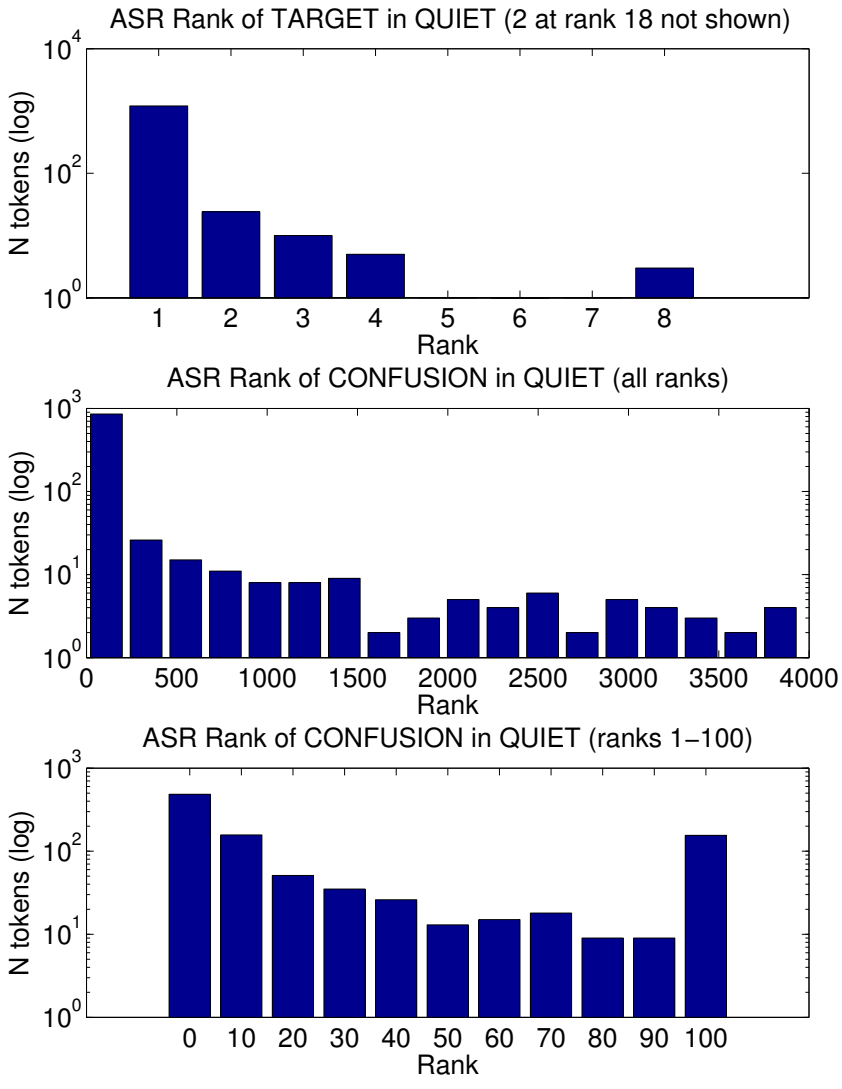
Automated procedures to identify the origins of misperceptions

- for **Reinterpretations**, use energetic masking (EM) model and missing data recogniser (bounded marginalisation, Cooke et al., 2001)
- for **Overrides**, analyse ASR word recognition likelihoods directly in the mixture (or simpler: just look at the babble constituents)

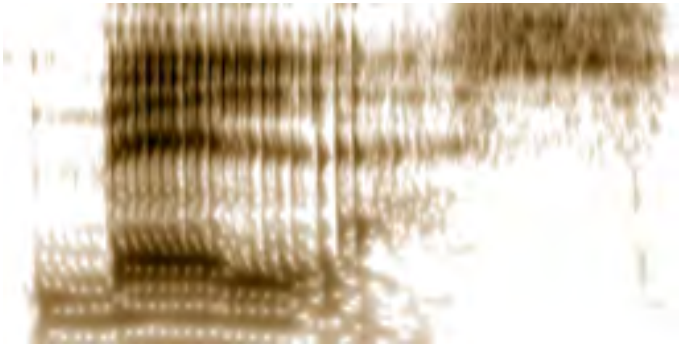
ongoing for **Blends**, use EM and glimpse decoder (Barker, Cooke & Ellis, 2005) to determine which assignment of evidence best matches reported confusion

- Also identify confusions based on close **acoustic similarity** using ASR rankings

ASR ranking in quiet



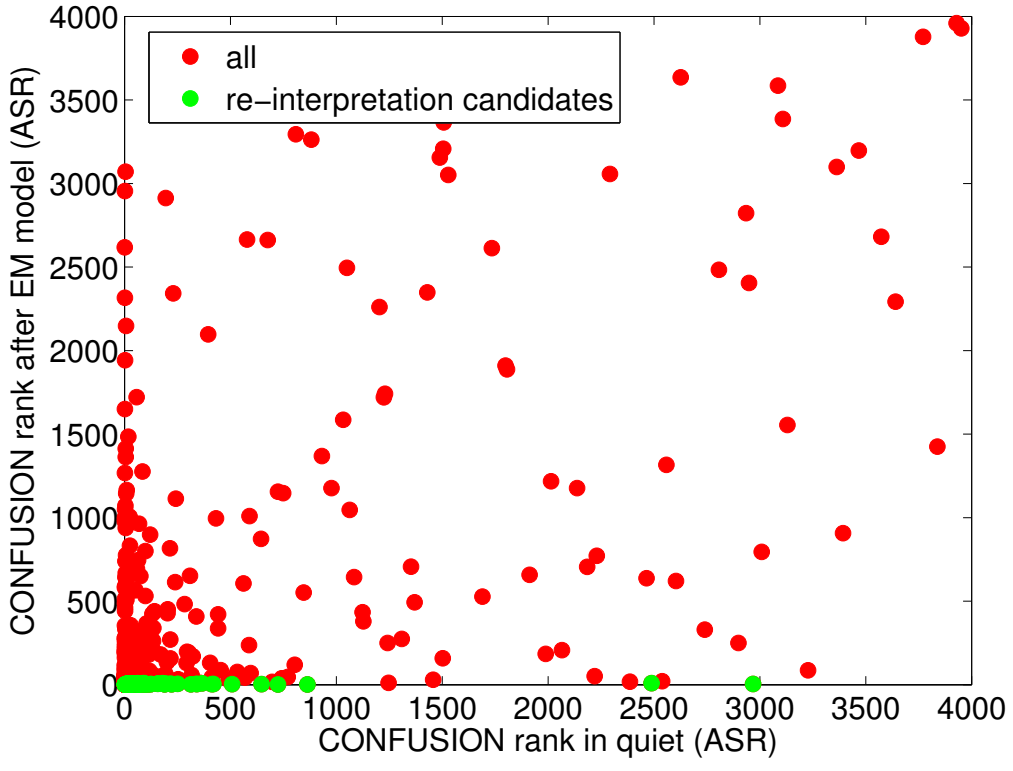
- word-based HMM recogniser using state-clustered triphones (3968 words)
- 55-dim auditory spectro-temporal excitation patterns



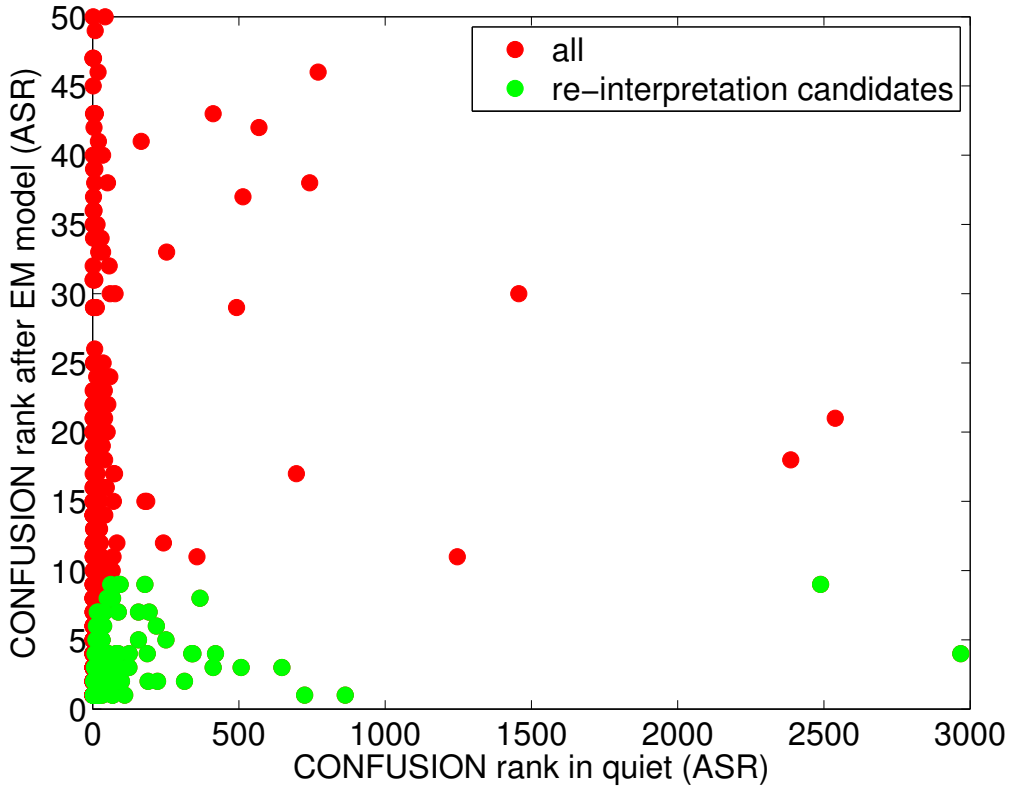
/belas/

Some confusions ranked highly in quiet (**acoustic similarity**)

Change in ranking of confusions following energetic masking model



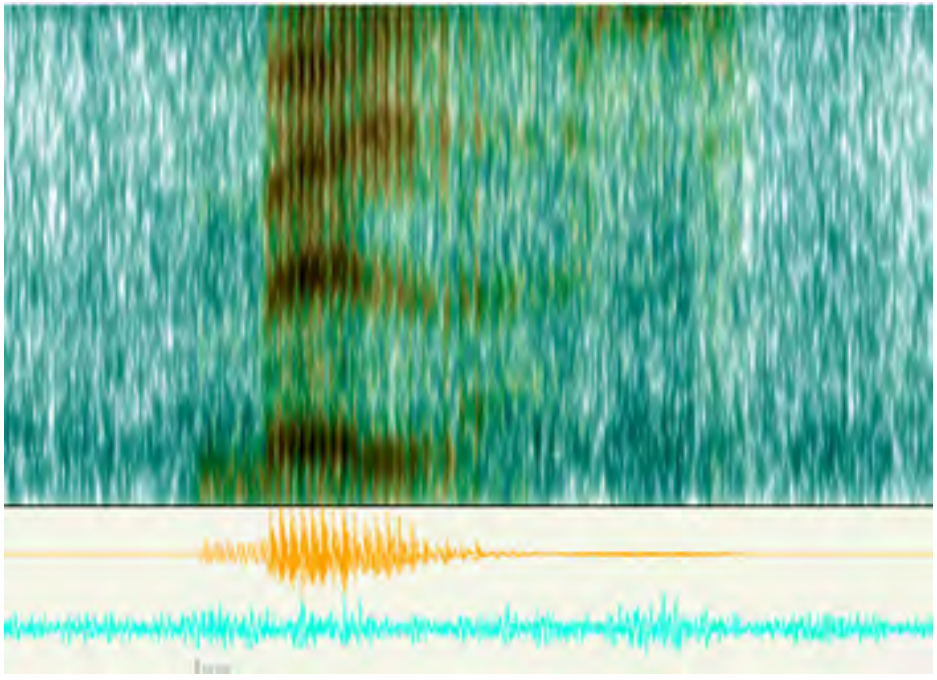
Zoomed



Working (albeit **arbitrary**) definition of reinterpretation candidates: confusion ranking more than halves after applying energetic masking model AND ranked in top 10

Example where energetic masking explains the confusion

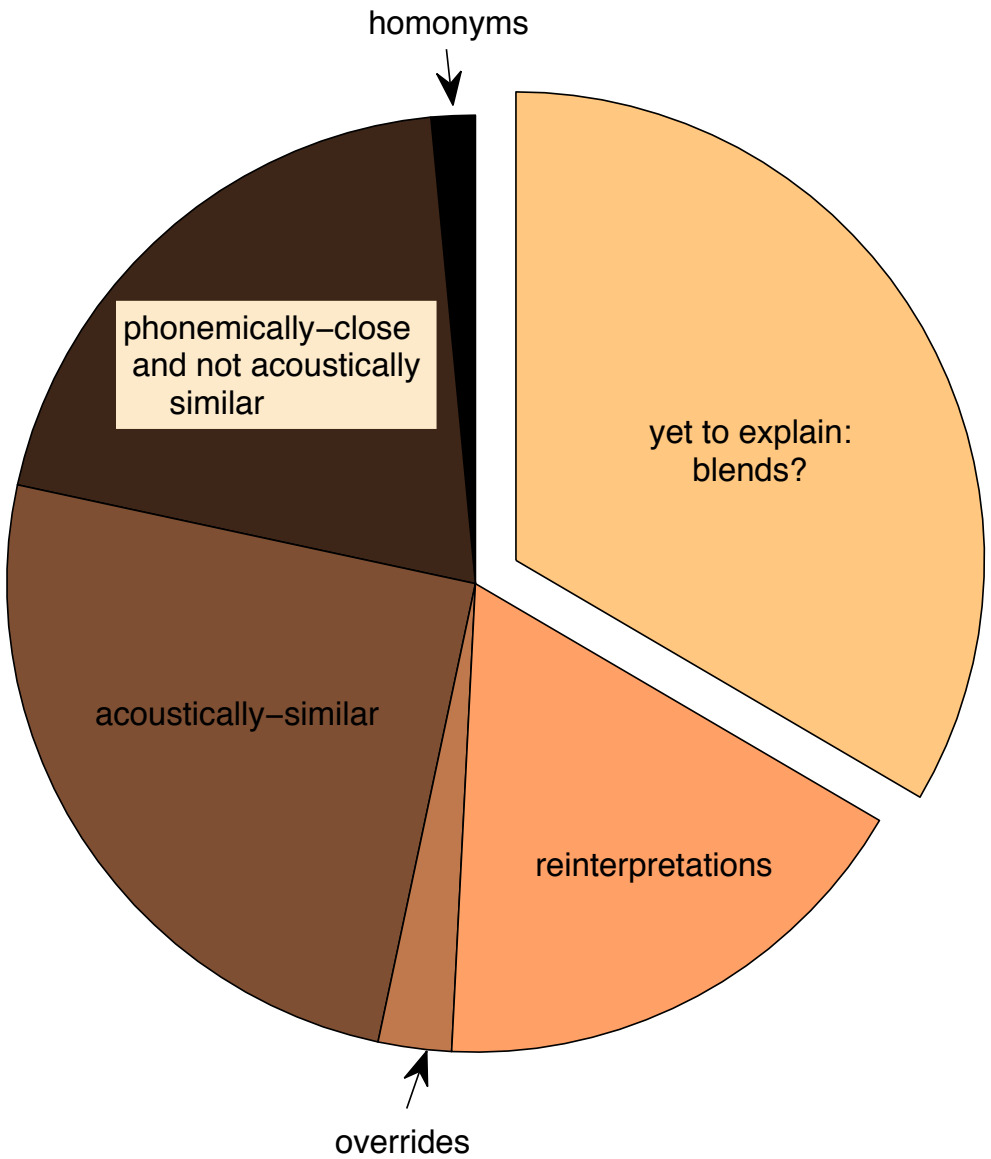
"velas" [candles] perceived as "pelo" [hair]



/belas/ ↦ /pelo/

condition	rank "velas"	rank "pelo"
quiet	1	725
mixture	-2466	341
EM model	11	1

Summary of error causes



Discussion

- Eliciting word errors in noise leads to unanticipated confusions
  - cf. VCVs or 2-endpoint continua

Discussion

- Eliciting word errors in noise leads to unanticipated confusions
  - cf. VCVs or 2-endpoint continua
- Misperceptions will be critical in evaluating end-to-end models of speech perception
  - Far more informative than WER, for instance

Discussion

- Eliciting word errors in noise leads to unanticipated confusions
  - cf. VCVs or 2-endpoint continua
- Misperceptions will be critical in evaluating end-to-end models of speech perception
  - Far more informative than WER, for instance
- Next steps: manipulate target and/or masker to disrupt misperception
  - e.g. Effect of changes in F0 or target/masker synchrony on auditory grouping

Endnote

- Elicitation in other languages imminent
  - English (Jon Barker)
  - Dutch (Odette Scharenborg)
- Intention to make these available to the community soon along with a microscopic modelling challenge!
- Thanks to Yan Tang, Bert Cranen and the Marie Curie ITN INSPIRE (*Investigating Speech Processing In Realistic Environments*)

Token_ID	Sent	Heard	Corr	NoiseType	SNR	Spe
92	ganado	manada	9	ssn	-4.61	s2
322	podemos	niños	9	ssn	-6.42	s3
327	clientes	dientes	9	ssn	-6.39	s2
346	trató	ratón	9	bmn1	-12.49	s4
433	motos	moda	9	bab8	0.73	s2
543	doblar	leche	9	bab4	0.02	s4
1041	hierba	hierro	9	bmn3	-5.71	s1
1291	visita	caros	9	bab4	-2.28	s3
1319	vino	chicos	9	bab4	-2.66	s3
1583	obtener	contener	9	bmn3	-6.67	s1
1960	criar	guiar	9	bmn3	-5.98	s3
2302	central	entrar	9	ssn	-4.22	s3
2415	doblar	temblar	9	bmn1	-9.55	s3
2558	encantan	encanto	9	ssn	-5.31	s4
2639	arde	arboles	9	bab8	-2.08	s2
2677	serie	ser	9	bmn1	-7.13	s2
2742	brecha	flecha	9	bmn1	-11.4	s2
2836	comiendo	fumar	9	bab4	-1.94	s3
3016	margen	angel	9	ssn	-4.76	s3
3342	cambien	cambio	9	bmn3	-3.48	s3
3785	habrá	acostumbrar	9	bab4	-0.79	s1
3977	vestido	vestir	9	ssn	-4.25	s1
4362	frontal	central	9	bab4	-2.28	s1
4490	salida	salir	9	ssn	-6.71	s4
4963	alerta	alegre	9	ssn	-4.80	s3
6115	nace	cama	9	bab8	-2.67	s1
6506	busco	curso	9	bmn3	-6.14	s1
6510	amistad	estar	9	bmn3	-7.25	s1
6576	timón	jamon	9	bab8	-0.38	s3
6713	variar	familiar	9	bab4	-1.68	s3
6790	poste	bosque	9	bab8	-2.03	s2
7317	grasa	casa	9	bmn3	-7.09	s3
7344	porqué	que	9	ssn	-4.73	s2
7546	nave	nada	9	bmn3	-3.14	s1
7820	siente	siempre	9	ssn	-4.61	s3
8536	cámara	cama	9	bmn3	-4.84	s1
8618	frena	cena	9	bmn3	-6.01	s2
8994	sincero	series	9	bab4	-2.47	s2
9017	sacado	sacar	9	bmn1	-12.82	s2
9240	pequeño	pequeñas	9	bab4	-1.66	s3
9670	hoyos	pollos	9	bab8	-0.07	s1
10134	cuota	puerta	9	bmn3	-7.13	s3
10503	distinto	destino	9	bmn3	-4.33	s4
10542	llegando	llegamos	9	bab4	0.03	s4
10595	puesto	cuesta	9	ssn	-6.17	s2
10761	dormido	dormir	9	bmn1	-7.48	s3