Low-level and high-level models of perceptual compensation for reverberation

Amy V. Beeston and Guy J. Brown

{a.beeston, g.brown}@dcs.shef.ac.uk
Department of Computer Science, University of Sheffield, UK

The University Of Sheffield.

Background

- Perceptual constancy allows us to compensate for our surroundings and overcome distortions of naturally reverberant environments.
- Prior listening in reverberant rooms improves speech perception [1, 2].
- Compensation is disrupted when reverberation applied to a test word and preceding context is incongruous [1].
- Here we develop low-level and high-level computational models of perceptual compensation in speech identification tasks.

Perceptual experiments

- Watkins demonstrated a 'sir/stir' category boundary shift for reverberated test words in response to reverberation distance of preceding context [1].
- He imposed the temporal envelope of 'stir' on 'sir' to give the impression of a 't' stop at one end of an 11-step interpolated continuum of test words.
- He recorded impulse responses (IRs) of a room (volume 183.6 m^3) at 'near' (0.32 m) and 'far' (10 m) distances, and independently reverberated test and context.
- The /t/ in a reverberated test word was more likely to be identified if preceding context speech was similarly reverberated [1].

- We replicated and extended Watkins' findings using natural speech, 20 talkers (male and female), and a wider range of consonants (/p/, /t/, /k/) [3].
- 80 utterances of form CW1 CW2 TEST CW3 from Articulation Index Corpus [4], each containing context words (CW) and TEST word SIR, SKUR, SPUR or STIR.
- Utterances were low-pass filtered (8th order Butterworth) to assess frequency-dependent characteristics of compensation. Results shown for 4 kHz condition.
- CW and TEST independently reverberated using Watkins' IRs [1] to give the impression of speech at different distances in a room e.g., near CW – far TEST.

	near-near				near-far				far-far			
	sir	skur	spur	stir	sir	skur	spur	stir	sir	skur	spur	stir
sir	19	0	0	1	18	0	0	2	16	1	1	2
skur	**0**	20	0	0	**3**	15	0	2	**0**	16	0	4
spur	**0**	1	18	1	**7**	2	10	1	**2**	1	14	3
stir	**0**	0	0	20	**8**	1	1	10	**1**	0	0	19

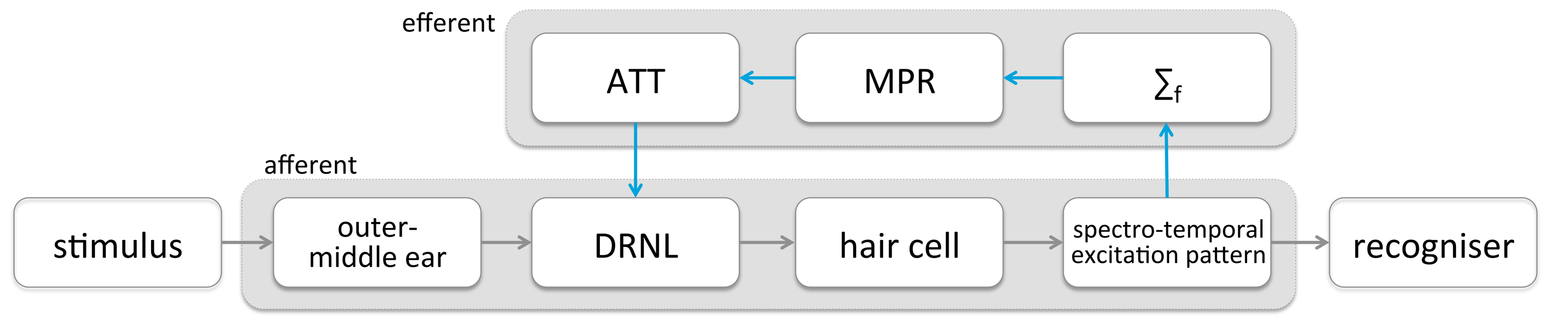
- Compensation for reverberation was observed: increased reverberation on TEST increased confusion rate, but errors reduced when CW similarly reverberated.
- Reverberation caused particular confusions to be made: most errors at near-far were test words mistaken for 'sir.'
- Compensation reduced mistaken 'sir' responses at far-far, but confusions persisted between 'skur', 'spur' and 'stir.'
- Comparable data to Watkins' ('sir' and 'not sir') resulted in a significant chi-squared value (Bonferroni corrected), with χ^2=8.007, p=0.023.

	# sir	# not sir
near-far	36	44
far-far	19	61

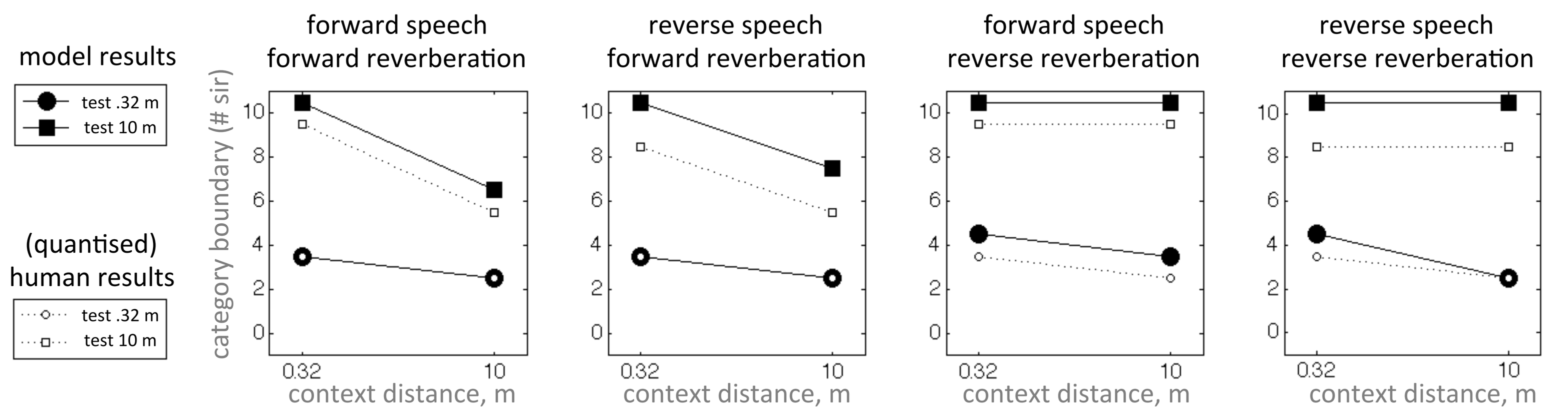
Funded by EPSRC. Thanks to Ray Meddis for the DRNL code, to Tim Jürgens for help with the phi-squared metric, and to Kalle Palomäki, Tony Watkins, Hynek Hermansky and Simon Makin for helpful discussion.

Low-level model

- Auditory efferent system has been implicated in controlling dynamic range [5].
- Mean-to-peak ratio (MPR) of wideband speech envelope updates attenuation (ATT) applied to nonlinear pathway of dual-resonance nonlinear (DRNL) filterbank [6].
- This helps to recover dips in the temporal envelope e.g., reverberated /t/.



- Good match to Watkins' sir-stir listener data using simple template recogniser [7].
- Effect of reverberation on test word: category boundary shifts up (more 'sir's).
- Compensation (forward reverberation cases): boundary shifts back (more 'stir's).
- Matching human listeners, compensation is abolished for reverse reverberation.

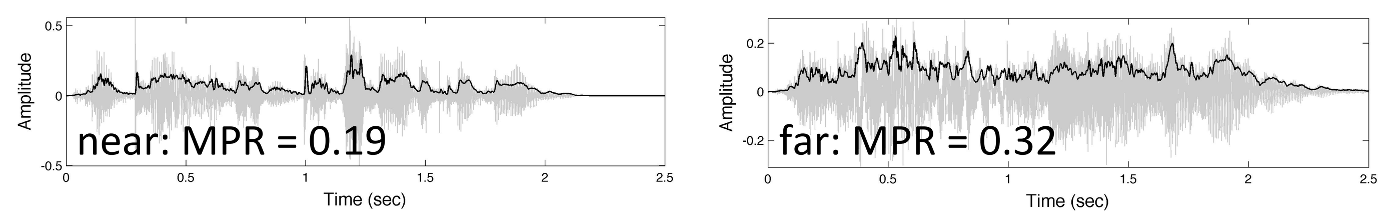


- Simplified model with efferent circuit engaged (at fixed ATT) in 'far' context cases.
- ASR features: 13 DCT-transformed auditory features + deltas + accelerations.
- Pearson's phi-squared metric denotes (here, lack of) similarity with human results by comparing each row of confusion matrices as a 2 x 4 contingency table [8].
- For identical distributions, $\Phi^2 = 0$. For non-overlapping distributions, $\Phi^2 = 1$.

	near-near					near-far					far-far				
	sir	k	p	t	Φ^2	sir	k	p	t	Φ^2	sir	k	p	t	Φ^2
sir	16	0	1	3	0.0564	5	12	0	3	0.4887	11	3	2	4	0.0731
skur	**0**	13	1	6	0.2121	**1**	12	3	4	0.1250	**3**	12	1	4	0.1143
spur	**0**	3	11	6	0.1565	**1**	14	5	0	0.4042	**1**	10	7	2	0.2558
stir	**0**	1	2	17	0.0811	**2**	4	3	11	0.1612	**5**	5	1	9	0.3060

MPR

- Mean-to-peak ratio (MPR) is tested as a metric to quantify reverberation.
- Reverberation fills dips in temporal envelope and dynamic range reduces.
- MPR increases with reverberation; inversely proportional to dynamic range.

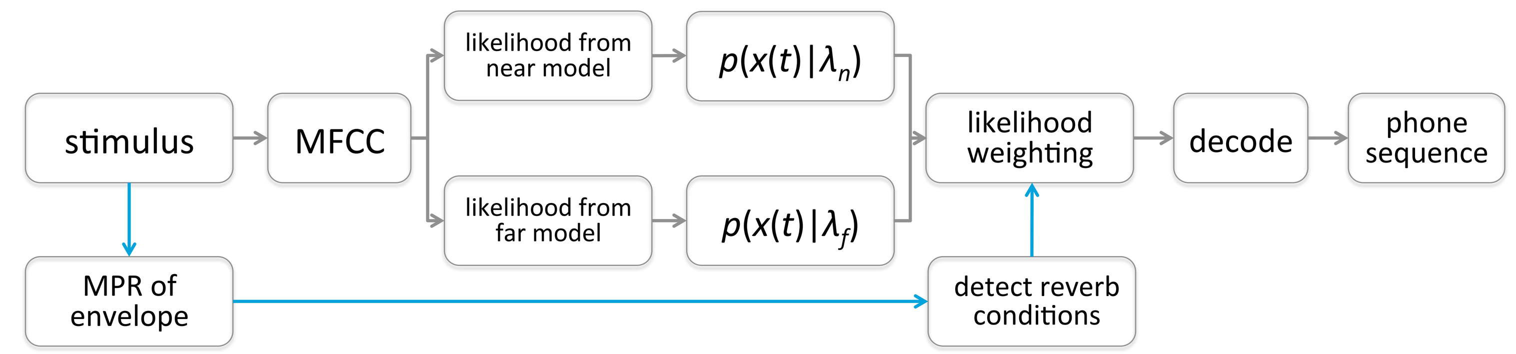


ASR

- Hidden Markov model (HMM) recogniser implemented using HTK [9].
- HMMs initially trained on TIMIT, then adapted to subset of AI corpus.
- 39 monophone models + silence model [10].
- AI corpus prompts expanded to phone sequences using CMU dictionary [11].
- Semi-forced alignment: recogniser identified TEST only.

High-level model

- Compensation for reverberation is viewed as an acoustic model selection process: analysis of speech preceding TEST informs selection of appropriate acoustic model.
- Performance is optimal when reverberation of context and test word match.
- Wrong model is selected in mismatched CW/TEST reverberation conditions: confusions increase.



- ASR features: 12 MFCCs + deltas + accelerations.
- Feature vectors for 'near' and 'far' reverberated utterances were concatenated for training to provide matching state segmentation to the likelihood weighting scheme
- Feature vectors subsequently split into separate 'near' and 'far' models for decoding.
- The combined near-far observation state likelihood is a weighted sum of parallel likelihoods in the log domain:

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

- $\alpha(t)$ adjusted dynamically using near/far classifier based on MPR metric.
- $\alpha(t)\rightarrow 0$ if reverberant; $\alpha(t)\rightarrow 1$ if dry.
- Model reproduces main confusions evident in human data ($\Phi^2 < 0.1$).

	near-near					near-far					far-far				
	sir	k	p	t	Φ^2	sir	k	p	t	Φ^2	sir	k	p	t	Φ^2
sir	16	0	0	4	0.0514	18	0	1	1	0.0333	14	1	2	3	0.0167
skur	**0**	19	0	1	0.0256	**3**	17	0	0	0.0531	**2**	16	0	2	0.0667
spur	**1**	0	17	2	0.0590	**5**	1	14	0	0.0583	**3**	0	16	1	0.0583
stir	**1**	1	1	17	0.0811	**8**	3	0	9	0.0513	**0**	0	0	20	0.0256

Discussion

- The high-level computer model replicates compensation for reverberation in the AI corpus speech identification task.
- Efferent model results are consistent with the proposal that auditory processes controlling dynamic range might contribute to reverberant 'sir/stir' distinction.
- Efferent model helps to recover dips in temporal envelope, but not to recover the more complex acoustic-phonetic cues for /p/, /t/, /k/ identification.
- Lack of training data may have contributed to poor performance of efferent-based model on AI corpus task (for the high-level model, we adapted the recogniser on the AI corpus test material).
- Future work will add frequency-dependent processing, since recent perceptual data suggests constancy occurs within individual frequency bands [12, 3]. We will also address recent findings of [13] concerning compensation with silent contexts.

1. AJ Watkins (2005). J. Acoust. Soc. Am. 118 (1) 249-262
2. EJ Brandewie & P Zahorik (2010). J. Acoust. Soc. Am. 128 (1) 291-299
3. AV Beeston, GJ Brown, AJ Watkins & SJ Makin (2011). Int. J. Audiology 50 (10) 771-772
4. J Wright (2005). Articulation Index. Linguistic Data Consortium
5. JJ Guinan (2006). Ear Hear. 27 (6) 589-607
6. RT Ferry & R Meddis (2007). J. Acoust. Soc. Am. 122 (6), 3519-3526
7. AV Beeston & GJ Brown (2010). Proc. Interspeech, Makuhari, 2462-2465
8. T Jürgens & T Brand (2009). J. Acoust. Soc. Am. 125 (5) 2635-2648
9. Hidden Markov model toolkit (HTK). http://htk.eng.cam.ac.uk
10. KF Lee & HW Hon (1989). IEEE Trans. Acoust., Speech, Signal Process. 37 (11) 1641-1648
11. Carnegie Mellon University (CMU) pronunciation dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict
12. AJ Watkins, SJ Makin & AP Raimond (2010). *In* Binaural processing and spatial hearing. Danavox Jubilee Foundation 371-380
13. JB Nielsen & T Dau (2010). J. Acoust. Soc. Am. 128 (5) 3088-3094