School of Psychology and Clinical Language Sciences

University of Reading

Room reflections, speech identification and the perceptual weighting of frequency bands

Anthony J Watkins | Andrew P Raimond | Simon J Makin

Background

- a speech message played several metres from the listener in a room is usually heard to have much the same phonetic content as it does when played nearby
- however, room reflections make the temporal envelopes of the speech very different at these distances
- this appears to be an instance of 'constancy', due to perception 'taking account' of the level of reflections in neighbouring 'context' sounds (Watkins, 2005a,b)
- other experiments have shown these constancy effects in speech from a noise excited vocoder
- the present experiments use the constancy effect to measure the perceptual weightings of the vocoder's frequency bands, asking whether the weighting pattern across bands is similar in the context and the test-words

Real-room impulse responses, RIRs

- real-room measurements with human-dummy heads, giving room-impulse responses (RIRs) :

measurement signal → dummy-head, speaker in mouth → real room → dummy-head listener, mike in ear → RIR

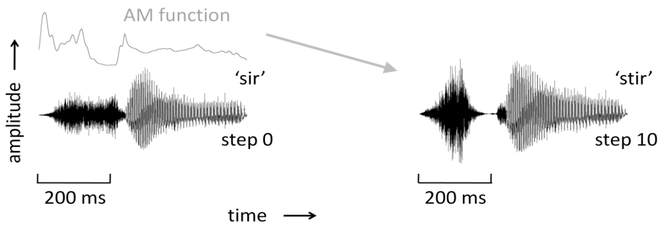
- RIRs used to effect real-room listening conditions :

'dry' recording, of speech → convolve with RIR → headphones → real-room listening

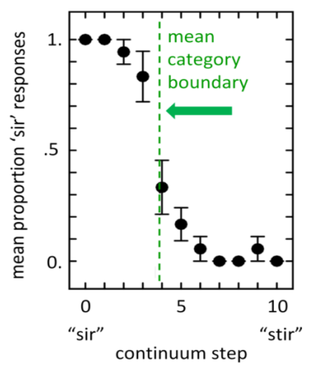
- distance between heads varies the level of room reflections:
  - early (50 ms) to late ratio; 18 dB at 0.32 m → 2 dB at 10 m (A-weighted energy decay rate; 60 dB per 960 ms at 10 m, and room volume = 183.6 m^3)

Test words and category boundaries

- listeners in 'virtual rooms', hearing RIR-processed sounds
- they identify test words from an 11-step continuum, formed by amplitude modulation (AM) of 'sir', giving 'stir':



- intermediate steps, (1-9) , by varying modulation depth
- played in a 'context' ; 'next you'll get __ to click on'
- listeners hear 'sir ' at lower steps, otherwise they hear 'stir



Sparse-NV speech

- speech processed with an 8-band noise-excited vocoder
- temporal envelope in each band from gammatone-filtered speech, (η=4, bandwidths= 'Cambridge ERBs')
- each envelope applied to a (similarly) gammatone-filtered noise
  - *n*=band number, and *n*=1,2,...,8
  - band centre-frequencies in kHz = 0.25 x $2^{(7/12)(n-1)}$

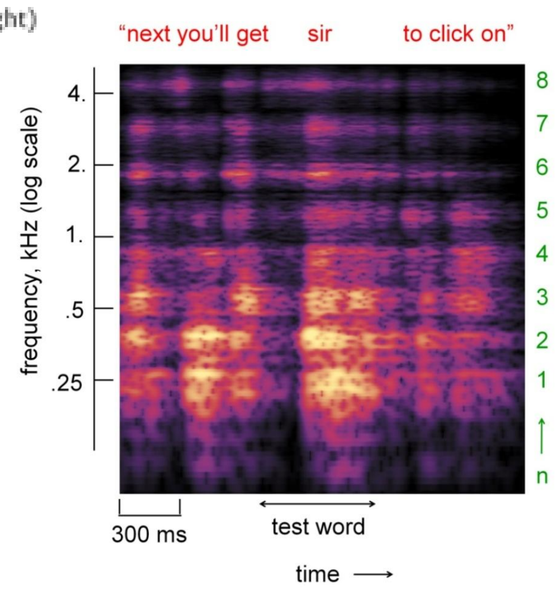
Grouping & sparse-NV speech

- individually, the vocoder's bands each sound like unintelligible noises
- but when the bands are all played together there is a grouping effect, and the speech-message is heard (Shannon, Zeng, Kamath, Wygonski, and Ekelid,1995)

Constancy & sparse-NV speech

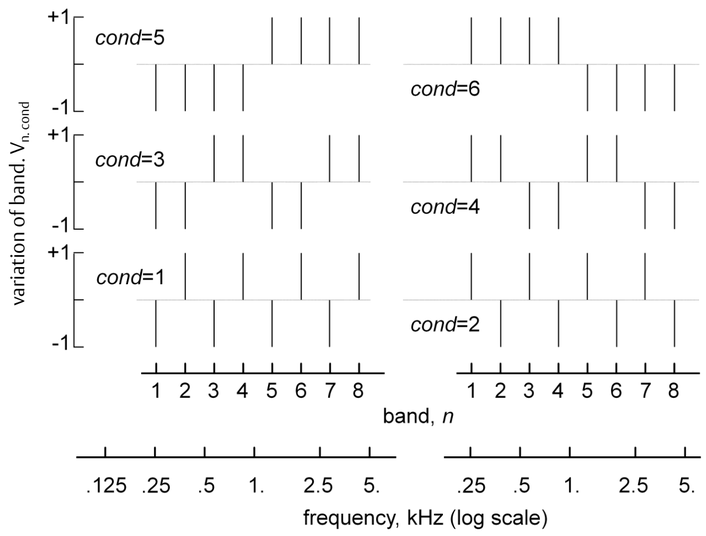
- increase level of reflections (distance) of test word
  - more 'sir' responses
  - category boundary shift, S_a
- increase distance of context as well → constancy effect:
  - fewer 'sir' responses
  - reduced category boundary shift, S_b

- the spectrogram (right) shows sparse-NV versions of the context and test-word used to demonstrate constancy effects

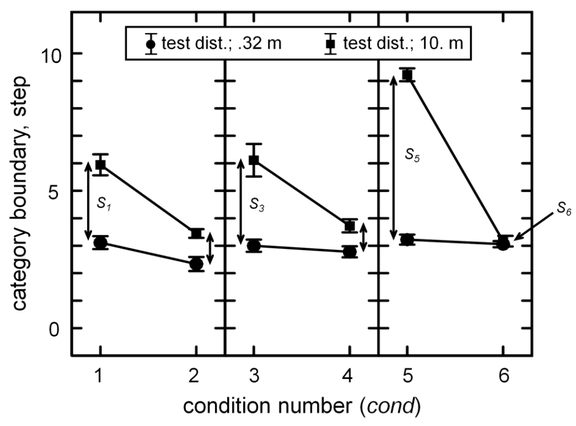


Expt. 1: Weighting of test-word's bands

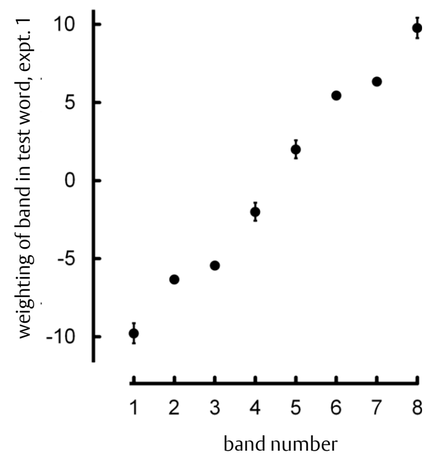
- context distance held at 0.32 m throughout
- distance varied between 0.32 m and 10 m in 4 of the test-word's bands, ($V_{n,\,cond}=+1$) while remainder held at 0.32 m ($V_{n,\,cond}=-1$) , according to condition (*cond*):



- category boundary shifts, S_{cond} vary across conditions:

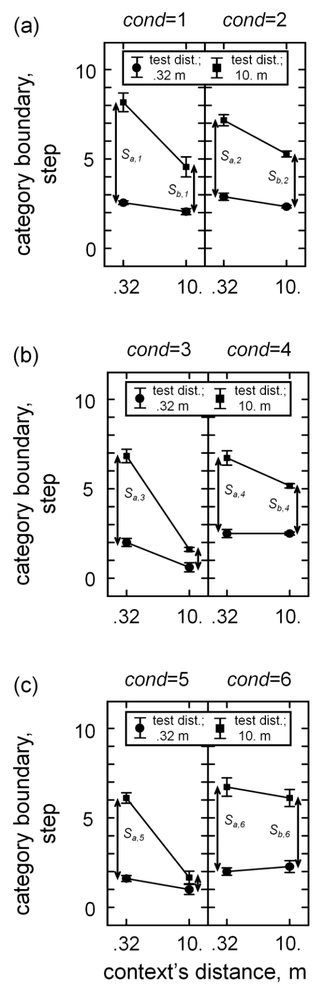


$$weighting\ of\ band\ n = \sum_{cond=1}^{cond=6} S_{cond}\, V_{n,\,cond}$$

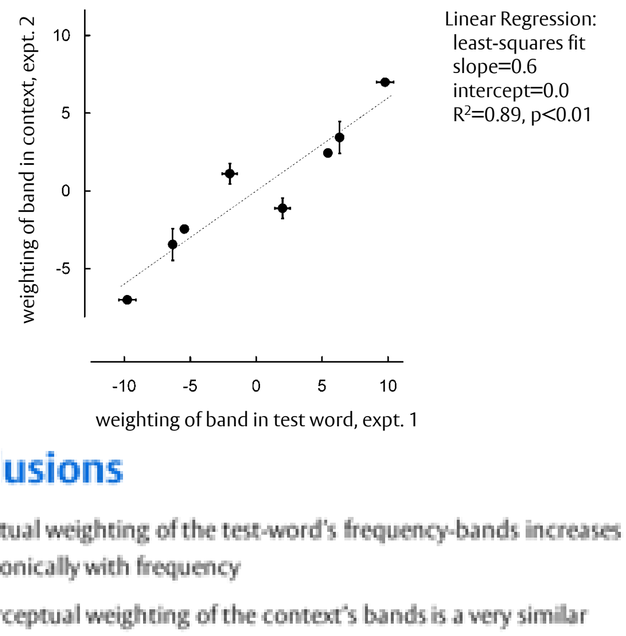


Expt. 2: Weighting of context's bands

- distance varied between 0.32 m and 10 m in all 8 of the test-word's bands
- distance varied between 0.32 m and 10 m in 4 of the context's bands, ($V_{n,\,cond}=+1$) while remainder held at 0.32 m ($V_{n,\,cond}=-1$) , according to the patterns shown for conditions (*conds*) of Expt. 1
- category boundary shifts when the context's distance is 0.32 m, $S_{a,\,cond}$ are generally reduced when the context's distance is10 m, (i.e., $S_{b,\,cond} < S_{a,\,cond}$), indicating constancy to a degree
- size of constancy effect, $S_{a,\,cond} - S_{b,\,cond}$ varies across conditions:



$$weighting\ of\ band\ n = \sum_{cond=1}^{cond=6} (S_{a,\,cond} - S_{b,\,cond})\, V_{n,\,cond}$$



Linear Regression:
least-squares fit
slope=0.6
intercept=0.0
R^2=0.89, p<0.01

Conclusions

- perceptual weighting of the test-word's frequency-bands increases monotonically with frequency
- the perceptual weighting of the context's bands is a very similar function of frequency
- this suggests that both functions arise through the perceptual weightings applied at a single stage of perceptual processing
- this might be the stage where the test-word's bands are grouped

References

1. Shannon, R.V, Zeng, F. Kamath, V. Wygonski, J. and Ekelid, M. (1995) Speech recognition with primarily temporal cues. *Science* **270** 303–304
2. Watkins, A.J. (2005a) Perceptual compensation for effects of reverberation in speech identification. *J. Acoust. Soc. Am.* **118** 249-262
3. Watkins, A.J. (2005b) Perceptual compensation for effects of echo and of reverberation in speech identification. *Acta acustica united with Acustica* **91** 892-901

Acknowledgements

- Thanks to Amy Beeston, Guy Brown, Peter Derleth , Kalle Palomäki and Hynek Hermansky for discussions
- Supported by an EPSRC grant to the first author

Further information

www.reading.ac.uk/~syswatkn