School of Psychology and Clinical Language Sciences

University of Reading

Spectral- and temporal-envelope room-acoustic cues in attentional tracking

Simon J Makin | Anthony J Watkins | Andrew P Raimond

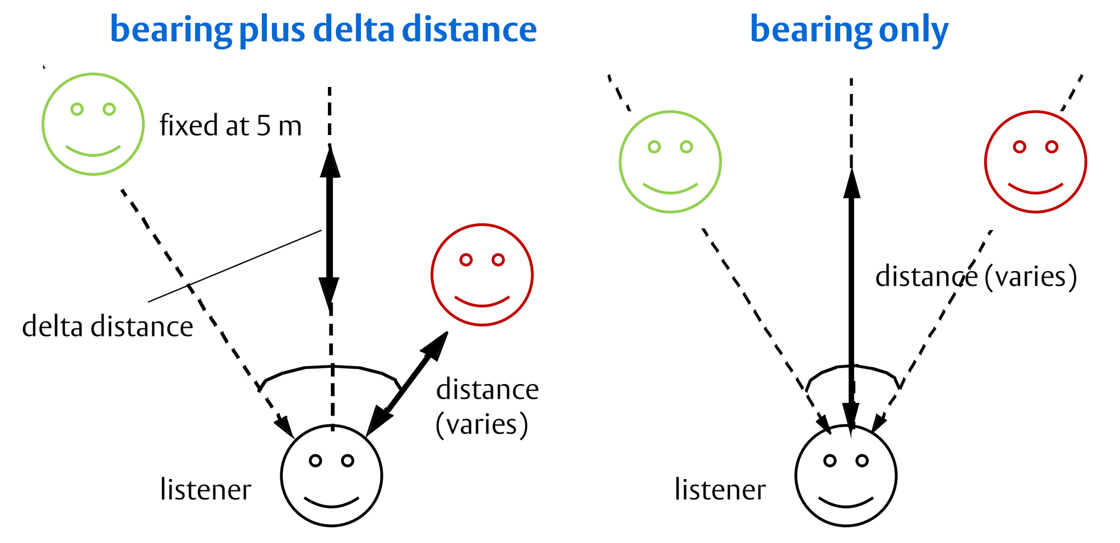
Introduction

- Listeners can selectively attend to a desired talker in the presence of interfering talkers, and a spatial position difference is known to aid this 'tracking'. For instance, much previous research has demonstrated the effectiveness of cues arising from a difference in bearing between the signals (e.g. Bronkhorst and Plomp, 1988).
- But there is reason to question the utility of such cues in real-room reverberation, as it degrades interaural localisation cues (e.g. Kidd et al., 2005). Other aspects of position in a room may play a part however. For instance, the decrease in Direct-to-Reverberant energy Ratio (DRR) with increasing distance between listener and source is a cue to perceived distance (Zahorik, 2002). More generally, reflections distort sounds' spectral- and temporal-envelopes, and this varies with location, potentially giving rise to position-specific 'grouping' cues.
- Listeners can also track talkers using differences in their voice characteristics (Darwin and Hukin, 2000), so this study asks how effectively cues arising from position (where both bearing and distance are varied) compete with cues arising from talker differences, while listening in real-room reverberation.

Methods

Experimental paradigm

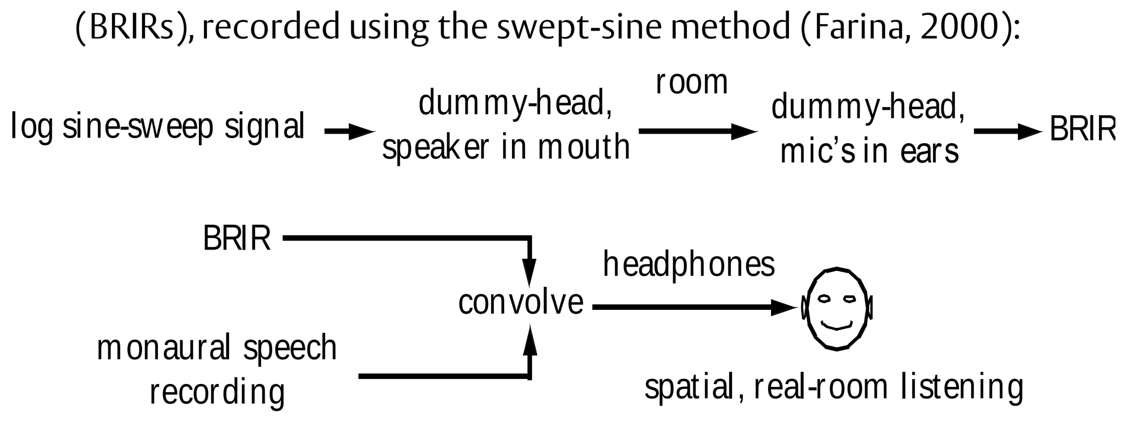
- A selective attention paradigm devised by Darwin and Hukin (2000) was used, where subjects hear two simultaneous sentences played in a (simulated) room:
- Target sentence: "**On this trial you'll get the word < > to select**"
  Distractor sentence: "**You'll also hear the sound < > played here**"
  recorded by two male talkers, along with two test words: ("**bead**" and "**globe**"), which were spliced into the < > position and time-aligned to be simultaneous.
- Listeners were asked to attend to the target sentence and indicate which test word they perceived as occurring in it. The two sentences and test words where individually spatialised such that talker and spatial attributes were in conflict. Listener response thus indicates which cue is more compelling.
- Listener's position was fixed. Speakers were placed at two locations in each condition, as shown:



- Distance was varied (0.65, 1.25, 2.5 and 5m), while equating overall rms levels (taking both channels together, so ILDs preserved) to eliminate level cues to distance, and two bearing differences (+/- 25^0 and +/- 5^0) were used.
- Mean probabilities of room-position response from 8 listeners.
- Stimuli were presented both dichotically and diotically to assess the contribution of binaural hearing. Diotic stimuli were the L or R channel presented to both ears and were level corrected to match dichotic stimuli.

Real-room reverberation

- Stimuli were spatialised using Binaural Room Impulse Responses (BRIRs), recorded using the swept-sine method (Farina, 2000):



- corrections were applied for frequency response of headphones and dummy-head speaker.

Cues to location in real-rooms

'Classic' localisation cues

- Different angles of incidence give rise to differences in the interaural time delays (ITD) and frequency-dependent interaural level differences (ILD) of the signals, leading to binaural interaction effects.
- Also, two sounds originating from different bearings will be differentially affected by head-shadowing, so that one will have a higher signal-to-noise ratio (SNR) in one ear, and vice-versa. Selectively attending to one or the other ear can thus effectively increase the SNR (so-called 'better-ear listening'). The ear with the better SNR could thus act as a cue to position without recourse to truly interaural processing.

Temporal-envelope distortion

- The amount of reverberant energy in a sound, relative to direct sound energy, increases with distance from source to listener, so that as a sound moves farther away in a room, it's temporal-envelope is increasingly distorted by the 'tails' which multiple reflections create at offsets:

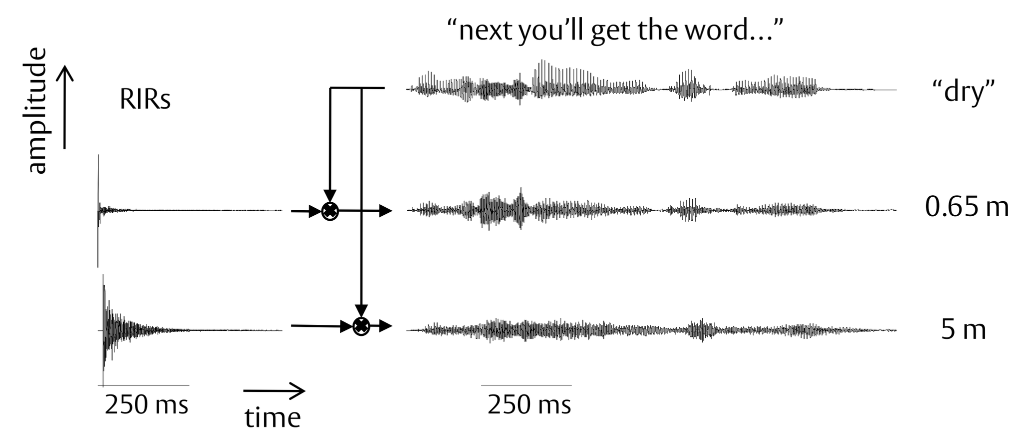


Fig.1. Three versions of the target sentence. 'Dry' speech at the top, convolved with a RIR recorded at 0.65m in the middle, and at 5m at the bottom. Corresponding (monaural) RIRs on the left. Note the increasingly distorted temporal-envelope: 'tails' 'smear' offsets and 'fill in gaps', reducing modulation depth.

- Such distortion is thus position-specific and could aid selective attention by functioning as a 'timbre difference' grouping cue.

Spectral-envelope distortion

- RIRs from different room-positions have different frequency responses. Convolving with the RIR will superimpose this frequency response on the speech spectrum, thus causing position-dependent 'colouration' of the long-term average spectrum:

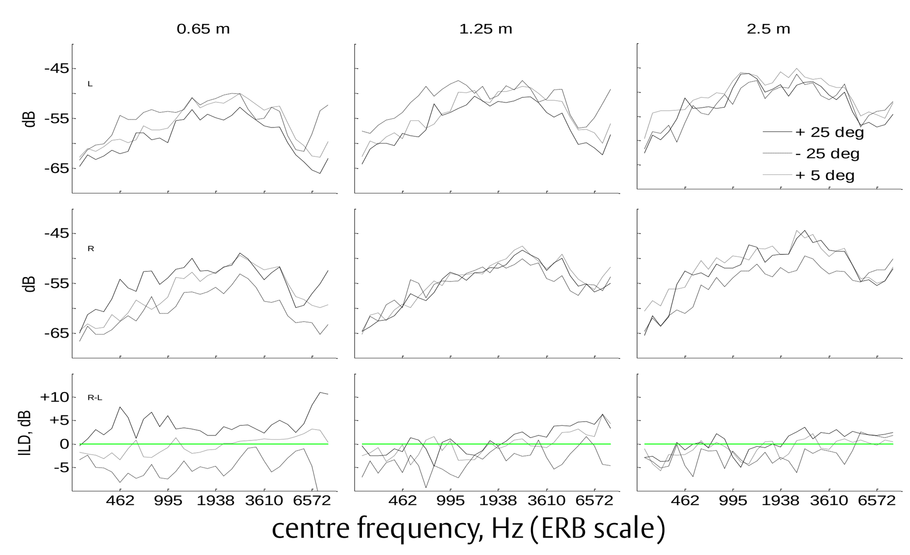


Fig.2. Auditory spectra (dB, re:1) obtained by processing BRIRs with auditory model (32 channel gammatone filterbank, equally spaced in ERB-rate). Solid lines are spectra for the +25^0 bearing, dashed lines for -25^0, and dotted lines for +5^0. Top row shows spectra for the left channel, middle row for the right, and bottom row is ILD (right–left). Note that each channel of each IR has a distinct spectrum, and that the difference between the ILDs for each bearing reduces as distance increases.

- Note the distinct spectra for each position, and the distinct patterns of ILD variation in dichotic conditions, which could act as timbre 'signatures'

Assessing relative contribution of cues

- To assess how much influence the reverberation-induced position cues have compared to the more traditional ITD and ILD cues, the BRIRs were processed to limit the cues available, as follows:
- To retain only ILD and spectral-envelope cues:

  BRIR → rotate components to cosine phase → window → time-align channels → 'spectral-envelope only' BRIR ('SO')
- to add temporal-envelope cues:

  BRIR → signal correlated noise → convolve with 'spectral-only' BRIR → 'spectral-plus-temporal-envelope' BRIR ('S+T')
- Also, in order to assess the role of 'better-ear listening' as opposed to truly binaural processing, a 'better-ear' analysis was conducted by using an auditory model to calculate Euclidean distances between the spectra of the two RIRs at each ear, with the ear receiving the most different signals in the dichotic condition then used in the better-ear analysis of the corresponding diotic data.

Results

Listeners' responses with real-room BRIRs

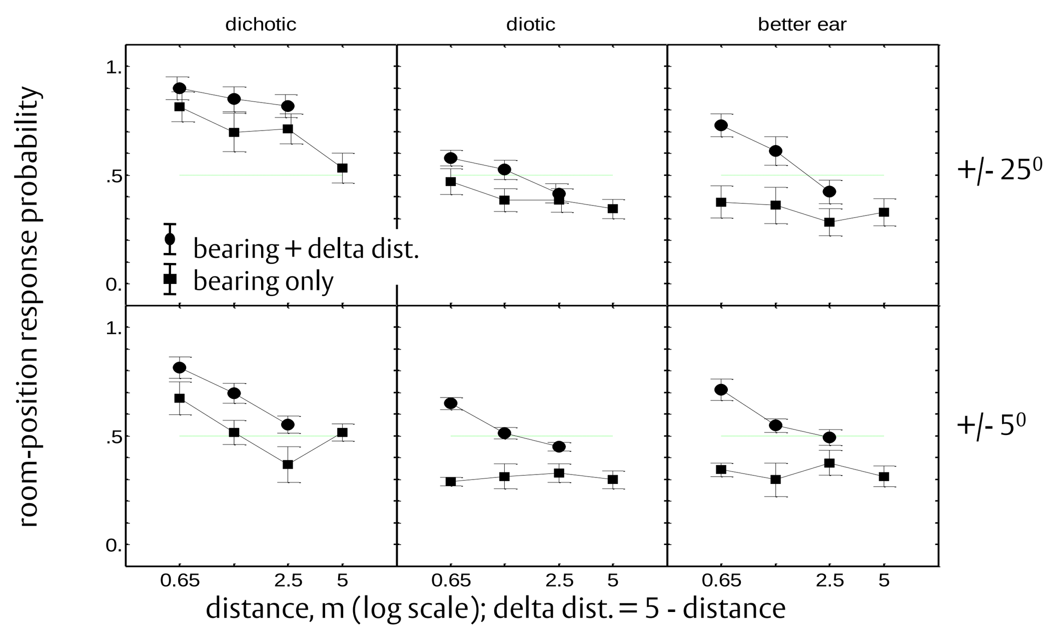


Fig. 3. Mean probabilities of room-position responses (n=8), in both 'bearing-only' (BO) and 'bearing plus delta distance' (BPDD) conditions. Top row is data for the large (+/- 25^0) bearing separation, bottom row for the small (+/- 5^0) separation. Data points near 1 indicate room-position is the dominant cue, those near 0 mean talker difference is dominant. Equal influence is indicated by the dotted line. Note that for delta distance data, *decreasing* distance means *increasing* separation. See previous for explanation of better-ear data.

- In dichotic conditions room-position is dominant. The influence of position is diminished in diotic conditions, but room-position can still be the dominant cue if there is a distance separation between talkers.
- There are prominent effects of distance separation in the diotic data, particularly when the bearing separation is small. It is possible this is due to temporal-envelope effects as the length of 'tails' in the temporal -envelope increases with distance.

Processed BRIRs: 'SO'

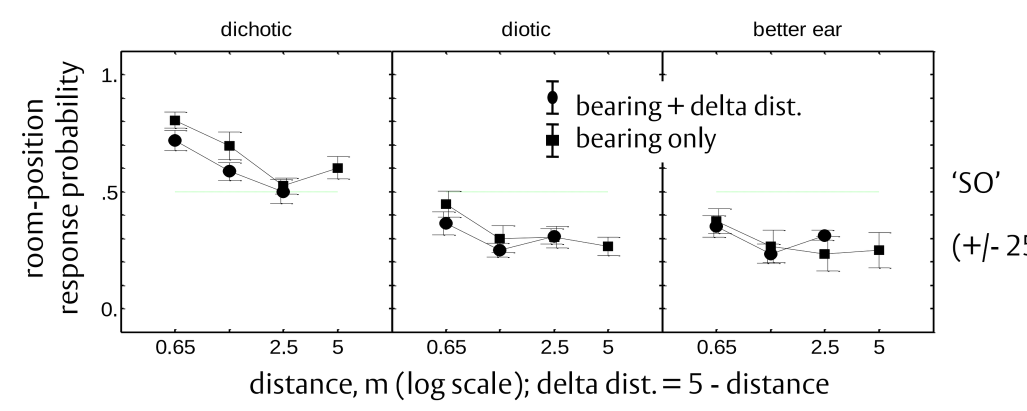


Fig. 4. Mean probabilities of room-position responses (n=8), using 'Spectral Only' BRIRs processed to limit cues to ILDs and spectral-envelope differences, in both BO and BPDD conditions, for the +/- 25^0 bearing separation.

- In dichotic conditions room-position is still dominant. This can only be due to ILD and spectral-envelope effects. There is no ITD information, and no 'better-ear' effect so the large binaural advantage seems mainly due to ILD. No longer any extra effect conferred by distance separation, presumably due to the absence of temporal-envelope differences.

Processed BRIRs: 'S+T'

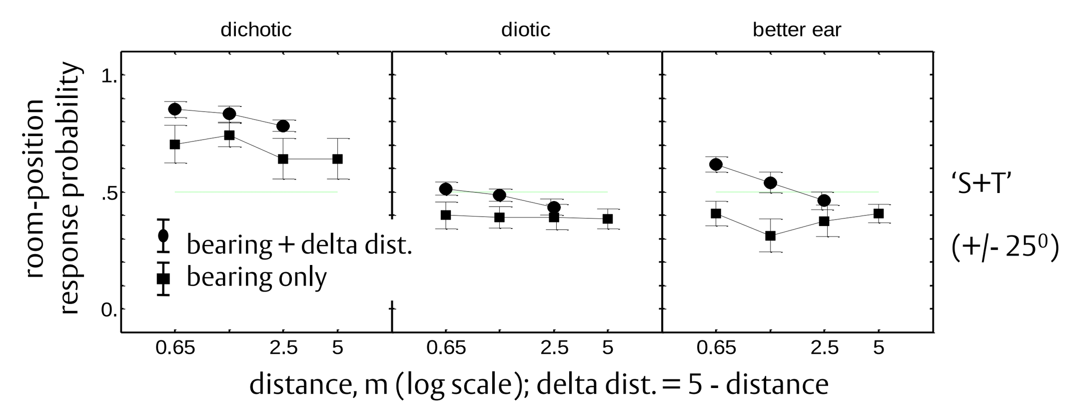


Fig. 5. Mean probabilities of room-position responses (n=8), using 'Spectral+Temporal' BRIRs processed to limit cues to ILDs, spectral– and temporal-envelope differences, in BO and BPDD configurations, for the +/- 25^0 bearing separation.

- There is a significant increase in room-position responses compared with 'SO' conditions, confirming the existence of temporal-envelope effects. The data is strikingly close to that in 'real-room' conditions, indicating that ITD effects play a relatively minor role.

Conclusions

- In dichotic conditions, room-position can dominate a same-sex talker difference. The binaural advantage seems primarily due to ILD.
- Position can also compete with a talker difference in diotic conditions, especially if there is a distance separation. Effects of distance seem due to temporal-envelope cues.
- ITD-based effects seem to play a relatively minor role in selective listening in real-room reverberation.

References

1. Bronkhorst, A. W. and Plomp, R. (1988) The effect of head-induced interaural time and level differences on speech intelligibility in noise. J. Acoust Soc. Am. **83** 1508-1516
2. Darwin, C. J. and Hukin, R. W. (2000) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. J. Acoust Soc. Am. **107** 970-977
3. Farina, A. (2000) Simultaneous measurement of impulse response and distortion with a swept-sine technique. 108th AES Convention, Paris, 18th-22nd Feb, 2000
4. Kidd, Jr., G., Mason, C. R., Brughera, A., and Hartmann, W. M. (2005) The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acust. Acta Acust.* **91** 526-526