An analysis of “speech glimpses” in realistic environments

Boston University College of Health
& Rehabilitation Sciences: Sargent College
Department of Speech, Language & Hearing Sciences

Virginia Best[1], Jörg M Buchholz[2], S Theo Goverts[3], H Steven Colburn[4]

1. Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA
2. Department of Linguistics, Macquarie University, Sydney, NSW, Australia
3. Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
4. Department of Biomedical Engineering, Boston University, Boston, MA, USA

BOSTON UNIVERSITY

BACKGROUND

A lot of effort is currently going into recording real acoustic environments [1], recreating them in the laboratory [2,3], generating naturalistic speech stimuli [4], and estimating realistic SNRs [5,6]. Here we make use of a framework that brings together these approaches to arrive at highly realistic speech-in-noise mixtures.

We analyzed the “speech glimpses” that are available in realistic mixtures. Our goals were to compare them to simpler, commonly used laboratory stimuli, and to provide a new perspective on the many sources of acoustic disruption that may hinder the understanding of speech in daily life.

METHODS

Realistic mixtures

- Speech stimuli were taken from the Everyday Conversational Sentences in Noise test (ECO-SiN; [4]). These sentences are extracted from real conversations conducted in noise.
- Noise stimuli were taken from the ARTE database [7]. We used six environments (office, church, living room, café, dinner party, food court).
- The ECO-SiN sentences were embedded in the ARTE noises at ecological SNRs using binaural room impulse responses at a distance of 1m in front of the listening position.

Laboratory mixtures

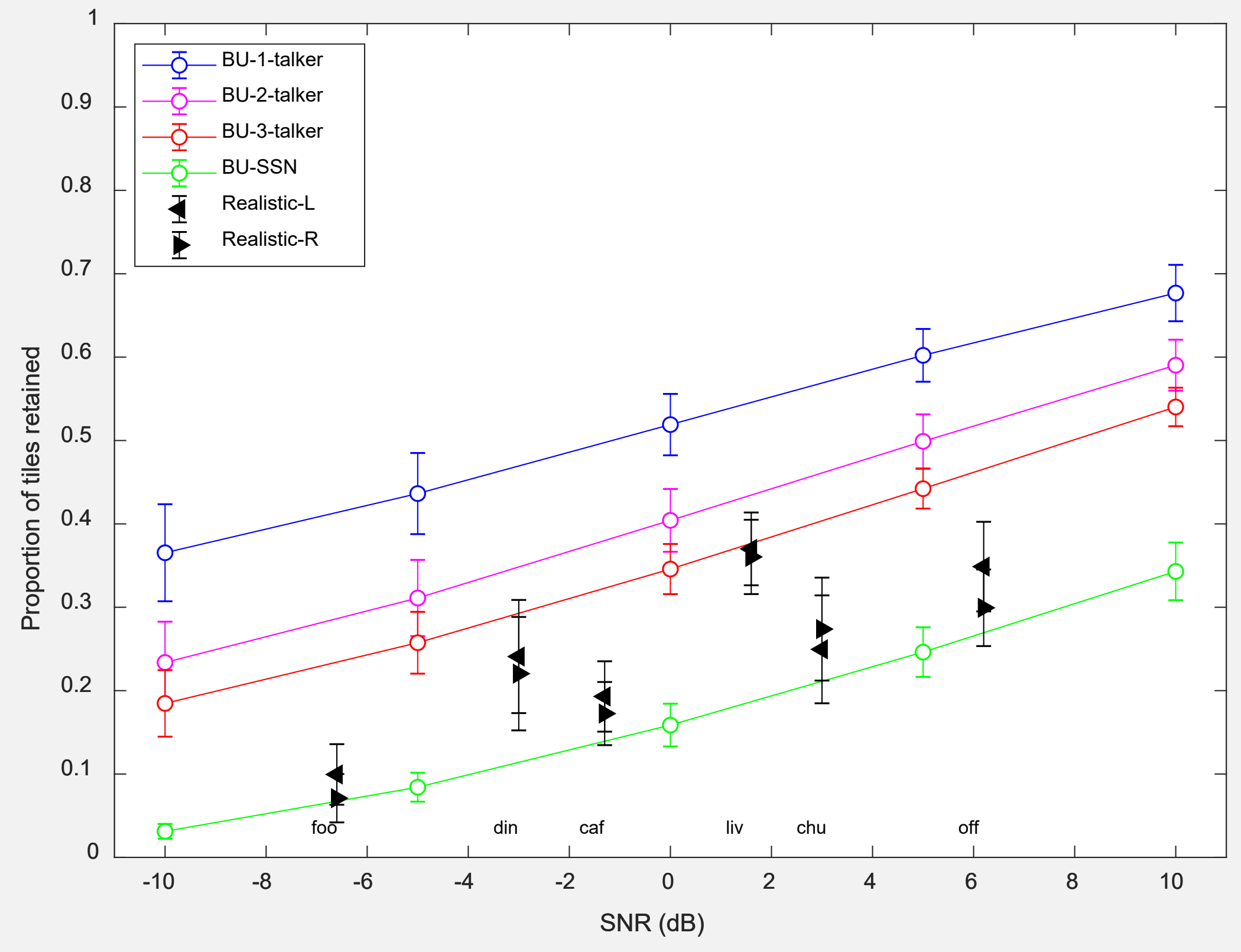
- Speech stimuli were taken from a matrix corpus (BU corpus; [8]). These have a fixed five-word structure and are clearly spoken.
- Target sentences were presented against one, two, or three competing masker sentences or a speech-shaped noise (SSN) masker.
- These mixtures were not spatialized.

Glimpsing

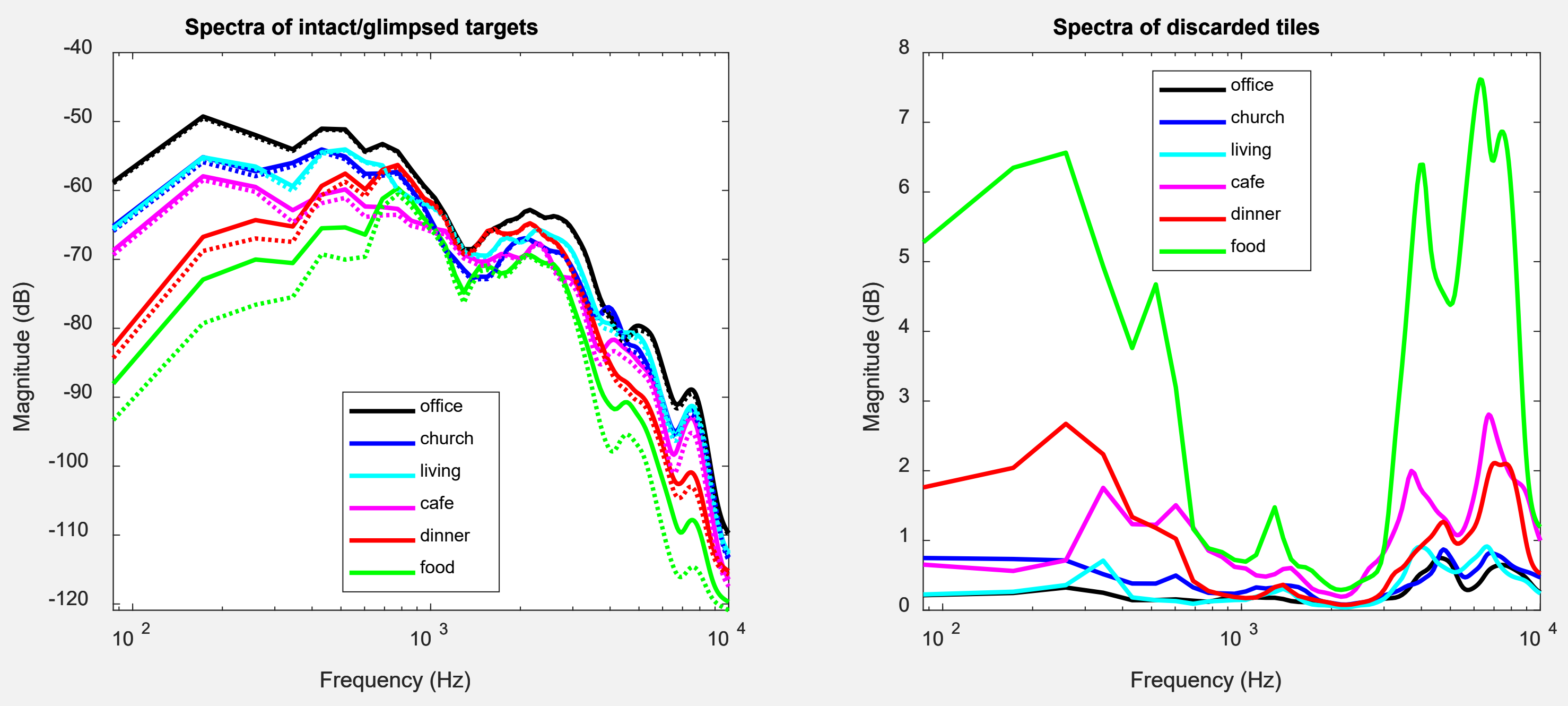
- Target glimpses were isolated using ideal time-frequency segregation ([9]) in which the mixture is divided into time-frequency “tiles” and only tiles for which the local SNR exceeds 0 dB are retained.
- Tiles were defined using 128 frequency channels logarithmically spaced between 80 Hz and 8 kHz, and 20-ms time windows with 50% overlap.

RESULTS

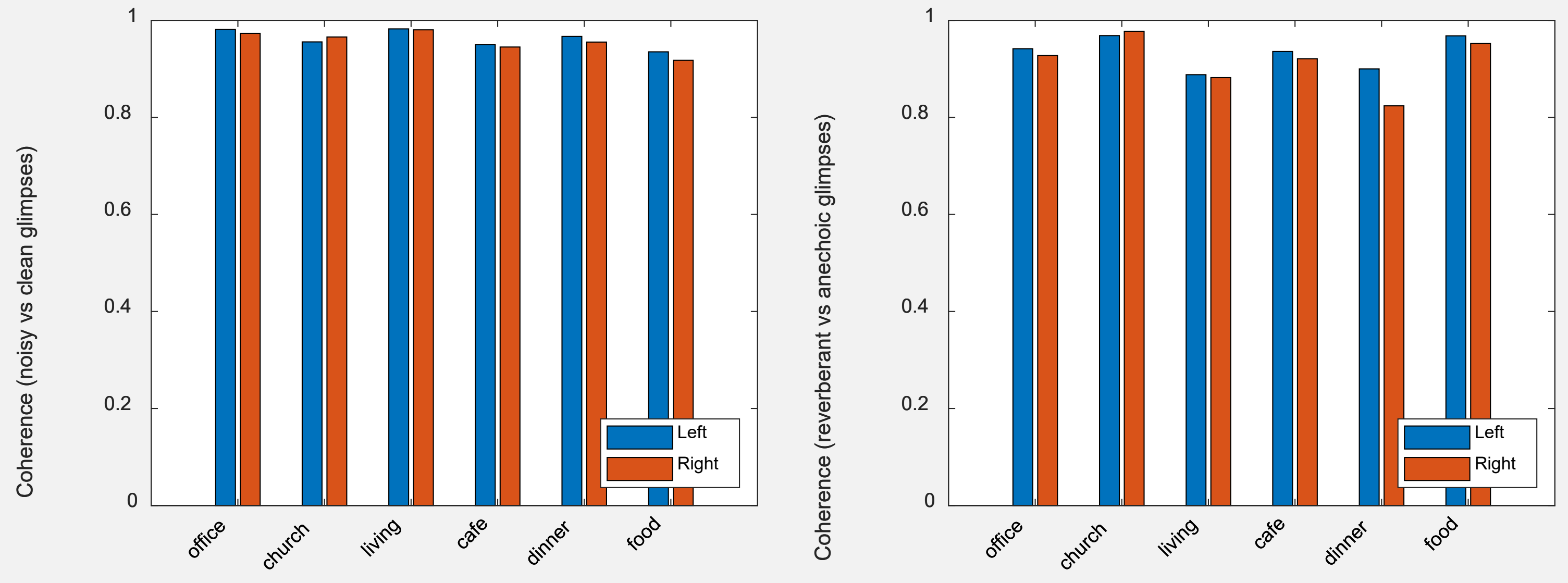
(A) At equivalent SNRs, realistic mixtures have fewer speech glimpses retained than laboratory speech-in-speech mixtures but more than laboratory speech-in-noise mixtures.



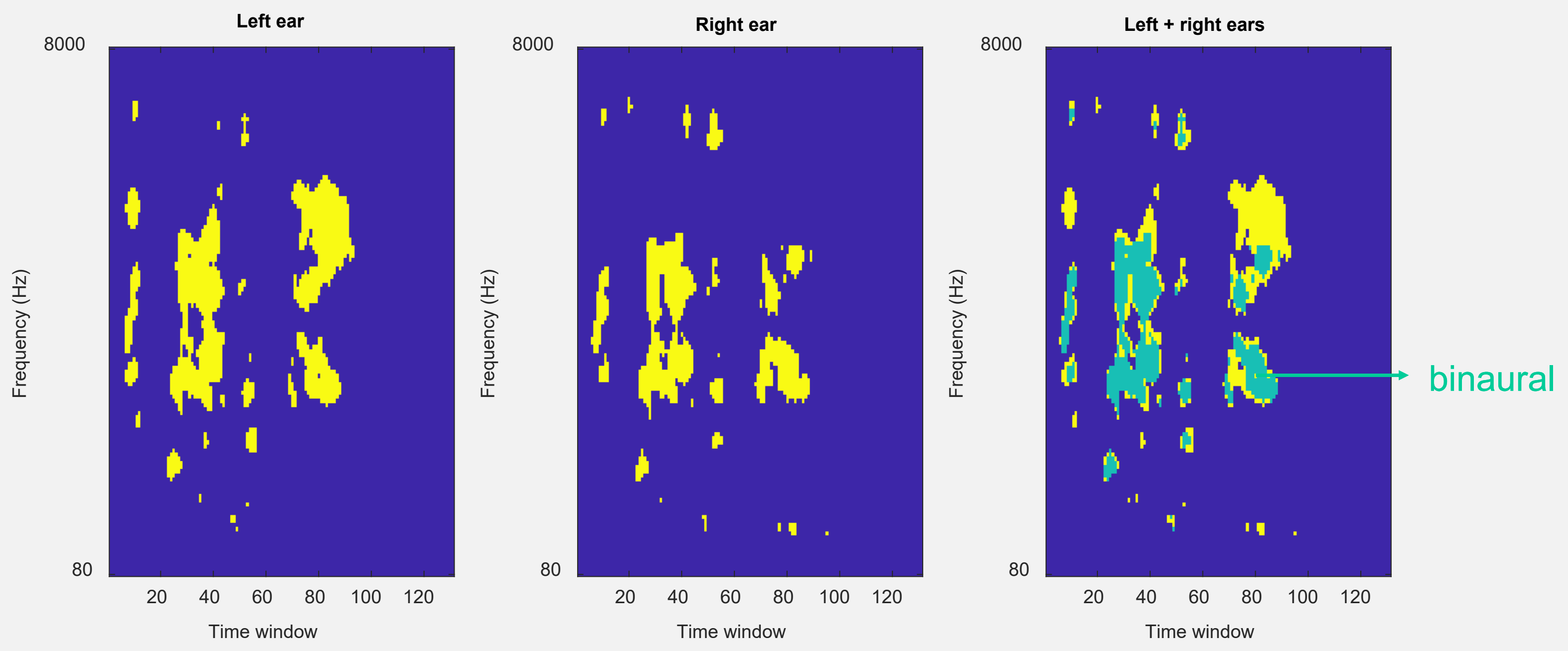
(B) In realistic mixtures, speech glimpses are primarily lost at low and high frequencies (equivalent to the previously reported SNR peak between 1-4 kHz [5]).



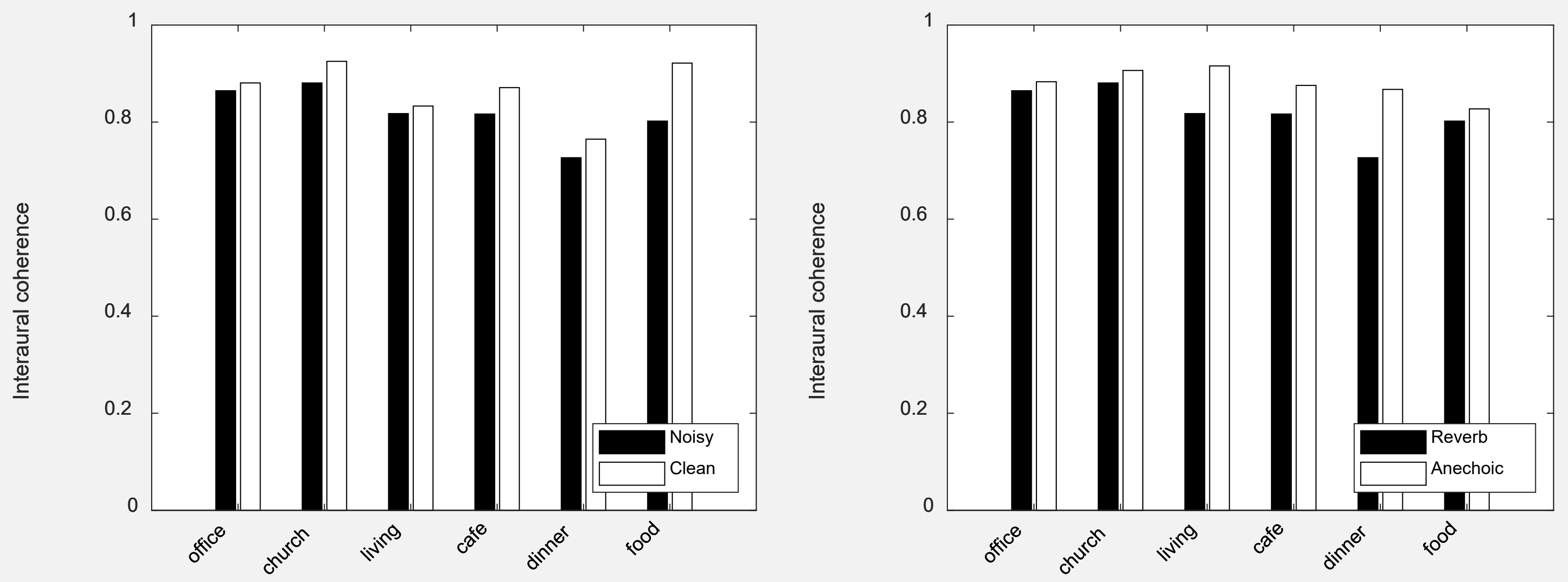
(C) Both noise and self-reverberation in realistic speech glimpses reduces their “quality” (as defined by their coherence with clean/anechoic glimpses).



(D) Realistic glimpse patterns are asymmetric; some glimpses are binaural while some occur only in one ear.



(E) Both noise and self-reverberation in binaural glimpses reduce their interaural coherence.



CONCLUSION

The acoustics of real environments differ from laboratory stimuli, and communication may be hindered by the number, quality, and binaural properties of the available speech glimpses.

REFERENCES

1. Goverts & Colburn (2020) Binaural recordings in natural acoustic environments: estimates of speech-likeness and interaural parameters. Trends Hear 24:2331216520972858.
2. Best et al (2015) An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment. Int J Audiol 54:682-6902.
3. Mansour et al (2021) Speech intelligibility in a realistic virtual sound environment. J Acoust Soc Am 149:2791-2801.
4. Miles et al (2020) Development of the Everyday Conversational Sentences in Noise test. J Acoust Soc Am 147:1562-1576.
5. Weisser & Buchholz (2019) Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions. J Acoust Soc Am 145:349–360.
6. Mansour et al (2021) A method for realistic, conversational signal-to-noise ratio estimation. J Acoust Soc Am 149:1559-1566.
7. Weisser et al (2019) The Ambisonic Recordings of Typical Environments (ARTE) Database. Acta Acust 105:695-713.
8. Kidd et al (2008) Listening to every other word: Examining the strength of linkage variables in forming streams of speech. J Acoust Soc Am 124:3793–3802.
9. Brungart et al (2006) Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J Acoust Soc Am 120:4007-4018.

Research reported in this publication was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Number R01DC015760. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.