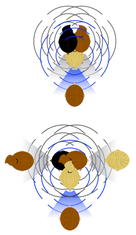
Interaction between talker head orientation, spatial separation, and extended high frequencies for speech-in-speech recognition

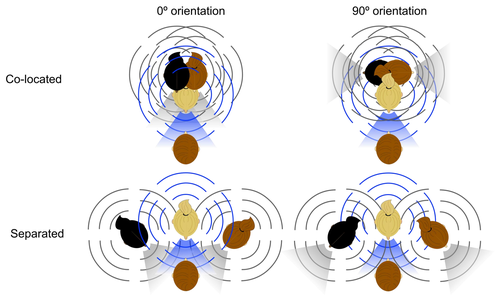
Rohit M. Ananthanarayana[1], Vahid Delaram[1], Allison Trine[1], Margaret K. Miller[2], G. Christopher Stecker[2], Emily Buss[3], Brian B. Monson[1]

1 Department of Speech and Hearing Science, College of Applied Health Sciences, University of Illinois Urbana-Champaign
2 Boys Town National Research Hospital, Omaha, Nebraska
3 Department of Otolaryngology/HNS, University of North Carolina at Chapel Hill

Introduction



- Speech-in-speech recognition experiments generally present stimuli as if target and masker talkers are facing the listener.
- In real-world situations, maskers are often rotated away from the listener, facing their own conversational partners.
- This target-masker head orientation mismatch provides cues to aid speech recognition, including cues at extended high frequencies (EHFs; > 8 kHz) due to the directional nature of EHFs in speech radiation[1].
- However, it is unclear how these EHF cues affect speech recognition when target and masker talkers are also spatially separated, as in realistic multi-talker situations.



Background

- A previous study[2] compared the benefits of non-facing masker head orientation (head orientation release from masking; HORM) and talker spatial separation (spatial release from masking; SRM) for speech recognition.
- Masker head orientation was either 0° or 60°, while target-masker spatial separation was either 0° or ± 54° azimuth.
- Results indicated that HORM was larger with co-located talkers but also observed for spatially separated talkers.
- In adults with normal EHF pure-tone thresholds, HORM in the co-located condition was comparable to SRM.
- Speech recognition performance in the non-facing masker condition was correlated with 16-kHz pure-tone thresholds.
- These data suggest that EHF cues are beneficial for speech recognition in realistic auditory scenes.

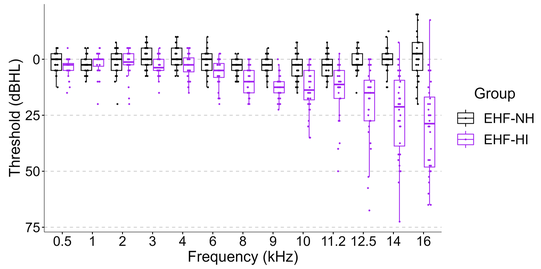
Current study

- We investigated EHF benefit in an auditory scene involving differences in talker head orientation and spatial location.
- EHF benefit was measured as the change in performance due to low-pass filtering speech stimuli at 8 kHz compared to presenting full-band stimuli.
- We hypothesized that the EHF benefit for speech-in-speech recognition would increase with differences in masker head orientation but reduce with talker spatial separation.

Methods

A. Participants

- 68 native English speakers (48 F, 17 M, 3 Other), age 18-49 years (mean 26.4 years) with clinically normal hearing.
- 36 participants had thresholds < 25 dB HL in both ears from 0.5-8 kHz and at EHFs (9-16 kHz; EHF-NH group).
- 32 participants had thresholds < 25 dB HL in both ears from 0.5-8 kHz but at least one elevated threshold at EHFs (EHF-HI group).



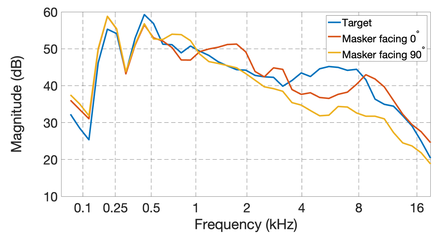
Average-ear pure tone thresholds

B. Stimuli

- Stimuli came from our publicly available corpus of anechoic recordings.
- Target speech: BKB sentences, female talker.
- Masker speech: narratives, two female talkers.

C. Conditions

- Spatial separation (Sep):
  - Target and masker co-located at 0° azimuth
  - Target at 0°, maskers at ±45° azimuth
- Masker head orientation (HO):
  - Facing the listener (0°)
  - Facing 90° away
- Filtering:
  - Full-band (FB)
  - Low-pass filtered at 8 kHz (LP8k)



ERB-scale long term average speech spectra of the target, facing masker and non-facing masker stimuli in the FB condition.

D. Procedure

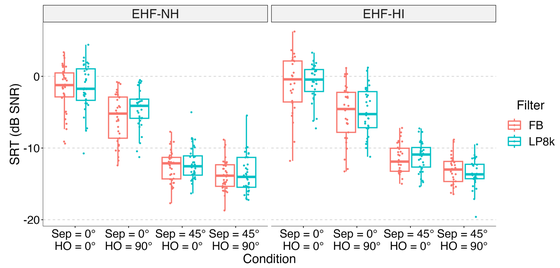
- Stimuli presented using loudspeaker array at 1-m radius.
- Masker level at 65 dB SPL, target level varied adaptively.
- Training block followed by eight experimental blocks (randomized order) with 32 trials each.

Methods (continued)

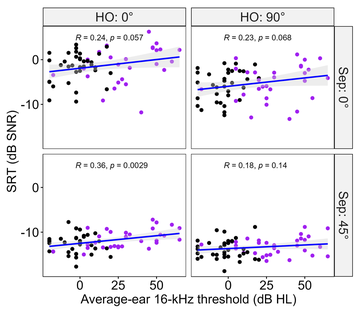
E. Analyses

- Speech reception threshold (SRT) – SNR required for 50% correct performance estimated for each condition by fitting psychometric function.
- Linear mixed effects (LME) models used to analyze effects of different conditions and EHF hearing thresholds on SRT.

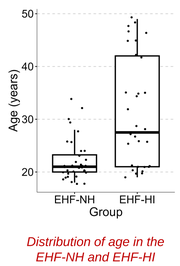
Results



SRTs in the 8 conditions grouped by hearing status



SRTs in FB conditions vs average-ear 16-kHz threshold



Distribution of age in the EHF-NH and EHF-HI groups

LME model outputs with EHF hearing status represented by (left) 'Grp' and (right) 'AvgEar16k'. 'Grp' compares EHF-NH vs EHF-HI and 'AvgEar16k' is the average-ear 16-kHz threshold.

	SRT	
Predictors	*Estimates*	*p*
(Intercept)	-5.916	**<0.001**
Filter [LP8k]	1.403	**0.001**
Sep [45]	-7.848	**<0.001**
HO [0]	4.204	**<0.001**
Grp [EHF-HI]	0.876	0.205
Filter [LP8k] × Sep [45]	-1.237	**0.038**
Filter [LP8k] × HO [0]	-1.203	**0.049**
Sep [45] × HO [0]	-2.845	**<0.001**
Filter [LP8k] × Grp [EHF-HI]	-1.181	0.060
Sep [45] × Grp [EHF-HI]	-0.343	0.580
HO [0] × Grp [EHF-HI]	-0.500	0.431
(Filter [LP8k] × Sep [45]) × HO [0]	1.248	0.139
(Filter [LP8k] × Sep [45]) × Grp [EHF-HI]	0.757	0.386
(Filter [LP8k] × HO [0]) × Grp [EHF-HI]	1.426	0.110
(Sep [45] × HO [0]) × Grp [EHF-HI]	0.982	0.264
(Filter [LP8k] × Sep [45] × HO [0]) × Grp [EHF-HI]	-1.019	0.409

	SRT	
Predictors	*Estimates*	*p*
(Intercept)	-6.034	**<0.001**
Filter [LP8k]	1.046	**0.004**
Sep [45]	-7.693	**<0.001**
HO [0]	4.058	**<0.001**
AvgEar16k	0.040	**0.007**
Filter [LP8k] × Sep [45]	-1.053	**0.037**
Filter [LP8k] × HO [0]	-0.725	0.159
Sep [45] × HO [0]	-2.748	**<0.001**
Filter [LP8k] × AvgEar16k	-0.016	0.255
Sep [45] × AvgEar16k	-0.024	0.074
HO [0] × AvgEar16k	-0.006	0.665
(Filter [LP8k] × Sep [45]) × HO [0]	0.994	0.163
(Filter [LP8k] × Sep [45]) × AvgEar16k	0.014	0.463
(Filter [LP8k] × HO [0]) × AvgEar16k	0.014	0.476
(Sep [45] × HO [0]) × AvgEar16k	0.027	0.160
(Filter [LP8k] × Sep [45] × HO [0]) × AvgEar16k	-0.017	0.530

Discussion

- Performance was better in the spatially separated than co-located conditions; performance was also better with the non-facing masker head orientation than facing.
- Performance was better when stimuli were full-band compared to low-pass filtered at 8 kHz; this EHF benefit was greatest in the spatially co-located non-facing masker condition, compared to spatially separated or facing masker conditions.
- SRM was larger than HORM, in contrast to a previous study[2] reporting similar magnitudes, possibly due to differences in head orientation angles and the nonlinear effects of directionality.
- Magnitudes of HORM and SRM were both reduced in presence of the other.
- There were no notable differences between the EHF-NH and EHF-HI groups; EHF benefit appeared lesser for EHF-HI listeners, but the difference was not significant.
- SRTs were correlated with average-ear 16-kHz threshold in the spatially co-located, facing masker condition.
- EHF-HI individuals were on average 8.9 years older than EHF-NH ($p<0.01$) and average-ear 16-kHz thresholds were significantly correlated with age ($r = 0.76$, $p<0.01$).
- With other predictors being the same, a linear mixed effects model with 16-kHz thresholds as the predictor of hearing status had lower AIC (2250.9) compared to EHF group (2255.3) or age (2253.3) as predictors.

Conclusions

- Extended high-frequency cues benefit speech-in-speech recognition in auditory scenes with realistic talker head orientations and spatial separation.
- EHF benefit is largest with spatially co-located talkers and maskers facing away from the listener; reduces with spatially separated talkers or maskers facing the listener.
- EHF pure-tone thresholds appear to affect utility of EHF cues, but their role is hard to dissociate from that of age.

References

1. Monson et al (2019). Ecological cocktail party listening reveals the utility of extended high-frequency hearing. Hearing Research, 381, 107773.
2. Braza et al (2022). Effect of masker head orientation, listener age, and extended high-frequency sensitivity on speech recognition in spatially separated speech. Ear and Hearing, 43(1), 90-100.

This study was supported by NIH grant number R01-DC019745 to BBM.