Extended High-frequency Cues to Phoneme Recognition: Insights from ASR

Paper #1125

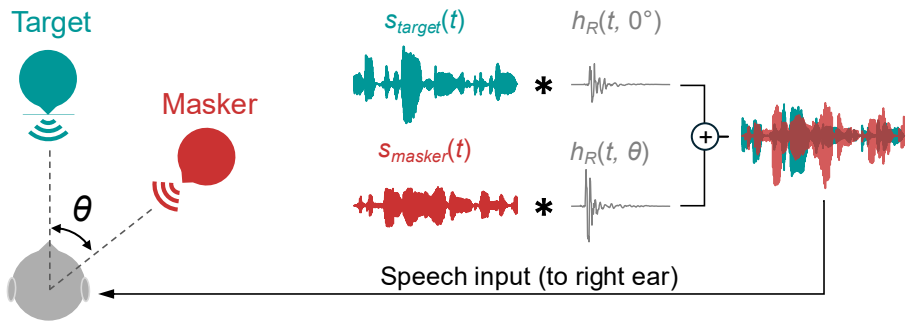
Zhe-chen Guo & Bharath Chandrasekaran
Department of Communication Sciences and Disorders, Northwestern University

1. Introduction

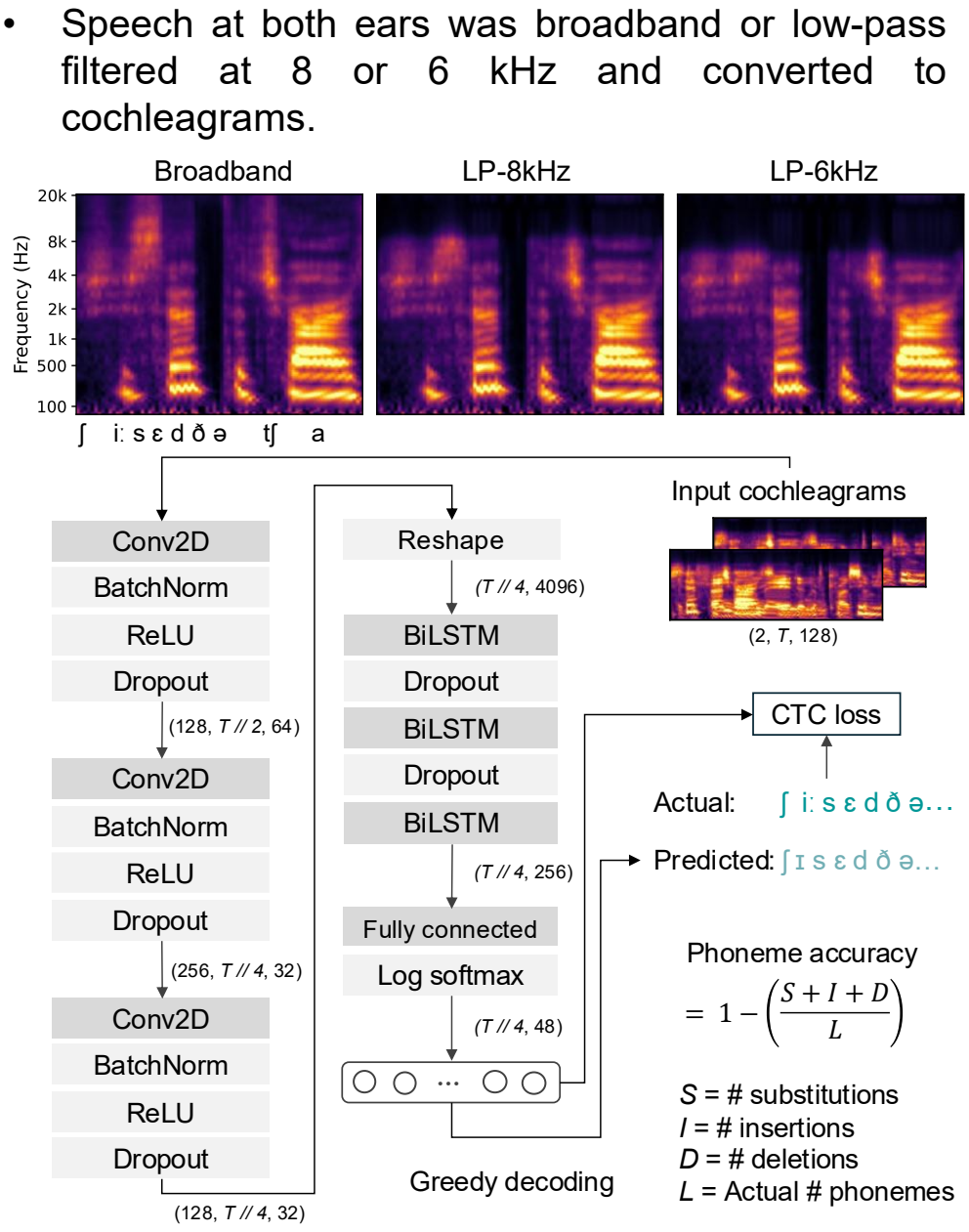
- **Extended high frequencies** (**EHFs**; >8 kHz) are often considered negligible for speech perception, excluded from audiometry and most ASR systems.
- EHF hearing improves speech perception in noise and predicts subjective hearing difficulties better than conventional audiograms (.25–8 kHz) [1–4].
- **EHF-audibility**: benefit arises directly from EHFs providing cues to **phoneme recognition**
  - Alternatively, it may indirectly reflect broader cochlear health [5–7].
- Evidence is limited: 1) unnatural stimuli (e.g., complete removal of lower frequencies) [8, 9]; 2) EHF effects modulated by spatial factors [10, 11].
- ASR models decoding phonemes from **cochleagrams**—a biologically-relevant speech representation—may provide useful insight [12].
- Broadband vs. low-pass filtered (at 8 and 6 kHz) speech in quiet and adverse spatial conditions.
- Do EHFs improve phoneme recognition? If yes, in what conditions? Are consonants more affected than vowels by lack of EHFs?

2. Experiment

- 13,636 recordings from British English speakers of VCTK corpus [13] (80% train, 10% val, 10% test).
- Target speech in quiet and synthesized spatial speech mixtures using head-related transfer functions [14] with separation θ = [±20°, ±45°, ±60°, ±120°] and target-to-masker ratio (TMR) = [+3, 0, −3, −6, −9, −12 dB SPL]



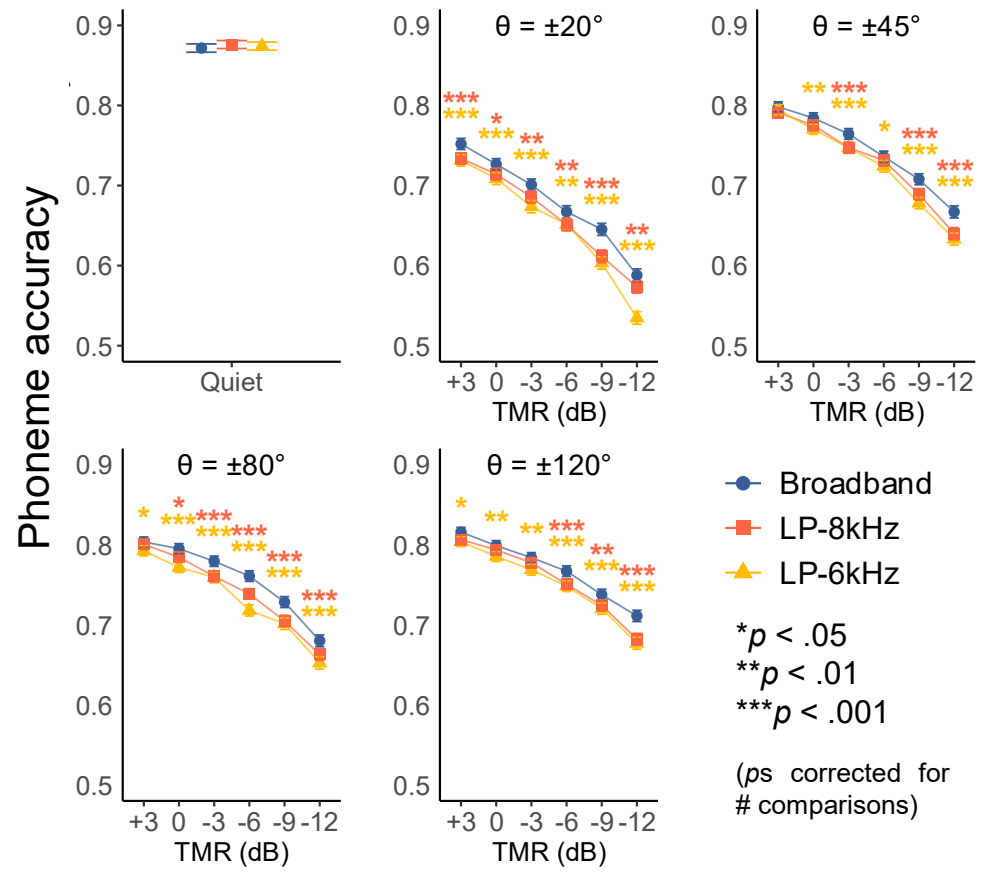
- Speech at both ears was broadband or low-pass filtered at 8 or 6 kHz and converted to cochleagrams.



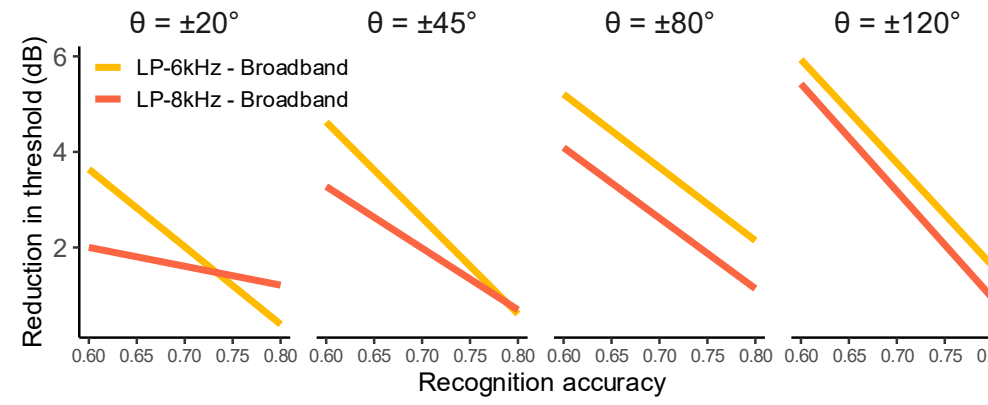
- Analysis of test-set results: 1) accuracy (broadband vs. LP-8kHz/6kHz); 2) posterior probability of each error type for each phoneme

3. Results

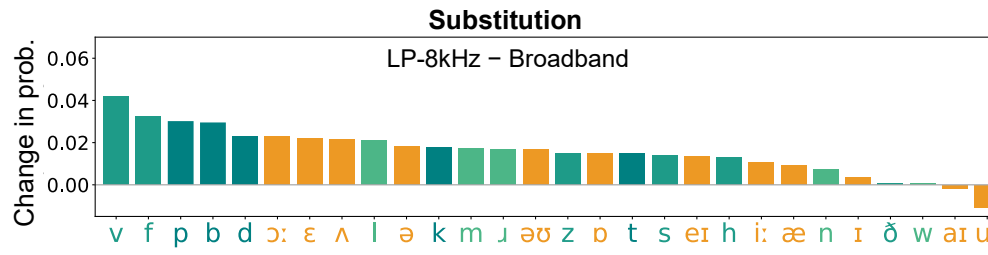
High-frequency cues above 8 and 6 kHz improved phoneme recognition in the presence of a masker, but not in quiet.



High-frequency cues reduced phoneme recognition thresholds, especially when recognition was difficult and when target-masker separation was large.



Removal of EHFs increased substitution errors more for consonants than vowels. No sig. difference in deletion and insertion errors.



4. Discussion

- Results from masked conditions suggested that EHFs provide direct cues to phonemes.
  - May partly explain the correlation between EHF hearing sensitivity and subjective hearing difficulties [2].
- Lack of EHF benefit in quiet highlights the need to also consider suboptimal listening situations.
- Removing EHFs affected /f, v/ the most, which have flat spectra with peaks close to 8 kHz [15, 16]
  - Also aligned with results from humans [3]
- Findings suggest reconsideration of EHFs in audiometric practice and ASR designs for spatially complex auditory environments.
- Future work: experiment with state-of-the-art ASR models (e.g., wav2vec 2.0) including EHFs.
  - Challenge: most benchmark datasets (e.g., LibriSpeech) use a 16-kHz sampling rate.

References:

The current work was supported by the Pat & Shirley Ryan Family Research Acceleration Fund awarded to B. C.