Spectral degradations in the TIMIT, QuickSIN, NU-6, and other popular bandlimited speech materials

Brian B. Monson[1] and Emily Buss[2]

1 Department of Speech and Hearing Science, College of Applied Health Sciences; Department of Biomedical and Translational Sciences, Carle Illinois College of Medicine; Neuroscience Program; University of Illinois Urbana-Champaign
2 Department of Otolaryngology/HNS, University of North Carolina at Chapel Hill

Introduction

The selection of speech materials for speech perception research directly affects measured behavioral outcomes. Many speech materials in use today were recorded decades ago using recording procedures that spectrally degraded the speech materials. For example, the use of low sampling rates (e.g., 16 or 22 kHz) was standard practice in speech research when some of the well-known and widely-used speech materials were recorded, resulting in materials that are bandlimited to 8 or 11 kHz. Additionally, transducers used for recording may have had low- and/or high-frequency roll-off or other variations in the frequency response that cut out or degraded spectral content. The result is **several speech corpora that do not represent the high-fidelity speech signals that listeners encounter in their everyday lives.**

Is the use of speech materials that are bandlimited to 8 kHz problematic? Several studies have demonstrated that extended high-frequency (EHF; >8 kHz) energy in speech is audible and useful for speech perception. EHFs in the 13-20-kHz band are audible for young, normal-hearing listeners. EHF cues in speech have been shown to support: (1) speech recognition in noise for adults and children, (2) consonant and vowel recognition when lower frequencies are either partially removed or entirely absent, (3) speech localization, (4) talker head orientation discrimination, and (5) subjective speech quality. In most studies, the utility of EHF speech cues was demonstrated by comparing the outcome measure for speech low-pass filtered at 8 kHz to that for full-bandwidth speech (Fig. 1). Moreover, EHF audiometric thresholds predict both EHF audibility in speech and speech-in-noise performance.

In this study, **we aimed to evaluate spectral degradations in commonly used speech materials** to ascertain potential effects on measured outcomes when using these materials for speech perception experiments.

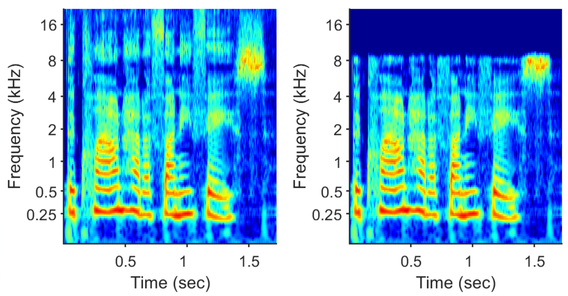

***Fig. 1**. Cochleogram of the sentence "The clown had a funny face," showing the spectral content for full-band speech (left) and speech bandlimited to 8 kHz (right).*

Methods

- Publicly available recordings (Table 1)
- A minimum of 50 samples (words or sentences) for each corpus, except the BEL corpus which had only 20 sentences from a male talker and 40 from female talkers
- Compared the long-term average speech spectra (LTASS) to anechoic full-band speech recorded at a sampling rate of 44.1 kHz using Class I precision microphones with flat frequency response out to 20 kHz (Monson et al., 2012; Monson et al., 2019)
- Speech materials manually edited to remove silence to prevent spectral content of noise floor from affecting LTASS
- NU-6 words were evaluated with the carrier phrase, "say the word" (excising the carrier phrase and recomputing the LTASS did not modify the primary features)
- LTASS computed using 2048-point fast Fourier transform (FFT) and a hanning window with 50% overlap; LTASS for the TIMIT/PRESTO computed using a 1024-point FFT because of its limited bandwidth
- To estimate expected between-talker variability, an individual LTASS was also calculated for each subject in the Monson et al (2012) recordings (10 male, 10 female); these spectra were used to characterize the range of LTASS values across talkers
- Each LTASS set to 65 dB SPL overall level

Acronym	Name	Talkers	Fs	Source/Ref.
TIMIT & PRESTO	Texas Instruments/Massachusetts Institute of Technology & Perceptually Robust English Sentence Test Open-set	Multiple male and female	16 kHz	Garofolo et al. (1993)
AzBio	Arizona Biomedical Institute Sentence Test	Multiple male and female	22 kHz	Spahr et al. (2012)
BKB-SIN	Bamford-Kowal-Bench Speech-In-Noise	One male		Bench et al. (1979); distributed by Etymotic
CNC	Consonant-Nucleus-Consonant	One male		Peterson, G. E., & Lehiste, I. (1962); Auditec
NU-6	Northwestern University Test No.6	One male		Auditec
Quick-SIN	Quick Speech-in-Noise	One female		Killion et al (2004); distributed by Auditec
SPIN-R	Speech-in-Noise Revised	One male		Bilger et al (1984); distributed by Cosmos
BEL	Basic English Lexicon	One male, two female	44.1 kHz	Rimikis et al (2013)
HINT	Hearing in Noise Test	One male	20 kHz	Nilsson et al (1994); distributed by Interacoustics
HIST	Hearing in Speech Test	One male	44.1 kHz	Levy et al (2015)
Monson	N/A	Multiple male and female	44.1 kHz	Monson et al (2012)
Monson BKB	N/A	One female	44.1 kHz	Monson et al (2019)

***Table 1.** Details of speech corpora analyzed in this study.*

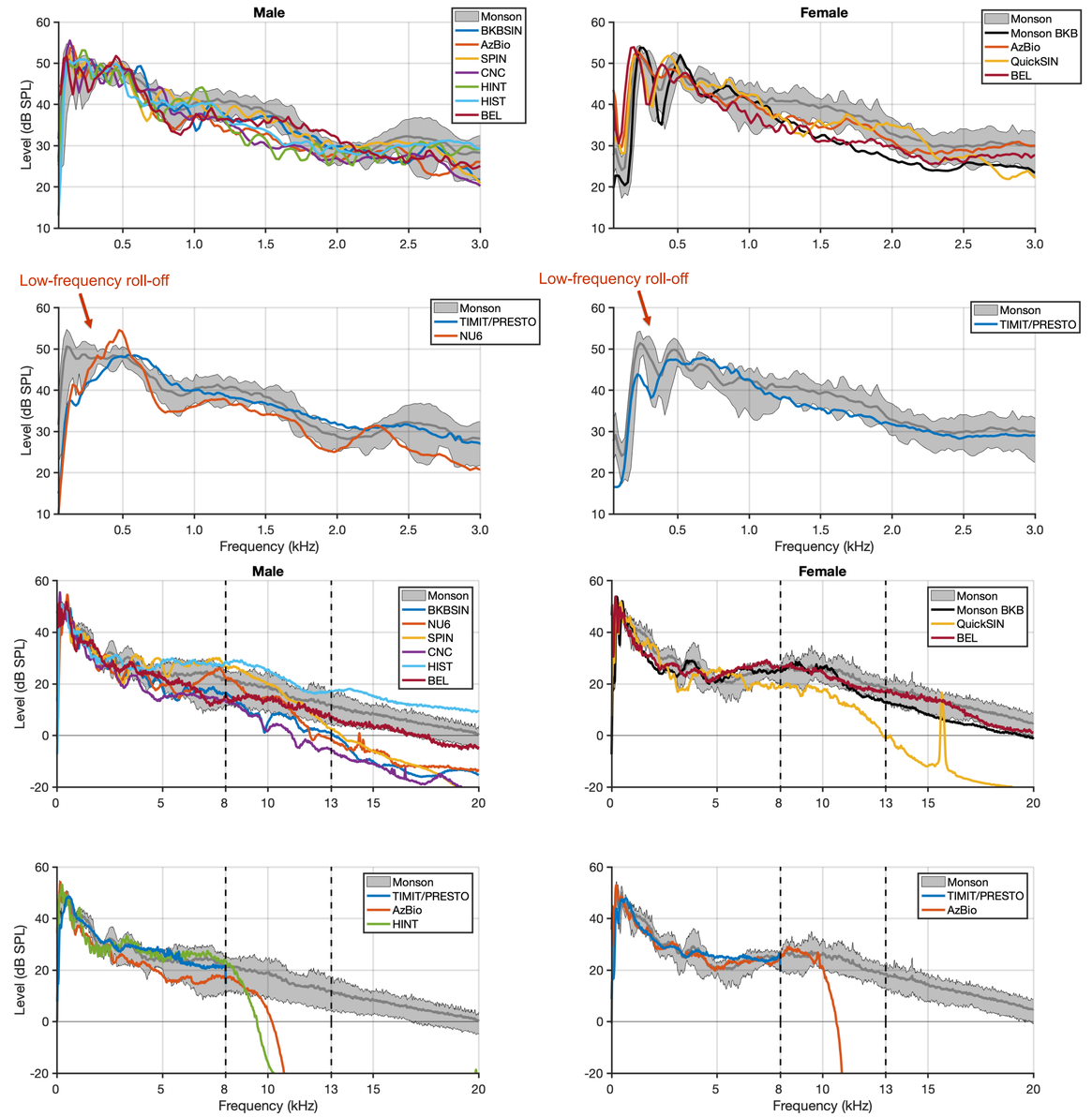

***Fig. 2**. LTASS zoomed-in view (upper 4 panels) and full-band view (lower 4 panels). Shading indicates range of levels across talkers in Monson et al (2012). Dotted lines show EHF boundary (8 kHz) and maximum audible low-pass filter cutoff frequency for speech (13 kHz).*

Conclusions

The use of decades-old speech materials that are degraded and exclude useful information is a practice worthy of reconsideration.

The use of corpora that exhibit both EHF and low-frequency degradations (i.e., TIMIT, NU-6, and QuickSIN) may be especially problematic, depending on the question of interest.

References

Monson, B. B. and Buss, E. (2022). "On the use of TIMIT, QuickSIN, NU-6, and other widely used bandlimited speech materials for speech perception experiments," JASA 152, 1639-1645.

Acknowledgements

This study was supported by NIH grant R01-DC019745.

ILLINOIS

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL