LISTENING PREFERENCE OF EMOTIONAL AND NEUTRAL SPEECH BETWEEN COMPRESSION SPEEDS

Petri Korhonen, Christopher Slugocki, Francis Kuk, Neal Ruperto
Widex Office of Research in Clinical Amplification (ORCA)

INTRODUCTION

The current study investigated how the emotional content of the speech materials influence the preference ratings for overall hearing aid sound quality when comparing amplitude compression algorithms. Specifically, the preference was examined using neutral and emotional speech. Speech in different emotion states is produced with distinct changes in speech production, and emotions are normally perceived through deviations from the normal/neutral state. One of the speech prosody features that conveys emotion is intensity contour. Amplitude compression, which provides time-varying gain so that more gain is available for soft level sounds, while less gain is provided during loud level sounds, may reduce the dynamic variation in the amplitude envelope of the input signal. Such changes in the amplitude envelope may distort the suprasegmental intensity contour cues of the speech. This may impede the listeners ability to use this information and may play a role in the acceptance of the sound. We hypothesize that the amplitude compression algorithm that better preserves the amplitude envelope will be preferred by the listeners, and that the preference is heightened for speech expressed with the intention to convey an emotion.

METHODS

Subjects

LISTENERS WITH HEARING IMPAIRMENT (HI):
N = 20 (10 females, 10 males)
Age: 52 - 83 yrs (mean = 69.7 yrs; SD = 10.0 yrs)
PTAs (4-freq): 48.6 and 49.8 dB HL (SD = 3.6) for the left and right ears respectively.
Symmetry of the hearing loss was within 5 dB from 250 Hz to 8000 Hz.
Hearing loss was sensorineural in nature.
18 listeners were regular HA wearers (mean experience = 21.8 yrs, SD = 17.9 yrs).

GROUP NH: LISTENERS WITH NORMAL HEARING:
N = 21 (12 females, 9 males)
Age: 51 - 78 yrs (mean = 61.7 yrs; SD = 6.7 yrs).
PTAs: < 20 dB HL at .5, 1 and 2 kHz, and < 30 dB HL at and beyond 4 kHz.
The four-frequency PTAs between the two ears was 14 dB HL (SD = 2.7 dB HL).

All participants were native speakers of English and exhibited normal cognitive function as assessed using the Montreal Cognitive Assessment (MoCA)(score ≥ 23).

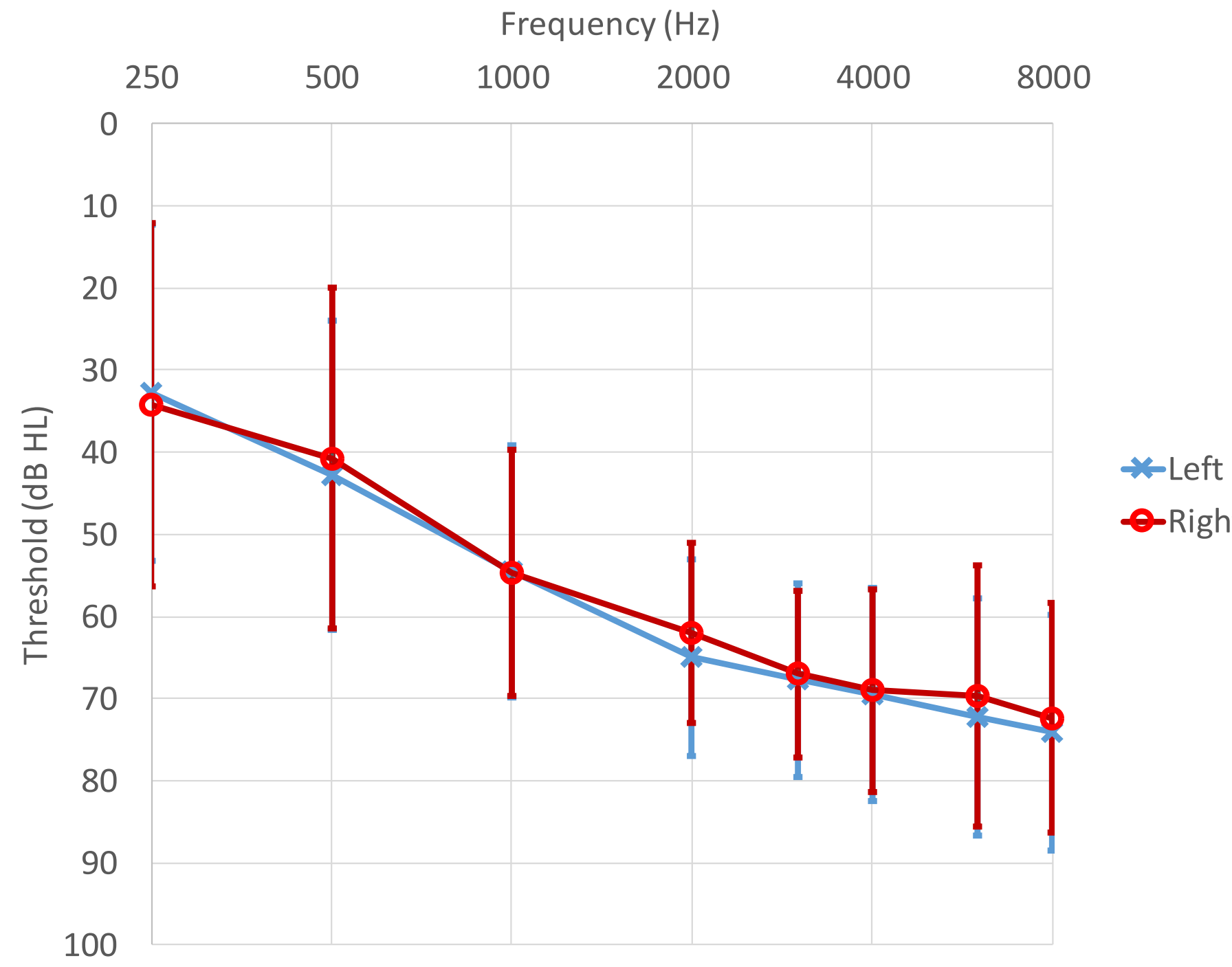

Figure 1. Average pure-tone thresholds of subjects with hearing-impairement. Error bars represent one standard deviation.

Stimuli

Two types of speech materials were used in the study: neutral speech and emotionally expressed speech. Five different neutral and five emotional speech passages were included. Sentences were taken from The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo 2018), European Broadcasting Union Sound Quality Assessment Material (EBU SQAM), and ORCA-US speech recordings. The emotions included: 'angry' (male), 'disgust' (male), 'fearful' (female), 'happy' (female), and 'surprise' (female). Neutral sentences included both male and female talkers.

The emotionally expressed speech materials had greater modulation variation than neutral speech materials across typical syllabic rates indicating greater dynamic variation when the talker was expressing an emotion (Figure 2).

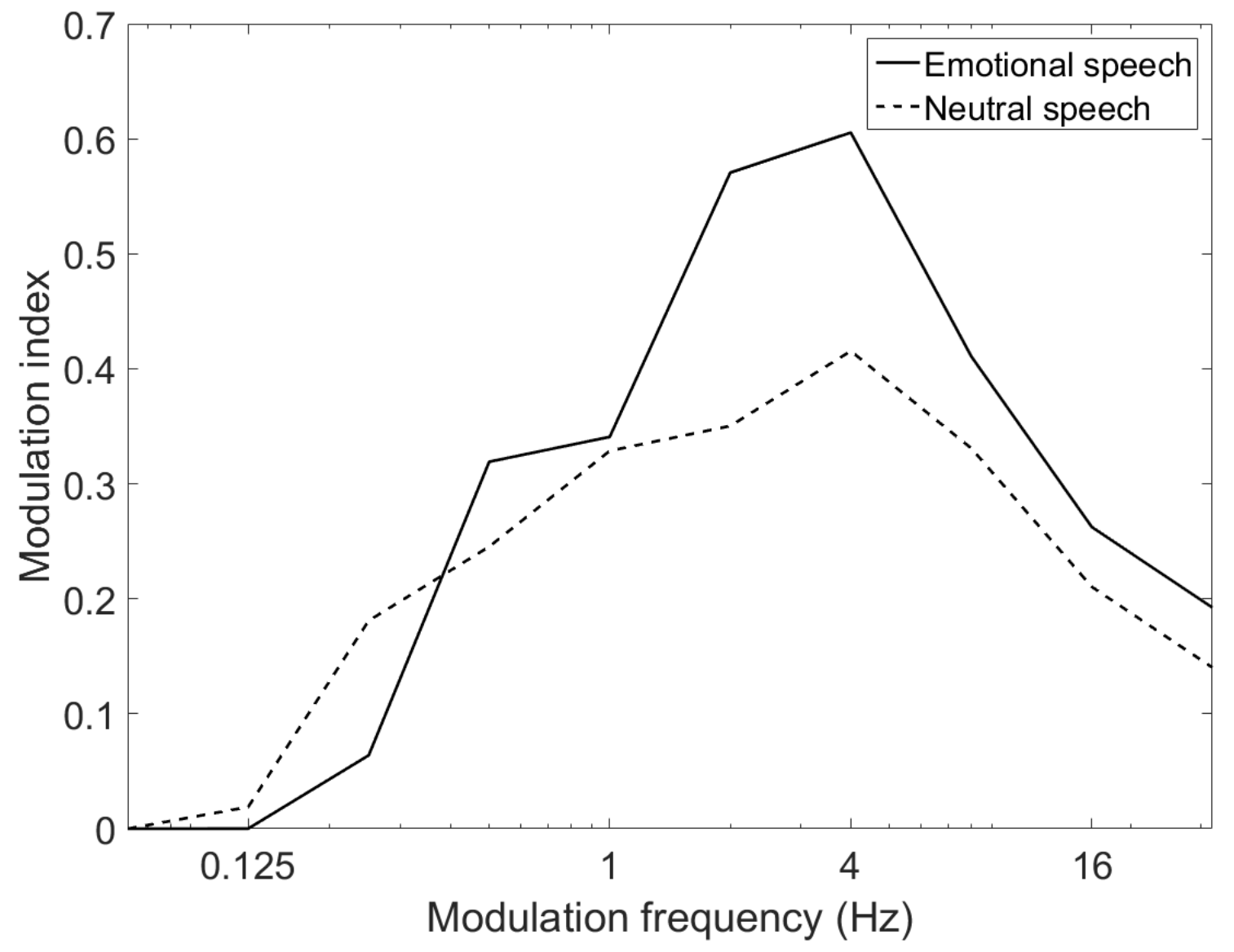

Figure 2. Modulation index of the unprocessed neutral and emotionally expressed speech materials derived from the broadband signal envelope (Korhonen et al 2019).

Hearing aids

HEARING AIDS:
15-channel wide dynamic range compression (WDRC) receiver-in-the-ear (RIC) hearing aids with a compression threshold as low as 0 dB HL. AD-converter sampling rate: 33.1 kHz, 18-bits resolution. The input dynamic range linear between 5 and 113 dB SPL. All the special features except feedback canceller were deactivated during the study. An omnidirectional microphone was used.

PROCESSING CONDITIONS:
Variable speed compressor (VSC) which includes two separate signal-processing branches, slow compressor (SC) and fast compressor (FC), operating simultaneously in parallel to determine the overall output gain. The attack time ≈ 1.5 sec, and the release time is ≈ 17 sec. The compression ratio in the SC block was hearing loss dependent and varied between 1.3 and 2.4 across frequencies. Compression threshold was 19 - 45 dB HL (frequency dependent) for the average hearing loss. The FC branch determines the gain based on the differences between the short- and long-term average input levels. When the short-term average input level is higher than the long-term average input level, gain determined by the FC block is reduced proportionally and vice versa. The attack time of the FC block is ≈ 12 ms, and the release time is ≈ 130 ms.

Fixed-speech fast acting compressor (FAC) with fixed attack time of 5 ms and a fixed release time of 50 ms. The CR and CT of the FAC match the values used by the SC of the VSC. This processing condition is only available in the developmental model of the hearing aid and is not a standard commercial feature.

FITTING:
Hearing aids were fitted using fully occluding instant-fit ear-tips to minimize the influence of any unintended direct sounds entering via leakages between the eartip and the ear canal. The NAL-NL2 rationale was used to set the target gain. All fittings were verified using a simulated speech mapping measure (SoundTracker) in the fitting software.

Procedures

Prerecorded stimuli was presented offline using insert earphones during data collection. The earphone frequency response was compensated for a flat response.

RECORDING
Test materials were recorded using KEMAR head and torso simulator with fully occluding earmolds. During the recording the speech stimuli were presented at 75 dB SPL. For the HI group the recordings were carried out individually for each subject using the hearing aids fitted for his/her hearing loss. For the NH listeners the hearing aid was programmed with flat 40 dB HL across frequencies.

DATA COLLECTION
Sound quality preference between the two compression settings was evaluated using single blinded paired comparison task. The order of compression settings were counterbalanced so that each processing condition (VSC or FAC) was presented first equal number of times. Participants were required to listen to the samples at least once and they were allowed to repeat listening to the samples as many times as they wished. Each sample was tested three times. Participants were asked to indicate their overall preference using a touch screen. During the data collection the earphone presentation level was the same as measured during recording to achieve "natural" level of amplification for HI group. For the NH group the earphone presentation level was set to 75 dB SPL to avoid potential loudness discomfort.

RESULTS

Subjective preference

One-way binomial test showed that 19 out of 21 NH listeners preferred VSC over FAC when listening to emotional speech ($p < 0.001$, one-sided), and 18 out of 21 when listening to neutral speech ($p < 0.001$, one-sided). 19 out of 20 HI listeners preferred VSC over FAC when listening to emotional speech ($p < 0.001$, one-sided), and 12 out of 20 when listening to neutral speech ($p = 0.251$, one-sided). A linear mixed effects model showed that the strength of preference for VSC over FAC processing was significantly affected by Hearing Group ($\chi^2(1) = 5.73$, $p = 0.017$) and Stimulus Type ($\chi^2(1) = 22.00$, $p < 0.001$) as qualified by their interaction ($\chi^2(1) = 4.73$, $p = 0.030$). This interaction reflects that HI group showed less of a preferrence for VSC when neutral speech was used as the stimuli (Figure 3).

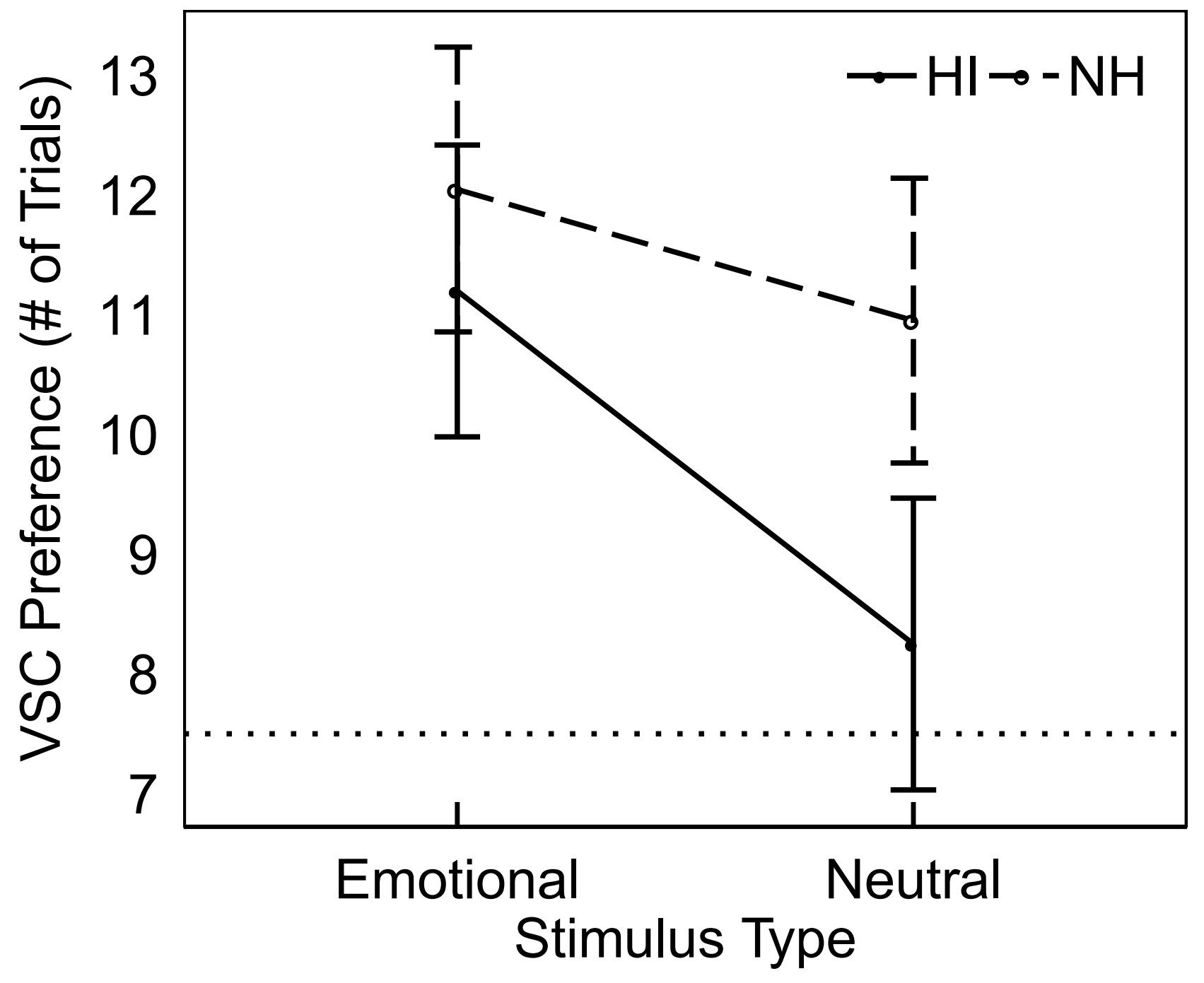

Figure 3. Preference for VSC over FAC when judging the sound quality of speech. Points represent marginal means and error bars represent 95% CIs of the means. The horizontal dotted line represents the expected outcome for "no preference."

Acoustic analysis of processed signal

Modulation spectrum of the broadband amplitude envelope was analyzed for the two types of stimuli (neutral and emotional) under the two processing conditions (VSC and FAC) averaged across all study stimuli and all study participants. Figure 4 displayed the modulation index of the stimuli used for NH and HI groups separately. For HI listeners the hearing loss was accounted for in the derivation of the modulation index so that minimum value for amplitude envelope was limited to the hearing threshold of the listener.

For both listener groups the modulation index was larger with VSC than with FAC across most modulation frequencies. This suggests that the VSC processing preserved the amplitude envelope variations better than FAC. The absolute difference in the average modulation depth between VSC and FAC processing conditions was greater for the NH group than for the HI group. Similarly, the modulation index was higher for emotional speech than for neutral speech at modulation frequencies between 0.5 and 16 Hz. The overall modulation indices were greater for the NH listeners than for the HI listeners.

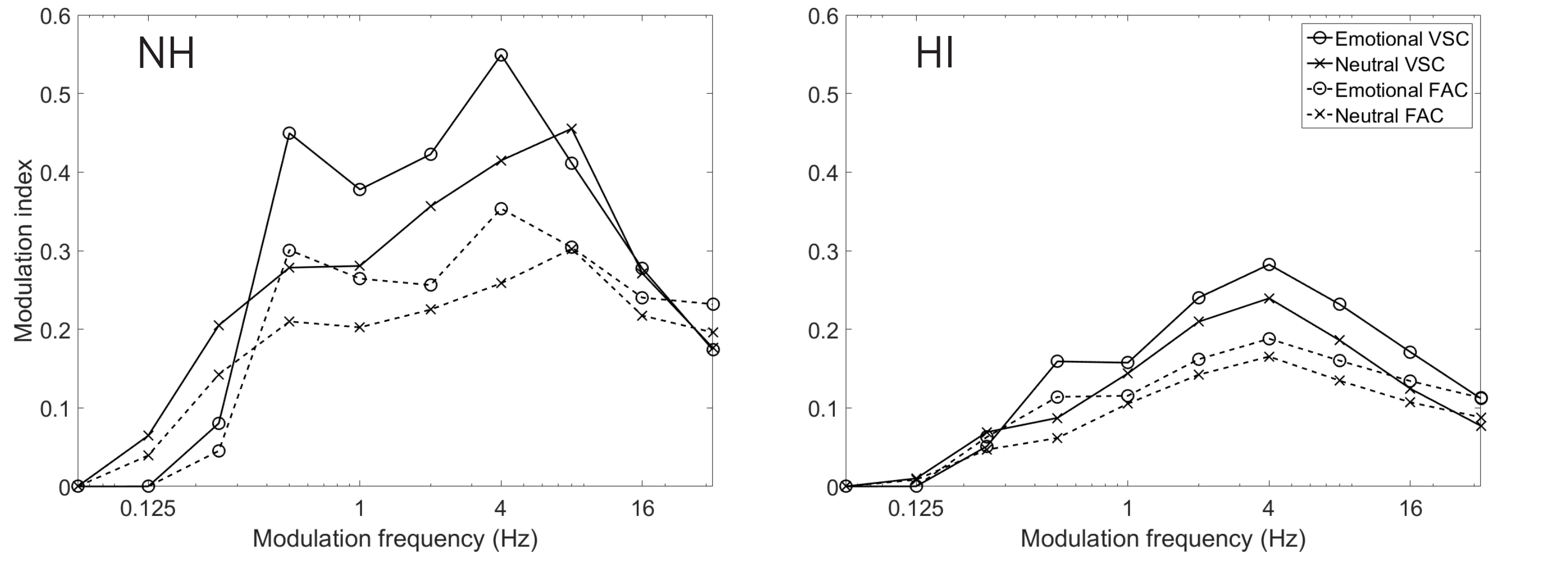

Figure 4. Modulation index for the processed study stimuli. Normal hearing (left) and hearing impaired averaged across listeners (right).

DISCUSSION

The current study demonstrated that the use of emotional speech materials may be more sensitive than neutral speech in differentiating between amplitude compression algorithms when using overall sound quality preference as a criterion. Data showed that the listeners preferred variable speed compression over fast acting compression particularly when listening to emotional speech. In the light of this data, clinicians should consider including emotionally valanced speech materials when evaluating listener preference for different hearing aid compression speed options in the clinic. Because emotionally valanced speech is part of real-life communication, its inclusion in clinical evaluation of hearing aid processing may have better ecological validity than the use of neutral speech.

REFERENCES

Korhonen P, Kuk F, Slugocki C. (2019) A Method to Evaluate the Effect of Signal Processing on the Temporal Envelope of Speech. Hear Rev 26(6):10-18.

Livingstone SR, Russo FA. (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.