Published in final edited form as:

J Am Acad Audiol. 2013 January ; 24(1): 26–36. doi:10.3766/jaaa.24.1.4.

Development, Reliability and Validity of PRESTO: A New High-Variability Sentence Recognition Test

Jaimie L. Gilbert¹, Terrin N. Tamati², and David B. Pisoni^{1,3}

¹Department of Psychological and Brain Sciences, Indiana University, Bloomington, 1101 East Tenth Street, Bloomington IN 47405

²Department of Linguistics, Indiana University, Bloomington, Memorial Hall 322, Bloomington, IN 47405

³Department of Otolaryngology – Head and Neck Surgery, Indiana University School of Medicine, 702 Barnhill Drive, Suite 0860, Indianapolis, IN 46202

Abstract

Background—There is a pressing need for new clinically feasible speech recognition tests that are theoretically motivated, sensitive to individual differences, and access the core perceptual and neurocognitive processes used in speech perception. PRESTO (Perceptually Robust English Sentence Test Open-set) is a new high-variability sentence test designed to reflect current theories of exemplar-based learning, attention and perception, including lexical organization and automatic encoding of indexical attributes. Using sentences selected from the TIMIT (Texas Instruments/Massachusetts Institute of Technology) speech corpus, PRESTO was developed to include talker and dialect variability. The test consists of lists balanced for talker gender, keywords, frequency, and familiarity.

Purpose—Investigate performance, reliability, and validity of PRESTO.

Research Design—In Phase I, PRESTO sentences were presented in multi-talker babble at four signal-to-noise ratios (SNRs) to obtain a distribution of performance. In Phase II, participants returned and were tested on new PRESTO sentences and on HINT (Hearing In Noise Test) sentences presented in multi-talker babble.

Study Sample—Young, normal-hearing adults (N=121) were recruited from the Indiana University community for Phase I. Participants who scored within the upper and lower quartiles of performance in Phase I were asked to return for Phase II (N=40).

Data Collection and Analysis—In both Phase I and Phase II, participants listened to sentences presented diotically through headphones while seated in enclosed carrels at the Speech Research Laboratory at Indiana University. They were instructed to type in the sentence that they heard using keyboards interfaced to a computer. Scoring for keywords was completed offline following data collection. Phase I data were analyzed by determining the distribution of performance on PRESTO at each SNR and at the average performance across all SNRs. PRESTO reliability was analyzed by a correlational analysis of participant performance at test (Phase I) and retest (Phase II). PRESTO validity was analyzed by a correlational analysis of participant performance on

Address Correspondence to: Jaimie L. Gilbert, Ph.D., Speech Research Laboratory, Department of Psychological and Brain Sciences, Indiana University, 1101 East Tenth Street, Bloomington, IN 47405, Phone: (812) 855-6204; Fax: 812-855-1300; jaigilbe@indiana.edu.

Portions of the research were presented at the American Academy of Audiology meeting in Chicago, IL, April 6 – 9, 2011; at the First International Conference on Cognitive Hearing Science for Communication in Linköping, Sweden, June 19 – 22, 2011; and at the Aging and Speech Communication Conference in Bloomington, Indiana, October 10 – 12, 2011.

PRESTO and HINT sentences tested in Phase II, and by an ANOVA analysis of within subject factors of Sentence Test, SNR, and a between subjects factor of Group, based on level of Phase I performance.

Results—A wide range of performance on PRESTO was observed; averaged across all SNRs, keyword accuracy ranged from 40.26 to 76.18 percent correct. PRESTO accuracy at re-test (Phase II) was highly correlated with Phase I accuracy ($r = .92, p < .001$). PRESTO scores were also correlated with scores on HINT sentences ($r = .52, p < .001$). Phase II results showed an interaction between Sentence Test type and SNR [$F(3, 114) = 121.36, p < .001$], with better performance on HINT sentences at more favorable SNRs and better performance on PRESTO sentences at poorer SNRs.

Conclusions—PRESTO demonstrated excellent test/re-test reliability. Although a moderate correlation was observed between PRESTO and HINT sentences, a different pattern of results occurred with the two types of sentences depending on the level of the competition, suggesting the use of different processing strategies. Findings from this study demonstrate the importance of high-variability materials for assessing and understanding individual differences in speech perception.

Keywords

Speech Recognition; Individual Differences; Talker Variability

Introduction

A persistent challenge in the field of audiology is the development of new theoretically motivated spoken word recognition tests that are practical and intuitive. One of the pressing issues deals with the development of clinically feasible tools that access the core, foundational, underlying neurocognitive processing strategies individuals use in everyday adverse listening conditions to understand speech and communicate successfully. Early tests of speech perception were designed to assess audibility of isolated speech sounds using phonetically balanced materials (Egan, 1948). Given that a majority of speech communication in daily life involves the use of meaningful sentences containing linguistically complex semantic and syntactic structures, sentence recognition tests were also developed (Hirsh et al, 1952). However, the theory behind the development of sentence recognition tests still focused primarily on the audibility of speech sounds in familiar words, e.g., the Hearing in Noise Test or HINT (Nilsson et al, 1994). The original purpose of HINT was to obtain a threshold level, which is well matched to audibility. However, over time, the HINT sentences have become the most commonly used, clinically standard material for obtaining speech recognition accuracy measures for clinical populations, particularly cochlear implant candidates (e.g., Fabry, 2008; Fabry, Firszt, Gifford, Holden, & Koch, 2009), deviating from the original theoretical motivation of the HINT.

Current theories of speech perception and spoken word recognition propose that episodic memory, lexical organization, and indexical properties all contribute to speech recognition success (e.g., Johnson, 1997; Pisoni, 1997; Goldinger, 1998; Luce and Pisoni, 1998; Pierrehumbert, 2006). Sentence tests in noise are useful in assessing speech recognition abilities because they require use of not only sensory but also cognitive resources (Akeroyd, 2008). This article describes the theoretical motivation behind the development of PRESTO (Perceptually Robust English Sentence Test Open-set), a new sentence recognition test, and reports results of preliminary studies on the reliability and validity of this novel assessment measure with young, normal-hearing listeners. Clinical adaptation of PRESTO will require further testing of individuals with a hearing loss. However, normal-hearing individuals may also display communication impairments if they have difficulty perceiving speech in adverse

listening conditions. Thus, it is also important to determine the range of individual differences present in the normal-hearing population. This study tested normal-hearing listeners to establish the feasibility of PRESTO before testing the more heterogeneous population of individuals with hearing loss who may differ in degree of hearing loss and specific frequencies where hearing loss is present.

Multiple Factors in Speech Perception

Sentence perception scores are difficult to compare across experiments, listeners, and stimulus materials because of the large number of factors and sources of variation that contribute to performance (for a recent review see Theunissen et al, 2009). A parallel to this dilemma also exists in the field of human memory. Jenkins (1979), and more recently Roediger (2008), have argued that subjects, events (stimulus materials), and encoding and retrieval processes all interact to influence the experimental outcomes of any given study of human memory. Roediger (2008) discussed the implication that experimental results in human memory research only hold true given the particular environment in which they were obtained, i.e., with all factors held constant.

Translating these observations from research on human memory to the field of speech perception, it can be stated with some degree of confidence that performance on any given sentence recognition test only holds true given the particular *target* (e.g., talker(s), sentential content), *background competition* (e.g., signal-to-noise ratio (SNR), white noise, speech-shaped noise, multi-talker babble), *individual perceiver* (e.g., native language, dialect, hearing status, cognitive resources) and *task goal* (e.g., keyword recognition, true/false or goodness judgments, voice identification, discrimination). The present study was designed to investigate speech recognition performance, specifically, spoken word recognition in high-variability, meaningful, English sentences.

Target Variability

The indexical properties of speech (or characteristics of the talker) have been found to be an important source of variance contributing to the perceptual processing of speech (Johnson and Mullennix, 1997). These properties are encoded automatically in parallel with early sensory, acoustic-phonetic information in the speech signal (Pisoni, 1997). However, adjustment to multiple talkers' voices also has been shown to impose an additional processing cost on spoken word recognition above and beyond effects observed under single-talker presentation conditions (Mullenix et al, 1989). Indexical properties of speech are implicitly encoded into memory (Palmeri et al, 1993; Nygaard et al, 1995), facilitating later word recognition based on perceptual learning of the talker's voice (Nygaard et al, 1994). These studies support the proposal that episodic encoding and memory of voice source information affects spoken word recognition performance by producing a talker-contingent (i.e., speaker-dependent) processing benefit in recognition when the same voice is used repeatedly.

Background Competition

Success in spoken word recognition tasks has also been found to vary with the type and level of signals in the listening environment involving spectral overlap (energetic masking) and/or perceptual interference (informational masking) (Pollack, 1975; Brungart, 2001a; Brungart et al, 2001b; Calandruccio et al, 2010). Competing speech causes the most interference in speech perception (Carhart et al, 1969; Carhart et al, 1975) because it produces both energetic and informational masking (Van Engen and Bradlow, 2007; Helfer and Freyman, 2009). Linguistic competition (Calandruccio et al, 2010) and variation in

talker regional dialect (Clopper and Bradlow, 2008) have also been found to influence speech recognition performance under more challenging SNRs.

Individual Differences

Given the same target speech signal presented in the same type and level of competition, all listeners do not recognize speech with the same degree of accuracy. Listeners vary substantially in their susceptibility to competing acoustic signals (Neff and Dethlefs, 1995; Richards and Zeng, 2001; Wightman et al, 2010). Investigating the interaction of cognition and audibility may provide new insights into the underlying basis for individual differences in speech perception abilities (see for example, Pisoni, 2000; Arlinger et al, 2009; Stenfelt and Rönnberg, 2009). Cognitive operations of interest include frontal cortex functions involving aspects of cognitive control and executive function, such as memory, learning, and attention.

Recently, elevated scores on two clinical subscales of executive function (Inhibit = response suppression and Working Memory = manipulation of objects in memory) were found in deaf children who use cochlear implants, when compared to normative scores (Beer et al, 2010). Additionally, working memory scores were found to be significantly correlated with children's speech perception in noise (Beer et al, 2010). Taken together, these findings suggest that differences in neurocognitive processes may underlie individual differences in speech perception. While no causal relation was documented, the findings of Beer et al (2010) suggest associations between cognition and speech perception. Another neurocognitive process – learning – may also be related to individual differences in speech perception. Individual differences in implicit (incidental) learning of visual sequences were found to be related to sentence recognition under degraded listening conditions (Conway et al, 2010). Thus, investigation of core, elementary neurocognitive processes may provide a novel approach to the study of individual differences in speech perception abilities. Another long-standing problem in the field of clinical audiology is that not all speech recognition tests reveal the same degree of individual differences (Wilson et al, 2001). For example, conventional low-variability single-talker speech recognition tests have been shown to be less sensitive to individual differences than perceptually robust high-variability tests (Gifford et al, 2008; Park et al, 2010).

Sentence Tests of Speech Recognition

PRESTO

PRESTO was constructed using sentence materials selected from the TIMIT speech corpus, a digital speech database with utterances produced by multiple talkers from various United States (U.S.) dialect regions (Garofolo et al, 1993). Speakers in the TIMIT speech corpus were identified as having one of eight different dialects, coinciding with a geographical region of the U.S. where they spent their childhood: New England, Northern, North Midland, South Midland, Southern, New York City, Western. Individuals who moved frequently during their childhood were classified into the dialectal category of Army Brat. Using the TIMIT corpus, Felty (2008) created sentence lists (18 sentences per list) that had equal numbers of keywords (76 keywords per list), male and female talkers (9 of each per list). Keywords varied in word familiarity and word frequency (Nusbaum et al, 1984). Lists were also balanced for the average keyword familiarity ($M = 6.9$) and average log keyword frequency ($M = 2.5$). Sentences in each PRESTO list also differed in length and syntactic structure. No sentence was repeated in any of the lists, and no talker was ever repeated within a single list. Each test list included speakers from at least five different U.S. dialect regions. Thus, PRESTO incorporated variability in words, sentences, talkers, and regional dialects.

HINT

The HINT was based on sentences originally compiled by Bench, Kowal, and Bamford (1979). The BKB materials were obtained from natural speech samples of hearing-impaired British children. Nilsson and colleagues (1994) adapted the BKB sentences for the HINT through a series of rigorous tests. For use in the United States, sentences were revised to remove phrases unique to British English, and were subsequently judged as “natural” by groups of American English speakers/listeners. In addition, sentences were selected or modified so that they were all the same length, and were equated for intelligibility at a fixed noise level. Lists were structured to ensure that phonemic distribution was equated across all lists.

HINT was originally designed as an adaptive speech recognition test based on the correct response of the entire sentence (see Nilsson et al, 1994). An intensity level measurement for the sentence speech reception threshold was determined by presenting HINT sentences in speech-shaped noise competition. However, current clinical use of HINT sentences routinely deviates from this original design; instead, HINT sentences are presented in quiet to obtain an accuracy measure, which is important for determining cochlear implant candidacy (e.g., Fabry, 2008). In the present study, HINT sentences were presented in multi-talker babble and responses were scored by keywords correct. This divergence from the original HINT and conventional protocol was intentional in order to directly compare performance on HINT and PRESTO under the same type and levels of competition, using the same methods of scoring.

Experiment Goals

The current experiment presents preliminary findings of performance on a new sentence recognition test, PRESTO. The theoretical motivation behind the development of PRESTO was to directly address the issue of multiple factors affecting speech perception performance. Specifically, PRESTO differs from all other available sentence recognition tests because: (1) talker-contingent learning cannot occur when the talker changes with every sentence in the list, and (2) target sentences contain talker, gender, and regional dialectal variability. The current experiment also explores the effect of competing multi-talker babble presented at different intensity levels on the perception two different types of sentences (PRESTO and HINT). The goals of the present study were to investigate the feasibility, reliability, and validity of PRESTO, and to assess individual differences by comparing performance within tests using an extreme groups design and across tests using the HINT, a widely-used, conventional, single-talker sentence recognition test.

The current study consisted of two phases. In Phase I, speech recognition scores for PRESTO sentences in 6-talker babble at 4 signal levels were obtained from 121 listeners. In Phase II, participants who scored in the upper or lower quartiles of this distribution were invited back to the laboratory and were retested on a new set of PRESTO sentences as well as a set of HINT sentences to assess reliability and validity of PRESTO. It was hypothesized that sentence variability on the PRESTO test would reveal individual differences as measured by the range and variance in score distribution. It was also predicted that accuracy on HINT would be greater than PRESTO because of the simpler more uniform sentence structure and the use of the same talker over repeated presentations of the test sentences.

Methods

Participants

Phase I—Participants were recruited from the Indiana University community and met the following inclusionary criteria: native speakers of American English with residency in the

U.S. before age 18, 18–39 years old, normal-hearing, and no significant history of hearing or speech disorders at the time of testing. Except for hearing status, inclusionary criteria were determined by self-report. Normal-hearing was defined as responding to pure tones presented at 25 dB HL at 250 – 8000 Hz in the right ear and in the left ear. Participants were allowed to miss a response to the first tone presented, due to possible misunderstanding of the test instructions. Tones were presented via the calibrated system (computer, headphones) that was also used to present the sentence stimuli. Data are reported from 121 participants (79 female, 42 male) that met inclusion criteria. Their mean age was 22.2 years (standard deviation $SD = 2.9$ years). Participants signed an Institutional Review Board (IRB) approved informed consent statement and were paid \$15 for 90 minutes of participation.

Phase II—Forty of the original Phase I participants returned to our laboratory within 9 months for another set of tests. They were retested on a new set of PRESTO sentences as well as a set of HINT sentences. Of these 40 subjects, 19 had scored within the upper quartile of the Phase I distribution (HiPRESTO Group), and 21 had scored within the lower quartile (LoPRESTO Group). Participants were paid \$35 dollars for the second phase of testing, including a \$15 bonus for completing the entire study.

Materials

Speech recognition performance in Phase I was measured using 10 PRESTO lists. Of the selected 180 utterances, 158 talkers produced one sentence and 11 talkers produced 2 sentences. Table 1 shows the dialect regions of the talkers used for the set of 180 sentences. Digital audio files containing the target sentences (equated in level) were mixed with random samples from a five-minute stream of 6-talker babble of three male talkers and three female talkers, all with General American dialects (with General American operationally defined as a non-marked, or supra-regional dialect, here operationally identified as speakers from New England, North Midland, South Midland or West). Four different signal-to-noise ratios (SNRs) were generated at +3 dB SNR (2 lists), 0 dB SNR (3 lists), –3 dB SNR (3 lists), and –5 dB SNR (2 lists), by holding the level of the target sentences constant and varying the level of the multi-talker babble. More lists were tested in the middle of the range of SNRs at 0 dB SNR and –3 dB SNR, because of the prediction that these SNR levels would be less susceptible to ceiling or floor effects.

Speech recognition performance in Phase II was assessed with 4 new PRESTO lists and HINT lists 1 to 4 (Nilsson et al, 1994). PRESTO sentences were produced by 69 different talkers (3 talkers produced 2 different utterances). Of these talkers, 12 also produced a sentence presented in Phase I, and 3 produced 2 sentences heard during Phase I. The dialect regions of the talkers who produced these 72 sentences are shown in Table 2. New sentence lists were presented to prevent any recall or practice effects from the sentences heard previously in Phase I. Both PRESTO and HINT sentences were mixed with multi-talker babble using the same methods employed in Phase I. One list of each sentence type was presented at each of the four SNRs (+3, 0, –3, and –5 dB SNR).

Procedures

Listeners were tested in groups of four or fewer in Phase I, but were tested individually in Phase II. Each participant was seated in front of a PowerMacG4 computer running MacOS 9.2 with diotic output to Beyer Dynamic DT-100 circum-aural headphones. Each target sentence was presented individually, and each sentence .wav file was equated to the same root-mean-square (RMS) amplitude level. Output levels of the target sentences were calibrated to be approximately 64 dB SPL. Computers were located in enclosed testing carrels in a quiet listening room used for speech perception experiments. Experimental programs were controlled by custom PsyScript 5.1d3 scripts. Pilot testing of procedures by

laboratory personnel suggested increased frustration and decreased effort on the task when listeners were presented with a continuous list of sentences at a difficult SNR. In an effort to maintain listener attention and effort throughout the entire experiment, the SNR levels were therefore randomized from trial to trial. As a result, each subject received the test sentences in a different random order.

Listeners typed the words they had heard into a dialog box displayed on the computer monitor. Partial answers and guessing were encouraged if listeners were unsure of their response. The experiment was self-paced, and all listeners heard each test sentence only once. Scoring was completed offline for keywords correct. Keywords in both PRESTO and HINT sentences were defined prior to testing (e.g., nouns, verbs, adjectives, and adverbs). Experimenters and experienced research assistants scored keywords correct according to a written documentation of scoring procedures. Any items for which there was uncertainty were discussed and a new rule for that situation was reached by consensus. Correct morphological endings were required for a response to be scored as correct. Homophones and responses containing minor spelling errors (within one keystroke of the correct letter) were scored as correct. In Phase II, SNRs were also randomized as in Phase I. Presentations were blocked according to test type; half of the subjects were tested on PRESTO first, and half were tested on HINT first.

In order to account for any differences in normality of the distributions, proportional accuracy scores were transformed to rationalized arcsine units (Studebaker, 1985) for statistical analyses. However, for clarity, the actual percent correct values are reported and plotted for better interpretation of means and standard deviations.

Results

Phase I PRESTO

Participants demonstrated a wide range of variability in word recognition accuracy on the PRESTO task. Averaged across all SNRs, mean PRESTO accuracy was 62.67 % ($SD = 5.87$ %). Overall performance ranged from 40.26 to 76.18 %. The range of performance across participants at each SNR and the averaged scores across all SNRs is shown in Figure 1. The means and standard deviations are listed in Table 3. Strong positive correlations were observed for all pairwise comparisons of accuracy for each SNR ($r = .61$ to $.80$, all $p < .001$), indicating that listeners maintained the same relative performance at different SNRs. Paired comparison t-tests revealed that listeners were more accurate at the better SNRs than at the poorer SNRs (all pairwise comparisons were significant, at $p < .001$).

PRESTO Reliability (Phase I vs. Phase II)

Different lists of sentences were presented in Phase II than in Phase I. However, in both Phase I and Phase II, participants were tested on speech recognition accuracy of high-variability speech, with lists containing the same range of linguistic and indexical characteristics. Data from the 40 participants who completed both phases of testing were analyzed to evaluate the test-retest reliability of PRESTO. Overall mean accuracy on the 4 PRESTO lists used in Phase II (1 list at each SNR) was 64.01% ($SD = 8.06\%$), with a range of performance from 46.71% to 80.26%. As in Phase I, listeners were more accurate at better SNRs (all pairwise comparisons significant, $p < .001$), and individual performance was strongly correlated between SNRs ($r = .59$ to $.73$, all $p < .001$). PRESTO overall accuracy was strongly correlated from Phase I to Phase II performance ($r = .92$, $p < .001$). Figure 2 shows a scatter plot of the overall percent keywords correct in the two test phases for each of the 40 participants (both groups). Performance was also correlated at each individual SNR across Phase I and Phase II ($r = .71$ to $.82$, $p < .001$).

Phase II HINT

Overall mean accuracy across the 4 HINT lists (1 list at each SNR) was 63.16% ($SD = 7.41\%$). Performance ranged from 36.03% to 74.26%. Individual performance was correlated across SNRs ($r = .37$ to $.55$, $p < .02$), with the exception that performance at +3 dB SNR was not correlated with performance at -5 dB SNR or at 0 dB SNR.

PRESTO and HINT Comparisons

Performance on PRESTO and HINT were significantly correlated ($r = .52$, $p = .001$). Figure 3 displays a scatter plot showing each participant's keyword accuracy on the two sentence tests. Phase II accuracy data were submitted to an analysis of variance (ANOVA) using within subject factors of Sentence Type (PRESTO, HINT), and SNR (+3 dB, 0 dB, -3 dB, and -5 dB), and a between subjects factor of Group (HiPRESTO, LoPRESTO). As expected, a significant difference was observed between the two extreme groups [$F(1, 38) = 37.65$, $p < .001$]. The HiPRESTO group performed better than the LoPRESTO group. A significant main effect of SNR was also observed [$F(3, 114) = 794.60$, $p < .001$] indicating decreased performance under more adverse listening environments. Performance on HINT type sentences was significantly better than performance on PRESTO type sentences [$F(1, 38) = 8.13$, $p = .007$]; however, several other factors interacted with the main effect of sentence type.

The interaction of SNR x Sentence Type [$F(3, 114) = 154.88$, $p < .001$] revealed that performance on each of the two sentence types depended on the level of the competition. In the two better listening conditions, scores on PRESTO were significantly lower than scores on HINT [+3 dB: $t(39) = 13.89$, $p < .001$; 0 dB: $t(39) = 6.52$, $p < .001$]. HINT scores at +3 dB SNR approached ceiling level ($M = 96.10\%$, $SD = 5.86\%$). In contrast, in the two poorer listening conditions, performance on PRESTO was better than performance on HINT (-3 dB: $t(39) = -5.04$, $p < .001$; -5 dB: $t(39) = -10.66$, $p < .001$]. Figure 4 shows box plots of the range of overall performance on HINT and PRESTO tests averaged across SNRs and at each individual SNR.

Listener group also interacted with Sentence Type [$F(1, 38) = 4.96$, $p = .032$]. As expected, the HiPRESTO participants were always better than the LoPRESTO participants on both Sentence Tests [HINT: $t(38) = 3.01$, $p = .005$; PRESTO: $t(38) = 8.95$, $p < .001$]. HiPRESTO participants scored significantly better on PRESTO ($M = 70.93\%$, $SD = 3.91\%$) than on HINT ($M = 66.52\%$, $SD = 4.89\%$) [$t(18) = 3.36$, $p = .003$]. However, LoPRESTO participants' performance was not significantly different across the two sentence tests (PRESTO: $M = 57.75\%$, $SD = 5.13\%$; HINT: $M = 60.12\%$, $SD = 8.07\%$).

Discussion

A group of 121 young, normal-hearing listeners who were tested in multi-talker babble displayed a wide range of performance on PRESTO, a new high-variability sentence recognition test. Phase I results established an average level of performance for young, normal-hearing listeners on PRESTO in multi-talker babble across a range of SNRs. PRESTO performance was also found to be reliable with a new set of sentences when re-tested on 40 of the original participants ($r = .92$). The strong test-retest reliability was found even with experimental design changes involving different lists and different numbers of lists.

An extreme groups analysis was carried out to compare individuals who had scored in the upper quartile of the Phase I distribution of PRESTO (HiPRESTO group) to individuals who had scored in the lower quartile of the Phase I distribution (LoPRESTO group). When retested, the HiPRESTO group consistently displayed better performance than the

LoPRESTO group. Results from Phase I and Phase II suggest that PRESTO provides a reliable measure of sentence recognition that can accurately measure performance over time, and consistently reveal individual differences in speech recognition performance.

To assess the validity of PRESTO, performance on this new high-variability test was compared to the same group of individuals' performance on HINT sentences, a conventional single-talker test of speech recognition, under the same type of competition. Performance on PRESTO and HINT sentences were significantly correlated, suggesting that the two tests measured closely related processes. However, the sentence tests interacted with the listening conditions. Participants performed better on HINT than PRESTO under more favorable conditions (+3, 0 dB SNRs), but they performed worse on HINT than PRESTO under poorer conditions (−3, −5 dB SNRs). Based on this crossover interaction, it is proposed that PRESTO and HINT materials measure related, but fundamentally different processes involved in speech recognition. The current results suggest that accuracy on HINT is related to a talker-contingent processing strategy. At more favorable listening levels, talker specific cues are readily available; however, at poorer listening levels, the cues upon which the listeners are relying (talker specific) are not as audible, leading to poorer accuracy. This effect would not occur with PRESTO sentences, because the talker is constantly changing, thus preventing the use of a processing strategy that relies upon talker-contingent learning.

Processing Strategies

There is now a rapidly growing body of evidence demonstrating that listeners encode, process, and store instance-specific episodic source details about a talker's utterance (Palmeri et al, 1993; Nygaard et al, 1995; Pisoni, 1997). Perceptual learning and adjustment to the vocal source occurs automatically, and listeners perform better on novel utterances from a repeated talker than novel utterances from an unfamiliar talker (e.g., Nygaard et al, 1994). For HINT sentences, listeners heard different utterances that were all produced by the same talker, providing an opportunity for perceptual learning of detailed, idiolectal, indexical source properties of speech. At the more favorable SNR, listeners were able to learn vocal source properties specific to the HINT talker. However, at more difficult SNRs, indexical characteristics of the talker's voice may not have been audible, preventing the encoding and use of talker-specific details. The availability of vocal source information under different conditions could account for the wider range of performance on HINT sentences observed at the extreme SNRs. In contrast, perception of the PRESTO sentences required greater adjustment and adaptation trial-by-trial. Because the talker changed from trial to trial, the impact of talker-specific perceptual learning was blocked, therefore preventing the listener from using a speaker-dependent strategy.

Thus, for HINT sentences, listeners could use an encoding strategy where they learned talker specific attributes; this strategy failed to be effective at poorer SNRs. On the other hand, for PRESTO sentences listeners were required to use a different encoding strategy that did not make use of talker-contingent learning. The strong correlation of individual performance on PRESTO across all SNRs also suggests that this processing strategy was consistent within listeners across different levels of competition. The present findings suggest that the flexible and robust processing strategy evoked by the high-variability PRESTO sentences is a defining core feature that is representative of processing of speech that occurs in real-world adverse listening conditions.

Differences between PRESTO and HINT

The design of the current experimental protocol was constructed to analyze differences in performance on PRESTO and HINT sentences based only on differences between the target utterances. The interaction between test type and SNR suggests that different processing

strategies were used to recognize the different types of sentences; however, PRESTO and HINT sentences differ on a number of dimensions other than just the number of target talkers.

One difference between HINT and PRESTO sentences is that the PRESTO lists included female talkers. Additional analyses revealed performance differences at -3 and -5 dB SNR, based on the gender of the talker who produced the sentence. Performance was best with female PRESTO talkers, lower for male PRESTO talkers, and worst with the one male HINT talker. Listeners were more accurate in perceiving the sentences produced by female talkers despite a greater contribution of energetic masking from the competition in the average female talker fundamental frequency range. Acoustic analysis of the multi-talker babble long-term power spectrum revealed that energy within the 150 – 300 Hz band (average female talker fundamental frequency range) was on average 3 dB greater than energy within the 75 – 150 Hz band (average male talker fundamental frequency range). Importantly, however, performance on PRESTO sentences produced by male talkers was significantly better than performance on HINT sentences produced by the one male talker. This suggests that PRESTO and HINT sentences were processed in fundamentally different ways, even when the effects of including female talkers in PRESTO were removed.

Another difference is the role of perceptual similarity and perceptual distinctiveness of the target talker in the multi-talker babble competition. The HINT talker and the competing talkers in the multi-talker babble all had General American English dialects. Hearing PRESTO talkers with non-standard dialects may have resulted in greater distinctiveness and discriminability between the target and competing voices (see for example, Van Engen and Bradlow, 2007). It is difficult to fully determine the extent to which regional dialect differences affected performance. While there were talkers from different dialect regions, it was not necessarily always true that a particular test sentence contains dialect-specific variants. However, subsequent analyses that included only General American English dialect speakers in PRESTO revealed the same pattern of results that were observed when the talkers from all dialect regions were combined together. This result suggests that performance differences observed between PRESTO and HINT were not related only to regional dialect similarity among the talkers.

In addition, previous research has reported that the effects of dialectal variability are larger at poorer SNRs (Clopper and Bradlow, 2008). Similarly, larger effects at poorer SNRs have also been reported for linguistic complexity (Calandruccio et al, 2010). Thus, it was originally predicted that performance on PRESTO would be poorer overall than performance on HINT sentences at more challenging SNRs, because of the increased sentence complexity and dialectal variation. However, contrary to this expectation, listeners performed better on PRESTO sentences at -3 and -5 dB SNR compared to their performance on the simpler HINT sentences at the same SNRs. This pattern may have also occurred because listeners used different processing strategies for encoding PRESTO and HINT sentences.

Speech is a complex, time-varying acoustic signal, with modulations and variability in the spectral and temporal domains that preclude simple stream segregation strategies based on general principles of gestalt perception (Remez et al, 1994; Remez, 2005). As discussed in the introduction, the speech recognition performance observed in any given experiment is a complex, multi-dimensional function of the target signal, background competition, individual perceiver, task, and the combined interactions among all of these factors. The current experimental results suggest that the inclusion of multiple sources of variability in testing materials, such as PRESTO, is a valid and novel method of assessing the core,

foundational processes underlying a listener's spoken word recognition skills that may have application in a wide range of listening situations.

Limitations

There are several limitations to the current study. The overall goal of developing the PRESTO test was to create a perceptually robust sentence recognition test that can be used to access the core, foundationally, underlying processing operations used in natural listening environments. A typical everyday listening environment would generally not include a change in talker after each sentence. However, the abrupt change of indexical source characteristics of the talker's voice in PRESTO requires the listener to engage in rapid adaptation and adjustment, which is an important defining core attribute of robust speech perception over a wide range of listening conditions.

The generalization of the current findings is also limited because only one type of competition was used—multi-talker babble. This type of competition was selected because it is representative of the most challenging adverse listening environments encountered in everyday speech communication situations. In addition, although the speech produced by the six talkers used in the background varied from trial to trial, it is possible that participants may have learned or adapted to repeated presentations of the same background talkers. The randomized nature of the sentence presentation conditions used in this study, although necessary for ensuring participant attention and effort at all SNRs, is also a limitation of this study. Because each listener received sentences and SNRs in a randomized/different order, it is difficult to perform more detailed analyses of the processing dynamics of perceptual learning over time. Finally, the conventional list structure used in clinical audiological assessments was not used in this study. Further studies are currently underway in our lab to continue developing PRESTO into a clinically useful measure.

Future Directions

Future studies will investigate list equivalency, performance on PRESTO by listeners with hearing loss, the influence of talker/listener dialect, the interaction between target variability and different types of competition, and comparisons to other conventional tests such as the Quick Speech in Noise (QuickSIN) (Killion et al, 2004) and Speech Perception in Noise (SPIN) (Kalikow et al, 1977) tests. The present results establish that high-variability sentence materials, although correlated with scores obtained from a conventional low-variability sentence test, measure fundamentally different processing strategies used to recognize speech, and that high-variability materials may be a better methodology to assess how an individual will actually perform in real-world adverse listening conditions. Precisely how much variability is needed in test materials used in the laboratory or clinic to access these core foundational processes still remains to be answered. PRESTO was considered to be a high-variability test because of the large number of talkers, wide range of sentence structures, keyword frequencies and familiarity, and talker regional dialects. Future studies will compare PRESTO to other sentence tests containing different degrees of variability, for example, the Az-Bio sentences which include only two male and two female talkers that are used repeatedly across test lists (Spahr and Dorman, 2004).

Conclusions

The current study provides some initial normative data on performance of PRESTO for a large group of young, normal-hearing listeners, establishes the validity of PRESTO in terms of correlation to performance on HINT, and demonstrates that PRESTO has excellent test-retest reliability. PRESTO was designed to assess a listener's ability to rapidly adapt to a variety of different talkers and different dialects using diverse sentence materials. Measurement of the ability to adjust to changes in talker from trial to trial in a test list may

also be an efficient method to quantitatively assess hearing impaired adults' perception of their hearing ability in everyday listening situations (Kirk et al, 1997). The development of new theoretically-motivated, clinically useful assessment methods that measure foundational neurocognitive skills underlying individual differences is necessary for more sensitive and valid assessment of speech recognition skills in listeners with hearing loss, (Gifford et al, 2008; Pisoni, 1998; Gifford et al, 2010). PRESTO represents an important step forward toward the development of a new class of clinical tests that differ fundamentally from conventional spoken word recognition tests used in the past. PRESTO was designed at the outset to access the core, foundational, underlying processes humans use to perceive speech robustly in adverse listening conditions, a skill that is used in everyday listening situations to maintain reliable speech communication between speakers and listeners.

Acknowledgments

We thank Luis Hernandez, Robin Canfield, Marissa Habeshy, Jillian Badell, Emily Garl, and Sushma Tatineni for their assistance on this project.

Preparation of this manuscript was supported in part by NIH NIDCD Training Grant T32DC00012 and NIH NIDCD Research Grant R01-DC00111 to Indiana University.

Abbreviations

HINT	Hearing In Noise Test
HiPRESTO	Group of participants in the upper quartile of performance distribution in Phase I
LoPRESTO	Group of participants in the lower quartile of performance distribution in Phase I
PRESTO	Perceptually Robust English Sentence Test Open-set
SNR	Signal-to-Noise Ratio
TIMIT	Texas Instruments/Massachusetts Institute of Technology

References

- Akeroyd MA. Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int J Audiol*. 2008; 47(Suppl. 2):S53–S71. [PubMed: 19012113]
- Arlinger S, Lunner T, Lyxell B, Pichora-Fuller MK. The emergence of Cognitive Hearing Science. *Scand J Psychol*. 2009; 50:371–384. [PubMed: 19778385]
- Beer, J.; Kronenberger, W.; Pisoni, DB. Executive function in everyday life: Implications for young cochlear implant users; Stockholm, Sweden. Paper presented at the 11th International Conference on Cochlear Implants and Other Implantable Auditory Devices; Jun. 2010
- Bench J, Kowal A, Bamford J. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *Brit J Audiol*. 1979; 13:108–112. [PubMed: 486816]
- Boulenger V, Hoen M, Ferragne E, Pellegrino F, Meunier F. Real-time lexical competitions during speech-in-speech comprehension. *Speech Commun*. 2010; 52:246–253.
- Brungart DS. Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am*. 2001a; 109:1101–1109. [PubMed: 11303924]
- Brungart DS, Simpson BD, Ericson MA, Scott KR. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J Acoust Soc Am*. 2001b; 110:2527–2538. [PubMed: 11757942]
- Calandruccio L, Dhar S, Bradlow AR. Speech-on-speech masking with variable access to the linguistic content of the masker speech. *J Acoust Soc Am*. 2010; 128:860–869. [PubMed: 20707455]

- Carhart R, Johnson C, Goodman J. Perceptual masking of spondees by combinations of talkers. *J Acoust Soc Am*. 1975; 58(Suppl. 1):S35.
- Carhart R, Tillman TW, Greetis ES. Perceptual masking in multiple sound backgrounds. *J Acoust Soc Am*. 1969; 45:694–703. [PubMed: 5776931]
- Clopper CG, Bradlow AR. Perception of dialect variation in noise: Intelligibility and classification. *Lang Speech*. 2008; 51:175–198. [PubMed: 19626923]
- Conway CM, Bauernschmidt A, Huang SS, Pisoni DB. Implicit statistical learning in language processing: Word predictability is the key. *Cognition*. 2010; 114:356–371. [PubMed: 19922909]
- Egan JP. Articulation testing methods. *Laryngoscope*. 1948; 58:955–991. [PubMed: 18887435]
- Fabry D. Cochlear implants and hearing aids: Converging/colliding technologies. *The Hearing Journal*. 2008; 61(7):10–16.
- Fabry D, Firszt JB, Gifford RH, Holden LK, Koch D. Evaluating speech perception benefit in adult cochlear implant recipients. *Audiology Today*. 2009; 21(3):36–43.
- Felty, R. Perceptually Robust English Sentence Test (Open-Set). Indiana University Bloomington; 2008. Unpublished manuscript
- Garofolo, JS.; Lamel, LF.; Fisher, WM.; Fiscus, JG.; Pallett, DS.; Dahlgren, NL. The DARPA TIMIT acoustic-phonetic continuous speech corpus. Philadelphia: Linguistic Data Consortium; 1993.
- Gifford RH, Shallop JK, Peterson AM. Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiol Neurotol*. 2008; 13:193–205. [PubMed: 18212519]
- Gifford RH, Dorman MF, Shallop JK, Sydlowski SA. Evidence for the expansion of adult cochlear implant candidacy. *Ear Hear*. 2010; 31:186–194. [PubMed: 20071994]
- Goldinger SD. Echoes of echoes: An episodic theory of lexical access. *Psychol Rev*. 1998; 105:251–279. [PubMed: 9577239]
- Helfer KS, Freyman RL. Lexical and indexical cues in masking by competing speech. *J Acoust Soc Am*. 2009; 125:447–456. [PubMed: 19173430]
- Hirsh IJ, Davis H, Silverman SR, Reynolds EG, Eldert E, Benson RW. Development of materials for speech audiometry. *J Speech Hear Disord*. 1952; 17:321–337. [PubMed: 13053556]
- Jenkins, JJ. Four points to remember: A tetrahedral model of memory experiments. In: Cermak, LS.; Craik, FIM., editors. *Levels of Processing in Human Memory*. Hillsdale, NJ: Erlbaum; 1979. p. 429–446.
- Johnson, K. Speech perception without speaker normalization: An exemplar model. In: Johnson, K.; Mullennix, JW., editors. *Talker Variability in Speech Processing*. San Diego, CA: Academic Press; 1997. p. 145–166.
- Johnson, K.; Mullennix, JW. *Talker Variability in Speech Processing*. San Diego, CA: Academic Press; 1997.
- Kalikow DN, Stevens KN, Elliot LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J Acoust Soc Am*. 1977; 61(5):1337–1351. [PubMed: 881487]
- Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*. 2004; 116(4):2395–2405. [PubMed: 15532670]
- Kirk J, Pisoni D, Miyamoto R. Effects of stimulus variability on speech perception in listeners with hearing impairment. *J Speech Lang Hear Res*. 1997; 40(6):1395–1405. [PubMed: 9430759]
- Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. *Ear Hear*. 1998; 19:1–36. [PubMed: 9504270]
- Mullennix JW, Pisoni DB, Martin CS. Some effects of talker variability on spoken word recognition. *J Acoust Soc Am*. 1989; 85:365–378. [PubMed: 2921419]
- Neff DL, Dethlefs TM. Individual differences in simultaneous masking with random-frequency, multicomponent maskers. *J Acoust Soc Am*. 1995; 98:125–134. [PubMed: 7608391]
- Nilsson M, Soli SD, Sullivan JA. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am*. 1994; 95(2):1085–1099. [PubMed: 8132902]

- Nusbaum, HC.; Pisoni, DB.; Davis, CK. Research on Speech Perception Progress Report No.10. Bloomington: Speech Research Laboratory, Indiana University; 1984. Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words; p. 357-376.
- Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. *Psychol Sci.* 1994; 5:42–46. [PubMed: 21526138]
- Nygaard LC, Sommers MS, Pisoni DB. Effects of stimulus variability on perception and representation of spoken words in memory. *Percept Psychophys.* 1995; 57:989–1001. [PubMed: 8532502]
- Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. *J Exp Psychol Learn Mem Cogn.* 1993; 19:309–328. [PubMed: 8454963]
- Park, H-Y.; Felty, R.; Lormore, K.; Pisoni, D. PRESTO: Perceptually Robust English Sentence Test: Open-set design, philosophy, and preliminary findings; Baltimore, Maryland. Poster presented at the 159th Acoustical Society of America Meeting; Apr. 2010
- Pierrehumbert JB. The next toolkit. *J Phonetics.* 2006; 34:516–530.
- Pisoni, DB. Some thoughts on “normalization” in speech perception. In: Johnson, K.; Mullennix, JW., editors. *Talker Variability in Speech Processing*. San Diego, CA: Academic Press; 1997. p. 9-32.
- Pisoni, DB. Development of new perceptually robust tests of speech discrimination. Los Angeles, CA: Invited talk presented for the American Academy of Audiology; Mar. 1998 p. 1998
- Pisoni DB. Cognitive factors and cochlear implants: Some thoughts on perception, learning, and memory in speech perception. *Ear Hear.* 2000; 21:70–78. [PubMed: 10708075]
- Pollack I. Auditory informational masking. *J Acoust Soc Am.* 1975; 57(Suppl. 1):S5–S5.
- Remez, RE. The perceptual organization of speech. In: Pisoni, DB.; Remez, RE., editors. *The Handbook of Speech Perception*. Oxford: Blackwell; 2005. p. 28-50.
- Remez RE, Rubin PE, Berns SM, Pardo JS, Lang JM. On the perceptual organization of speech. *Psychol Rev.* 1994; 101:129–156. [PubMed: 8121955]
- Richards VM, Zeng T. Informational masking in profile analysis: Comparing ideal and human observers. *J Assoc Res Otolaryngol.* 2001; 02:189–198. [PubMed: 11669393]
- Roediger HL, Fiske S. Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology.* 2008; 59:225–254.
- Spahr AJ, Dorman MF. Performance of subjects fit with the Advanced Bionics CII and Nucleus 3G cochlear implant devices. *Arch Otolaryngol Head Neck Surg.* 2004; 130:624–628. [PubMed: 15148187]
- Stenfelt S, Rönnerberg J. The Signal-Cognition interface: Interactions between degraded auditory signals and cognitive processes. *Scand J Psychol.* 2009; 50:385–393. [PubMed: 19778386]
- Studebaker GA. A “rationalized” arcsine transform. *J Speech Hear Res.* 1985; 28:455–462. [PubMed: 4046587]
- Theunissen M, Swanepoel DW, Hanekom J. Sentence recognition in noise: Variables in compilation and interpretation of tests. *Int J Audiol.* 2009; 48:743–757. [PubMed: 19951143]
- Van Engen KJ, Bradlow AR. Sentence recognition in native- and foreign-language multi-talker background noise. *J Acoust Soc Am.* 2007; 121:519–526. [PubMed: 17297805]
- Wightman FL, Kistler DJ, O’Bryan A. Individual differences and age effects in a dichotic informational masking paradigm. *J Acoust Soc Am.* 2010; 128:270–279. [PubMed: 20649222]
- Wilson RH, McArdle RA, Smith SL. An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. *J Speech Lang Hear Res.* 2007; 50:844–856. [PubMed: 17675590]

Figure 1.

Box plot of performance distribution in Phase I of keyword accuracy on PRESTO at each presentation condition (+3 dB SNR, 0 dB SNR, -3 dB SNR, -5 dB SNR) and overall mean score across SNRs. The far left of the box denotes the 25th percentile, the line denotes the 50th percentile, and the far right of the box denotes the 75th percentile. The whiskers extend to 1.5 times the interquartile range covered by the box. Open circles denote data points between 1.5 and 3 times the interquartile range. Filled circles denote data points greater than 3 times the interquartile range.

Figure 2.
Scatter plot of the accuracy scores on PRESTO from Phase I and Phase II.

Figure 3.
Scatter plot of the accuracy scores on PRESTO and HINT obtained in Phase II.

Figure 4.

Box plot of performance distribution in Phase II of keyword accuracy on PRESTO (dark grey) and HINT (light grey) at each presentation condition (+3 dB SNR, 0 dB SNR, -3 dB SNR, -5 dB SNR) and overall mean score across SNRs. The far left of the box denotes the 25th percentile, the line denotes the 50th percentile, and the far right of the box denotes the 75th percentile. The whiskers extend to 1.5 times the interquartile range covered by the box. Open circles denote data points between 1.5 and 3 times the interquartile range. Filled circles denote data points greater than 3 times the interquartile range.

Table 1

Total Number of Talkers For Each Dialect Region (Garofolo, 1993) in 10 PRESTO Lists for Phase I

Dialect Region	Number of Sentences
New England	14
Northern	26
North Midland	27
South Midland	26
Southern	32
New York City	17
Western	29
Army Brat	9

Table 2

Total Number of Talkers For Each Dialect Region (Garofolo, 1993) in 4 PRESTO Lists for Phase II

Dialect Region	Number of Sentences
New England	4
Northern	7
North Midland	10
South Midland	14
Southern	15
New York City	4
Western	12
Army Brat	6

Table 3

Means and Standard Deviations on PRESTO in Multi-Talker Babble for 121 Listeners

Signal-to-Noise Ratio	Mean Accuracy	Standard Deviation
Overall Mean	62.67	5.87
+3 dB	87.05	6.17
0 dB	70.81	6.17
-3 dB	55.00	7.02
-5 dB	37.59	6.60