Recognition of time-compressed speech does not predict recognition of natural fast-rate speech by older listeners

Sandra Gordon-Salant[a] **and Danielle J. Zion**
Department of Hearing and Speech Sciences, University of Maryland, College Park, Maryland 20742
sgsalant@umd.edu, dzion@umd.edu

Carol Espy-Wilson
Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742
espy@umd.edu

Abstract: This study investigated whether recognition of time-compressed speech predicts recognition of natural fast-rate speech, and whether this relationship is influenced by listener age. High and low context sentences were presented to younger and older normal-hearing adults at a normal speech rate, naturally fast speech rate, and fast rate implemented by time compressing the normal-rate sentences. Recognition of time-compressed sentences over-estimated recognition of natural fast sentences for both groups, especially for older listeners. The findings suggest that older listeners are at a much greater disadvantage when listening to natural fast speech than would be predicted by recognition performance for time-compressed speech.

PACS numbers: 43.71.Bp, 43.71.Lz, 43.71.Gv [AC]
Date Received: August 2, 2014 **Date Accepted:** August 8, 2014

1. Introduction

Time-compressed (TC) speech is often used in the laboratory to test auditory processing capabilities of listeners. One common application is to investigate the effects of a decline in speed of processing and working memory abilities with age (e.g., Wingfield *et al.*, 1985; Vaughan *et al.*, 2006; Wingfield *et al.*, 2006). Results generally show that older listeners are at a disadvantage compared to younger adult listeners for fast speech rates, particularly when the speech materials contain reduced linguistic and semantic cues (Wingfield *et al.*, 1985; Gordon-Salant and Fitzgibbons, 1993, Vaughan *et al.*, 2006) or are syntactically complex (Wingfield *et al.*, 2006). This age-related difference in recognition of TC speech is exacerbated in noise, perhaps because the cognitive processing resources (i.e., processing speed, attention, working memory) required to process fast speech with few contextual cues in noise are more limited among older than younger listeners (Tun, 1998). The difficulties in accurately understanding fast speech, especially in noise, may contribute to some of the problems that older listeners with normal hearing encounter in everyday listening situations; that is, they may experience problems in understanding the speech of individuals who speak at a fast rate. However, it is unclear whether recognition performance patterns observed for TC speech accurately reflect listener performance for natural fast-rate speech. The principal question addressed in this investigation is whether recognition of TC speech, often used in the laboratory setting, predicts recognition of natural fast-rate speech encountered in the real-world, and whether such predictions vary with listener age.

[a]Author to whom correspondence should be addressed.

Natural fast speech is characterized by several acoustic attributes, although findings vary somewhat across studies. Gay (1978) examined changes in vowel formants and duration in consonant-vowel-consonant (CVC) utterances (/p_p/) produced by four speakers at normal and fast speech rates, and observed that the principal acoustic change was a reduction in vowel duration. Crystal and House (1982) compared recordings of read passages by naturally fast and naturally slow talkers, and reported that the faster talkers reduced the pause duration between read sentences and the frequency of stop consonant releases relative to the slower talkers to accomplish a 25% faster speaking rate. Other investigations note that unstressed vowels are more reduced in duration than stressed vowels in fast speech (Janse *et al.*, 2003). In particular, schwa deletion appears to be a consistent finding in analyses of fast speech and "casual" speech (Davidson, 2006); this schwa deletion is attributed to increased gestural overlap necessitated by the dynamics of rapid speech production. One broad generalization gleaned from these prior studies is that acoustic changes in natural fast speech affect selected speech segments, rather than all speech segments in a consistent manner. It also appears that the specific speech segments altered with natural fast speech may depend on the type of speech sample recorded (isolated monosyllabic words involving specific articulatory movements, target multisyllabic words in a carrier sentence, passages of read text, etc.). In contrast, computer algorithms used to create TC speech (e.g., the Pitch Synchronous Overlap and Add, or PSOLA algorithm) preserve the temporal fine structure pattern and spectral information in speech, but in a reduced time scale. Indeed, Schneider *et al.* (2005) noted that these algorithms create speech with minimal stimulus degradation, other than the shift in time scale.

One prior study compared processing of natural fast-rate speech and TC speech by young adult listeners (Janse, 2004). Stimuli were multisyllabic nouns with a word-initial plosive consonant embedded in sentence fragments, in which the target word-initial plosive did not occur elsewhere in the sentence. The fast-to-normal ratio of the fast-rate speech (both natural and time compressed) was 0.72, which represents a modest increment in speech rate. The listeners' task was to press a button when they detected a specified word-initial phoneme. Listeners demonstrated a longer processing time for natural fast-rate speech than for TC speech, which was interpreted as reflecting the impact of less precise articulation in natural fast speech. These findings suggest the strong possibility that recognition accuracy for naturally produced fast speech is poorer than for TC speech. If this prediction is verified, then performance patterns observed by younger and older listeners for recognition of TC speech may substantially over-estimate recognition accuracy for natural fast speech encountered in everyday communication situations. The goals of the present study were to determine if recognition accuracy was higher for TC speech than natural fast-rate speech, when each type of fast speech was presented at the same rate, and if these differences were greater for older than younger listeners.

2. Method

2.1 Participants

Two listener groups participated in this experiment. The first group was 13 young normal hearing (YNH) participants (2 male, 11 female), 19–22 yrs of age (mean age: 20.4 yrs) and the second group was 12 older normal hearing (ONH) participants (3 male, 9 female), 66–76 yrs of age (mean age: 68.8 yrs). All participants were native speakers of American English who had normal hearing sensitivity, as defined by pure-tone air conduction thresholds $\leq$25 dB hearing level (HL) (ANSI, 2010) at octave frequencies from 250 to 4000 Hz. The pure tone thresholds in each ear were not significantly different between the two groups at all frequencies except for 4000 Hz, where mean thresholds for the older listeners (right ear, RE: 13.6 dB HL, left ear, LE: 17.27 dB HL) exceeded mean thresholds for the younger listeners (RE: 6.9 dB HL, LE: 5.38 dB HL) ($p < 0.01$). Nevertheless, all of these thresholds were still clearly within the range of normal hearing. Participants were compensated for their listening time.

2.2 Stimuli

Original IEEE (IEEE Audio and Electronics Group, 1969) and anomalous IEEE (Herman and Pisoni, 2000) sentences were recorded by a male talker at a normal and a natural fast rate. The original IEEE sentences (referred to here as High Probability, or HP) provide contextual cues to the listener (e.g., "The sink is the thing in which we pile dishes."), while the anomalous, or anomalous probability (AP) sentences are syntactically correct but semantically meaningless (e.g., "He pressed the bid of the funny ripe bench."). These stimuli were recorded by a 25-yr-old male talker who is a native speaker of American English with a general American dialect. Sentences were recorded in a quiet room onto a Dell PC with a Shure SM63 microphone with Marantz PMD661 solid state recorder set to a 44 100 Hz sampling rate. The talker was instructed to speak at a natural, conversational rate for the normal speech rate recordings. The normal rate recordings were then time compressed to 50% (i.e., fast-to-normal ratio = 0.5) and played back to the talker to provide a target rate for the natural fast speech. After recording was completed, original audio files were digitized onto laboratory computers and edited using Adobe Audition version 1.5 software. Editing involved splicing individual sentences from the original recording and saving each as a separate audio file. Overall duration analysis of the normal-rate and natural fast recordings revealed that, on average, the fast sentences approximated 40% time-compression (fast-to-normal ratio = 0.6), rather than the target 50% time-compression. Therefore, a 40% time-compression ratio was chosen for the TC rate condition, and these sentences were generated by modifying the normal rate recordings using the PSOLA method in Praat (version 5.3; Boersma and Weenick, 2013). The root-mean-square (rms) level was equated across all natural and TC sentences to within ±1 dB, and a 1000 Hz calibration tone was created to be equal in rms to the stimuli. Four-talker (two male, two female) babble was used for the background noise conditions; this babble was also equated in rms and calibrated using the 1000 Hz calibration tone.

Twelve test lists (6 HP, 6 AP) of 10 sentences each were created and written as separate tracks (1 list per track) to a CD-R. Each sentence contains 5 keywords, for a total of 50 keywords per list. The sentences in each list were recorded in three rates: Natural normal rate, natural fast rate, and 40% TC (relative to normal rate), for a total corpus of 360 sentences.

2.3 Procedures

During the experiment, participants were seated in a sound attenuating booth. Sentence stimuli and background babble were played back on separate channels of a CD player (Tascam CD-200) and routed through an audiometer (Interacoustics AC40) where they were separately attenuated, mixed, and delivered monaurally to a single ER-3A insert earphone. Sentences were played at a level of 65 dB sound pressure level (SPL), and babble was played at a level of 55 dB SPL, to create a +10 dB signal-to-noise ratio (SNR). This SNR level was chosen as a result of pilot testing to avoid floor effects. There were 12 conditions altogether: 2 contexts (HP, AP) × 2 environmental conditions (quiet, noise) × 3 speech rates (normal, TC, fast).

Calibration was completed prior to data collection for each participant. Data collection for each participant began with a list of normal rate, HP sentences in quiet, which was the easiest of the 12 listening conditions. Earlier pilot testing revealed ceiling performance for this condition, regardless of the order it appeared during testing. Therefore, it was chosen as the first condition in the test order to familiarize listeners with the target talker and general test procedure. The order of presentation of the remaining 11 conditions, as well as list assignment to condition, was randomized for each participant. Participants were instructed to repeat each target sentence aloud, with guessing encouraged when they were unsure of any part of the sentence. Total test time for each participant was approximately 1 h. This protocol was approved by the University of Maryland Institutional Review Board for Human Subjects Research.

2.4 Data analysis

Participants' verbal responses were scored for keywords identified correctly for each list, so that each condition yielded a percent correct score (number of keywords correct out of 50). Analysis of variance (ANOVA) was computed separately for quiet and noise data, following arcsine transformation, using a repeated-measures design with two within-subjects factors (rate and context) and one between-subjects factor (group), with an alpha level of 0.05. Where significant main effects and interactions were observed, *post hoc* pairwise comparisons were calculated with Bonferroni correction for multiple comparisons.

3. Results

3.1 Quiet

Mean recognition scores of the two groups for the HP and AP sentences at three rates in quiet are displayed in Fig. 1 (panel A). It is clear that recognition scores are considerably lower for naturally fast speech than for either 40TC speech or normal-rate speech, particularly for the sentences with reduced contextual cues (AP). Repeated-measures ANOVA confirmed a significant context × rate interaction [$F(2,46) = 82.364$, $p < 0.001$] as well as a significant context × group interaction [$F(1,23) = 10.71$, $p = 0.003$]. *Post hoc* analysis of the context × rate interaction revealed that each pairwise comparison among the three rates in both contexts was significant ($p \leq 0.006$) except for HP normal rate vs 40TC, where performance was near ceiling. For comparisons that were significant, performance was best for the normal rate speech, significantly poorer for the 40TC rate, and poorest with natural fast speech. These results are consistent with the predicted poorer performance for naturally fast speech than for

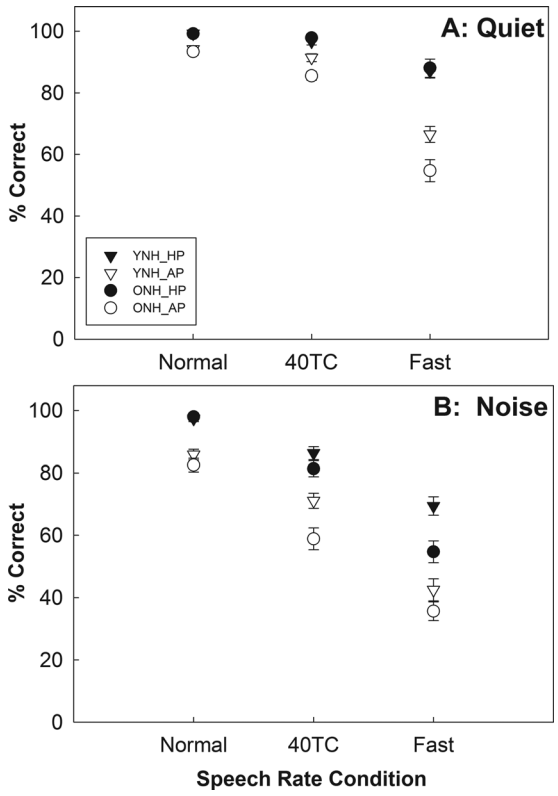

Fig. 1. Mean recognition scores of the two groups (YNH; ONH) for the IEEE original (HP [filled symbols]) and anomalous (AP [open symbols]) sentences at the three speech rates in quiet (A) and in noise (B). Error bars represent standard error.

TC speech, even in quiet, for both listener groups. *Post hoc* analysis of the context × group interaction revealed that older listeners performed significantly more poorly than the younger listeners for AP sentences ($p < 0.001$), but not HP sentences ($p > 0.05$), consistent with previous research (Gordon-Salant and Fitzgibbons, 2001).

3.2 Noise

Mean recognition scores of the two groups for the HP and AP sentences at three rates in noise are displayed in Fig. 1 (panel B). A similar pattern of poorer recognition of naturally fast speech than TC speech, as observed in quiet, is seen here, except that performance levels are considerably lower. Repeated measures ANOVA revealed a significant rate × group [$F(2,46) = 3.597$, $p < 0.05$)] interaction and a significant main effect of context [$F(1,23) = 254.503$, $p < 0.001$). *Post hoc* pairwise comparisons revealed that the rate effect was significant ($p < 0.001$) for both groups and was the same as that described for the results obtained in quiet: Performance was highest for normal rate, significantly poorer for 40TC, and poorest for natural fast speech. The source of the interaction was that the YNH listeners outperformed the ONH listeners for 40TC ($p = 0.012$) and fast speech ($p = 0.005$), but there was no difference between groups for the normal rate ($p = 0.52$). This suggests that the older listeners were more detrimentally affected by the fast speech rate, especially naturally fast speech, compared to the younger listeners, despite equivalent performance for normal-rate speech. The main effect of context was attributed to significantly higher recognition of HP sentences than AP sentences in the noise conditions, as expected.

3.3 Relationship between 40TC vs natural fast speech

The final analysis of the data was an examination of the extent to which recognition of TC speech is related to recognition of natural fast speech. Figure 2 plots the individual speech recognition scores of listeners in both groups, in both context conditions, separately in quiet (panel A) and noise (panel B). Pearson product-moment correlations revealed that in quiet, recognition performance for 40TC speech and natural fast speech was significantly correlated both for HP ($r = 0.403$, $p < 0.05$) and AP ($r = 0.524$, $p < 0.01$) sentences for all listeners. The HP correlation in quiet may be less meaningful because many listeners performed at ceiling, especially for TC speech. In noise, correlations were significant for HP ($r = 0.651$, $p < 0.001$) and AP ($r = 0.569$, $p < 0.01$) sentences for data collapsed across younger and older listeners. It is notable that all individual data points except one appear above the linear prediction line (slope of 1.0) in both panels of the Fig. 2, indicating that recognition of TC materials was consistently higher than recognition of natural-fast speech materials for both listener groups, in both contexts, and in both environmental conditions (quiet and noise).

4. Discussion

The over-arching goal of this study was to determine if recognition performance for TC speech over-predicts recognition performance for natural fast speech at comparable fast speech rates. Analyses of the mean speech recognition scores of the two listener groups support this hypothesis: Both listener groups consistently exhibited significantly poorer recognition scores for natural fast-rate speech than for TC speech. This was true in quiet and noise, and for both high-context and low-context sentences. Although the correlation data indicate a relationship between recognition scores for 40TC speech and natural fast speech, the individual data also show that nearly all listeners exhibited higher recognition scores for 40TC speech materials than for natural fast rate speech. These data, obtained from normal-hearing listeners, strongly indicate that listening to natural fast speech represents a significant communication challenge in everyday listening scenarios.

Acoustic differences in TC versus natural-fast speech may underlie the observed differences in recognition performance. Preliminary acoustic analyses of the natural fast recordings in this study show not only shorter overall phoneme (consonant and vowel) durations, but also deletions and distortions, especially of stops and affricates. These changes are associated with co-articulation, which produces overlap of

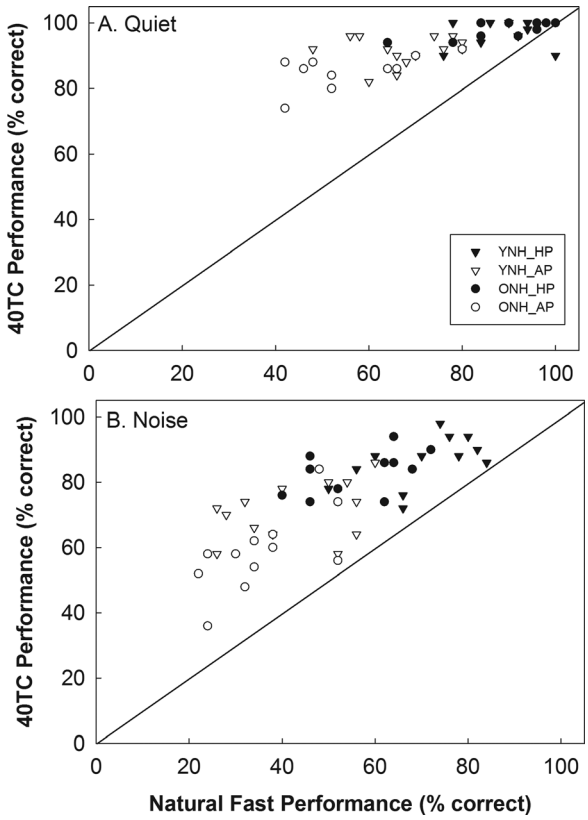

Fig. 2. Correlation data for TC (40TC) versus natural fast speech recognition performance of the two groups (YNH; ONH) for the IEEE original (HP [filled symbols]) and anomalous (AP [open symbols]) sentences in quiet (A) and in noise (B).

speech segments, and lenition, in which there is undershoot in reaching the target phoneme. A detailed report of the acoustic characteristics of the natural fast speech is forthcoming. These acoustic analyses generally confirm that, at least for the talker and fast speech rate used in the present study, there are numerous segmental distortions in natural rapid speech, unlike in TC speech.

Aside from acoustic differences between TC and natural-fast speech, performance disparities may also depend on the amount of time compression applied. Brungart *et al.* (2007) used TC to accelerate slow and conversational audio-visual speech to a fast rate and found performance was equivalent with natural fast speech and 30% TC conversational speech, but worse with 66% TC slow speech. Furukawa *et al.* (2004) used 40%, 50%, 60%, and 65% TC speech, with and without contextual cues, to evaluate the effects of age on rapid speech recognition in older adults with normal hearing and with mild to moderate hearing loss. They found no differences in performance between groups at the slowest TC ratio (40%) but reported that a hearing impaired group performed significantly worse than the normal hearing group, regardless of contextual cues, at the three faster rates. These authors suggest that neither of the older groups was able to benefit from the use of context in the rapid speech materials. In contrast, Gordon-Salant and Fitzgibbons (1993) used 30%, 40%, 50%, and 60% TC low probability Revised Speech Perception in Noise (R-SPIN) sentences (Bilger *et al.*, 1984), and found main effects of age, hearing loss, and time-compression ratio, when recognition scores for undistorted speech were used as a covariate. Although the 40% TC speech used in this study may be considered relatively slow in terms of rapid speech, it approximated the average duration of the natural fast sentences in this study and clearly underestimated the listening difficulty

encountered with natural fast speech, especially for the older listeners. It is possible, however, that other rates of natural-fast speech or variable-rate speech throughout the message could produce different performance patterns between younger and older listeners.

Taken together, the results of this study suggest that even when overall duration is equal, listening to natural fast speech is more difficult than listening to artificially TC speech. This is true even when listening conditions are favorable, with no background noise and use of speech materials that provide contextual cues. In less favorable conditions, older listeners, who are known to have difficulty processing rapid speech information, may be at an even greater disadvantage despite normal hearing sensitivity. Previous research using TC speech materials may not capture the actual processing difficulty encountered when listening to natural rapid speech, especially for older listeners.

Acknowledgments

This work was supported by a grant from the UMD-ADVANCE Program (funded by NSF). The authors thank Erin Pickett and Sarah Brown for their assistance with recording and processing of stimuli and Douglas Brungart and Julie Cohen for their advice and assistance.

References and links

ANSI (**2010**). ANSI S3.6-2010, American National Standard Specification for Audiometers (Revision of ANSI S3.6-1996, 2004) (American National Standards Institute, New York).

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., and Rzeckowski, C. (**1984**). "Standardization of a test of speech perception in noise," J. Speech, Lang., Hear Res. **27**, 32–48.

Boersma, P., and Weenink, D. (**2013**). "Praat: doing phonetics by computer [Computer program]. Version 5.3.51," retrieved 2 June 2013 from http://www.praat.org/.

Brungart, D. S., van Wassenhove, V., Brandewie, E., and Romigh, G. (**2007**). "The effects of temporal acceleration and deceleration on AV speech perception." *Auditory-Visual Speech Processing (AVSP 2007)*, Kasteel Groenendaal, Hilvarenbeek, The Netherlands.

Crystal, T. H., and House, A. S. (**1982**). "Segmental durations in connected speech signals: Preliminary results," J. Acoust. Soc. Am. **72**, 705–716.

Davidson, L. (**2006**). "Schwa elision in fast speech: Segmental deletion or gestural overlap?," Phonetica **63**, 79–112.

Furukawa, I., Vaughan, N., and Storzbach, D. (**2004**). "Can context diminish the effects of rapid speech recognition in older listeners?," J. Acoust. Soc. Am. **116**, 2524.

Gay, T. (**1978**). "Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am. **63**, 223–230.

Gordon-Salant, S., and Fitzgibbons, P. (**2001**). "Sources of age-related recognition difficulty for time-compressed speech," J. Speech Hear. Res. **44**, 709–719.

Gordon-Salant, S., and Fitzgibbons, P. J. (**1993**). "Temporal factors and speech recognition performance in young and elderly listeners," J. Speech Hear. Res. **36**, 1276–1285.

Herman, R., and Pisoni, D. B. (**2000**). "Perception of elliptical speech by an adult hearing-impaired listener with a cochlear implant: Some preliminary findings on coarse-coding in speech perception," in Research on Spoken Language Processing Progress Report No. 24 (Speech Research Laboratory, Indiana University, Bloomington, IN), pp. 87–112.

IEEE Audio and Electronics Group (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Janse, E. (**2004**). "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech," Speech Commun. **42**, 155–173.

Janse, E., Nooteboom, S., and Quené, H. (**2003**). "Word-level intelligibility of time-compressed speech: Prosodic and segmental factors," Speech Commun. **41**, 287–301.

Schneider, B. A., Daneman, M., and Murphy, D. R. (**2005**). "Speech comprehension difficulties in older adults: Cognitive slowing or age-related changes in hearing?," Psychol. Aging **20**, 261–271.

Tun, P. A. (**1998**). "Fast noisy speech: Age differences in processing rapid speech with background noise," Psychol. Aging **13**, 424–434.

Vaughan, N. E., Storzbach, D., and Furukawa, I. (**2006**). "Sequencing versus nonsequencing working memory in understanding of rapid speech by older listeners," J. Am. Acad. Aud. **17**, 506–518.

Wingfield, A., McCoy, S. I., Peelle, J. E., Tun, P. A., and Cox, L. C. (**2006**). "Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity," J. Am. Acad. Aud. **17**, 487–497.

Wingfield, A., Poon, L. W., Lombardi, L., and Lowe, D. (**1985**). "Speed of processing in normal aging: Effects of speech rage, linguistic structure, and processing time," J. Gerontol. **40**, 579–585.