Department of Veterans Affairs

Journal of Rehabilitation Research and Development
Vol. 39 No. 5, September/October 2002
Pages 559–566

Time-expanded speech and speech recognition in older adults

Nancy E. Vaughan, PhD; Izumi Furukawa, MA; Nirmala Balasingam, MSEE; Margaret Mortz, PhD; Stephen A. Fausti, PhD

Department of Veterans Affairs Rehabilitation Research and Development, National Center for Rehabilitative Auditory Research, Portland Veterans Administration Hospital, Portland, OR; Intel Corporation, Folsom, CA; Computer Science and Engineering, Washington State University, Spokane, WA

Abstract—Speech understanding deficits are common in older adults. In addition to hearing sensitivity, changes in certain cognitive functions may affect speech recognition. One such change that may impact the ability to follow a rapidly changing speech signal is processing speed. When speakers slow the rate of their speech naturally in order to speak clearly, speech recognition is improved. The acoustic characteristics of naturally slowed speech are of interest in developing time-expansion algorithms to improve speech recognition for older listeners. In this study, we tested younger normally hearing, older normally hearing, and older hearing-impaired listeners on time-expanded speech using increased duration and increased intensity of unvoiced consonants. Although all groups performed best on unprocessed speech, performance with processed speech was better with the consonant gain feature without time expansion in the noise condition and better at the slowest time-expanded rate in the quiet condition. The effects of signal processing on speech recognition are discussed.

Key words: *clear speech, speech recognition, time expansion.*

This material was based on work supported by the American Academy of Audiology.

Address all correspondence and requests for reprints to Nancy E. Vaughan, PhD; Rehabilitation Research and Development, National Center for Rehabilitative Auditory Research (NCRAR), VA Medical Center, PO Box 1034, Portland, OR 97207; 503-220-8262, ext. 56030; fax: 503-273-5021; email: vaughann@ohsu.edu.

INTRODUCTION

Many older Americans suffer from difficulty understanding spoken language in everyday communicative situations in spite of adequate hearing sensitivity [1]. As the number of older Americans increases and life expectancy increases, an understanding of the difficulties experienced by older listeners in daily communication situations will be needed to contribute to viable solutions. At present, these difficulties are not well understood. Age-related speech understanding difficulties are not highly correlated with degree and configuration of hearing loss and are often greater than pure tone thresholds would predict [2–4]. Further, hearing aids do not always provide the expected benefits. Although reduced hearing sensitivity may contribute to these difficulties, another potential explanation involves age-related cognitive changes that may affect the ability to efficiently process speech. One such consistent change is cognitive slowing [5]. It has been shown that older adults are more adversely affected by increased speech rates than are younger adults [6–8]. The dynamic nature of speech requires rapid processing to keep pace with incoming information. When talkers slow their speech, intelligibility is improved for older listeners and for hearing impaired listeners [9–11]. The purpose of the current study was to investigate the effects of computerized slowing of speech (time expansion) on speech recognition in older listeners. The speech produced when a talker intentionally tries to improve intelligibility by

speaking slowly and clearly, but without exaggeration, is called "clear speech" [10,12]. Unfortunately, neither can we depend on talkers to use clear speech consistently, nor do we know which acoustic features of clear speech are critical for improving speech recognition. A computerized method of time expansion would make improved listening conditions available for more listeners through hearing aids or other listening devices.

Clear Speech

The improvement in intelligibility produced by clear speech is a robust phenomenon. An average difference of 17 percent between clear speech and conversational speech intelligibility occurred when talkers were instructed to use clear speech under controlled conditions [9]. In adverse listening conditions (noise and reverberation), that clear-conversational difference increased to 20 percent for normal hearing adults and 26 percent for hearing-impaired listeners [13]. Slowing is a consistent feature of clear speech, but specific acoustic changes occur in addition to insertion of pauses and lengthening of durations of individual speech sounds [12]. Such changes include less reduction of vowels, release of stops and final consonants, and increased root mean square (rms) intensities for obstruent sounds produced by restricting airflow such as stops, affricates, and fricatives [10]. Analysis of the speech of talkers who are naturally more easily understood than other talkers in experimental conditions reveals acoustic-phonetic characteristics similar to speakers who have been instructed to produce clear speech [14]. Hence, slowing the rate of speech appears to occur as a consequence of acoustic modifications to individual phonemes, which improve intelligibility of speech for hearing-impaired listeners and older listeners.

Time-Expansion Algorithms

Time-expansion methods that employ acoustic modifications to slow the rate of speech are called nonuniform algorithms (for a review of methodology, see Nejime and Moore [15]). Uniform methods simply increase the length of the speech signal by the regular insertion of silent intervals. Nonuniform algorithms vary widely in their methods of selection and modification of acoustic features. Two recent methods involved complex frequency computations using short-term Fourier transforms and waveform expansion for parts of the speech waveform that exceeded a certain power threshold (mostly vowels) [15–16]. No significant improvement in sentence recognition was achieved with these methods either for hearing-impaired listeners or for listeners with simulated cochlear hearing loss [15–16].

Intelligibility of nonsense syllables was successfully improved by increasing the consonant-vowel ratio [17]; however, this method used manual identification of consonants and vowels rather than automatic identification of specific acoustic properties. Acoustic characteristics of individual phonemes are not invariant [14,18,19], which causes difficulty in automating the process. However, classes of phonemes (stops, fricatives, etc.) do demonstrate reliable acoustic features that could be identified by an automatic software function [19].

It appears from previous work that one could make nonuniform time expansion more effective by using appropriate acoustic modifications derived from clear speech. In the current algorithm, our goal was to incorporate a feature of clear speech by increasing the duration and intensity of unvoiced consonants. These acoustic modifications will increase audibility of generally weak phonemes for older hearing-impaired listeners and will slow the rate of speech to provide extra processing time for all older listeners.

METHODS AND MATERIALS

Participants

Thirty-six adults comprising three groups of 18 participants each were recruited for this study. Older participants were recruited from the surrounding community, and younger participants were recruited primarily from the university campus. The three groups consisted of—

- Younger normally hearing (YNL) adults (m [mean] = 26 years, SD [standard deviation] = 3.7).
- Older normally hearing (ONL) adults (m = 67.3 years, SD = 4.1).
- Older hearing-impaired (OHI) adults, who had mild to moderate sensorineural hearing loss (70.3 years, SD = 4.1).

For purposes of this study, we defined normal hearing as pure tone thresholds of 25 dB hearing level (HL) or better at octave frequencies from 250 Hz to 4,000 Hz and interoctave frequencies of 1,500 Hz and 3,000 Hz. We defined mild to moderate sensorineural hearing loss as thresholds between 25 dB HL and 65 dB HL at two or more of the tested frequencies with air-bone gaps no

greater than 10 dB. High-frequency pure tone averages for 1,000, 2,000, and 4,000 Hz (PTA2) were –1.3 dB HL (SD = 4.1) in the YNL group, 5.4 dB HL (SD = 4.3) in the ONL group, and 28.2 dB HL (SD = 5.0) in the OHI group. Hearing sensitivity was bilaterally symmetrical; i.e., interaural differences were less than 10 dB for each participant. We obtained speech reception thresholds (SRT) also through a loudspeaker to establish a presentation level for the experimental speech materials in the sound field.

Instrumentation

All audiometric testing took place in a standard double-walled, sound-treated chamber. We obtained pure tone air and bone conduction thresholds using a Grason-Stadler model GSI 10 clinical audiometer with TDH 50 earphones and a Radio Ear B71 bone oscillator. Speech materials were presented from compact disks through a calibrated loudspeaker controlled by the GSI 10 clinical audiometer.

Procedures

Practice Session

We used four separate paragraphs of 10 sentences each (not used in the experimental conditions) for practice before the test session. Practice paragraphs consisted of one paragraph (10 sentences) in quiet and one paragraph in noise at a normal rate and two paragraphs (one in quiet and one in noise) at the most extreme rate and gain conditions (1.4× expansion with gain). Feedback was given during the practice session to help familiarize the listeners with time-compressed speech.

Experimental Session

Speech was presented at 25 dB SL in reference to the sound field SRT with the listener seated at 0° azimuth at a distance of 1.5 m from a single loudspeaker. For the noise conditions, the speech was mixed with 12-talker babble and amplitude-normalized at +4 dB signal-to-babble ratio before presentation through the single loudspeaker. The listeners were instructed to repeat the sentences exactly as they heard them and to guess the word or words if they were unsure.

Scoring

We scored sentences according to the number of key words correct per paragraph. The key words used for scoring were identified by Cox et al. to equalize the test paragraphs in terms of intelligibility [20]. The number of key words per sentence varied from one to three, but 25 key words were consistently in each 10-sentence paragraph. Paragraph scores were used for the analysis.

Experimental Speech Materials

All speech materials were prerecorded, digitized, and stored on compact disks. The female talker was a native speaker of American English with no pronounced regional dialect. We chose the Connected Speech Test (CST) for optimal face validity because of its structure and content [20]. The CST comprises paragraphs consisting of 10 topically related sentences. Sentences were presented both in noise and in quiet at three rates of speech—one normal rate (172 words per minute) and two time-expanded rates at 1.2 times slower and at 1.4 times slower, creating 12 combinations of conditions. At each rate, sentences were presented with and without increased consonant intensity (consonant gain). We used 12 paragraphs for this study, resulting in one complete paragraph (10 sentences) per condition.

The noise background was a 12-talker babble that was digitized from a magnetic tape recording obtained from a commercial source (Auditec of St. Louis) and mixed with the speech files at a +4 dB signal-to-babble ratio. All speech files (with and without babble) were then amplitude normalized.

Time-Expansion Algorithm

We subjected the speech materials to a time-expansion algorithm for processing. The algorithm identified unvoiced consonants using measures of peak energy, average energy, and overall energy fluctuations. We used a rule-based combination of those three features for simple classification of each speech frame into silence, unvoiced consonant, voiced consonant, or vowel. Of interest in this study was the modification of unvoiced consonants. No modifications were made to other phonemes in this algorithm. Once we identified an unvoiced consonant, we prolonged the duration by inserting copies of segments of the consonant at that point of the speech signal. Increasing the intensity of the unvoiced consonants was optional, so if that feature was used, the energy in each temporal frame was multiplied by a factor of 2. We controlled overall amplitude by fading (ramping) the edges of the frames after the additional energy was

added. Processing was accomplished in the time domain without complex frequency computations.

RESULTS

We recognized that the 25 dB HL limit for normal-hearing sensitivity in this study may have allowed for differences in sensitivity between the older normally hearing and the younger normally hearing listeners. It is important, therefore, to note that average threshold differences between the two normally hearing groups in this study exceeded 10 dB only at 3,000 and 4,000 Hz (**Figure 1**). The average difference at 3,000 Hz was 11.7 dB, and at 4,000 Hz, it was 15.4 dB.

The unprocessed speech (normal rate) elicited better speech recognition scores from all three groups than either of the slower rates of speech (**Table 1**). All groups scored better in quiet than in babble regardless of rate of speech or consonant gain condition. As expected, the YNL group had higher average scores than either of the older groups in all test conditions.

A four-way analysis of variance (ANOVA) was conducted to examine the effects of the four independent variables on speech recognition scores: group (YNL, ONL, OHI), rate (normal, 1.2× expansion, 1.4× expansion), Noise (babble, quiet), and consonant gain (present or absent). All scores were arcsine-transformed before statistical analysis. We performed the analysis using the NCSS 2000 Statistical System for Windows [21]. The four-way ANOVA revealed significant main effects of group, rate, noise, and gain. Two-way interactions that reached significance were group by noise and rate by noise. Rate and noise and consonant gain showed the only significant three-way interaction. **Table 2** displays the ANOVA values for these results.

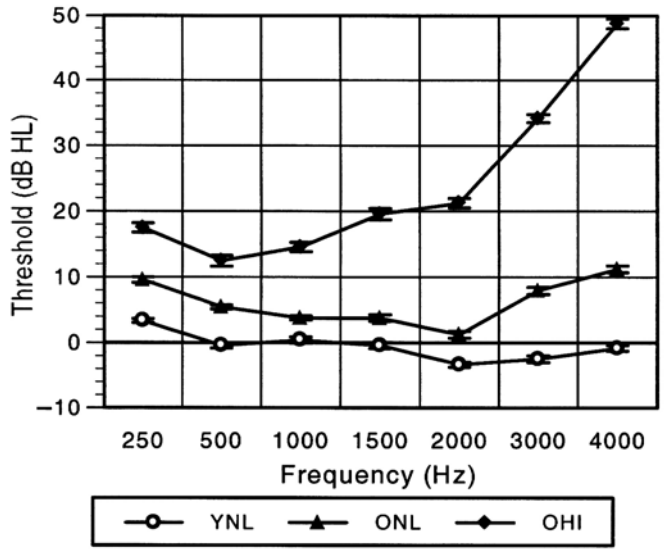

Figure 1.
Means and standard errors of better ear pure tone thresholds at each frequency for younger normally hearing (YNL), older normally hearing (ONL), and older hearing-impaired (OHI) groups.

The post hoc comparisons by means of Newman-Keuls multiple comparison test revealed that the YNL group scored significantly better than either of the two older groups and that the two older groups did not differ from each other when all conditions were combined. Similarly, the normal rate of speech produced better scores than either of the two slower rates (1.2× and 1.4×), and the slower rates were not significantly different from each other overall. The consonant gain effect was not as strong as the rate and group effects, but combined across all conditions, speech recognition scores were better when consonant gain was implemented than when it was not.

We conducted further post hoc testing to investigate the nature of the significant two-way interactions of noise and group and of noise and rate. In noise, the performance of both of the older groups (ONL, OHI) was equally decreased, while the younger group (YNL) had the best performance. In quiet, average scores were quite similar among the three groups, but the post hoc comparisons found a statistically significant difference between the younger group and the older normally hearing group. The older hearing-impaired group did not differ statistically from either of the other two groups. The interaction between rate and noise revealed that the slowest rate (1.4×) produced the poorest scores in quiet, and the intermediate rate (1.2×) produced the same scores as those at the normal rate. The noise condition differentiated all three rates. Performance was better at the slowest rate (1.4×) than at the intermediate rate (1.2×) and the normal rate produced the best scores on average. In noise, it was more effective to slow the speech rate to a greater degree, but in quiet, too much slowing had an adverse effect.

To explore the three-way interaction among rate, noise, and gain, we conducted a separate two-way ANOVA for each of the three rates with noise and gain as independent variables and arcsine-transformed scores as the response variable. Results showed that the

Table 1.
Mean speech recognition scores and standard deviations (SD) in percent for younger normally hearing (YNL), older normally hearing (ONL), and older hearing-impaired (OHI) groups in quiet and in noise conditions at normal rate of speech and at two time-expansion rates.

Rates of Speech	YNL Group (SD)		ONL Group (SD)		OHI Group (SD)	
	Quiet	Noise	Quiet	Noise	Quiet	Noise
Normal	99.7 (1.1)	98.5 (2.3)	99.3 (1.9)	92.2 (7.0)	99.5 (1.4)	92.3 (6.2)
1.2×	100.0 (0)	94.2 (5.3)	99.5 (1.4)	86.2 (7.6)	99.2 (2.0)	84.2 (9.5)
1.4×	99.3 (1.5)	96.3 (4.7)	97.0 (3.6)	87.5(10.4)	97.2 (3.8)	85.5(10.8)

Table 2.
ANOVA results for comparison of effects of four variables (group, rate, noise, consonant gain) on time-compressed Connected Speech Test (CST) scores.

Variable	df	*F* Ratio	*p* Value
Group	2, 395	46.52	0.00000*
Rate	2, 395	16.73	0.00000*
Noise	1, 395	374.06	0.00000*
Consonant Gain	1, 395	9.07	0.00277*
Group × Rate	4, 395	0.73	0.57010
Group × Noise	2, 395	18.09	0.00000*
Group × Gain	2, 395	2.45	0.08757
Rate × Noise	2, 395	14.56	0.00000*
Rate × Gain	2, 395	0.59	0.55747
Noise × Gain	1, 395	1.11	0.29375
Group × Rate × Noise	4, 395	0.36	0. 83628
Group × Rate × Gain	4, 395	0.52	0.72114
Group × Noise × Gain	2, 395	0.28	0.75650
Rate × Noise × Gain	2, 395	19.51	0.00000*
Group × Rate × Noise × Gain	4, 395	0.82	0.51404

*Term significant at alpha = 0.05
df = degrees of freedom

interaction of gain and noise was present at the normal rate and at the slowest rate (1.4×) of speech. At the normal rate, the addition of consonant gain produced significantly better scores in noise (F[1,143] = 5.20, p = 0.024) (**Figure 2**), but it did not make a difference in quiet without time expansion. Alternatively, at the slowest rate (1.4×), the addition of consonant gain improved performance in quiet (**Figure 3**), but degraded it in noise (F[1,143] = 18.13, p = 0.00000). Although both gain (F[1,143] = 6.29, $p < 0.05$) and noise (F[1,143] = 231.85, p = 0.00000) were significant main effects at the intermediate rate (1.2×), the effects were independent of each other.

CONCLUSIONS AND DISCUSSION

Our investigation addressed the efficacy of a nonuniform time-expansion algorithm for improvement of speech recognition in older listeners. The unique feature of the time-scaling method in this study was the identification and modification of unvoiced consonants as a strategy for time expansion. Although this strategy did not improve speech recognition over that of unprocessed speech, this study suggests that noise combined with signal processing, changes in the natural prosody, and higher-level processing deficits play a role in recognition of processed speech.

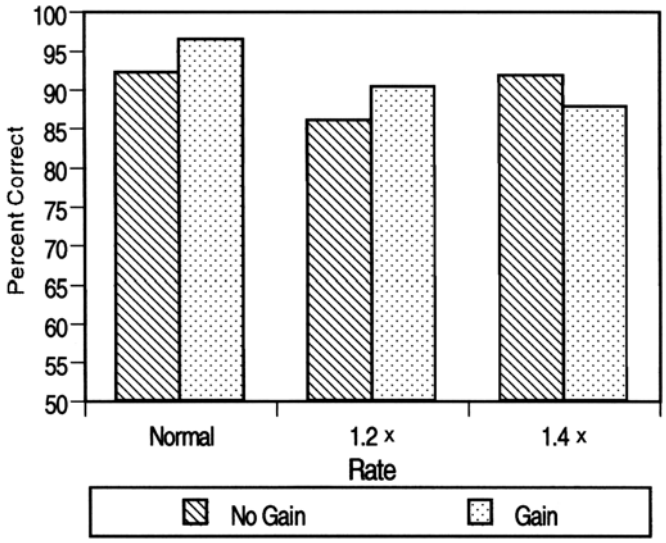

Figure 2.
Effects of consonant gain in noise at each rate (normal, 1.2×, 1.4×) across groups.

Slowing the rate of speech has been expected to be beneficial to older listeners, especially in adverse listening conditions, such as noise backgrounds. In this study, both older normally hearing and hearing-impaired listeners were equally disadvantaged in noise regardless of speech rate. In quiet, the younger listeners' performance was statistically superior but the differences among all three groups would not likely be of clinical concern (see **Table 1**). Background noise and distortions caused by the acoustic modifications of time expansion may have had an additive effect on speech recognition in this study. Older listeners have been found to be more sensitive to the effects of multiply degraded speech than younger listeners are, especially when one of the degraded conditions was time compression [22]. The results of this study suggest that temporal distortions of speech in either direction—time compression or time expansion—may be detrimental to speech recognition when combined with other types of distortion, such as noise.

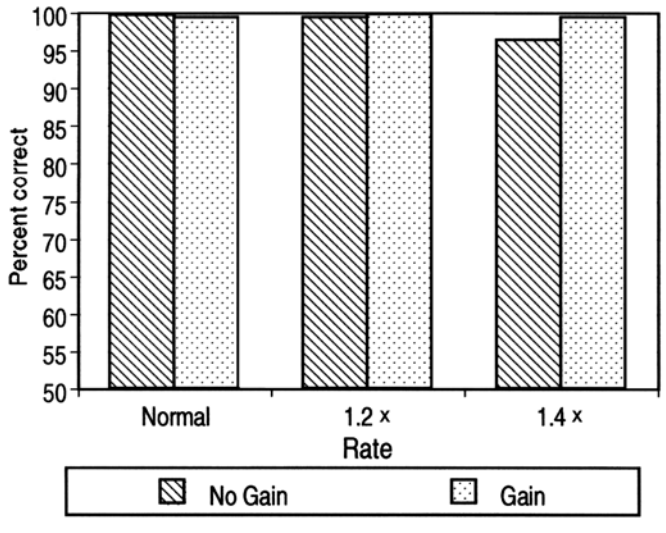

Figure 3.
Effects of consonant gain in quiet at each rate (normal, 1.2×, 1.4×) across groups.

Aside from these additive effects, another source of distortion in the speech processed by the algorithm in this study was a change in the prosody of the talker's utterances. The importance of prosody for older listeners was demonstrated in another recent study [23]. Older listeners performed better on speech recognition when silences were inserted in sentences at natural phrase boundaries than when silences were simply added at regular intervals. It is possible that modifying only one class of phonemes (unvoiced consonants) without concurrent changes in other phonemes, as seen in clear speech, degrades rather than enhances intelligibility because it interferes with the natural prosody of speech.

Increasing the intensity of unvoiced consonants (consonant gain) was shown in this study to have a positive effect under certain conditions. In an earlier study, Gordon-Salant found that increased consonant-vowel ratio (CVR) improved recognition of nonsense syllables in all conditions for older listeners, more than increased consonant duration or a combination of both modifications [17]. Increased CVR was accomplished by an increase in consonant energy by 10 dB relative to the energy in the accompanying vowel, which was calculated as a gain factor of 3.16. This was a higher consonant gain than was achieved in the present study (gain factor of 2), where the positive effects of consonant gain were seen in noise at the normal rate of speech and at the moderately slowed rate (1.2×), but in quiet, consonant gain was helpful only at the slowest rate. Furthermore, in noise when consonant gain was combined with the most time expansion (slowest rate) in the present study, the effects were detrimental to speech recognition. One explanation for differences in the effects of consonant gain between the current study and that of Gordon-Salant et al. may be related partially to the difference in the type of speech materials: nonsense syllables versus meaningful sentences [17]. Higher

level processing for semantic and syntactic analysis is required for sentence recognition. Although contextual cues in sentences might be considered helpful for speech recognition, processing demands are less for nonsense syllables. Processing resources may be adequate for the demands of distorted speech at the lower levels of auditory functions but may result in inadequate resources for higher level processing of semantic and syntactic content of sentences.

Automatic processing techniques that incorporate acoustic characteristics of clear speech with as little distortion as possible would benefit a large number of older veterans. Particularly, for those veterans for whom hearing aid amplification alone does not provide optimal treatment for speech understanding problems in daily communication.

REFERENCES

1. Tyberghein J. Presbycusis and phonemic regression. Acta Otorhinolaryngol Belg 1996;50:85.
2. Crandell C, Henoch M, Dunkerson, K. A review of speech perception and aging: Some implications for aural rehabilitation. J Acad Rehabil Audiol 1991;24:121–32.
3. Gaeth JH. A study of phonemic regression in relation to hearing loss [dissertation]. Evanston (IL): Northwestern University; 1948.
4. Plomp R. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. J Speech Hear Res 1986;29:146–54.
5. Wingfield A. Speech perception and the comprehension of spoken language in adult aging. In: Park DC, Schwarz N, editors. Cognitive aging. Philadelphia (PA): Psychology Press/Taylor & Francis; 2000. p. 175–95.
6. Gordon-Salant S, Fitzgibbons PJ. Selected cognitive factors and speech recognition performance among young and elderly listeners. J Speech Lang Hear Res 1997;40:423–31.
7. Tun P. Fast noisy speech: Age differences in processing rapid speech with background noise. Psychol Aging 1998;13:424–34.
8. Vaughan N, Letowski T. Effects of age, speech rate, and type of test on temporal auditory processing. J Speech Lang Hear Res 1997;40:1192–1200.
9. Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. J Speech Lang Hear Res 1985;28:96–103.
10. Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. J Speech Lang Hear Res 1986; 29:434–46.
11. Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. J Speech Lang Hear Res 1989;32:600–3.
12. Schum D. Intelligibility of clear and conversational speech of young and elderly talkers. J Am Acad Audiol 1996;7:212–18.
13. Payton KL, Uchanski RM, Braida LD. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. J Acoust Soc Am 1994;95:1581–92.
14. Bond Z, Moore T. A note on the acoustic-phonetic characteristics of inadvertently clear speech. Speech Commun 1994;14:325–27.
15. Nejime Y, Moore B. Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. J Acoust Soc Am 1998; 103:572–76.
16. Uchanski R, Choi S, Braida L, Reed C, Durlach N. Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. J Speech Lang Hear Res 1996; 39:494–509.
17. Gordon-Salant S. Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing. J Acoust Soc Am 1986;80(6):1599–607.
18. Sharma A, Kraus N, McGee T, Carrell T, Nicol T. Acoustic versus phonetic representation of speech as reflected by the mismatch negativity event-related potential. Electroencephalogr Clin Neurophysiol 1993;88:64–71.
19. Stevens K, Blumstein S. The search for invariant acoustic correlates of phonetic features. In: Eimas P, Miller J, editors. Perspectives on the study speech. Hillsdale (NJ): Lawrence Erlbaum Associates, Inc.; 1981. p. 1–35.
20. Cox RM, Alexander GC, Gilmore C. Development of the Connected Speech Test (CST). Ear Hear 1987;8:119S–26S.
21. Hintze JL. Number Cruncher Statistical system. In: 2000 ed. Kaysville (UT): NCSS; 2000.
22. Gordon-Salant S, Fitzgibbons PJ. Recognition of multiply degraded speech by young and elderly listeners. J Speech Lang Hear Res 1995;38:1150–56.
23. Wingfield A, Tun P, Koh C, Rosen M. Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech. Psychol Aging 1999; 14:380–89.

Submitted for publication June 26, 2001. Accepted in revised form February 5, 2002.