Published in final edited form as:

J Speech Hear Res. 1991 October ; 34(5): 1180–1184.

Recognition of Synthetic Speech by Hearing-Impaired Elderly Listeners

Larry E. Humes, Kathleen J. Nelson, and David B. Pisoni
Indiana University, Bloomington

Abstract

The Modified Rhyme Test (MRT), recorded using natural speech and two forms of synthetic speech, DECTalk and Votrax, was used to measure both open-set and closed-set speech-recognition performance. Performance of hearing-impaired elderly listeners was compared to two groups of young normal-hearing adults, one listening in quiet, and the other listening in a background of spectrally shaped noise designed to simulate the peripheral hearing loss of the elderly. Votrax synthetic speech yielded significant decrements in speech recognition compared to either natural or DECTalk synthetic speech for all three subject groups. There were no differences in performance between natural speech and DECTalk speech for the elderly hearing-impaired listeners or the young listeners with simulated hearing loss. The normal-hearing young adults listening in quiet out-performed both of the other groups, but there were no differences in performance between the young listeners with simulated hearing loss and the elderly hearing-impaired listeners. When the closed-set identification of synthetic speech was compared to its open-set recognition, the hearing-impaired elderly gained as much from the reduction in stimulus/response uncertainty as the two younger groups. Finally, among the elderly hearing-impaired listeners, speech-recognition performance was correlated negatively with hearing sensitivity, but scores were correlated positively among the different talker conditions. Those listeners with the greatest hearing loss had the most difficulty understanding speech and those having the most trouble understanding natural speech also had the greatest difficulty with synthetic speech.

Keywords

speech recognition; synthetic speech; hearing impaired; elderly; context

The present study addresses a problem arising from the convergence of two trends, one technological and the other demographic. The technological trend has to do with the increased use of computer-synthesized speech in today's society. Examples include talking calculators, recorded telephone services, travel consultants, weather information, and telemarketing messages. The demographic trend involves the rapid growth of the segment of the American population over the age of 65 years. Currently comprising about 11%–12% of the U.S. population, individuals over age 65 have been projected to comprise 20%–25% of the U.S. population by the year 2025 (NCHS, 1986). Approximately 30% of those over the age of 65, moreover, have a significant hearing loss (Herbst, 1983; NCHS, 1977, 1986). This hearing loss, usually a high-frequency sensorineural hearing loss (Corso, 1959; Spoor, 1967), reduces the ability to understand natural speech (CHABA, 1988). Computer-synthesized speech is considered to be a degraded or impoverished speech signal, lacking

many of the important redundancies inherent in natural speech that make the latter more readily understood (Logan, Greene, & Pisoni, 1989; Pisoni, Nusbaum, & Greene, 1985). One of the questions posed in the present investigation was whether the impoverished speech signal generated by contemporary speech synthesizers further exacerbates the speech-understanding difficulties of the hearing-impaired elderly.

Another question raised in this investigation is whether or not the difficulties understanding speech experienced by the hearing-impaired elderly are attributable primarily to their hearing loss. Two approaches to this issue were pursued here. In one approach, a group of young normal listeners with a hearing loss simulated by the introduction of a spectrally shaped background noise served as a comparison group. These individuals have been found to mimic the performance of *young* listeners with sensorineural hearing loss on a variety of psychoacoustic and speech-recognition tasks (Dubno & Schaefer, 1989; Fabry & Van Tasell, 1986; Humes, Dirks, Bell, & Kincaid, 1987; Humes, Espinoza-Varas, & Watson, 1988; Zurek & Delhorne, 1987). The other approach was to perform correlational analyses on the data from the elderly hearing-impaired listeners to determine the contribution of peripheral hearing loss to their speech-recognition difficulties. Whereas the former approach is based on group comparisons of average data, the latter approach is based on the performance of individual subjects in the hearing-impaired group.

Finally, this study examined the effects on performance of reducing stimulus/response uncertainty. Uncertainty was reduced by presenting the monosyllabic test items in a closed-set rather than an open-set response format. The reduction of uncertainty by restricting the set of possible words is a form of context (Miller, Heise, & Lichten, 1951). We examined the ability of the various groups to make use of context in their encoding of the more difficult computer-synthesized speech signals.

Method

Subjects

The subjects in this study included 30 individuals, 10 in each of three groups. The 10 subjects (5 male, 5 female) comprising Group HI (hearing-impaired elderly) ranged in age from 66–73 years. Each subject in this group had a bilaterally symmetrical high-frequency sensorineural hearing loss attributed to presbycusis. The mean air-conduction thresholds for the test ears of these subjects are provided in Figure 1. All subjects also had normal tympanograms and an acoustic reflex in the test ear in response to the contralateral presentation of a 1000-Hz tone at 100 dB HL (ANSI, 1969).

The remaining 20 subjects were normal-hearing young adults ranging in age from 19–22 years. All 20 subjects had pure-tone air-conduction thresholds less than 15 dB HL for octave intervals from 250 through 8000 Hz and normal immittance measurements bilaterally (tympanograms and acoustic reflexes). Half the subjects were in group NH (normal hearing) and were tested in quiet, whereas the other half, group MN (masked normals), were tested in a background of spectrally shaped noise designed to simulate the average hearing loss of group HI. The mean masked thresholds produced in a separate group of 3 normal-hearing subjects by the spectrally shaped noise are also shown in Figure 1. The average hearing losses for groups HI and MN are in good agreement.

For the elderly hearing-impaired subjects, the test ear was the ear with the lowest speech-recognition threshold. If both ears had equivalent SRTs, then the test ear was selected arbitrarily. For the other two groups of subjects, the test ear was chosen arbitrarily with about half the test ears in each group being left ears and half being right ears.

All subjects in this study were native speakers of English in good physical health at the time of testing.

Materials and Apparatus

The stimulus materials used were re-recordings of the Modified Rhyme Test (MRT; House, Williams, Hecker, & Kryter, 1965) used by Logan et al. (1989). The MRT materials used in this study consisted of 300 monosyllabic consonant-vowel-consonant English words divided into 50 sets of six rhyming words each. One of the six rhyming words was selected from each of the 50 sets to form six randomized lists of 50 words of equivalent difficulty (Logan et al., 1989). Five of the six 50-word lists were used with a different list assigned to each condition. The natural-speech samples were audio recordings originally produced by Dennis Klatt at MIT in 1979. In addition, two MRT word lists were produced by each of two text-to-speech systems, DECtalk 1.8 (Perfect Paul voice) and Votrax Type 'n Talk. These two samples of synthetic speech represent the extremes in quality of text-to-speech systems with DECtalk representing high-quality synthetic speech and Votrax representing low-quality speech (Logan et al., 1989; Pisoni et al., 1985). Two lists were required for the synthetic materials so that they could be administered in both an open and a closed response format.

The vowel /a/ was recorded at the beginning of each audiotape by the same “talker” as the test materials. The calibration vowels had an rms amplitude equal to that of the test items and were used to specify level at playback. Approximately 4 s of silence separated each of the monosyllabic words on the tapes. Further details regarding the generation of the stimulus materials can be found in Logan et al. (1989).

The audiotapes were played back using a cassette tape deck (Sansui, Model D-W9). The output from the tape deck was routed to an amplifier/mixer and delivered to a TDH-39 headphone mounted in an MX-41/AR supra-aural cushion. For the subjects in group MN, a digital audiotape of the spectrally shaped noise was output from a separate tape deck (Panasonic SV-3500) and then mixed with the speech signal prior to amplification and transduction. A speech presentation level of 75 dB SPL, as measured in an NBS-9A coupler, was used throughout this study.

Procedures

After subject selection, brief instructions were presented orally to the subjects, followed immediately by more detailed, written instructions. Each subject listened to a short sample of the synthetic speech, which described the procedures used in the upcoming task. This served to familiarize them with the speech quality of the system. The speech material heard during the familiarization task was also accompanied by a printed verbatim script of the speech sample, on which subjects could follow the narrative.

Testing was initiated following instructions and familiarization. Each of the 50 monosyllabic words comprising a test was presented in sequence with approximately 4 s of silence between successive words. A carrier phrase was not used. For open-set speech recognition, subjects were presented with a response sheet containing 50 numbered blank lines and were instructed to write the word heard in the blank. For closed-set speech identification, the response sheet contained 50 numbered rows of six rhyming monosyllabic words. Subjects were instructed to circle the word in each row that they heard on each trial. For both tasks, subjects were encouraged to respond to each trial, even if they had to guess. The order of presentation of each of the five listening conditions was random.

Results and Discussion

Figure 2 shows the speech-recognition scores obtained from the three groups of subjects for each of the three “talkers.” The individual percent-correct scores of each of the subjects were converted to proportions and arcsine transformed prior to statistical analysis. A mixed-model two-factor analysis of variance (ANOVA) was then performed on the transformed scores. This analysis revealed significant ($p < .01$) effects of group [$F(2, 27) = 16.43$] and talker [$F(2, 54) = 152.6$] and a significant group-by-talker interaction [$F(4, 54) = 6.41$]. The latter interaction necessitated further analyses of the simple effects of group and talker. First, separate ANOVAs were performed on the data from each group to examine the effect of talker. For each group, the effect of talker was significant [Group NH: $F(2, 18) = 156.0$, $p < .001$; Group HI: $F(2, 18) = 17.9$, $p < .001$; Group MN: $F(2, 18) = 148.3$, $p < .001$]. Post hoc analyses, however, revealed a slightly different effect of talker across the three groups of subjects. For the young normal-hearing subjects, recognition of natural speech was better than DECTalk speech and the latter was superior to Votrax speech. For the other two groups of subjects, Votrax speech again resulted in significantly poorer speech recognition, but there were no differences between natural and DECTalk speech. Thus, for the listeners in this study with sensorineural hearing loss, either actual or simulated, DECTalk speech yielded performance roughly equivalent to that of natural speech whereas Votrax speech resulted in significantly reduced performance.

To examine the effects of group, three additional ANOVAs were performed, one for each talker. The effect of group on speech recognition was found to be significant for each talker [Natural: $F(2, 27) = 14.7$, $p < .001$; DECTalk: $F(2, 27) = 19.0$, $p < .001$; Votrax: $F(2, 27) = 12.6$, $p < .001$]. Student-Newman-Keuls post hoc analyses ($p < .05$) revealed the same pattern of results for each talker. The young normal-hearing subjects listening in quiet outperformed both of the other groups, but there was no difference in performance between the elderly hearing-impaired listeners and the young listeners with simulated hearing loss. Thus, the speech-recognition performance of the hearing-impaired elderly in this study was accurately mimicked by the young normal-hearing listeners with simulated hearing loss. This suggests that the primary factor underlying the word-recognition difficulties in quiet of the elderly hearing-impaired subjects in this study is the loss of audibility and accompanying loudness recruitment associated with sensorineural hearing loss. When these effects are simulated in a group of young listeners lacking cochlear pathology, equivalent performance is observed. This suggests that neither age-related nor pathology-related factors are needed to account for the word-recognition performance in quiet of the hearing-impaired elderly in this study.

Aside from the obvious trends in the mean data, the error bars in Figure 2, representing one standard deviation above the mean, also reveal that the variability within the elderly hearing-impaired subjects was much greater than that of the other groups. Much of this variability, however, appears to be associated with individual differences in hearing sensitivity. Correlations between speech-recognition performance and high-frequency, pure-tone average (average hearing threshold at 1000, 2000, and 4000 Hz) for the elderly hearing-impaired subjects were negative and significant ($p < .01$) for each talker. Specifically, the correlation coefficients between speech recognition and high-frequency pure-tone average were -0.73 , -0.75 , and -0.80 for natural, DECTalk and Votrax speech, respectively. Thus, as hearing loss increased, word-recognition performance in quiet decreased and approximately 50%–60% of the variance in performance could be explained by this factor alone.

Speech-recognition performance was also correlated across talkers for the subjects in each group. No significant correlations were observed within the two groups of young normal-hearing subjects (NH and MN). This was due, most likely, to the narrow range of

performance observed for these two groups of subjects. For the elderly hearing-impaired subjects, however, strong positive and statistically significant ($p < .01$) correlations were observed among the scores from the various talkers. The correlation between performance with natural speech and synthetic speech was 0.96 and 0.90 for DECTalk and Votrax, respectively, whereas the correlation between performance with the two synthetic speech systems was 0.85. Hence, those elderly hearing-impaired listeners having the most difficulty with natural speech were most likely to be the same ones having the greatest difficulty with synthetic speech.

Finally, the effect of reducing stimulus/response uncertainty on perception of synthetic speech was examined by comparing the performance on the open-set recognition task to that for an equivalent closed-set identification task. The difference in performance between open-and closed-set formats was calculated by subtracting the score for the open response format from the score for the closed format. The mean values for each talker and subject group are shown in Figure 3. All values are positive, indicating that a net gain in recognition performance was achieved by changing from an open to a closed response format. This finding was expected given the reduction in uncertainty associated with the change in context (Miller et al., 1951). The gain in arcsine-transformed units was then determined for each subject and talker and subjected to a mixed-model ANOVA. This analysis revealed a significant effect of talker only [$F(1, 27) = 16.3$; $p < .01$]. The effect of subject group and the group-by-talker interaction were not significant ($p > .05$). The gain in performance from open to closed response format was greatest for the poor-quality Votrax speech. Interestingly, the three groups of listeners in this study did not differ in the size of the effect. The elderly hearing-impaired subjects gained as much from the additional contextual information available in the closed response format as the two younger groups.

In summary, several conclusions can be drawn from these data. First, for the hearing-impaired elderly listeners in this study, high-quality synthetic speech (DECTalk) was perceived as well as natural speech. Second, much of the word-recognition difficulty experienced in quiet by the 10 hearing-impaired elderly subjects in this study was due to their hearing loss. Third, those impaired listeners in this study who had difficulty understanding natural speech were also the ones who tended to have the most difficulty understanding synthetic speech. Thus, it is not necessary to perform a separate assessment of speech-understanding abilities using synthetic speech. Finally, the hearing-impaired elderly listeners in this study gained as much from the addition of contextual information and the reduction of uncertainty as did the younger listeners.

Acknowledgments

This work was supported, in part, by grants to authors Humes and Pisoni from the NIH.

References

- American National Standards Institute. Specifications for audiometers. ANSI; New York: 1969. ANSI S3.6-1969
- CHABA. Speech understanding and aging. *Journal of the Acoustical Society of America*. 1988; 83:859–895. [PubMed: 3281988]
- Corso J. Age and sex differences in pure-tone thresholds. *Journal of the Acoustical Society of America*. 1959; 31:498–507.
- Dubno JR, Schaefer AB. Frequency resolution for broadband-noise masked normal listeners. *Journal of the Acoustical Society of America*. 1989; 86:S122.
- Fabry DA, Van Tasell DJ. Masked and filtered simulation of hearing loss. Effects on consonant recognition. *Journal of Speech and Hearing Research*. 1986; 29:170–178. [PubMed: 3724111]

- Herbst, KRG. Psychological consequences of disorders of hearing in the elderly. In: Hinchcliffe, R., editor. Hearing and balance in the elderly. Churchill-Livingston; Edinburgh, Scotland: 1983. p. 174-200.
- House AS, Williams CE, Hecker MHL, Kryter KD. Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*. 1965; 37:158–166. [PubMed: 14265103]
- Humes LE, Dirks DD, Bell TS, Kincaid GE. Recognition of nonsense syllables by hearing-impaired listeners and by noise-masked normal hearers. *Journal of the Acoustical Society of America*. 1987; 81:765–773. [PubMed: 3584685]
- Humes LE, Espinoza-Varas B, Watson CS. Modelling sensorineural hearing loss. I. Model and retrospective evaluation. *Journal of the Acoustical Society of America*. 1988; 83:188–202. [PubMed: 3343439]
- Logan JS, Greene BG, Pisoni DB. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*. 1989; 86:566–581. [PubMed: 2527884]
- Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*. 1951; 41:329–335. [PubMed: 14861384]
- NCHS. Prevalence of selected impairments. 1977. p. 81-1562. DHHS Pub. No. (PHS)
- NCHS. Current estimates from the National Health Interviews Survey, U.S. 1986. p. 86-1588. Vital and Health Statistics, Series 101985No. 160. DHHS Pub. No. (PHS)
- Pisoni DB, Nusbaum HC, Greene BG. Perception of synthetic speech generated by rule. *Proceedings of the IEEE*. 1985; 73:1665–1676.
- Spoor A. Presbycusis values in relation to noise-induced hearing loss. *International Audiology*. 1967; 6:48–57.
- Zurek DM, Delhorne LA. Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. *Journal of the Acoustical Society of America*. 1987; 82:1548–1559. [PubMed: 3693695]

FIGURE 1.

Mean air-conduction pure-tone thresholds for the elderly hearing-impaired listeners in quiet (filled circles) and a small group of normal-hearing young adults ($N=3$) in a background of spectrally shaped noise adjusted to match the hearing loss of the elderly listeners (unfilled circles). Error bars represent one standard deviation above the mean values of the elderly hearing-impaired listeners.

FIGURE 2.

Mean percent-correct speech-recognition scores for the natural (NAT), DECtalk (DEC/O), and Votrax (VOT/O) talkers. “/O” refers to open-set response format. NH = normal-hearing young adults; HI = hearing-impaired elderly adults; and MN = masked, normal-hearing young adults. Error bars represent one standard error above the mean.

FIGURE 3.

Mean intelligibility gain in percent correct for each of the three subject groups and each talker. Intelligibility gain represents the improvement in score from the open to the closed response format. Error bars represent one standard error above the mean.