AGING VOICES AND SPEECH INTELLIGIBILITY: IMPLICATIONS FOR COMMUNICATION BY OLDER TALKERS

Huiwen Goy[1], M. Kathleen Pichora-Fuller[1] and Pascal van Lieshout[1,2]
[1]Department of Psychology, University of Toronto Mississauga
[2]Department of Speech-Language Pathology, University of Toronto

1 Introduction

The voice changes with normal aging. For instance, older adults may have more extreme values on measures of irregularities in fundamental frequency (jitter) and intensity (shimmer) than the voices of younger adults [1]. Listeners perceive such voices to be rougher [2] and perform more poorly on word recognition when speech has been synthetically jittered [3]. Given that some older adults have voices that naturally contain more of these irregularities, it is possible that their voices would be more difficult to understand in challenging noisy situations, and when their communication partners are experiencing age-related changes in auditory processing. To investigate the effects of voice quality on intelligibility, we recorded the speech of three older talkers who had different amounts of vocal jitter and shimmer. We predicted that intelligibility would relate to the voice quality of the talkers.

2 Experiment 1

2.1 Method

2.1.1 Talkers

Three older adult females between 68 and 74 years of age were selected from a database of recordings of healthy adults [1]. They were native Canadian English speakers who reported that they were in good health. They had pure-tone audiometric thresholds ≤ 25 dB HL from 250 to 3000 Hz, and inter-aural differences in thresholds ≤ 15 dB from 250 to 8000 Hz. These talkers were selected based on the percentile ranks of their values of jitter (local) and shimmer (local) within their age and gender group (see Table 1).

	Jitter (%)		Shimmer (%)	
Talker	Mean	Rank	Mean	Rank
Worst	0.85	16	4.87	12
Mid	0.29	65	3.36	25
Best	0.17	90	1.01	90

Table 1: Values of three voice acoustic measures and percentile ranks for the three selected talkers. Worst = poorest voice quality; Mid = medium voice quality; Best = best voice quality.

2.1.2 Stimuli recording

Talkers recorded the Northwestern University No. 6 (NU6) word recognition test items [4] in a single-walled IAC booth using a Sennheiser Linear E825S microphone placed 5 cm away from the lips. Tucker-Davis Technologies System III hardware was used. Items were monosyllabic target words following the carrier phrase "*Say the word* ____". For the recording sessions, sentences were presented on a monitor at a rate controlled by the experimenter. Talkers were instructed to "read the sentence aloud in your normal, most comfortable voice" and they took frequent breaks. Prior to each recording session, a sample of recordings from the Words in Noise (WIN) test [5] spoken by a professional talker was played to demonstrate an appropriate speaking rate; five practice sentences were spoken by each talker. The four NU6 lists were recorded over two days in different orders on each day to yield four tokens of each of the 200 sentences. During editing, the RMS energy of each sentence was equated to 0.05 Pa using a custom MATLAB program.

2.1.3 Listeners

Listeners were 16 younger adults (M=18.4 years, SD=0.6) who had learned English before the age of 5 years in an English-speaking country and had pure-tone thresholds ≤ 20 dB HL from 250 to 8000 Hz, with no significant inter-aural difference in thresholds. Listeners gave informed consent and received course credit for participating.

2.1.4 Procedure

Listeners were tested on four NU6 lists, each spoken by a different talker (3 females + the professional WIN talker), while seated in a double-walled IAC booth. Sentences were presented monaurally over TDH-50P earphones at 70 dB SPL and mixed with multi-talker babble from the WIN test at +1 dB SNR. Participants were instructed to report the last word of each sentence and guessing was encouraged with no time limit on responding. The combination and order of NU6 lists and talkers was counterbalanced across listeners. Before data analysis, listener scores were transformed from raw scores to rationalized arcsine units (RAU) [5], and all post-hoc t-tests were conducted with Holm-Bonferroni correction. Acoustic measures of stimuli were obtained using the Praat speech analysis program [6].

2.2 Results

Surprisingly, the talker with the poorest voice quality was as intelligible as the talker with the best voice, and both were more intelligible than the talker with the medium-quality voice, while the professional talker was the least intelligible of all (Figure 1). This pattern of results was confirmed by a within-subjects ANOVA, showing a significant effect of talker, $F(3, 45) = 130.5$, $p < .001$. The mean correct word recognition scores for the talkers with the best and worst voices did not differ significantly ($p = .31$); their scores were different than that of the talker with the medium-quality voice (p's < .001), and the score for the professional talker was different from all other talkers (p's < .001).

2.3 Acoustic measurements

Although stimuli intensity had been equated at the sentence level for all talkers, the distribution of energy within sentences differed among talkers (Table 2).

	Sentence	Carrier		Target word	
Talker	Rate (syll/sec)	F_0 (Hz)	Int (dB)	F_0 (Hz)	Int (dB)
Pro	3.4 (0.2)	247 (11)	70.4 (0.5)	170 (16)	67.5 (1.3)
Worst	2.3 (0.2)	172 (8)	68.4 (0.9)	192 (35)	68.6 (1.4)
Mid	3.2 (0.3)	173 (13)	68.2 (1.3)	171 (23)	68.8 (1.2)
Best	3.0 (0.3)	190 (17)	68.7 (1.8)	187 (41)	68.4 (1.4)

Table 2: Mean values for acoustic measures of sentences (S.D.'s). Int = intensity; Pro = professional talker.

Specifically, targets spoken by the professional talker were about 1 dB less in intensity than for other talkers while the mean intensity of her carrier phrases was about 2 dB higher. Relative to the mean F_0 of the carrier phrase, the mean F_0 of target words was lowered for the professional talker but raised for the talker with the worst voice quality.

2.4 Discussion

Surprisingly, listeners performed very poorly with the professional talker, and better than expected with the talker who had the poorest voice quality. The professional talker's lower intensity on target words and the higher F_0 for the talker with the worst voice quality may have contributed to these results. Experiment 2 investigated these possibilities.

3 Experiment 2

3.1 Method

We equated the intensity of all target words and replaced the four talkers' carrier phrases with the standard NU6 carrier phrase of the talker recorded by Auditec. Listeners were 16 young adults (*M*=19.2 years, SD=2.7) who had similar characteristics as listeners in Experiment 1 and were naïve to this task. The procedure was the same as in Experiment 1.

3.2 Results

The pattern of results was similar to that of Experiment 1 (Figure 1); a within-subjects ANOVA showed a significant effect of talker, $F(3, 45) = 54.3$, $p < .001$. Correct word recognition scores for the talkers with the best and worst voices were not significantly different ($p = .17$) and their scores were higher than for the talker with the medium-quality voice (p's < .01); the score for the professional talker was different from all other talkers (p's < .001). A second ANOVA with 'experiment' as a between-subjects factor and 'talker' as a within-subjects factor showed that listeners performed slightly better in Experiment 1 (*M*=62.8%, *SD*=18.9%) than in Experiment 2 (*M*=58.5%, *SD*=16.8%), due to a decrease of 10 percentage points for the talker with the best voice in Experiment 2. There were significant main effects of 'Experiment', $F(1, 30) = 4.39$, $p = 0.04$, and 'talker', $F(3, 90) = 164.7$, $p < .001$, and a significant interaction between these factors, $F(3, 90)$, $p = .04$. Listeners achieved significantly higher scores for the talker with the best voice in Experiment 2 than in Experiment 1 ($p = .006$), but none of the results for other talkers differed significantly between experiments (p's > .4).

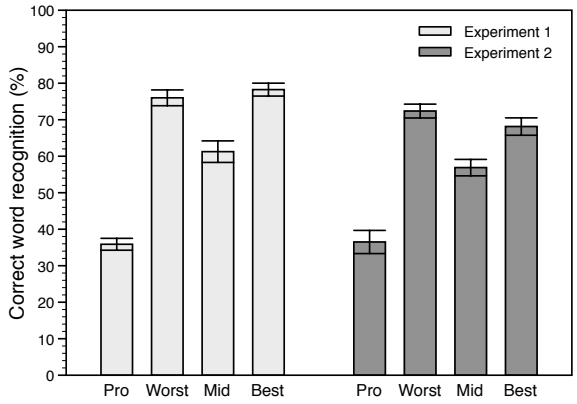

Figure 1: Mean correct word recognition scores of listeners for four talkers in two experiments.

3.3 Acoustic measurement of talkers

The talker with the poorest voice quality produced the longest duration target words of all talkers, with especially long consonant durations and long transitions between vowels and consonants (Table 3).

	Word	Consonant		Vowel steady-state portion	
Talker	Dur (ms)	Dur (ms)	Peak int (dB)	Dur (ms)	Mean int (dB)
Pro	471	63	63.1	101	72.9
Worst	717	203	68.7	98	73.9
Mid	596	143	68.7	114	72.8
Best	644	173	70.3	107	72.4

Table 3: Mean acoustic measures of target words. Dur = duration.

3.4 Discussion

Talker differences in intelligibility were not caused by differences in the intensity of target words or an emphasis on target words by F_0. The durations of transitions between vowels and consonants were longest for the talker with the poorest voice quality, followed by the talker with the best voice, with the professional talker having the shortest transitions. Since portions of the speech signal that contain change supply the most information for speech recognition [8], the differences in transition duration may have played a key role in determining talker intelligibility in this study.

4 Conclusions

Age-related changes in the voice may negatively affect speech communication, but results from this study suggest that talkers may compensate for poorer voice quality through articulation and speech rate adaptations.

Acknowledgements

This study was funded by an NSERC grant awarded to K.P.-F. (RGPIN 138472) and a CRC grant awarded to P.vL. (950-213162). We thank James Qi, Jessica Banh and Rajbir Deo for their help in experimental setup and data collection.

References

[1] H Goy, DN Fernandes, MK Pichora-Fuller, and P Van Lieshout, "Normative voice data for younger and older adults," J. Voice **27**, 545-555 (2013).

[2] J Hillenbrand, "Perception of aperiodicities in synthetically generated voices," J. Acoust. Soc. Am. **83**, 2361-2371 (1988).

[3] MK Pichora-Fuller, BA Schneider, E MacDonald, HE Pass, and S. Brown, "Temporal jitter disrupts speech intelligibility: A simulation of auditory aging," Hear. Res. **223**, 114-121 (2007).

[4] TW Tillman, and R Carhart, *An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6 (SAM-TR-66-55)*, Brooks Air Force Base, TX: USAF School of Aerospace Medicine (1966).

[5] RH Wilson, "Development of a speech in multitalker babble paradigm to assess word-recognition performance," J. Am. Acad. Audiol. **14**, 453–470 (2003).

[6] GA Studebaker, "A "rationalized" arcsine transform," J. Speech Hear. Res. **28**, 455-462 (1985).

[7] P Boersma, and D Weenink, Praat: doing phonetics by computer (University of Amsterdam, The Netherlands)

[8] CE Stilp, and KR Kluender, "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Nat. Acad. Sci. **107**, 12387-12392 (2010).