Effect of Within- and Between-talker Variability on Word Identification in Noise by Younger and Older Adults

Huiwen Goy1, Kathy Pichora-Fuller1,2, Pascal van Lieshout1,2,3,4, Gurjit Singh1,3, Bruce Schneider1

1 Department of Psychology, University of Toronto, Mississauga, Canada
2 Department of Speech Pathology, University of Toronto, Toronto, Canada
3 Toronto Rehabilitation Institute, Toronto, Canada
4 Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Canada

huiwen.goy@utoronto.ca, kpfuller@utm.utoronto.ca, p.vanlieshout@utoronto.ca, gurjit@psych.utoronto.ca, bschneid@utm.utoronto.ca

1. Introduction

Talkers alter their speech in noisy environments yet most speech-in-noise testing uses recorded in quiet. Sentences from a common test (SPIN-R) were recorded by a new talker in different talking conditions and the original and new materials were used to test word identification accuracy in younger and older adults. Inter- and intra-talker differences affected performance. Intelligibility was better for materials heard in noise when the materials were spoken in noise, or when the talker was asked to speak loudly, especially for older listeners. The most likely acoustical explanation for the differences seems to be increased intensity and duration in the production of the sentence-final target words.

Index Terms: Aging, Lombard Effect, speech in noise perception and production

Speech produced in noisy environments has different characteristics than speech produced in quiet. The reflex by which a talker increases vocal effort as the amount of noise in the environment increases is termed the Lombard Effect [1]. Although the Lombard Effect is well-known, standardized tests of speech intelligibility in noise have been developed using speech materials recorded in quiet conditions, with background noise being added later. The purpose of the present study was to investigate the extent to which the results of speech intelligibility testing might differ for younger and older adults with good audiograms if more ecologically appropriate spoken versions of the materials were employed.

The best known aspect of the Lombard Effect is that the intensity of speech increases when the talker is listening to noise; however, numerous experiments have confirmed that there are changes in the acoustical characteristics of speech produced in noise besides changes in intensity, including changes in duration and spectral properties. These changes occur across talkers, various types of speech material, and different levels and types of noise. When male and female adults recited a passage in quiet and when listening over earphones to multi-talker, traffic, or broadband noise, the mean intensity of speech increased as the noise level increased, and the mean pitch shifted to a higher frequency [2]. For two females reciting digits in quiet and in three levels of white noise, as the noise level increased, vocal intensity increased, and there was a significant increase in high-frequency energy and an increase in word duration [3]. For males and females producing single words in quiet and in an 85 dB SPL white noise environment, vocal level increased when they spoke in noise [4]. Additionally, there was an increase in vowel duration leading to an increase in overall word duration, and an increase in vowel pitch. In summary, speaking in noise leads to increased vocal intensity, pitch, duration, and more energy in higher frequencies.

Speech produced in noise has been shown to be more intelligible than speech produced in quiet [5]. Analyses of the recordings showed that the intensity and duration of the words and sentences were greater for speech produced in noise compared to speech produced in quiet. It has been hypothesized that the Lombard Effect may compensate for the talker’s increased difficulty in monitoring his or her own speech in a noisy environment [6]. In turn, these changes in speech production may aid understanding for a listener in a noisy environment.

Despite the well-documented changes in speech production when talkers are in a noisy environment and the likely consequences of these changes for speech intelligibility, there have been few studies which have taken into account how changes in voice characteristics might influence the results of tests of speech understanding in noise. Usually studies have used stimuli recorded in quiet and presented them in a noisy listening environment, creating an artificial situation. For one experiment, a female recorded 50 low-context sentences from the Speech Perception in Noise test (SPIN) [7] in quiet, in multi-talker babble, and in broadband noise, and the sentences were equated for intensity level [8]. There were three noise environments for the listener: multi-talker babble at 0, -5, and -10 dB signal-to-noise ratio (S:N). When young adults listened to the recordings at -5 dB S:N, correct word identification was 15% higher on average for speech produced in noise than for speech produced in quiet, but when listening at the 0 and -10 dB S:N there were no significant differences in performance depending on the recording condition. Speech produced in babble was more intelligible than speech produced in wideband noise, but only in the -5 dB S:N listening environment. Thus, the differences in intelligibility due to the Lombard Effect seem to be greatest in intermediate rather than in more extreme listening conditions where performance is closer to ceiling or floor. It may be worth noting, however, that a male recorded the original SPIN sentences and that the test is standardized for lists of 50 sentences with 25 high-context and 25 low-context [9] whereas [8] used a female talker and only a subset of low-context sentences.
Difficulty understanding speech in noise is a hallmark of auditory aging, even when older listeners have clinically normal audiometric thresholds in the speech range [10]. Older listeners need over a 2 dB S:N advantage to achieve the same word identification performance on the SPIN-R test as younger adults [11], with the age-related differences in speech understanding being greater for words in low-context than in high-context sentences. It is possible that the older listeners are particularly disadvantaged when the speech heard in noise was not produced in noise. Perhaps the apparent age-related differences in speech intelligibility would be reduced if the older adults had the benefit of listening to speech produced under the Lombard Effect.

In the present study, we tested the word identification performance of younger and older adults with good audiograms in an intermediate S:N using full 50-sentence SPIN-R lists, when the speech materials were spoken by the original talker and by a new talker who imitated the original recordings under conditions chosen to elicit variations in speech production as in the Lombard Effect. We expected that, compared to performance on the standardized test, word identification would be better when listeners heard the speech produced by the new talker in noise, and that the improvement would be greatest when the type of noise heard by the listeners matched the noise heard by the talker. We also expected that older adults might show greater improvement than younger adults, especially for words in low context.

2. Methods

2.1. Participants

Participants were recruited from the university and local community. They were 16 younger adults (mean age = 19.9 years, SD = 1.7) and 16 older adults (mean age = 69.0 years, SD = 4.1). All spoke English as a first language, and had audiometric thresholds of ≤ 25 dB HL at frequencies from .25 to 3 kHz in the test ear. Each participant was paid $10/hour for their participation and gave informed consent. This study was approved by the university’s ethics review board.

2.2. Stimuli recording

2.2.1. Stimuli

The stimuli used were the eight lists of the SPIN-R test, each with 50 sentences categorized into 25 high-context and 25 low-context sentences. In high-context items the sentence-final target word is highly predictable from the beginning of the sentence (e.g., “The farmer baled the hay.”), whereas in the low-context items the target word is not so predictable from the preceding context (e.g., “Tom discussed the hay.”). Each sentence is paired with multi-talker babble. The lists are equated for average speech level and all lists were presented with speech at 70 dB SPL in multi-talker babble at 0 dB S:N.

2.2.2. Talking conditions

List 1 spoken by the original talker was used for comparisons with a new male talker (DF), who recorded lists 1 to 8. DF was a graduate student in linguistics who was in his 20s and spoke Canadian English as his first language. He was selected from a larger pool of talkers as the best match to the talker who recorded the original SPIN-R test in terms of fundamental frequency, speaking rate, and overall intensity (see section 2.2.3).

The speech of the new talker was recorded using a Tucker-Davis Technologies System III, Time-Frequency Representation (TFR) recording software by Avava Innovations Inc., and a Sennheiser E845S table microphone. The talker was seated in a sound-attenuated booth; the distance from the microphone to his mouth was eight inches. The lists were spoken with the talker’s ears occluded by earphones (Sennheiser HD265 Linear). Lists 1, 2, 7 and 8 were spoken in quiet. For lists 1 and 2, the instructions were to “speak normally”, and to use the original talker as a guide for intonation and rate of speaking. For list 7, DF was instructed to “speak clearly”. For list 8, he was instructed to “speak loudly”. Lists 3 to 6 were spoken in noise; DF was informed that some background noise would be played while he was speaking, but no instructions were given to speak over the noise, or to pay attention to it. Lists 3 and 4 were spoken while the talker was listening to multi-talker babble over earphones, played at 62 and 66 dB SPL respectively, and lists 5 and 6 were spoken while listening to speech spectrum noise at 62 and 66 dB SPL respectively.

All lists were digitized at a sampling rate of 24 kHz. The new recordings were edited using the Praat speech analysis program [12] to match the original recordings for soundfile length. The original list 1 and the new lists were low-pass filtered at 6 kHz to reduce differences due to recording technique. The RMS level of all lists sentences was matched.

2.2.3. Voice analysis of new and original talkers

List 1 was used to compare the talkers. For the new talker, the average F0 was 121.7 Hz and the average speaking rate was 4.3 syllables/sec, compared to 120.5 Hz and 4.3 syllables/sec for the original talker.

2.3. Experiment procedure

Each participant was tested on the eight lists of sentences (eight talking conditions) in two one-hour sessions. The order of conditions for each participant was pre-determined, with each participant starting on a different list to balance practice effects across talking conditions. Participants were familiarized with the task using twelve sentences not used in the experiment. The sentences were presented monaurally at 70 dB SPL under TDH-50P earphones, while the participant was seated in a double-walled sound-attenuating booth. Each sentence was presented with multi-talker babble at 0 dB S:N. The participant was instructed to report the last word of each sentence, immediately after it was heard, but there was no time limit on responding. Guessing was strongly encouraged.

3. Results

Figure 1 illustrates the performance of younger and older adults in the eight talking conditions, in a listening environment of multi-talker babble at 0 dB S:N. Listeners identified words more accurately in the high-context than low-context conditions. Younger adults identified words more accurately in noise than did older adults. In the low-context conditions, the differences in performance between younger and older adults were more pronounced. For both age groups, word identification for the new talker’s voice was better than for the original talker. In general, word identification was best for talking conditions with louder noise environments (conditions 4 and 6), when the type of noise in the talking condition matched the noise in the listening environment.
This description of word identification performance was confirmed by an Analysis of Variance (ANOVA) with age (young vs old) as a between-subjects factor, and talking condition (8 types) and context (high vs low) as within-subjects factors, which showed main effects of age, $F(1,30) = 17.03, p < 0.001$, context, $F(1,30) = 531.02, p < 0.001$, and talking condition, $F(7,210) = 173.99, p < 0.001$. A Student-Newman-Keuls (SNK) multiple comparisons test confirmed that performance in all conditions with the new talker was better than with the original talker and that performance in conditions 3, 6, 4 and 8 was better than in all other conditions ($p = 0.05$), although performance did not differ across this set of conditions ($p = 0.05$). There were significant interactions of age and context, $F(1,30) = 24.83, p < 0.001$, and age and talking condition, $F(7,210) = 3.260, p < 0.01$.

Figure 2 illustrates the difference in performance between the various speaking conditions with the reference condition being the new talker speaking with a ‘normal’ voice in quiet. Overall, the word identification performance of older adults was more affected by the talker condition than was the performance of younger adults. Compared to the reference condition, the differences due to talking condition were greatest for the original SPIN recording, followed by the new talker’s clear speech in quiet and his speech in speech spectrum noise at 62 dB SPL. Finally, the differences due to talking condition were more evident for words in low-context sentences than high-context sentences.

This description was confirmed by another ANOVA. There were main effects of age, $F(1,30) = 15.411, p < 0.001$, and talking condition, $F(6,180) = 123.66, p < 0.001$, as well as a significant interaction between talking condition and context, $F(6,180) = 32.76, p < 0.001$. SNK tests indicated that compared to the reference condition, scores were reduced the most for original SPIN recording, followed by clear speech in quiet and speech spoken in speech spectrum noise at 62 dB SPL, but the scores increased with an equivalent amount of improvement for all other conditions ($p = 0.01$).

Figure 1: Mean percent correctly identified target words, for 16 younger adults (YA) and 16 older adults (OA), in 8 talking conditions. Results are categorized into high- and low-context condition. Standard error bars are included.

4. Discussion

There were differences in speech intelligibility due to both between-talker and within-talker differences. Older adults were more affected than younger adults by both types of talker differences when listening to speech in noise. Thus, in everyday life, age-related differences in speech understanding would be attenuated by the Lombard Effect.

Since the sentence lists were controlled for overall level and duration, these finding seem more likely to be attributable to differences in the target word than to more global differences between the lists of sentences. The target words in different lists might differ critically in vowel contrast, or in duration, pitch, level, or a more general shift in energy across the spectrum.

4.1. Vowel contrasts between talkers

It has been shown that a talker’s vowel contrasts initially increase as noise increases, possibly in an attempt to maintain clear speech under adverse speaking conditions, but the contrasts decrease at the highest noise levels [13]. The $F_1$ and $F_2$ frequencies of the five vowels spoken in target words by the original talker and the new talker in the same words in List 1 were compared (Fig. 3). There is no obvious difference in vowel contrast between the talkers which might explain the difference in intelligibility between them. In fact, the two talkers resemble each other more closely than the average values of male talkers from two different studies producing the same vowels in isolated words.

4.2. Word-level differences between lists

Overall, compared to the target words spoken by the original talker in List 1 in quiet, the target words spoken by the new talker in List 2 in quiet were higher in pitch, longer in duration, and slightly more intense, which may partly explain why the speech of the new talker was more intelligible (Table 1). In particular, the new talker appears to have greater energy in the target words, especially in the 2-4 kHz and 5-6 kHz frequency ranges (Figure 4).

Figure 2: Differences in percent correct word identification in conditions 1 and 3-8, as compared to condition 2 (new talker speaking “normally” in quiet) for 16 younger adults (YA) and 16 older adults (OA). Results are categorized into high and low context conditions in each talking condition.
5. Conclusions

The results from this study support the findings by [8] that word identification improves when the speaking environment is matched with the listening environment. Small inter- and intra-talker differences in one or more acoustical dimensions may lead to a large difference in intelligibility. Older adults benefit more than younger adults from the Lombard Effect, for SPIN-R sentences presented at 0 dB S.N.

6. Acknowledgments

This project was funded by the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada. We wish to thank James Qi, for technical help.

7. References