The Influence of Linguistic Content on the Lombard Effect

Rupal Patel
Kevin W. Schell
Northeastern University, Boston

Purpose: The *Lombard effect* describes the tendency for speakers to increase pitch, intensity, and duration in the presence of noise. It is unclear whether these modifications are uniformly applied across all words within an utterance or whether information-bearing content words are further enhanced compared with function words. In the present study, the authors investigated the influence of linguistic content on acoustic modifications made to speech in noise.

Method: Sixteen speaker–listener pairs engaged in an interactive cooperative game in quiet, 60 dB of multitalker noise, and 90 dB of multitalker noise. Speaker productions were analyzed to examine differences in fundamental frequency (F_0), intensity, and duration of target words in sentences across noise conditions.

Results: Proportional increases in $F_{0,}$ intensity, and duration were noted for all word types as noise increased from quiet to 60 dB. From quiet to 90 dB, content words that referred to agents, objects, and locations were disproportionately elongated compared with function words. Additionally, agents were further enhanced by increased F_0.

Conclusions: At moderate noise levels, most word types appear to be uniformly boosted in F_0, intensity, and duration. As noise increases, linguistic content shapes the extent of the Lombard effect, with F_0 and duration serving as primary cues for marking information-bearing word types.

KEY WORDS: Lombard effect, prosody, acoustic modifications, speech in noise, linguistic content

To communicate effectively in noise, a speaker must modify the acoustic properties of the speech signal. Many of us have experienced the need to alter our speaking style in a crowded restaurant or on the subway. These modifications were initially described by Lombard (1911) as an increase in intensity mediated by an automatic regulating device. He argued that regulation of the speech signal resulted from a feedback loop that allowed the speaker to self-monitor. In other words, as background noise increases, a speaker must increase vocal intensity in order to monitor the speech signal. It is now agreed that additional acoustic properties of speech, including fundamental frequency (F_0), formant frequencies, and duration, are also altered to varying degrees depending on the competing noise type and level (Brown & Brandt, 1972; Junqua, 1993, 1996; Lane & Tranel, 1971; Letowski, Frank, & Caravella, 1993; Pittman & Wiley, 2001; Rivers & Rastatter, 1985; Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988).

One possible explanation for altering multiple acoustic properties of the speech signal is that the speaker is not merely maintaining the ability to self-monitor but is also attempting to optimize information transfer to his/her listener (Lane & Tranel, 1971). Lombard speech has been noted to improve speech intelligibility (Dreher & O'Neill, 1957) and increase

1092-4388/08/5101-0209

communicative effectiveness (Letowski et al., 1993) in the presence of noise. In contrast, shouting or "loud speech" reduces intelligibility because of distortion of the speech signal when the signal-to-noise ratio remains constant (Pickett, 1957).

In addition to improving intelligibility, the acoustic parameters altered during Lombard speech such as pitch, loudness, and duration (collectively referred to as *prosody*) may also be adjusted to convey linguistic stress and intention (Cutler, 1994; Rivers & Rastatter, 1985). To date, most investigations on Lombard speech have measured changes to the speech signal across various noise conditions (Letowski et al., 1993; Pittman & Wiley, 2001; Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988) but have not controlled for the linguistic content of the utterances produced. Although the entire utterance may need to be modified to some extent, it is not clear whether the degree of modifications required varies by word type. Given the observation that speakers alter their speech to convey their intention to listeners (cf. Dreher & O'Neill, 1957; Lane & Tranel, 1971; Summers et al., 1988), the Lombard effect may be enhanced for semantically salient content words compared with function words, which tend to be less informative. This finding would have clinical implications for the design of intervention strategies and technologies that optimize communication transfer in noise. For example, current speech synthesizers in assistive communication aids do not incorporate models of speech in noise and are thus ineffective in everyday noise environments. Similarly, clients who need to be heard and understood in noise would benefit from strategies that incorporate the notion of linguistic content on communication effectiveness.

Previous work suggests that the acoustic properties of Lombard speech may be influenced by various factors, including the nature of the speech task, the noise type, and the noise level. Most studies to date have relied on dictated speech samples or read speech consisting of single words, consonants in sentences, target words in carrier sentences, and nonsense sentences (Brown, Brandt, & John, 1972; Holmberg, Hillman, Perkell, & Gress, 1994; Junqua, 1996; Kalikow, Stevens, & Elliott, 1977; Pittman & Wiley, 2001; Speaks & Jerger, 1965; Summers et al., 1988). Lombard speech has been studied in varying types of noise, including broadband noise, traffic noise, white noise, pink noise, and, more recently, multitalker noise (Brown & Brandt, 1972; Dreher & O'Neill, 1957; Junqua, 1993; Letowski et al., 1993; Pickett, 1957; Pitman & Wiley, 2001; Rivers & Rastatter, 1985). Although these tasks provide experimental control across noise conditions, they lack the naturalness of spoken communication. Moreover, the magnitude of the Lombard effect appears to be "governed by the premium on intelligible communication" (Lane & Tranel, 1971, p. 682). On the one end are tasks such as reading word lists, which have little, if any, intelligibility premium compared with communicative scenarios, in which a speaker and listener engage in a cooperative task. Thus, if speakers are not engaged in a communicative interaction, intelligibility may not be a concern, and the acoustic changes yielded may not be comparable to those made in natural conversations. For example, Amazi and Garber (1982) noted that adult speakers increase vocal intensity to a greater degree for conversational speech than for single-word tasks.

In an attempt to analyze speech modifications in naturalistic communicative scenarios, Rivers and Rastatter (1985) studied the effect of multitalker noise on the production of stressed and non-stressed words in spontaneous speech. Speech stimuli consisted of picture cards taken from the Peabody Language Development Kit (Dunn, Horton, & Smith, 1968). The noise conditions included quiet, 90 dB of white noise, and 90 dB multitalker noise. Across noise conditions, the average F_0 for stressed words increased by 62 Hz from the quiet condition for both males and females while the average F_0 for non-stressed words increased by only 33 Hz for males and 25 Hz for females. The authors also noted that multitalker noise was more disruptive to speech than white noise resulting in a greater change in F_0.

While Rivers and Rastatter (1985) examined the effect of linguistic stress on F_0, Pittman and Wiley (2001) sought to identify intensity and durational changes to speech in quiet, 80-dB wideband noise, and 80-dB multitalker babble. Their stimuli, however, lacked the naturalness of Rivers and Rastatter's (1985) stimuli in that they used 50 low-predictability sentences taken from the Speech Produced in Noise (SPIN) test (Kalikow et al., 1977). They found an average increase in intensity of 14.5 dB and an average increase in duration of 77 ms from quiet to the noise conditions. These modifications represent average changes in intensity and duration across all words within the utterance. Thus, it remains unclear whether information-bearing content words are modified to a greater extent than function words. Table 1 summarizes related studies that have explored modifications to pitch, intensity, and/or duration in noise.

The present study elaborated on Rivers and Rastatter's (1985) premise of using pictorial stimuli to elicit spontaneous sentence-level productions. Although these authors studied the impact of noise on linguistic stress, in the present study we sought to determine whether acoustic modifications to speech produced in noise were uniformly applied to all words or whether content words were enhanced to a greater extent than function words. Presumably, content words carry a greater burden of the intelligibility premium than do function words and thus may be disproportionately boosted in noise in order to

Table 1. Summary of related studies that have explored modifications to pitch, intensity, and/or duration in noise.

Authors	Stimuli	Noise conditions (dB)	Microphone type; placement	F_0	Intensity	Duration
Brown & Brandt (1972)	Read sentences	Quiet, 87, 107	Condenser; not reported	—	4.2-dB average increase from quiet to 107-dB noise condition	—
Dreher & O'Neill (1957)	Read words and sentences	Quiet, 70, 80, 90, 100	Altec 21-C condenser; button behind the corner of the mouth out of the breath stream	—	Average increase of 9.1 dB for words and 6.1 dB for sentences between quiet and 100-dB condition	—
Junqua (1993)	Read single words	Quiet and 85-dB wide band noise (WBN)	Not reported	—	For females, 12.6-dB average increase; for males, 18.2-dB average increase between quiet and 85-dB WBN	—
Pick et al. (1989)	Spontaneous speech	Quiet and 90 dB SPL WBN	Plantronics MS 50; T61 attached to headphones 4.5 cm from lips out of breath stream	—	Average increase of 10 dB for the naïve group between quiet and 90-dB WBN	—
Pittman & Wiley (2000)	Read single words; 50 target words	Quiet and 80 dB SPL WBN and multitalkerbabble (MTB)	Shure SM10A; 1 in. from lips out of breath stream	—	14.5-dB average increase from quiet to 80 dB WBN and MTB condition	88-ms average increase in WBN; 66-ms average increase in MTB
Rivers & Rastatter (1985)	Spontaneous speech	Quiet and 90 dB SPL WBN, MTB	Electro Voice No. 423 A; 5 cm from the lips	Males: nonstress → average increase of 33 Hz between quiet and 90 dB MTB stress → average increase of 62 Hz between quiet and 90 dB MTB Females: nonstress → average increase of 25 Hz between quiet and 90 dB MTB stress → average increase of 62 Hz between quiet and 90 dB MTB	N/A	—

(Continued on the following page)

Table 1 *Continued.* Summary of related studies that have explored modifications to pitch, intensity, and/or duration in noise.

Authors	Stimuli	Noise conditions (dB)	Microphone type; placement	F_0	Intensity	Duration
Letowski et al. (1993)	Read speech ("Grandfather Passage")	70, 90 dB SPL WBN, traffic noise, MTB	ACO condenser; 12 in. in front of the lips	For females, average increase of 18 Hz between quiet and 90 dB For males, average increase of 28.5 Hz between quiet and 90 dB	7.4-dB increase	Reported no systematic changes in speech rate
Summers et al. (1988)	Read words (2 participants)	Quiet, 80, 90, 100 dB SPL	Electrovoice condenser microphone; 4 in. from the lips	Average increase of 16.1 Hz between quiet and 100 dB noise	Average increase of 6.9 dB between quiet and 100 dB noise	Average increase of 101.5 ms between quiet and 100 dB noise

optimize communication. Furthermore, if linguistic content influenced the extent of the acoustic modifications, which prosodic cues were used to convey these contrasts?

Method

Participants

Sixteen adult (8 male, 8 female speakers; M = 22.25 years of age) monolingual speakers of American English participated in the study. All participants had adequate or corrected visual function and no known history of speech and language disorders based on self-report. All participants passed a pure-tone hearing screening with thresholds at or below 25 dB in at least one ear at 250, 500, 1000, 2000, and 4000 Hz.

Materials and Apparatus

An interactive computer game was designed to elicit relatively spontaneous speech with minimal cueing while maintaining experimental control over spoken utterances. Each speaker and listener pair engaged in a cooperative game presented on two computer monitors located in two separate rooms. The speaker communicated with the listener via a headset microphone (Shure SM-10A). Multitalker noise (Auditec, St. Louis, MO) was routed from a portable CD player (Sony CFD-E75), calibrated through an audiometer (Grason-Stadler GSI-16), and presented to the speaker via supra-aural headphones (Telephonics TDH-50P). The listener heard the multitalker noise through built-in audiometer monitors (GSI-16) and the speaker's speech signal through a separate monitor system (Altec-Lansing ACS90). Multitalker noise and speech output were measured and calibrated using a sound-level meter (Quest Technologies, Model 1700) positioned at the listener's ear to maintain consistency across sessions. The intensity of the noise delivered via speakers was carefully monitored, and the highest noise level did not exceed 90 dB SPL at the participant's ear.

Procedure

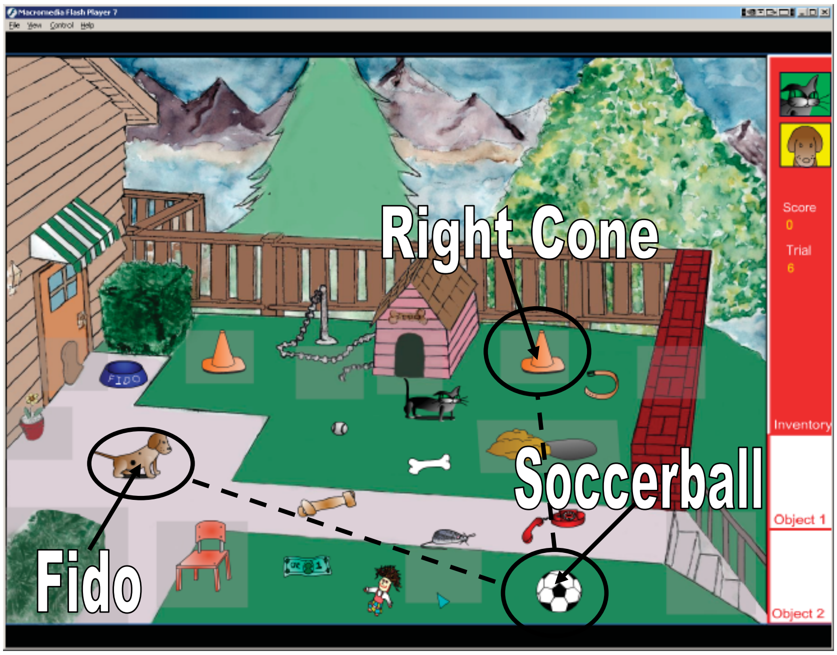
The speaker–listener pair occupied separate rooms of a sound-treated audiometric booth for the duration of the experiment, and each member of the pair was not visible to the other. This arrangement was chosen because it required the listener to rely only on acoustic cues to decode the speaker's commands. One research assistant was present in each room of the sound-treated booth for the duration of all sessions in order to answer questions and provide further instruction, if necessary.

The goal of the game was for the speaker to effectively instruct the listener to perform a series of actions with animated characters and objects on the screen. The game consisted of two animated agents (Fido the dog; Silo the cat) performing a number of actions with one or more objects. The scene consisted of 10 objects that could be moved to 1 of 12 different locations. To control for type of stimuli across noise conditions, the speaker first viewed an animation. Each animation consisted of an agent (e.g., Fido) picking up an object (e.g., soccer ball) and placing it at a location (e.g., the right cone). After viewing the animation, the speaker instructed the listener to perform the viewed action (e.g., "Make Fido pick up the soccer ball and put it by the right cone"). Figure 1 provides a screen shot of the game interface. Speakers were required to communicate at least these essential elements of the action frame in order to elicit the appropriate response from the listener. The listener's response was considered correct if it reflected accurate understanding of the agent, the object, and the general location, which was defined by a translucent bounding box. Thus, accurate understanding of the preposition was irrelevant in scoring (e.g., the listener putting the soccer ball *by* the right cone or putting it *on* the right cone were scored equivalently). After issuing the command, the speaker could view the listener's actions in real time in order to offer corrective feedback, if necessary. If the listener did not follow the directions accurately, speakers were instructed to repeat the entire utterance. Although repeated utterances were necessary to establish conversational naturalness, they were not included in the present analysis due to sparse and insufficient data.[1]

To study the impact of linguistic content on the acoustic modifications, we established operational definitions for content and function words. For the given task, agents, objects, object modifiers, locations, location modifiers, and verbs—words that either conveyed contrastive meaning or were essential for eliciting the correct listener action—served as content word types. In some cases, compound words or word pairs were considered as a single linguistic unit. For example, if a speaker used an action phrase such as "pick up," it was coded as a single verb unit given that the speaker's intent was to convey a single action. Words that would carry little lexical meaning in the present context—such as articles, prepositions, pronouns, and conjunctions—were considered *function words* (e.g., "of," "the," "a," "and," etc.). These words were grouped together given that no specific subset of function words was hypothesized to be particularly information-bearing in the present task. Thirty utterances were elicited in each of the three noise conditions. The stimuli were randomized and counterbalanced between the two animated characters across all noise conditions. Table 2 provides a list of words

[1] Only 27 repairs were observed in a total of 1,440 trials produced across all 15 speaker–listener pairs. In fact, some speakers did not produce any repair tokens at all.

Figure 1. Screenshot of the interactive game in which the speaker instructs the listener to "Make Fido pick up the soccer ball and put it by the right cone."



within each word type that were used to construct the target stimuli.

Participants first received instruction and orientation to the game consisting of a demonstration phase in quiet (≤40 dB SPL). This ensured that each participant understood the experimental task. Following the demonstration, three phases of 30 trials each were completed, with a short break between phases. During each phase, the speaker and listener were exposed to one of three noise levels (quiet, 60 dB multitalker noise, and 90 dB multitalker noise). The order of the quiet, 60-dB noise condition, and 90-dB noise condition was counterbalanced across speakers. To control for practice effects, the order of the 30 target stimuli were randomized across phases for each speaker. At the conclusion of the experiment, a hearing screening was re-administered as a safety precaution to ensure that exposure to noise during the experiment was not harmful. Across all participants, there were no significant hearing threshold shifts pre- and postexperiment.

Acoustic Analysis

A total of 1,440 trials (30 Trials × 3 Noise Conditions × 16 Speakers) were analyzed to examine acoustic modifications in F_0, intensity, and duration to the various word

Table 2. Content words used to construct target stimuli and possible function words.

Agent	Object	Object modifier	Location	Location modifier	Verbs	Function words
Fido	Baseball	Big	Chain	Left	Bring	A
Silo	Bone	Small	Chair	Right	Have	By
	Bowl		Cone		Make	If
	Collar		Doghouse		Pick up	In
	Doll		Hedge		Put	It
	Dollar		Hole			Of
	Mouse		House			On
	Phone		Lawn			The
	Soccer ball		Pot			And
			Stairs			
			Wall			

types across noise conditions. Audio recordings were made directly to a computer at a sampling rate of 22050 Hz. In each noise condition, 30 trials were elicited. Recordings from each speaker were first segmented into 90 individual trials using Adobe Audition software. The Praat speech analysis software package (Boersma & Weenik, 2000) was then used to manually label the beginning and end of individual words within each trial.

Once all the files were labeled, the Praat system was used to generate a series of time-stamped relative intensity values (dB) and frequency values (Hz) across the duration of each syllable within each phrase. A customized software program was then used to operate on the Praat-generated pitch and intensity values in order to calculate the following acoustic features: peak F_0, peak intensity, and word duration.[2] *Peak F_0* was defined as the highest nominal frequency point within a labeled selection. Similarly, *peak intensity* was defined as the highest nominal amplitude point within a labeled selection. *Word duration* was calculated as the length of the labeled selection.

Manual correction of the automatically generated F_0 values was required on 152 of the 1,440 recorded samples because the pitch-tracking algorithm reported octave jumps or pitch breaks that could not be verified auditorily. Manually adjusting the upper and lower F_0 limits and frame duration parameters in Praat typically led to improved F_0 tracking. These new F_0 values were verified through visual and auditory inspection and were confirmed using direct calculation of the pitch period from the waveform. When Praat-derived F_0 values continued to be judged as errors (this occurred in 32 productions), these values were replaced by manually derived values obtained from the waveform.

Interjudge Reliability

Interjudge reliability for labeling the beginning and end of each word was calculated for a random sample of 10% of the utterances. The reliability between labelers for syllable duration was $r = .989$ ($M = 0.008$ s, $SD = 0.001$ s). Based on the relabeled duration values, we recomputed all intensity and F_0 values for this sample. The mean difference between the first and second measurement was 2.1 Hz ($SD = 0.9$ Hz) for peak F_0 and 0.71 dB ($SD = 0.4$ dB) for peak intensity.

Results

A mixed linear model was used to examine the effects of noise type (60 dB, 90 dB) and word type (function words, verbs, agent, object, location, object modifiers, and location modifiers) on the change in a given acoustic variable from its baseline value (i.e., in quiet) for each given speaker. Noise type, word type, and the baseline value (for each word type) were within-subject factors in each model. We also examined the effect of one between-subjects factor—gender (male/female)—on the acoustic changes. The same model was used for each of the three acoustic variables: peak F_0, peak intensity, and syllable duration. The response variables were all continuous: fundamental frequency in Hz, intensity in dB, and duration in seconds. We used a within-subject correlation structure of compound symmetry in each model. This structure provided the lowest value of Akaike's information criterion for the models that were estimated. We used the F statistic to test the null hypothesis with $\alpha = .05$.

Considering each acoustic parameter separately, pairwise contrasts were conducted on the conditions of interest. Specifically, the mean of function words (referred to as *functors* in Tables 3 and 4 and Figures 2, 3, and 4) was contrasted with the agent, object, location, and the mean of the verbs, object modifiers, and location modifiers. These contrasts were performed for the change in acoustic feature from baseline (quiet) to 60 dB and from baseline to 90 dB (see Tables 3 and 4). Thus, in total, 12 contrasts were carried out for each acoustic variable. The t statistic was used to test the contrast between word types. To account for multiple comparisons, the Bonferroni correction factor was used to adjust the alpha level to .004. All analyses were performed using SAS Version 9.1 using the Proc Mixed procedure.

Syllable Duration

Statistically significant main effects in syllable duration were found for word type, $F(6, 90) = 15.49$, $p < .0001$, and noise type, $F(1, 15) = 75.81$, $p < .0001$. The two-way Word Type × Noise Type interaction, $F(6, 90) = 10.13$, $p < .0001$, was also statistically significant. For all word types, male and female speakers did not differ in how they changed syllable duration from baseline to the two noise conditions. Figure 2 illustrates the change in duration for each word type and noise type across all speakers. On average, syllable duration increased by 17.2 ms from baseline to 60 dB, compared with 56.9 ms from baseline to 90 dB. The change in syllable duration across noise types was more pronounced for some word types than others. For example, syllable duration for the location increased by 134.1 ms from baseline to 90 dB, whereas it only increased by 7 ms for function words. Although the location was 52.2 ms longer than function

[2] Initial analyses were conducted on average and peak measures of F_0 and intensity. Peak and average measures for both features yielded the same statistical differences. Since both measures provided redundant results, we chose to report on peak measures because they appeared to be more representative of the acoustic modifications evident in the data set and thus provided the clearest results. Averaging over average values diluted the absolute differences between word classes, and average values also tended to be less representative in cases where there were multisyllabic or compound words that had strong and weak syllables.

Table 3. Contrasts between function words and content word types that were significant at the $p < .004$ (indicated by *) level for each acoustic feature from baseline to the 60-dB noise condition.

Variable	Functors vs. verbs	Functors vs. agent	Functors vs. object	Functors vs. location	Functors vs. object modifiers	Functors vs. location modifiers
Syllable duration						
Peak F_0						
Peak intensity					*	*

words from baseline to 60 dB, this change in duration was not statistically significant ($p = .0082$) at the adjusted alpha level. In contrast, the change in syllable duration from baseline to 90 dB for function words was 114.4 ms shorter than for the agent ($p < .0001$), 79.1 ms shorter than for the object ($p = .0003$), and 127 ms shorter than for the location ($p < .0001$).

Peak F_0

Statistically significant main effects in the change in peak F_0 from baseline to each noise condition were found for word type, $F(6, 90) = 31.70$, $p < .0001$, and noise type, $F(1, 15) = 286.50$, $p < .0001$. A statistically significant Word Type × Noise Type interaction was found, $F(6, 90) = 11.10$, $p < .0001$. Gender did not impact the degree of change in peak F_0 from baseline to the two noise conditions across word types. See Figure 3 for the average change in peak F_0 by word type and noise type. From baseline to 60 dB, the mean change in peak F_0 across all word types was 14.5 Hz, compared with 54.7 Hz from baseline to 90 dB. The change in peak F_0 varied by word type but only for the baseline to 90 dB condition. Statistically significant differences of change in peak F_0 were found between function words and the agent ($p = .0004$), object modifiers ($p < .0001$), and location modifiers ($p < .0001$). Across all speakers, the change in peak F_0 for function words was 19.7 Hz lower than the agent but 44.8 Hz higher than object modifiers and 54.7 Hz higher than location modifiers.

Peak Intensity

Statistically significant main effects in the change in peak intensity from baseline to each noise condition were found for word type, $F(6, 90) = 159.84$, $p < .0001$, and noise type, $F(1, 15) = 637.50$, $p < .0001$. The two-way Word Type × Noise Type interaction, $F(6, 90) = 26.87$, $p < .0001$, was also statistically significant. Once again, gender did not impact the degree of change in peak intensity from baseline to the two noise conditions across word types. Figure 4 shows average changes in peak intensity across speakers by word type and noise type. From baseline to 60 dB, the mean change in peak intensity across word types was 6.9 dB compared with 15.7 dB from baseline to 90 dB. The change in peak intensity from baseline to the two noise conditions differed by word type. From baseline to 60 dB, statistically significant differences were found between function words and object modifiers ($p < .0001$) and between function words and location modifiers ($p = .0002$). The change in peak intensity for function words was 6.9 dB louder than object modifiers and 7.1 dB louder than location modifiers. Similarly, from baseline to 90 dB, statistically significant differences of change in peak intensity were only noted between function words and object modifiers ($p < .0001$) as well as location modifiers ($p < .0001$). The change in peak intensity for function words was 17.4 dB louder than location modifiers and 15.9 dB louder than object modifiers.

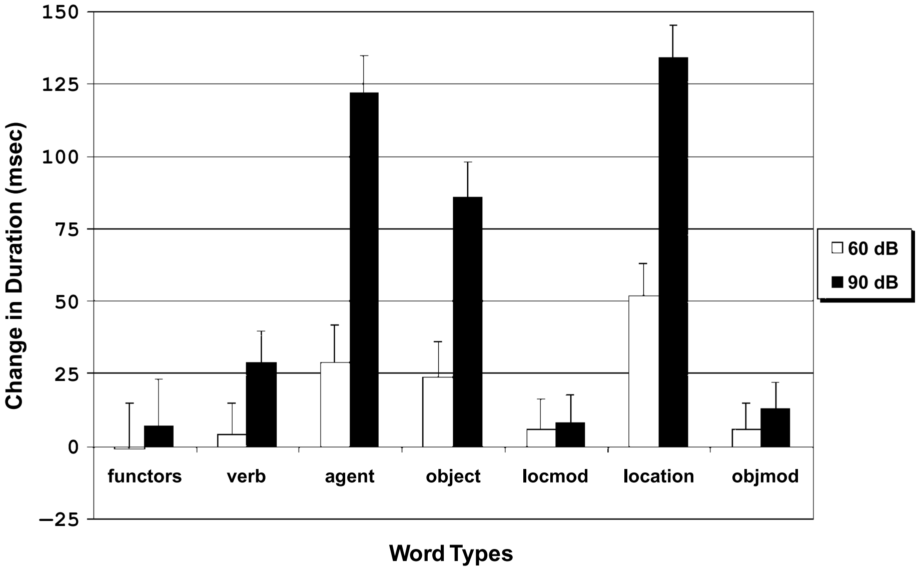
Discussion

Speech produced in noise is characterized by increases in F_0, intensity, and duration, collectively referred to as the *Lombard effect* (Brown & Brandt, 1972; Junqua, 1993, 1996; Lane & Tranel, 1971; Letowski et al., 1993; Lombard, 1911; Pittman & Wiley, 2001; Rivers & Rastatter, 1985; Summers et al., 1988). The present study

Table 4. Contrasts between function words and content word types that were significant at the $p < .004$ (indicated by *) level for each acoustic feature from baseline to the 90-dB noise condition.

Variable	Functors vs. verbs	Functors vs. agent	Functors vs. object	Functors vs. location	Functors vs. object modifiers	Functors vs. location modifiers
Syllable duration		*	*	*		
Peak F_0		*			*	*
Peak intensity					*	*

Figure 2. Change in duration (ms) from quiet to 60 dB and 90 dB noise across each word type. Locmod = location modifier; objmod = object modifier.



sought to determine whether the acoustic changes associated with the Lombard effect would be influenced by linguistic content in that task-based information-bearing content words would be enhanced to a greater degree than function words. Similar to previous studies, results of the present investigation indicated that all three acoustic measures (F_0, intensity, and duration) increased as the noise level increased. A comparison between the quiet and 60-dB noise conditions revealed a relatively proportional increase in all three acoustic features across most word types. In other words, the acoustic profile of content words and function words was maintained but merely shifted upward with increases in noise. Given that participants almost always used either "bring" or "pick up" as the verb term in all utterances, the extent of increase in F_0, intensity, and duration was limited compared with other content word types. Additionally, object modifiers and location modifiers were not raised to the

Figure 3. Change in peak fundamental frequency (Hz) from quiet to 60 dB and 90 dB noise across each word type.

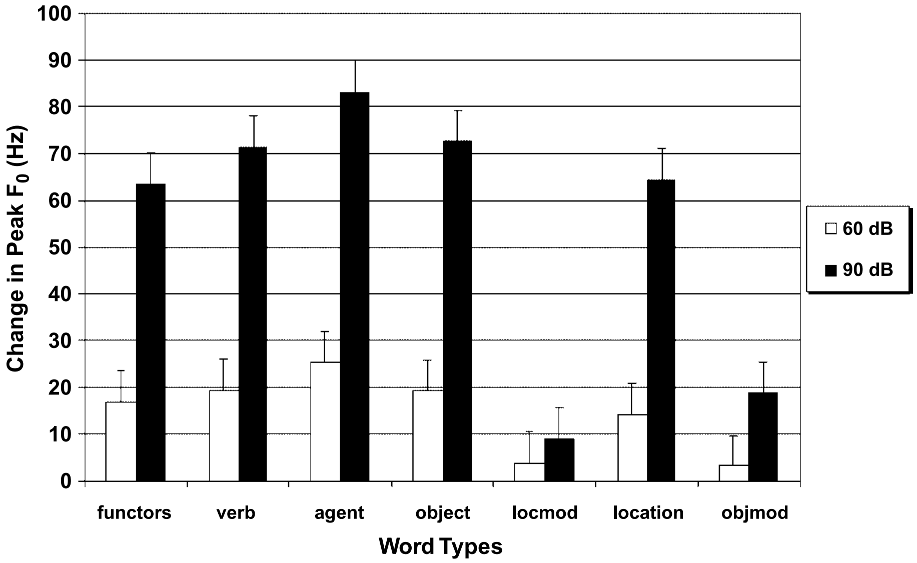
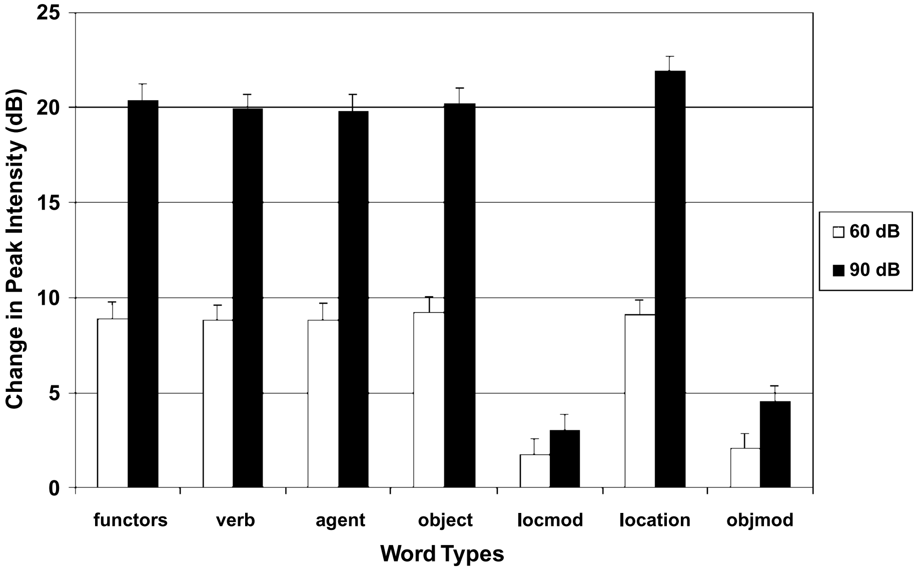


Figure 4. Change in peak intensity (dB) from quiet to 60 dB and 90 dB noise across each word type.



same extent as the other content word types. In fact, increases in F_0, intensity, and duration were greater for function words compared with these modifier terms. One explanation may lie in the fact that the present task required conveying spatial relationships, and thus prepositional words such as "in" and "on," which are conventionally thought of as function words, may in fact have been information-bearing in the given context. Although target objects only needed to be placed in the vicinity of the target location (indicated by a shaded bounding box; see Figure 1), some speakers preferred to specify the exact placement using prepositional terms. Additionally, most speakers established a template that was used to instruct the listener throughout the session (i.e., "Make *agent verb* the *object modifier object* and put it in the *location modifier location*"). Given the small number of objects and possible locations in the scene, object and location modifiers were often redundant, and thus their information content was relatively minimal. Future extensions of this work may benefit from subdivision of the function words into categories such as prepositions, articles, and conjunctions coupled with appropriately designed new task scenarios in order to study the impact of noise on these word classes.

A major methodological difference between the present study and previous work relates to the naturalness of the speaking task. Previous work has relied on read speech (cf. Brown et. al., 1972; Dreher & O'Neill, 1957; Kalikow et al., 1977; Speaks & Jerger, 1965; Summers et al., 1988) or on spontaneous speech production elicited using picture stimuli (Rivers & Rastatter, 1985). The present study aimed to optimize the tradeoffs between experimental control and naturalness by designing anquired to communicate to listeners about a shared visual context.

As the listening environment became more adverse in the 90-dB condition, linguistic content appeared to play a greater role in influencing the acoustic modifications. For example, content words that presumably carried a significant burden of the linguistic message such as agents, objects, and locations were elongated to a greater degree than less informative content words and function words. Because animations differed primarily in terms of the agent, object, and location, as long as the speaker indicated these essential elements of the action frame, the listener could follow the instruction accurately.

Interestingly, the agent word type was further highlighted by an increase in peak F_0 between the quiet and 90-dB condition. Other word types were not enhanced to the same extent as the agent. In fact, the F_0 of modifier terms was decreased in relation to function words as noise level increased. Average increases in F_0 found in the present study are consistent with the changes in F_0 of stressed and nonstressed words recorded by Rivers and Rastatter (1985).

In terms of intensity, we found an 18-dB increase in average peak intensity for males and females between the quiet and 90-dB conditions. Although these findings are consistent with some previous studies (Junqua, 1993; Pittman & Wiley, 2001), read speech has been noted to have resulted in smaller increases (3–11 dB) in intensity between quiet and 90-dB noise conditions (Brown & Brandt, 1972; Dreher & O'Neill; Letowski et al., 1993; Pick, Siegal, Fox, Garber, & Jearney, 1989; Summers, et al., 1988). The difference noted between read and more natural speech suggests that perhaps speakers can ignore background noise when they are reading words compared with when they are engaged in a natural task (Lane & Tranel, 1971). The present findings suggest that the intensity contour across content and function word types within a sentence was maintained but shifted upward by as much as 19 dB as noise level increased. Perhaps the physiological changes required to increase intensity cannot be easily manipulated on a word-by-word basis in connected speech. Increases in intensity result from adduction of the vocal folds, leading to a buildup of subglottal pressure (Kent & Read, 2002). This change in subglottal pressure is gradual and may not be easily modified within an utterance (Winkworth & Davis, 1997).

Findings of the present study suggest that duration is enhanced further for information-bearing content words compared with less informative content words and function words in the loud noise condition. In other words, speakers prolong the duration of content words to a greater degree in noise compared with the relative change for function words. The average increase in duration between quiet and 90 dB of 56.9 ms found in the present study is consistent with Pittman & Wiley (2001), who used a read speech task. Summers et al. (1988), however—who also used a read speech task—reported an average increase of 101.5 ms. The contrasting results of these two studies using similar speech tasks suggests that read speech may significantly affect the temporal parameter of speech produced in noise, thus leading to results that are not representative of natural speech.

Limitations and Future Directions

Although the design of this study allowed for the elicitation of relatively spontaneous speech, the task was highly controlled and repetitive. The speaking task differed from conversational speech in that speakers tended to follow a template sentence structure for each spoken command across all noise conditions. Additionally, some speakers often formulated their instructions as they watched the animation resulting in prolonged pauses and exaggerated elongation of function words. Future work may explore methodologies to elicit increasingly natural productions with diverse sentence types in order to yield more generalizable results.

Little to no change in the dependent measures was noted between the quiet and 60-dB conditions for function versus content word types. Perhaps the quiet and 90-dB conditions would be sufficient to study the influence of linguistic content on acoustic modifications in noise. Furthermore, in addition to analyzing F_0, intensity, and duration, future studies may also examine additional acoustic features such as spectral tilt (Junqua, 1996; Pittman & Wiley, 2001) as well as consider alternate aggregate statistics such as the mean and slope of F_0 and intensity across words within an utterance.

The results of the present study may have been somewhat influenced by the operational definitions used to group content and function word types. It is plausible that some function words were, in fact, highly informative for the given task. For example, prepositions such as "in" and "on" may have carried greater linguistic significance in the present spatial task than they typically would in natural spoken conversation. Similarly, although modifiers are typically thought of as content words, they carried little information in the present task. Future studies would need to consider refining the operational definitions of content and function words by taking into account the information content of the word, its grammatical role, and the situational context of the communication task. Additionally, it may be fruitful to analyze acoustic modifications to specific function word categories. For example, animations could be designed to introduce ambiguities that require the speaker to specify a given article or preposition. In fact, future extensions may consider designing various animation scenarios that systematically assess whether all word classes undergo similar acoustic modifications when they play a contextually salient role in the task.

Last, given that the current paradigm did not yield a sufficient number of communication repairs, future studies may be designed to deliberately elicit repair tokens by introducing more adverse listening conditions or by including a confederate listener. These modifications would help tease apart interactions between intelligibility and linguistically salient Lombard modifications.

A potential clinical application of the present findings would be to incorporate the notion of linguistic content into intervention strategies for clients who have difficulty being heard and understood in noisy environments. For example, speakers with Parkinson's disease may benefit from targeted practice in noise simulations, in which they are instructed to emphasize (using intensity, F_0, and durational cues) information-bearing content words to a greater degree than less informative words. This strategy may require less effort and thus may be more efficient.

Our findings also have implications for developing natural spoken dialogue systems. Current speech synthesizers do not incorporate models of speech in noise and are thus ineffective in noisy environments. Speech synthesizers are purely static output devices that are unaware of their acoustic environment. Simply turning up the volume and playing the output of the synthesizer louder often leads to distortion and further degradation of intelligibility. This has serious consequences for individuals with severe communication disorders who must rely on speech synthesis to speak on their behalf. On the basis of our empirical data, we have begun an initial effort to design an adaptive speech synthesizer that listens to the ambient noise level and modifies its speaking style to compensate for background noise in human-like ways (Patel, Everett, & Sadikov, 2006).

Acknowledgments

We would like to express our gratitude to Elyes Yaich for developing the interactive game interface, Howard Cabral for his assistance with the statistical analysis, and the participants for their time. This research was supported, in part, by Grant 0509935 from the National Science Foundation and by the Northeastern University Research Development and Scholarship Fund.

References

Amazi, D. K., & Garber, S. R. (1982). The Lombard sign as a function of age and task. *Journal of Speech and Hearing Research, 25,* 581–585.

Boersma, P., & Weenik, D. (2000). Praat: A system for doing phonetics by computer, Version 3.4. *Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam*. Retrieved July 27, 2005, from www.praat.org

Brown, W. S., & Brandt, J. F. (1972). The effect of masking on vocal intensity during vocal and whispered speech. *Journal of Auditory Research, 12,* 157–161.

Brown, W. S., Brandt, J. F., & John, F. (1972). The effect of masking on vocal intensity during vocal and whispered speech. *Journal of Auditory Research, 12,* 157–161.

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua, 92,* 81–104.

Dreher, J. J., & O'Neill, J. J. (1957). Effects of ambient noise on speaker intelligibility for words and phrases. *The Journal of the Acoustical Society of America, 29,* 1320–1323.

Holmberg, E. B., Hillman, R. E., Perkell, J. S., & Gress, C. (1994). Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in SPL across repeated recordings. *Journal of Speech and Hearing Research, 37,* 484–495.

Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America, 93,* 510–524.

Junqua, J. C. (1996). The influence of acoustics on speech production: A noise induced stress phenomenon known as the Lombard reflex. *Speech Communication, 20,* 13–22.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61,* 1337–1351.

Kent, R. D., & Read, C. (2002). *Acoustic analysis of speech* (2nd ed.). New York: Singular.

Lane, H. L., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research, 14,* 677–709.

Letowski, T., Frank, T., & Caravella, J. (1993). Acoustical properties of speech produced in noise through supra-aural headphones. *Ear and Hearing, 14,* 332–338.

Lombard, E. (1911). Le signe de l'élévation de la voix [The sign of the rise in the voice]. *Maladies Oreille, Larynx, Nez, Pharynx, 27,* 101–119.

Patel, R., Everett, E., & Sadikov, E. (2006). Loudmouth: Modifying text-to-speech synthesizer in noise. In *Proceedings of Association for Computing Machinery, SIGACCESS Conference on Computers & Accessibility* (pp. 227–229). New York: Association for Computing Machinery.

Pick, H. L., Siegal, G. M., Fox, P. M., Garber, S. R., & Jearney, J. K. (1989). Inhibiting the Lombard effect. *The Journal of the Acoustical Society of America, 85,* 894–900.

Pickett, J. M. (1957). Perception of vowels heard in noises of various spectra. *The Journal of the Acoustical Society of America, 29,* 613–620.

Pittman, A. L., & Wiley, T. L. (2001). Recognition of speech produced in noise. *Journal of Speech, Language, and Hearing Research, 44,* 487–496.

Rivers, C., & Rastatter, M. P. (1985). The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults. *The Journal of Auditory Research, 25,* 37–45.

Speaks, C., & Jerger, J. (1965). Method for measurement of speech identification. *Journal of Speech and Hearing Research, 44,* 487–496.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analysis. *The Journal of the Acoustical Society of America, 84,* 486–490.

Winkworth, A. L., & Davis, P. J. (1997). Speech breathing and the Lombard effect. *Journal of Speech, Language, and Hearing Research, 40,* 159–169.

Received November 2, 2006

Accepted June 28, 2007

DOI: 10.1044/1092-4388(2008/016)

Contact author: Rupal Patel, Department of Speech Language Pathology and Audiology, Northeastern University,
360 Huntington Ave., Room 102 FR, Boston, MA 02115.
E-mail: r.patel@neu.edu.