Available online at www.sciencedirect.com
SCIENCE DIRECT
Speech Communication 48 (2006) 1079–1093
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation

Felicia Roberts [a,*], Alexander L. Francis [b], Melanie Morgan [a]

[a] Department of Communication, Purdue University, Beering Hall, Room 2114, West Lafayette, IN 47907, USA
[b] Department of Speech, Language, and Hearing Science, Heavilon Hall, Room B-11, West Lafayette, IN 47907, USA

Received 16 January 2005; received in revised form 31 January 2006; accepted 1 February 2006

Abstract

The forms, functions, and organization of sounds and utterances are generally the focus of speech communication research; little is known, however, about how the silence between speaker turns shades the meaning of the surrounding talk. We use an experimental protocol to test whether listeners' perception of trouble in interaction (e.g., disagreement or unwillingness) varies when prosodic cues are manipulated in the context of 2 speech acts (requests and assessments). The prosodic cues investigated were inter-turn silence and the duration, absolute pitch, and pitch contour of affirmative response tokens ("yeah" and "sure") that followed the inter-turn silence. Study participants evaluated spoken dialogues simulating telephone calls between friends in which the length of silence following a request/assessment (i.e., the inter-turn silence) was manipulated in Praat as were prosodic features of the responses. Results indicate that with each incremental increase in pause duration (0–600–1200 ms) listeners perceived increasingly less willingness to comply with requests and increasingly weaker agreement with assessments. Inter-turn silence and duration of response token proved to be stronger cues to unwillingness and disagreement than did the response token's pitch characteristics. However, listeners tend to perceive response token duration as a cue to "trouble" when inter-turn silence cues were, apparently, ambiguous (less than 1 s).

Keywords: Silence; Prosody; Pausing; Human conversation; Word duration

1. Introduction

Speech Communication research tends to address the forms, functions, and organization of sounds and utterances. Little is known, however, about how the silence between speaker turns shades the meaning of the surrounding talk. Systematic exploration of the spaces between utterances can provide insight into this fundamental organizing principle of human interaction and, conceivably, could then be integrated into speech processing research. The current study, which is focused on perceptions of inter-turn silence in particular prosodic contexts, aims to contribute in this direction. Hirschberg (2002) concludes that issues of understanding (for both humans and machines) can be most "naturally"

* Corresponding author. Tel.: +1 765 494 3323; fax: +1 765 496 1394.
E-mail addresses: froberts@purdue.edu (F. Roberts), francisa@purdue.edu (A.L. Francis), melanie.morgan@cla.purdue.edu (M. Morgan).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.02.001

solved by recourse to intonational resources (p. 39). We concur with that sentiment and push our understanding a step further by examining cues at the intersection of silence and intonation—a dynamic space from which human listeners derive meaning.

The model of conversation we use to develop our approach is well known and widely accepted among communication researchers.[1] Based on empirical study of tape recordings of naturally occurring interaction, Sacks et al. (1974) outline a model that captures the intuition that silence is a great deal more than an absence of speech. Findings from qualitative conversation analytic research have demonstrated that when silence follows certain speech acts (e.g. invitations or requests) it is indicative of possible trouble in the interaction (e.g. invitation about to be declined or request denied) (Davidson, 1984; Pomerantz, 1984). This "trouble" surrounding silence is empirically available through speakers' routine practice of producing "subsequent versions" (Davidson, 1984, p. 104) of their requests or invitations following silence. The following is a simplified transcription of an actual conversation in which such an activity occurs (based on Davidson, 1984, p. 104)

A: Well did you want me to just pick you- get into Robinson's so you could buy a little pair of slippers?
(silence)
A: I mean or can I get you something?

Essentially, the silence following the proposal or request occasions the speaker's review of prior talk to find some way to make it more clear, understandable, or perhaps acceptable to the listener. With this and other examples, Davidson (1984) demonstrates how speakers faced with silence at a place where response is expected reformulate their talk to display that they understand their hearer to be reluctant, not hearing, or for some other reason slow to respond. This detection of trouble by the speaker is thus attuned to what is *not* said, to the lack of uptake after particular types of speech acts which require response. We submit that this approach to detection of trouble in terms of linguistic properties, including discourse acts, fits with the proposal by Batliner et al. (2003) to move toward annotation of (formal) linguistic properties which "can be used ... in combination with other knowledge sources to find *trouble in communication*" (p. 118). Silence, we propose, is within the realm of formal linguistic cues, but we maintain as well that silence must be understood in the context of surrounding talk, both in terms of social actions occasioned by preceding talk and prosodic features of that talk.

Previous research suggests that a silence duration of approximately 1 s is oriented to by speakers as troubles-indicative (Jefferson, 1989). In an exhaustive examination of 289 pages of transcription (in which silences had been subjectively timed), Jefferson (1989) identified 170 interactional sequences where the silences indicated some problematic moment. Sixty-two percent of these troubles-indicative silences ($n = 106$) were between 900 and 1200 ms. Silences longer than that were generally filled in some way (i.e., with non-verbal activity such as scanning documents or writing something down).

Although evidence suggests 1 s as a rough metric, it is also true that troubles-indicative silences occur in all lengths in natural conversation. In fact, Davidson's (1984) research concerning reformulation of a speech act following silence has instances of much smaller gaps. Although these were timed subjectively, relative to the speed of the surrounding talk, it is nonetheless clear that silence is a sequentially relevant and dynamic phenomenon in conversation.

This insight has been examined from a cognitive perspective by researchers working within the "Feeling of Knowing" (FOK) paradigm (Hart, 1965; Smith and Clark, 1993) or the "Feeling of Another's Knowing" (FOAK) paradigm (as identified by Brennan and Williams, 1995). These approaches have examined delaying of response (as well as other prosodic cues, fillers, and lexical hedges) in terms of the production and detection of *uncertainty*. Latency to respond is conceptualized as an inability to find an answer (Glucksberg and Mccloskey, 1981); the outward appearance of this searching/monitoring process (Nelson, 1993) is thus characterized as "uncertainty." Swerts and Krahmer (2005) take this research a step further by combining audio and visual cues so that facial expressions are also examined as signaling uncertainty.

Whether audio or audiovisual cues are used, the common thread in the FOK and FOAK literature is the use of factual questions to elicit answers which

[1] This assertion is based on Roberts and Robinson (2004). They note that conversation analysis (CA) has gained the attention of communication scholars in the last decade as evidenced by publication of CA studies in the leading communication journals and by the methodological debates that the approach has engendered.

are then used as stimuli in examinations of uncertainty. What these efforts have yet to make explicit is that uncertainty must be more broadly construed; affective orientations such as "unwilling" or "uninterested" or "not in agreement" are interpersonally relevant shadings of uncertainty, and could become manifest in the context of speech acts where requests and opinions are at stake. The uncertainty in these cases is not necessarily about "not knowing" or "unable to retrieve" but instead concerns something more ambiguous. Thus, previous research cannot help us understand the role of response latency in the context of utterances other than those which embody factual searches.

Furthermore, when delay has been examined, the pauses are categorized subjectively ("long" and "short" by Brennan and Williams, 1995, p. 389; "present" or "absent" in Swerts and Krahmer, 2005, p. 84). What this means is that delay is being treated as a general cue, which fits with insights from early descriptive studies, but the lack of precision in measurements makes it hard to compare findings across studies or to apply findings in environments where greater exactness is required (for example, in speech processing domains).

Despite the general sense that latency of response is related to uncertainty or even deception (see Anderson, 1999, for an overview) it has also been shown to signal thoughtfulness (Burgoon et al., 1995). Thus as Levinson (1983) points out, it is unwise to view silence as meaning only one thing. Certainly, as described by Davidson (1984) and Pomerantz (1984), silence takes on meaning in the context of the structure of the interaction so far. Therefore, the current study examines the shape of responses following the silence, where "shape" means prosodic characteristics of the response token. Our concern is to see if the shape of the response token in any way attenuates or accentuates the perceptions engendered by the inter-turn silence.

In sum, there has been no systematic experimental study to test whether listeners' perception of "trouble" in an interaction varies with length of silences, with type of speech act, and with prosody of responsive gestures. Of additional concern for the design of more naturalistic interactive machines, we do not have the kind of precise measurements of inter-turn silence that scientists working on speech processing would find reliable and relevant. While the subjective approach to timing inter-turn silences is wholly adequate for purposes of understanding sequences of social actions, more precise knowledge of relevant silence lengths may help bridge this particular gap between the work of discourse analysts and that of speech scientists.

Thus, the purpose of the present study is to establish an accurate baseline for examining human perception of the valence of silence in interaction; by "valence" we mean the perception of a negative or positive weight associated with the inter-turn silence. In particular, we test whether silence is associated with lack of enthusiasm or weakness of agreement as measured by subjective reaction on a 6-point scale. In addition to assessing the valence of varying silence lengths overall, we examine the effect of silence in relationship to common speech acts and in relationship to the intonation and duration of response tokens following silence. In this way, we aim to account for both pragmatic and prosodic contexts which may shape the perception of inter-turn silences.

To investigate these phenomena we devised three experiments; the basic approach to each experiment was the same (described in Section 2) and we used the same stimuli across the study with specific manipulations for each experiment. The first experiment isolated the effects of inter-turn silence; the second examined the relative contribution of prosodic cues in the response token; and the third brought together these findings to examine the salience of inter-turn silence in relation to response token prosody.

2. Methods

2.1. Overview

For this study, undergraduate students listened to audio recordings of simulated telephone calls between 2 female friends (6 target dialogues and 3 distractor dialogues to divert subjects' attention from focus of the study). The constructed dialogues concerned relatively mundane themes (flyers for a school function, new furniture, going to the gym) and each one ended with the caller either formulating a request in terms of that topic (e.g. getting a ride to pick up the flyers) or offering an opinion on the topic at hand (e.g. reporting that the flyers look good). The call recipient answered in the affirmative for both speech acts ("Sure" for the request and "Yeah" to display agreement with the opinion).

Upon hearing the recipient's response to the request or assessment, the study subjects had 8 seconds to rate, on a six-point scale, their perception

of the speaker's enthusiasm for assenting to the request or assessment embedded in the dialogue (see Appendix A for a sample of dialogues and related questions). Ratings higher on the scale indicated a perception of greater willingness on the call recipient's part to comply with the request or agree with the assessment. The stimuli were counter-balanced in terms of thematic content, speech act, inter-turn silence and other prosodic features.

Approximately 10 min were required for study participation and the subjects were compensated for their time.

2.2. Construction of stimuli

Dialogues were performed by undergraduate Theater majors; these students were roughly the same age as the target population for subject recruitment. The conversations simulated telephone calls among peers and were based on transcriptions of actual telephone calls among friends of an age similar to the study population (see Roberts and Robinson, 2004; Schegloff, 1979). Relying on real telephone openings increased the verisimilitude of the stimuli and helped to convey a feeling of friendly, peer interaction. Because responses to requests and assessments likely vary depending on the social relationship of the interactants, it was important to convey as much as possible that the people in the conversations were friends of long-standing. Such familiarity is generally conveyed by features such as omission of the caller's name, using an informal register, and/or truncating the greeting sequence (Hopper, 1992; Schegloff, 1979).

The call recipient—that is the actress who would be responding to the request or assessment—was directed to make her response tokens agreeable, but not overly enthusiastic. The actress was instructed to keep her voice in her normal register and to respond quickly, with no sense of hesitation. In other words, the actress was to sound willing and agreeable in an everyday sort of way to her "friend's" mundane request or assessment.

The same actors performed all target dialogues and maintained the same caller-call recipient roles. Thus, we controlled for any variation in response that might be elicited by voice quality of the requester/responder. By varying the thematic context (flyers, furniture, gym) we could test the effect, if any, of context and also add some topical variety to simulate the possibility of different types of calls between the friends.

Stimuli were recorded to digital audio tape (DAT) (Sony TCD-D8) using a hypercardiod electret condenser microphone (Audio-Technica D1000HE) coupled via an ART Studio V3 in a sound-isolating booth (IAC, Model #403A). Recorded tokens were redigitized using a High Density Linear A/D-D/A converter, Shure Mixer Amplifier and Praat 4.2.21 running on a Dell Optiplex/Windows XP computer. Tokens were digitized at 22.05 kHz and peak amplitude-normalized to within 90% of the 16-bit maximum quantization range prior to editing.

Once the dialogues were appropriately edited they were randomized with distractor dialogues. Stimuli were counterbalanced for presentation of all conditions and practice effects were controlled for by presenting a reverse presentation order to some participant groups.

3. Experiment 1: The interaction of silence and speech act

3.1. Identification and insertion of neutral response tokens

Our main concern in Experiment 1 was to isolate the effect of silence; we therefore controlled for possible confounding from the acoustic qualities of the actor's slightly different response token pronunciations (across the different contextual conditions recorded) by identifying median response tokens and copying them across the relevant stimuli.

From all of the response tokens recorded (3 "sures" for the requests and 3 "yeahs" for the assessments) the pitch range and direction of pitch change were calculated by hand, based on measurements in Praat 4.2.21 (Boersma and Weenink, 2001) using the default autocorrelation method (Boersma, 1993). The "sure" token and the "yeah" token which fell in the middle in terms of these parameters were chosen as the response tokens for the study stimuli. The median "sure" token used in the stimulus dialogues was 335 ms long with a falling contour from 325 Hz to 213 Hz. The stimulus "yeah" token was 300 ms long with a rise–fall contour (278–302–214 Hz). Once chosen, these median tokens were then digitally edited into the corresponding response slots in the dialogues. These tokens became the default or neutral token upon which additional manipulations were done in Experiments 2 and 3.

Once the neutral response token for each dialogue was edited in, silences were inserted between the focal speech act and the response token. Three lengths of silence were used to test the interaction between silence and speech act: 0 ms (no lag time), 600 ms, and 1200 ms. These lengths were chosen to provide a baseline simulation of no gap between request/assessment and response and then equal increments leading up to the proposed limits of a standard maximum of silence (Jefferson, 1989). Silence was taken from other dead space in the dialogue (i.e. it was not a machine-produced silence) to best maintain the natural acoustic environment. These natural silences were spliced to the end of phonation on the request/assessment utterance as visually apparent from the sound wave.

3.2. Study participants

One hundred and eleven (111) undergraduate American English speaking students, participated in Experiment 1. Sixty-seven females and 43 males, ranging in age from 19 to 39 ($M = 20.34$, SD = 2.16) were recruited from undergraduate introductory communication courses; participation was voluntary. Although the groupings of students were convenience samples, neither their demographic profiles nor their mean scores on like items were significantly different.

3.3. Statistical procedures

The study employed a $3 \times 2 \times 2$ mixed factorial design to investigate factors contributing to perception of response favorability. The between subjects factor was silence length (0 ms gap, 600 ms gap, and 1200 ms gap), while the within subjects factors were speech acts (requests and assessments) and thematic context (flyers, furniture, and exercising). Theme was incorporated into the study design to increase generalizability across contexts (Jackson, 1992). Within subject factors were counterbalanced to control for order effects.

Prior to conducting analyses examining the core research questions and hypotheses, the effect of thematic context and sex of listener were examined for any possible differences in response scores. Although sex of listener is of interest for the development of research on silence, as is theme or topic of conversation, these were not central concerns of the present study; thus, if listener judgments were equivalent across sex of listener and thematic context, then collapsing across these variables would be justified in further analyses to increase the power to detect differences in those factors of central theoretical interest for the current study.

To evaluate the relationship between thematic context and silence length within each speech act, six one-way ANOVAs were conducted. To examine the effect of sex of listener on perception, two separate tests were conducted for each speech act.

The approach to the evaluations of theme and sex of listener differed because subjects received each theme, but did not receive all possible theme combinations across all silence length and all act conditions. This reduced subject fatigue by reducing the number of target stimuli for each subject from 18 to 6. This still allowed us, however, to systematically investigate each factor. Thus, we ran individual F-tests for the first set of analyses (on theme). To protect family-wise error rates, we used the Bonferoni criteria.

We were able to reduce the number of individual tests for the second analysis because theme was not a variable under consideration.

3.3.1. Effect of theme

Results of the ANOVAs for the analysis of theme were not significant for any of the silence lengths: 0 ms gap, $F(2,108) = 1.39$, $p = .254$; 600 ms gap, $F(2,108) = .749$, $p = .475$; 1200 ms gap, $F(2,108) = .59$, $p = .557$. Results for assessments: 0 ms gap, $F(2,108) = .312$, $p = .733$; 600 ms gap, $F(2,108) = 1.39$, $p = .254$; 1200 ms gap, $F(2,108) = 1.39$, $p = .254$. These analyses thereby justified collapsing across each theme for each speech act. Means and standard deviations for each of the analyses are presented in Table 1. Since the results indicated no differences in response favorability due to thematic context, subsequent analyses of other factors were collapsed across this variable.

Table 1
Effect of theme on speech act

Silence length	Flyer		Furniture		Gym	
	Mean	SD	Mean	SD	Mean	SD
Requests						
0 ms	4.50	.86	4.14	1.13	4.45	1.03
600 ms	3.05	.98	3.29	.84	3.17	.83
1200 ms	2.18	1.04	2.10	.76	2.33	.87
Assessments						
0 ms	3.87	.90	3.67	1.02	3.74	1.13
600 ms	3.21	1.10	2.84	1.03	3.10	.80
1200 ms	2.26	.64	2.00	.79	2.21	.91

Table 2
Effect of sex of listener on scores for requests and assessments

Silence length	Males		Females	
	Mean	SD	Mean	SD
Requests				
0 ms	4.51	1.18	4.24	.92
600 ms	3.12	.85	3.18	.92
1200 ms	2.44	.83	2.10	.70
Assessments				
0 ms	3.77	1.02	3.76	1.02
600 ms	3.02	1.04	3.09	1.00
1200 ms	2.21	.81	2.13	.80

3.3.2. Effect of sex of listener

For the analysis of the effect of the sex of the listener, we conducted separate tests for each speech act: we first conducted a repeated measures analysis of variance with silence length as the within subjects factor, sex as the between subjects factor, and ratings of perception of willingness to comply with the request as the dependent measure. The means and standard deviations for subjective rating scores for favorability ratings are presented in Table 2.

The results for the ANOVA indicated no significant main effect for sex in participants' perceptions of silences following requests: $F(1,108) = 2.07$, $p = .15$; nor was there a significant interaction effect for sex by silence length $F(2,216) = 2.08$, $p = .13$; likewise, there was no significant main effect for sex in participants' perceptions of silences following assessments: $F(1,108) = .001$, $p = .97$; nor was there a significant interaction effect for sex by silence length $F(2,216) = .20$, $p = .81$.

Since the results indicated no differences in response favorability due to sex of study participant, subsequent analyses of other factors were collapsed across this variable.

3.4. Results: different silence lengths elicit significantly different judgments of the valence of the silence

To examine the interaction of silence and speech act, we conducted two separate tests, one for each speech act. First, a repeated measures analysis of variance was conducted with silence length as the within subjects factor; ratings of perception of willingness to comply with the *request* was the dependent measure. The means and standard deviations for participant ratings of favorableness are presented in Table 3. The results for the ANOVA indicated a significant main effect for length of silence $F(2,109) = 187.99$, $p < .001$. Follow up tests were conducted using paired t-tests; Holm's sequential Bonferroni procedure was used to control for Type I error (Green et al., 1997).

Table 3
Ratings on perception of willingness to comply with requests

Silence length	Mean	SD
0 ms	4.34	1.03
600 ms	3.16	.76
1200 ms	2.24	.89

When comparing silence lengths across requests, all results were significant: 0 ms gap versus 600 ms gap, $t(110) = 10.50$, $p < .001$; 600 ms gap versus the 1200 ms gap $t(110) = 18.82$, $p < .001$; 0 ms gap versus the 1200 ms gap $t(110) = 11.33$, $p < .001$.

Next, a repeated measures analysis of variance was conducted with silence length as the within subjects factor; ratings on perception of agreement with the assessment was the dependent measure. The means and standard deviations for subjective rating scores are presented in Table 4. The results for the ANOVA indicated a significant main effect for length of silence $F(2,109) = 94.01$, $p < .001$. Again, paired t-tests (Bonferroni-corrected) were used to control for Type I error.

When comparing silence lengths across assessments all results were significant: 0 ms gap versus 1200 ms gap, $t(110) = 13.40$, $p < .001$; 0 ms gap versus the 600 ms gap $t(110) = 6.23$, $p < .001$; as well as the 600 ms gap versus the 1200 ms gap $t(110) = 9.16$, $p < .001$.

In order to examine differences in perceptions of silence lengths between requests and assessments, paired samples t-tests (Bonferroni-corrected) were conducted. The results indicated that participants' perception of the zero ms gap was different for requests and assessments $t(110) = 5.42$, $p < .001$. There were no differences however, in participant's perceptions of requests versus assessments for the remaining two silence lengths: 600 ms gap $t(110) = 1.01$, $p = .31$; 1200 ms gap $t(110) = 1.24$, $p = .22$. Means and standard deviations are presented in Table 5.

Table 4
Ratings of perception of agreement with assessments

Silence length	Mean	SD
0 ms	3.75	1.02
600 ms	3.05	1.01
1200 ms	2.17	0.80

Table 5
Comparison of perceived valence of silence across speech acts (request vs. assessments)

Silence length	Requests		Assessments	
	Mean	SD	Mean	SD
0 ms	4.34	1.03	3.75	1.02
600 ms	3.16	.89	3.05	1.01
1200 ms	2.24	.77	2.17	.80

3.5. Discussion

Findings strongly indicate that listeners hear inter-turn silences in a way that affects their perception and judgment of the interactional tone. Subjects in the listening task perceive inter-turn silence, when presented in the context of friendly dialogues, as indicative of lack of willingness or weakness of agreement. Uniformly, across speech acts, as silence length increased to 600 and then 1200 ms, judgments of the interactant's willingness/enthusiasm to comply with requests or agree with assessments decreased significantly.

Despite a consistent and increasing negative valence on perceptions associated with increasing silence lengths, the results indicated that only at the 0 ms gap did participants' perceptions differ for requests and assessments (Mean = 4.34 and 3.75, respectively; $t(110) = 5.42$, $p < .001$). This indicates that for the zero second gap, judges rated the silence after assessments as more problematic (i.e., the scores are significantly lower on the 6 point scale). Thus, in a baseline condition, the "yeah" as a token of agreement with an assessment may actually be perceived as a less agreeable response than the "sure" token for the requests. In other words, the semantics of the lexical items could be influencing judgments. However, as none of the other silence lengths were tempered by speech act, we conclude that silence length is the more salient feature for listeners.

What this finding suggests is that while speech act (or semantics of response) may be relevant when followed by very short intervals of silence, as silence length increases, the effect of the silence begins to overshadow. The fact that, so far, it appears that speech act affects perception of silence only when there is no gap, motivates further exploration to see whether prosodic properties of the lexical item make a difference—particularly since prosody of response was controlled for in the Experiment 1. Experiment 2, then, focuses attention on the talk following the inter-turn silence.

4. Experiment 2: The effect on listener judgments of prosodic features of response tokens

Experiment 2 explored the salience of several possible prosodic cues to unwillingness and disagreement. Because listener judgments of unwillingness and disagreement are likely affected by both inter-turn silence and the acoustic properties of the response token, in this experiment we explored prosody while controlling for inter-turn silence length. We were thus able to examine the interaction of prosody with speech acts to determine which cues would be good candidates for a final (third) experiment where silence and prosody could be explored together.

For this experiment, pitch and duration characteristics of the response tokens were manipulated. Pitch was examined in terms of both contour (direction of pitch change) and absolute pitch (a shift in pitch across the utterance while the contour remains unchanged). Duration was examined in terms of word duration (i.e., "stretching" the response token).

Although absolute pitch in English has not been systematically studied at the discourse level, there is some evidence that rising intonation is associated with perceptions of uncertain/incorrect answers (Brennan and Williams, 1995; Smith and Clark, 1993). This is supported by evidence that a fine distinction between "uncertainty" and "incredulity" (as perceived in the context of a rise–fall–rise contour) is available in English from frequency properties, particularly pitch range (Hirschberg and Ward, 1992). Because of the enormous variety of possible contours that could be studied, and since we were also interested in the effect of absolute pitch, we chose to address only one contour, low-rising, given previous research suggesting that such a contour might contribute to listener perceptions of uncertainty/reluctance (Brennan and Williams, 1995; Smith and Clark, 1993).

For response token duration, existing empirical evidence is less encouraging, but not sufficiently negative in the particular affective domain studied here (i.e., reluctance) to rule it out altogether. Duration did not prove to be a key factor in distinguishing uncertainty from incredulity in English (Hirschberg and Ward, 1992) nor did it figure in the production of a surprised reaction in the use of "bitte" as a repair initiator in German (Selting, 1996). Although this was not a perceptual study, Selting's (1996) acoustic analysis suggests that speakers do not orient to durational cues for encoding or performing surprise.

Despite the lack of evidence so far for duration as a distinguishing cue at the discourse level, we nonetheless pursued our examination of this feature for the current study because introspection and casual observation tends to favor an interpretation that word duration (stretching an affirmative response) leaves a patina of doubt or reluctance on an otherwise affirming lexical choice.

4.1. Manipulations of response tokens

Absolute pitch, pitch contour, and duration of response token were all manipulated in Experiment 2. The stimuli from Experiment 1 were maintained except that all inter-turn silence lengths were equalized (220 ms), set to the mean length of gap from the initial, natural performance of the dialogues. The only manipulations were thus on one actor's response tokens (''sure'' following requests and ''yeah'' following assessments).

4.1.1. Absolute pitch: pitch shift

For this manipulation, the natural (or original) response token pitch contour was maintained, but the fundamental frequency of the whole word (''yeah'' and ''sure'') was raised 70 Hz using Praat 4.2.21. We term this a ''pitch shift'' to indicate that segmental and pitch contour information are held constant, but the word is produced at an overall higher fundamental frequency. Since there was no a priori basis for determining an appropriate magnitude for this pitch shift, we piloted two pitch levels: 70 Hz above normal and 100 Hz above normal, both of which seemed natural enough in isolation. However, in the context of the dialogues it became clear that 100 Hz above the actor's baseline production (260 Hz) was too high to sound natural. We therefore used only the 70 Hz manipulation for this experiment.

4.1.2. Pitch contour

The variety of possible pitch contours that could be examined in English was overwhelming for the current study, so we chose to simply test the contour opposite of the one that was produced by the actor in her natural interpretation of the dialogue. Since her baseline reading was a high falling contour (indicating a standard affirmative stance), we decided to alter the contour for this experiment by inverting it to a low rising contour, one that might convey uncertainty as established in previous research (Brennan and Williams, 1995; Smith and Clark, 1993).

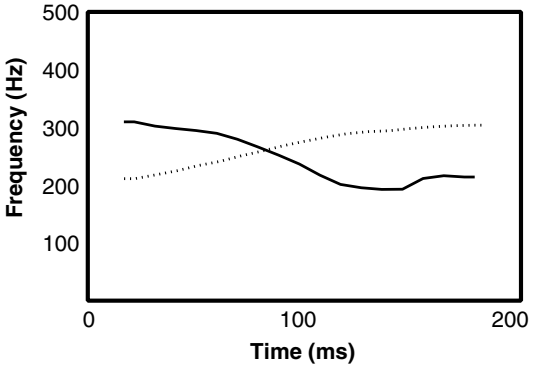

Fig. 1. Reflection of f0 contour: solid (falling) line shows natural syllable's f0 contour, broken line shows resynthesized ''rising'' contour.

To construct the rising contour, we replaced the original pitch contour with a new one that was the inverse or reflection of the original (Fig. 1). This was created by identifying three points along the original contour and reflecting them through the line of the average fundamental frequency of the original contour. For example, if the initial point were determined to be 34 Hz above the average, then the initial point of the reflected contour would be 34 Hz below the original contour.

4.1.3. Duration

Duration was scaled automatically using the Praat 4.2.21 duration tier manipulation standard settings.

Two durations in addition to the natural duration were used in Experiment 2: Duration 1 indicates the normal word duration as originally (and naturally) produced by the actor (300 ms and 335 ms, respectively, for ''yeah'' and ''sure''). The other durations were roughly twice and thrice the natural duration of the core voiced elements of the word. To maintain naturalness, the nucleus and margins around it (the transition into and off of the nucleus) were included for ''stretching.'' Thus, for Duration 2 (''double'') the resulting durations were 460 ms and 560 ms respectively for ''yeah'' and ''sure.'' For Duration 3 (''triple'') the resulting durations were 630 and 760 ms respectively for ''yeah'' and ''sure''. Duration manipulations were fully crossed with pitch manipulations.

4.2. Study participants

Seventy-one (71) undergraduate American English speaking students, participated in the study.

Twenty-nine women and forty-two men, ranging in age from 16 to 41 ($M = 21.36$, $SD = 3.74$) were recruited from undergraduate liberal arts courses. Participation was voluntary and participants received compensation for their efforts. The task was identical to Experiment 1 (see Section 2.1).

4.3. Statistical procedures

In a $3 \times 3 \times 2$ mixed factorial design, duration was a between groups factor; pitch type (Natural contour, Pitch Rise, Pitch Shift) and speech act type (Request and Assessment) were the within group factors. Even though theme was not a significant factor in initial testing (see Section 3.3), we continued to use all three themes in a counterbalanced manner to increase generalizability across contexts (Jackson, 1992). The same distractors, order of presentation, and manner of presentation were used from Experiment 1.

4.4. Results: duration of response token seems salient for listeners (when inter-turn silence is normalized)

Results from all responses were collected and first examined using a repeated measures analysis of variance. Ratings of perception of willingness or agreement was the dependent measure. All means and standard deviations appear in Appendix B.

The results for the ANOVA indicated a significant main effect for pitch, $F(1,68) = 12.33$, $p = .001$) and duration $F(2,68) = 53.44$, $p < .01$. However, as Fig. 2 shows, there was no significant interaction of pitch type with speech act, $F(1,68) = .059$, $p = .80$.

Conversely, there was a significant interaction effect between response token duration and speech act, $F(2,68) = 30.19$, $p < .01$. An examination of this interaction indicated a general decline in scores as response token duration increased (Fig. 3), with the source of the interaction associated with the natural duration (the unmanipulated token). Indeed, paired t-tests (collapsed across pitch type) reveal that the only significant difference in mean scores between requests and assessments was at the natural duration (Mean difference = 1.06, $SD = .74$, $p < .01$). Actual means can be seen in Appendix B.

These results are similar to the finding of Experiment 1 that speech act was an important factor only at the shorter intervals of inter-turn silence. It appears that speech act (or the semantics of the lexical item) only plays a role when other cues (here response token duration) are not clearly indicative for listeners. Since listeners are asked to make judgments on scales of willingness or agreement, they are likely listening for relevant cues to push them to one end or the other of that scale. Apparently, response token duration is the stronger cue.

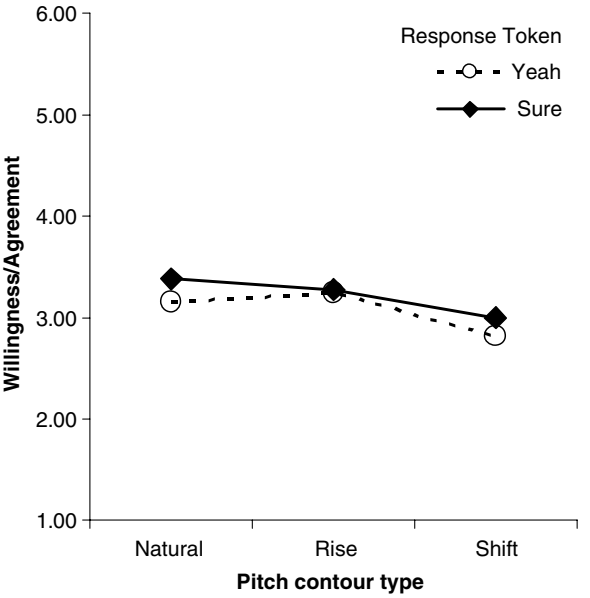

Fig. 2. Interaction of pitch type and speech act.

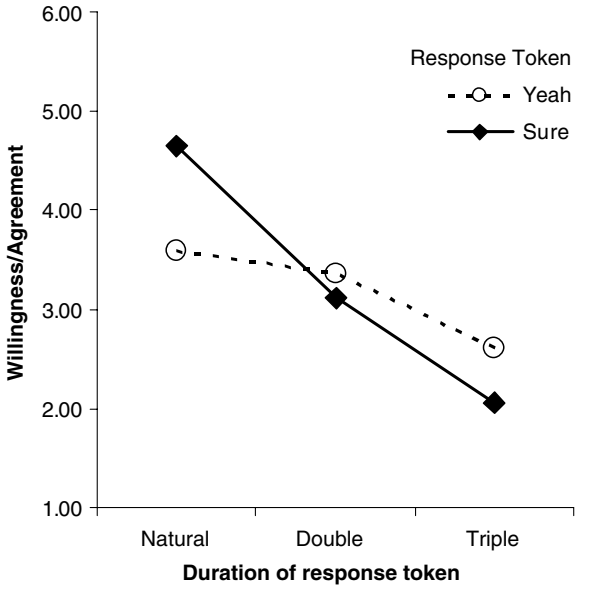

Fig. 3. Interaction of response token duration and speech act.

With the variation in speech act accounted for primarily by duration, we then averaged scores across speech acts to assess more clearly the relative contributions of pitch and duration. The ANOVA revealed a significant interaction effect of these two variables, $F(2, 68) = 5.11$, $p = .009$.

Fig. 4 characterizes this interaction in that, overall, as duration of the response tokens increased, judgments of willingness and agreement declined, regardless of pitch type. This indicates that the increased duration of the response token was most systematically associated with negative judgments. Post-hoc tests (Bonferroni corrected) revealed that each pitch type was significantly different from its like pitch type across each level of response token duration. (Actual means can be seen in Appendix B.) The only exception to this was for the rising pitch contour, the means for which were not significantly different when comparing it in the two longer response token durations (Mean difference = .57, $p = .11$).

4.5. Discussion

Although at the discourse level it is unwise to equate a single prosodic feature with a single perceptual effect, the sum total of the findings for this experiment point to duration as the salient prosodic cue when listeners are making judgments about willingness and agreement. Further exploration of pitch shift may also be warranted since it was viewed most positively by judges in the natural word length condition, but most negatively in the most stretched duration condition.

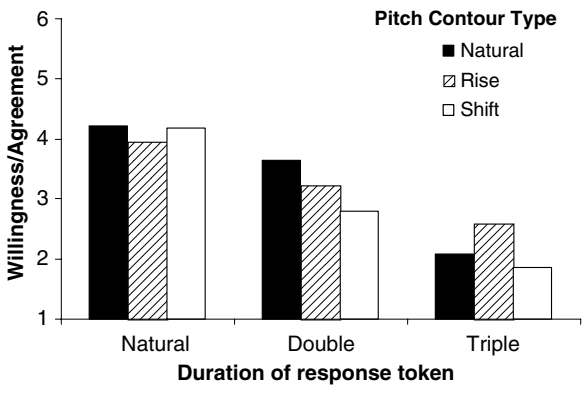

Fig. 4. Interaction of response token duration and pitch type.

5. Experiment 3: The effect on listener judgments of inter-turn silence and response token duration

Experiment 3 explored the relative contributions of inter-turn silence and response token duration as cues to unwillingness and disagreement. Based on results from Experiments 1 and 2, it became clear that this category of listener judgments is affected by both inter-turn silence and the acoustic properties of the response token, especially word duration. In this final experiment we examined these two prosodic features together to determine the relative salience of these cues for listeners.

For this experiment, inter-turn silence length manipulations and response token duration manipulations were used from the earlier experiments. Thus, stimuli were identical across all experiments in terms of variables investigated, as well as speakers' vocal qualities, themes used, speech act types, and other discourse level features.

Despite the lack of evidence in other research for duration as a distinguishing cue at the discourse level, we nonetheless pursued our examination of this feature for the current experiment based on findings from Experiment 2 that the duration of the response token does affect perceptions of unwillingness and disagreement.

5.1. Manipulations of response tokens

No new manipulations were introduced in this Experiment. The stimuli from Experiment 1 were maintained with their inter-turn silences "intact", that is, as manipulated for that phase of the study. Response tokens from Experiment 2 were inserted in place of the original default tokens so that the actor's response tokens ("sure" following requests and "yeah" following assessments) were "stretched" in the same manner as they had been in Experiment 2 (see Section 4.1).

5.2. Statistical procedures

In a $3 \times 3 \times 2$ mixed factorial design, duration (natural, double, and triple) was a between groups factor; silence length (0 ms gap, 600 ms, 1200 ms) and speech act type (request and assessment) were the within group factors. Even though theme was not a significant factor in initial testing (see Section 3.3.1), we continued to use all three themes in a counterbalanced manner to increase generalizability

across contexts (Jackson, 1992). The same distractors, order of presentation, and manner of presentation were used from Experiments 1 and 2.

5.3. Study participants

One hundred and twenty-three (123) undergraduate American English speaking students, participated in the study. Seventy-eight women and forty-three men, ranging in age from 18 to 55 ($M = 21.47$, SD = 4.75) were recruited from undergraduate liberal arts courses. Participation was voluntary and participants received compensation for their efforts. The task was identical to Experiments 1 and 2 (see Section 2.1).

5.4. Results: silence is the more salient cue, but response token duration is in a trading relationship with it

Results from the ANOVA indicated significant main effects for silence length, $F(1,115) = 268.59$, $p < .01$, and for duration, $F(2,115) = 14.03$, $p < .01$. However, there was no main effect for speech act, $F(1,115) = .143$, $p = .706$ or for sex of listener, $F(1,115) = .077$, $p = .782$. A two-way interaction effect was evident for silence length and response duration, $F(2,115) = 11.18$, $p < .01$. The remaining two-way and three-way interactions were not statistically significant.

With these findings, we collapsed across speech acts and across sex of listener to further examine the relationship between inter-turn silence and response token duration.

Fig. 5 suggests that inter-turn silence is more salient, overall, for listeners, than the effect of word duration. For each response token duration, raters perceived a decrease in willingness/agreement as the inter-turn silence length increased.

Interestingly, in both the no gap (0 ms) condition and the longest gap condition (1200 ms), listeners judged the normal word duration and the doubled word duration as nearly the same. For the normal duration, the mean = 3.94, SD = 1.09 and the word duration that was twice normal the mean = 3.90, SD = 1.07. In other words, when the answer comes immediately and affirmatively after the request or assessment, only the longest response token (triple the normal length) elicited distinctly negative judgments (Mean = 2.84, SD = .64). This mean was significantly different in the no gap condition (as determined in a post hoc test, Bonferroni corrected; mean difference = 1.10, $p < .001$).

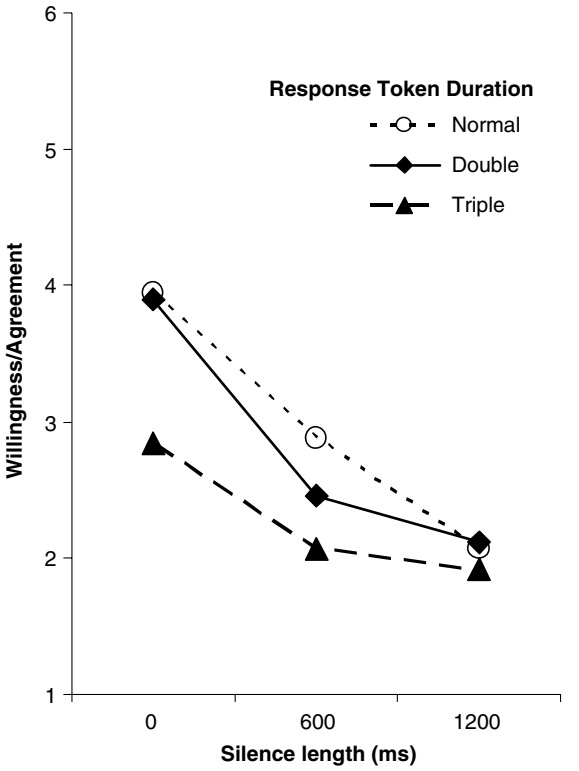

Fig. 5. Interaction of silence length and response token duration.

In the longest inter-turn silence condition (1200 ms) listeners also did not appear to discriminate when the response token durations were normal or twice normal duration (Mean = 2.07, SD = .69; Mean = 2.02, SD = .64, respectively). However, the difference between those scores and the mean ratings for the longest word duration were not significantly different.

At the 600 ms silence length, all means for the 3 response token durations were statistically significantly different in post hoc tests. At the 600 ms silence interval, the mean difference between normal and double response token duration was .43, $p = .003$; the mean difference between double and triple duration was .38, $p = .008$.

5.5. Discussion

These results show a clear salience of inter-turn silence over response token duration as an indicator of "trouble" in discourse; although both cues are also clearly useful to listeners. When inter-turn silence is above the critical one second threshold, listeners rate

the discourse as uniformly problematic, regardless of response token duration. In contrast, when inter-turn silence is very short (i.e., there is an immediate response to the request/assessment), listeners treat normal and moderately long response token durations as relatively unproblematic. However, there is evidence that listeners use response token duration as a cue to unwillingness or disagreement in the context of no silence when the token is sufficiently long.

The most interesting case is when inter-turn silence is ambiguous, neither short enough to promote a clear perception of willingness/agreement, nor long enough to promote a clear perception of reluctance/hesitation. In the 600 ms inter-turn silence case, listeners are forced to turn to the less prominent cue of response token duration, and it is here that this cue makes the most obvious differences.

Overall, when the inter-turn silence cue is not salient (the 0 ms condition) listeners rely on the longest response token duration for making judgments of (un)willingness and (dis)agreement. Although the same trend is interpretable from the means from the longest inter-turn silence, the effect of the long word duration is attenuated somewhat (it was not statistically significantly different); we would argue, therefore, that inter-turn silence is the stronger cue.

It is in the middle ground, at 600 ms of inter-turn silence, that the silence cue may be somewhat ambiguous and may lead listeners to focus more on the response token duration cue. This is likely a fertile context in which to further explore prosodic cues and other features of interest. In this ambiguous length of 600 ms, listeners may well be "lead" to rely on other cues of interest in making judgments about the interactional work at hand.

6. Conclusions

Selting (1996) notes that marking utterances prosodically provides listeners with cues that some kind of special inference may be relevant. The current study has explored a particular set of inferences that listeners may make at the intersection of inter-turn silence, speech act, and prosody. The purpose of this study was to examine the valence of inter-turn silence in conjunction with prosodic cues we hypothesized as relevant to the perception of (un)willingness or (dis)agreement in conversation. Overall, findings strongly indicate that listeners hear inter-turn silences in a way that affects their perception and judgment of the interactional tone. In Experiment 1, when prosodic features were controlled, listener judgments of willingness and agreement consistently declined as inter-turn silence increased. In Experiment 3, when duration of the response token was manipulated, silence proved to be the more robust cue.

Despite the strong support for silence as troubles-indicative, Experiment 3 also points to how these discourse level cues may function in a manner similar to the trading relations observed among acoustic cues to phonetic segments (Repp, 1982). At the discourse level, when the silence length was closer to half a second, and therefore neither clearly troublesome nor obviously trouble-free, listeners were able to use durational cues of the response token to make judgments about the tone of the interaction. This finding underscores the challenges and complexities of exploring prosodic cues in that trade-offs between them, even at the discourse level, are relevant for listeners.

We wondered if sex of the listener might affect response to the silence lengths. Based on a wide variety of studies, across several disciplines, the notion of gendered styles of interaction is well accepted (see Holmes and Meyerhoff, 2003, for an overview of current approaches). However, for this study, in terms of perception of the valence of silences between female friends, findings across the experiments indicated that sex of the listener was not a significant factor. This can now be investigated further as different gender dyads are examined. For the present, it appears that male and female listeners judge inter-turn silence in the same way when the interactants are female.

The issue of theme, included here to increase generalizability across contexts, also remains to be examined in greater detail. The themes used in the current study were mundane; they were different enough to evoke different schemata, yet not different in their intensity. Had we examined the difference between asking for a ride to the store versus a ride to see a dying relative, we may well have detected some differences in response to the themes. For the moment, however, it appears that everyday thematic contexts do not affect perception of the valence of silence following routine speech acts.

In the domain of speech act, findings were consistent in that the precipitating speech act (request and assessment) and the affirmative response ("sure" and "yeah") did not interact significantly with silence or other prosodic features, except in the no gap conditions. However, scores for the assessment condition were consistently lower. This finding is likely due to

a limitation of the study in that the response tokens, while prosodically similar, were semantically different. "Sure" has a more narrow scope (it cannot agree with an opinion unless that opinion is stated as a yes/no question) whereas "yeah" could function as a rejoinder to both requests and assessments. Because of its broader scope, "yeah" may be perceived as a weaker form or as more ambiguous, which would explain the overall lower scores for this condition. Future studies might examine the use of just one of the response tokens (e.g., "yeah" which can be responsive to a variety of speech acts) to further investigate the weight of the prosodic features carried on the lexical item.

As a contribution to speech processing research, this study opens the door for a shift in attention from silence as a mere "off" state[2] in conversation to conceptualizing silence as a dynamic moment in which interactants are at once attending to what's next and reviewing what has just occurred. In conjunction with research on uncertainty, this study suggests more focused attention on the work that goes on in between the time an utterance is launched and the response is heard.

More broadly, our findings fit into a line of work that examines prosodic resources that listeners draw on for signaling problems in communication. Krahmer et al. (2002) found that problems with a preceding bit of talk (by a machine) were signaled as disconfirming by human speakers from (inter alia) longer "no" responses and longer delays between the problematic machine utterance and the human's response. In fact, listeners, in perception tests, were able to draw on the prosodic shape of the "no" alone to determine its import as a go-ahead or go-back signal. Shimojima et al. (2002), also show that listeners draw on prosodic and temporal features for "settling" meanings. In their examination of echoic responses they demonstrate how intonation and duration are used to secure some bit of information as "common ground" (p. 123).

These two studies, in concert with the present findings, indicate how speaker/listeners deploy, and are sensitive to, prosodic features of talk for deciding whether or not information is correct and/or understood in common. For the current project, a somewhat murkier area of "correct" and "understood" is tackled in that affective orientations of willingness and agreement are at stake. This is in concert with a move in dialog systems research away from flawless recognition of speech sounds to more concern with analyzing (emotional) tone as the target analytic job of the machine (see the Special Issue of Speech Communication, 2003, Vol. 40). While emotional tone may be a less likely concern in the more information intensive environments of human–machine interaction, they are nonetheless crucial, particularly as machines are used in settings such as medical contexts where such affective orientation to suggestions, advice, and recommendations may be relevant. If listeners can classify "no" as positive or negative without recourse to preceding context (Krahmer et al., 2002), can machines? And in a more challenging vein, can machines also classify affirmative, but tenuous responses? Or does this represent a limit for human–machine interactions? Perhaps machines can only be reliable when simple correctness is as stake.

As noted in a review by O'Shaughnessy (2003), automatic speech recognition "exploits only the most rudimentary knowledge about human production and perception phenomena" (p. 1272). In fairness, our knowledge of human perception/production may itself still be only rudimentary. Nonetheless, O'Shaughnessy's assessment indicates room for further systematic exploration of the more subtle aspects of interactional competence. In this way, they may be better understood in general and then applied as needed in the design of a more naturalistic human-computer interface. As current and previous research shows, "naturalness" may come less from segmental accuracy of speech, and more from discourse level prosodic resources that humans draw on so effortlessly in their everyday interactions.

Appendix A. Example of a request and an assessment

(1) ((Telephone Rings))
A: Hello?
B: Rachel?
A: Yeah,
B: Hey it's me.
A: Hey how's it goin.

[2] In most engineering approaches to speech synthesis and recognition, issues of silence have traditionally been handled by Brady's (1969) stochastic 6-state model for conversation. This model conceptualizes conversation as alternating intervals of silence and speech. While this is understandable from an acoustic standpoint, it disregards the fact that silence between speaker turns is dynamic (Sacks et al., 1974); conversational turn-taking is probably best described as a sequential, not stochastic process (Wilson and Zimmerman, 1986).

B: Good. I just called Kinkos,
A: uh huh
B: And the call-out flyers are ready. Can you give me a ride over there?
B: Sure!

Question: How willing is Rachel to give her friend a ride?

(2) ((Telephone Rings))
A: Hello?
B: Rachel?
A: Yeah,
B: Hey it's me.
A: Hey how's it goin.
B: Good. I just saw Kim's new furniture. It looks pretty good!
A: Yeah!

Question: How much does Rachel agree with her friend about the furniture?

Appendix B

Mean scores for Requests and Assessments: Three Durations and Three Pitch Types

	Assessments		Requests	
	Mean	SD	Mean	SD
Duration 1[a]				
Natural[b]	3.77[*]	1.06	4.68[*]	1.08
Rising contour	3.41[&]	1.29	4.50[&]	0.96
Pitch shift	3.60[@]	1.08	4.77[@]	1.10
Duration 2[c]				
Natural	3.46	1.14	3.83[**]	1.09
Rising contour	3.58	1.01	2.88	1.48
Pitch shift	3.04	1.12	2.63[**]	1.09
Duration 3[d]				
Natural	2.32	0.98	1.84[#]	0.75
Rising Contour	2.76	1.33	2.56[#,+]	1.19
Pitch shift	1.92	0.98	1.80[+]	0.76

[a] Actor's natural production of word length for "yeah" and "sure".
[b] Actor's default contour and pitch.
[c] Manipulation of word duration to twice the natural length.
[d] Manipulation of word duration to thrice the natural length.
[*] Means significantly different at $p < .01$.
[**] Means significantly different at $p < .01$.
[#] Means significantly different at $p < .01$.
[&] Means significantly different at $p < .01$.
[+] Means significantly different at $p < .01$.
[@] Means significantly different at $p < .01$.

References

Anderson, P., 1999. Nonverbal Communication: Forms and Functions. Mayfield Publishing, Mountain View, CA.

Batliner, A., Fischer, K., Hubera, R., Spilkera, J., Nötha, E., 2003. How to find trouble in communication. Speech Commun. 40, 117–143.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. Inst. Phonetic Sci. 17, 97–110.

Boersma, P., Weenink, D., 2001. Praat: a system for doing phonetics by computer. Retrieved May 22, 2001, from University of Amsterdam, Institute of Phonetics Sciences Web site: http://www.praat.org (confirmed April 15, 2005).

Brady, P.T., 1969. A model for generating on–off speech patterns in two-way conversation. Bell Syst. Technol. J., 2445–2472.

Brennan, S.E., Williams, M., 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. J. Memory Language 34, 383–398.

Burgoon, J.K., Buller, D.B., Guerrero, L.K., 1995. Interpersonal deception: IX. Effects of social skill and nonverbal communication on deception success and detection accuracy. J. Language Social Psychol. 14, 289–311.

Davidson, J., 1984. Subsequent versions of invitations, offers, requests, and proposals dealing with potential or actual rejection. In: Atkinson, J.M., Heritage, J. (Eds.), Structures of Social Action: Studies in Conversation Analysis. Cambridge University Press, Cambridge, pp. 102–128.

Glucksberg, S., Mccloskey, M., 1981. Decisions about ignorance: knowing that you don't know. J. Experiment. Psychol. Human Learn. Memory 7, 311–325.

Green, S.B., Salkind, N.J., Akey, T.M., 1997. Using SPSS for Windows: Analyzing and Understanding Data. Prentice Hall, New Jersey.

Hart, J.T., 1965. Memory and the feeling-of-knowing experience. J. Education. Psychol. 56, 208–216.

Hirschberg, J., 2002. Communication and prosody: Functional aspects of prosody. Speech Commun. 36, 31–43.

Hirschberg, J., Ward, G., 1992. The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise–fall–rise intonation contour of English. J. Phonet. 20, 241–251.

Holmes, J., Meyerhoff, M., 2003. The Handbook of Language and Gender. Blackwell Publishing, Oxford.

Hopper, R., 1992. Telephone conversation. Indiana University Press, Bloomington.

Jackson, S., 1992. Message Effects Research: Principles of Design and Analysis. Guilford Press, New York.

Jefferson, G., 1989. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In: Roger, D., Bull, P. (Eds.), Conversation: An Interdisciplinary Perspective. Multilingual Matters, pp. 166–196.

Krahmer, E., Swerts, M., Theune, M., Weegles, M., 2002. The dual of denial: two uses of disconfirmations in dialogues and their prosodic correlates. Speech Commun. 36, 133–145.

Levinson, S.C., 1983. Pragmatics. Cambridge University Press, Cambridge.

Nelson, T., 1993. Metacognition: Core Readings. Prentice Hall.

O'Shaughnessy, D., 2003. Interacting with computers by voice: automatic speech recognition and synthesis. Proc. IEEE 91, 1272–1305.

Pomerantz, A., 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In: Atkinson, J.M., Heritage, J. (Eds.), Structures of Social Action: Studies in Conversation Analysis. Cambridge University Press, Cambridge, pp. 57–101.

Repp, B.H., 1982. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychol. Bull. 92, 81–110.

Roberts, F., Robinson, J.D., 2004. Interobserver agreement on first-stage conversation analytic transcription. Human Commun. Res. 30, 376–410.

Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Schegloff, E.A., 1979. Identification and recognition in telephone conversation openings. In: Psathas, G. (Ed.), Everyday Language: Studies in Ethnomethodology. Irvington, New York, pp. 23–78.

Selting, M., 1996. Prosody as an activity-type distinctive cue in conversation: the case of so-called 'astonished' questions in repair initiation. In: Couper-Kuhlen, E., Selting, M. (Eds.), Prosody in Conversation. Cambridge University Press, Cambridge.

Shimojima, A., Katagiri, Y., Koiso, H., Swerts, M., 2002. Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses. Speech Commun. 36, 113–132.

Smith, V.L., Clark, H.H., 1993. On the course of answering questions. J. Memory Language 32, 25–38.

Swerts, M., Krahmer, E., 2005. Audiovisual prosody and feeling of knowing. J. Memory Language 53, 81–94.

Wilson, T.P., Zimmerman, D.H., 1986. The structure of silence between turns in two-party conversation. Discourse Process. 9, 375–390.