Identifying a temporal threshold of tolerance for silent gaps after requests

Felicia Roberts
Brian Lamb School of Communication, Beering Hall, Purdue University, West Lafayette, Indiana 47907
froberts@purdue.edu

Alexander L. Francis
Department of Speech, Language, and Hearing Sciences, Purdue University, Heavilon Hall, West Lafayette, Indiana 47907
francisa@purdue.edu

Abstract: This study addresses whether there is a threshold, some particular length of silent gap between two speakers' turns, at which negative social attributions emerge. The effect of such inter-turn silence was tested by constructing dialogues where responses to requests were identical and affirmative so that study participants' ($n = 380$) ratings about "willingness" would be colored by lag time, not semantics. 100 ms intervals between 200 and 1200 ms were tested in a between groups design. There was a notable drop-off in ratings at 600 ms and a statistically significant difference in ratings between 700 and 800 ms.

PACS numbers: 43.72.Kb [DO]
Date Received: February 8, 2013 **Date Accepted:** April 11, 2013

1. Introduction

Research on the detection of emotion and attitude in the speech signal has developed rapidly in the past 15 years (for a detailed review, see Schuller *et al.*, 2011). The work has evolved toward the study of cues which have immediate interpersonal significance, putting relationship rather than individuals at the center of our thinking (Ranganath *et al.*, 2013). Indeed, for speech technology purposes, Scherer (2003) argues for greater understanding of vocal modifications that signal "interpersonal stance" rather than very rare or highly intense (and therefore socially regulated) emotional expressions in the voice (pp. 242–243). Similarly, Rilliard *et al.* (2009) pursue a multimodal, audiovisual approach to the study of attitude and prosody, advancing further the exploration of the audio-visual interplay between speaker-hearers. Perhaps most intriguing from an acoustics standpoint, Okada *et al.* (2012) examine how higher order pitch relationships (consonance and dissonance) can be detected in agreeable vs disagreeable conversation. In sum, acoustic and related speech recognition research is taking account of the interpersonal contexts of language use, studying how attitude, not just emotion, is expressed.

The shift to dialogically relevant cues to attitude is a crucial development. Understandably, these explorations have focused on the presence of features in the signal; however, important interpersonal information is also transmitted through the absence of signal, particularly through silence in conversation. Thus, in this study, we motivate a specific question about perceptions of "willingness" or "unwillingness" to comply with requests as cued by different lengths of inter-turn silence. In Scherer's terms, the "interpersonal stance" at stake is one of supportiveness (Scherer, 2003, p. 243). To address the question of silence as a relevant cue to interpersonal support, we first describe the turn-taking model for conversation that provides the theoretical and empirical grounding for our research design and predictions about inter-turn silence and interlocutors' stance.

Nearly 40 years ago, Sacks *et al.* (1974) described several "grossly apparent facts" about conversation, and from those observations derived what has become the most widely used, empirically grounded model of turn-taking for studies of communication (Roberts and Robinson, 2004). Their model addresses the way in which turns at talk are constructed, ordered, and distributed among speakers. Most important for the current study is their identification of the "transition relevance place" or the moment when one unit of talk comes to an end and a next unit begins (initiated by the same or another speaker.) This deceptively simple moment is actually the locus of intense interpersonal activity, and such transitions are accomplished effortlessly, seamlessly, and without notice billions of time a day across the globe. Indeed, the natural transition time between speakers is relatively consistent, approximately 250 ms, across diverse languages and cultures (Stivers *et al.*, 2009.) Furthermore, it has long been known from qualitative research that when there is a brief gap after a turn where response is expected (e.g., after requests or invitations), a first speaker will treat this as some form of trouble, evidenced in his or her pursuit of response (Davidson, 1984; Pomerantz, 1984).

Using experimental methods to explore this phenomenon in typologically distinct languages, Roberts *et al.* (2011) found that negative judgments increase with increasing lengths of silence, regardless of language background. Thus there is clearly something generalized across cultures in both routine (non-delayed) timing of responsive turns (Stivers *et al.*, 2009) and in negative perception of delays in turn transition (Roberts *et al.*, 2006; Roberts *et al.*, 2011.) The similarity of judgments across languages and cultures suggests that there may be something basically perceptual, not language-specific, about the assessment of inter-turn silent delays of response. However this assumption is based on evidence from only three silence conditions (no gap, 600 ms, and 1200 ms), precluding a truly informed extrapolation to a possible decision curve. Importantly, if there is a discontinuous decision curve, identifying the location of that discontinuity could help in refining theories of how cognitive processing and acoustic cues are related in the detection of an interpersonal stance such as supportiveness.

In sum, prior research tells us that turn-transition gaps matter and that increasing gap length is relevant cross-linguistically, but the question remains: how long does a silent gap have to be before people perceive that there is trouble in the interaction? Is there a threshold after which negative judgments take hold or is the response to silent gaps simply a continuous function of increasing silence duration? To answer these questions, we tested every 100 ms interval from 200 to 1200 ms, isolating the effect of the silence by constructing dialogues where responses to requests were identical and affirmative such that compliance was lexically clear and specific. Judgments about "agreeability" would thus be colored by lag time, not by the semantics of the response. While lack of supportiveness is not the only possible attribution to be made when gaps are present, it is a salient dimension, and well grounded in both qualitative and quantitative research, as noted above. By focusing on this particular attribution, we could best control for the effect of speech act and a variety of semantic complexities inherent in discourse based research.

2. Method

2.1 Participants

Three-hundred and eighty undergraduate students were compensated $2.00 each for completing a 6 min listening task. Ages ranged from 18 to 32 (mean = 21.22) with 196 men and 184 women participating. This experiment was conducted in accordance with a protocol approved by the Human Research Subjects Protection Program of Purdue University.

2.2 Procedure

In a between groups design, participants were recruited by classroom. Each classroom ($n = 22$) received the same recorded instructions: Listeners were told they would hear

"several telephone conversations among a group of friends" and that each friend "was just relaxing at home." Each simulated telephone call was approximately 10 s long and ended with the caller formulating a request, invitation, or assessment to which the call recipient responded affirmatively. There were 15 such conversations: 11 were distracters and 4 (the requests) were targets for analysis. Each of these target dialogues had the identical affirmative response edited into it (see Sec. 2.3).

Following each dialogue, study participants answered a related question on a six-point Likert-type scale to indicate their perception of a call recipient's willingness or agreement or enthusiasm for, respectively, the request, invitation, or assessment embedded in the dialogue. There was only one scale for each dialogue and each was presented with the same polarity (i.e., lower ratings were always left-anchored.) Thus, we privileged one possible interpretation of the inter-turn silences (e.g., willingness or lack thereof) rather than, for example, friendliness or comprehension. Our aim was not to assess the various colorings that the gaps might engender, but simply to address one particular and salient dimension. Because we were targeting the request sequences to control for the effect of speech act, higher ratings reported here indicate stronger perceptions of willingness to comply with a request.

2.3 Construction of stimuli

The conversations used as stimuli were simulated telephone calls based on the transcription of an actual telephone call among friends of an age similar to the study population (Roberts and Robinson, 2004). To control for any effect from asymmetries of relationship, it was important to convey that the conversationalists were friends of long-standing. This was conveyed by features such as omission of the caller's name, using an informal register, and truncating the greeting sequence (Hopper, 1992; Schegloff, 1979).

To control for possible confounding from the acoustic qualities of the various productions by the actors in the original dialogues, the identical affirmative response token ("sure") was edited into each target conversation as were the silent gaps (described below). This token was selected from all of the "sure" tokens produced naturally during dialogue recording as the one with median duration, pitch range, and direction of pitch change (335 ms in length, with a generally falling f_0 contour from 325 to 213 Hz; see Roberts *et al.*, 2006, for additional detail). Because the same trained actors were used across all of the dialogues, and because the length of each target dialogue was identical (32 words), there was very little variation in speech rate within or across the conversations. Dialogues ranged from 148 to 152 wpm, with the selected "sure" token being drawn from one with a rate of 152 wpm.

Once the median token was chosen and edited into all request dialogues, silence, taken from other dead space in the dialogue (i.e., not machine-produced silence) was then inserted between the end of the request utterance and the affirmative response, thus producing controlled lengths of inter-turn silence. Durations from 200 to 1200 ms in 100 ms steps were inserted between the target requests and responses. For distracter dialogues, there was no manipulation of the actors' naturally produced timing or intonation. Inter-turn gaps preceding the final assent in each distracter dialogue ($n = 11$) ranged from 0 ms (no gap) to 684 ms, with a mean of 190 ms. All measurements and editing were done by the first author using Praat (Boersma and Weenink, 2001).

2.4 Experimental design

Eleven inter-turn silence intervals (from 200 to 1200 ms) were tested with inter-turn silence length (*gap*) as a repeated measure within and between groups. Adjacent gap lengths (e.g., 300 and 400 ms) were repeated within a group and each measure was repeated between successive groups. In other words, group A received stimuli with 200 and 300 ms gaps, group B received stimuli with 300 and 400 ms gaps, group C received stimuli with 400 and 500 ms gaps, and so on. Although different groups were presented different gap lengths, all groups heard all of the same conversations. To control for

order effects, classrooms were paired on the gap length conditions and each classroom in the pair was presented the same stimuli but in reverse order (i.e., A_1 received 200 and 300 ms gaps, and A_2 received 300 and 200 ms gaps).

In order to ensure that listeners did not notice that the only difference between dialogues was in terms of silence gap duration, and to allow for repeated testing of the same duration in the same immediate phonetic and lexical context within a listener group, different dialogue frames were used. For example, for group A, the first exposure to the 300 ms gap condition was in a conversation about going to the store, and the request was asking for a ride "over there." The second exposure to the 300 ms gap in group A was in a conversation about posting flyers for a school function and the request was also for a ride "over there" to pick up the flyers.

Despite the repetition of the same request formulation on each of the four targets, a debriefing questionnaire indicated that participants overwhelmingly (97%) thought the study was about "tone of voice" or other attributions such as "hiding your true feelings" or "differences between men and women" that were not the targeted phenomena. The debriefing questionnaire was completed by a convenience sample ($n = 123$; 32%), but all gap length conditions were sampled.

3. Results

Prior to core analyses, an omnibus one-way analysis of variance was used to test for any differences between the samples from different classrooms that were exposed to the same gap conditions. Results predictably revealed an overall between-groups effect of silence, $F(19,\ 740) = 17.868$, $p < 0.001$, but none of the *post hoc* comparisons (Bonferroni correction) between the reversed order administrations of the same gap lengths (i.e., between A_1 and A_2) were statistically significant. Thus, we combined ratings across classrooms for the targeted gap lengths.

Capturing the main effect of gap length, Fig. 1 shows that as the length of an inter-turn silence increases, ratings of perceived willingness decline. The effect of gap is significant, $F(19,\ 740) = 32.13$, $p < 0.001$, $\eta^2 = 0.300$, and is consistent with prior research (Roberts *et al.*, 2006; Roberts *et al.*, 2011).

The specific aim of the current study was, however, to determine if response to silent gaps between requests and affirmative responses was simply a continuous function of increasing silence. To address this question, we compared means of temporally adjacent gap conditions (in 100 ms increments). Figure 1 presents this analysis by using a solid line to indicate the slope of change between adjacent means and error bars which indicate 95% confidence intervals of the adjacent means.

To explore whether each 100 ms step increase in gap length was significant, Bonferroni-corrected pairwise comparisons were examined and revealed no significant differences between ratings for the 200 ms gap and those for any gap duration below 600 ms; however, the difference between ratings for the 200 ms gap and the 600 ms gap were significant ($M_{\text{diff}} = 0.42$, $p = 0.004$). This suggests that the intervening 100 ms increments up to 600 ms may not be particularly salient for extracting social meaning, at least in the mundane context of requests between friends and in the absence of other secondary cues.

Examining the territory in the vicinity of the 600 ms gap, it appears that the distinction in ratings for the 500 and 600 ms gap lengths is relatively minor. The mean difference is 0.11 (n.s.) and the slope of the line is shallow ($y = 0.0011$). However, on the other side of the 600 ms gap, the ratings drop further and more quickly from 600 to 700 ms. The change is not statistically significant, but the mean difference ($M_{\text{diff}} = 0.22$) is twice that of the 500–600 ms comparison and the slope is steeper as well ($y = 0.0022$). It is clear in the interval from 700 to 800 ms that listener ratings of "willingness" drop off significantly ($M_{\text{diff}} = 0.39$, $p < 0.001$). The mean difference between 800 and 900 ms is similar and worth noting ($M_{\text{diff}} = 0.35$, $p = 0.279$), but after 900 ms, the effect of the gap flattens out considerably and there are no significant pairwise comparisons between temporally adjacent gap lengths.

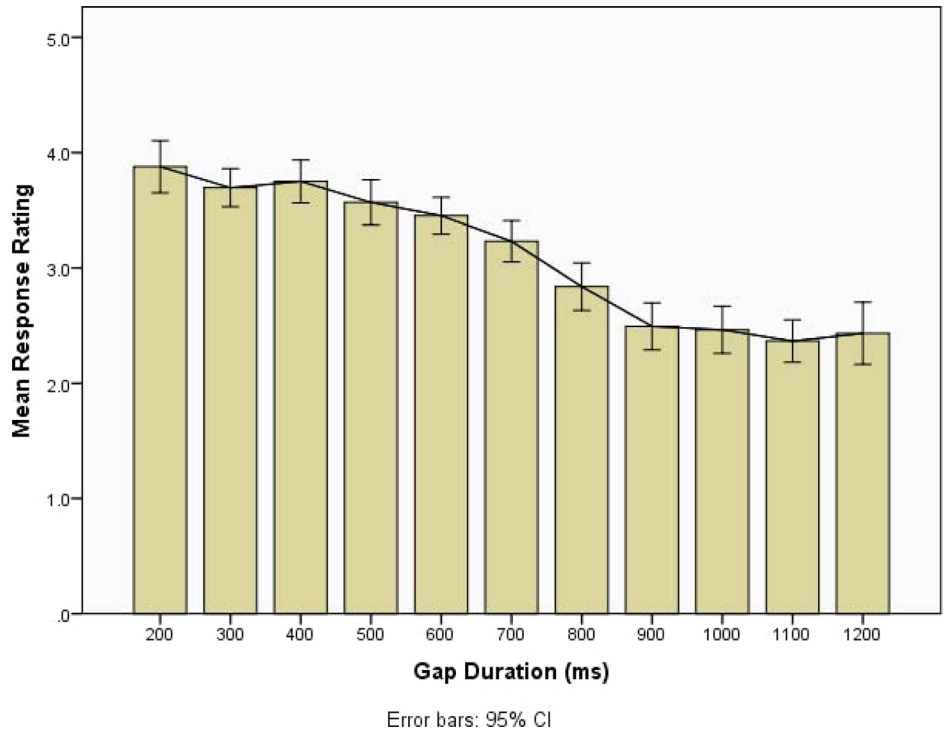

Fig. 1. (Color online) Mean listener ratings of "willingness" to comply with a request, according to duration of silent gap between requests and affirmative response. Judgment scale ranged from 1 "not willing" to 6 "very willing." The dark line links mean values and represents slope of the decision curve. The dotted line represents comparison of responses to adjacent gap durations. Error bars indicate standard error of the mean.

4. Discussion

Following developments in speech perception research oriented to the study of cues to interpersonal stance, the current experiment was designed around the idea that inter-turn silence provides an entry point for examining the dynamic relationship between acoustic cues and social organization. To do this, we gathered raters' judgments about another's willingness to help a friend with a mundane task. The investigation centered on an acoustic property of naturally occurring interaction—the absence of signal—which can be a cue to higher level discourse meaning. In the stimuli constructed for the present study, we were able to provide precise and replicable measurements for the long-observed relationship between "trouble" in conversation and inter-turn silence (Davidson, 1984; Jefferson, 1989; Pomerantz, 1984.)

Results from the manipulations of gap length suggest that there is a non-linear relationship between this cue and the social perception of a stance akin to "reluctance." In the absence of other vocal or visual cues, there is clearly a critical range for the effect of the inter-turn silence after requests: perceived willingness is consistently higher before 500 ms, begins to drop after 600 ms, and then clearly and significant steps down from 700 to 800 ms. After 900 ms, in the absence of other cues, there appears to be a simple floor effect. These findings coincide with results from an observational study that measured gap durations in actual conversation. That research indicates that a recipient who hears less than 700 ms of silence would most likely be unable to guess whether a response will be positive or negative on the basis of silence alone, whereas a gap of 700 ms or more could allow the person to anticipate that a negative response is forthcoming (Kendrick and Torreira, 2012). The addition of other acoustic or visual cues could further erode perceptions of willingness, and this remains an area

for future research. For example, the effect of speech rate should be considered, as perception of gaps as "too long" will likely vary relative to the surrounding talk. (However, for a discussion of the challenges of pursuing such research cross-linguistically, see Roberts *et al.*, 2011.)

With the caveat that 600 ms should not be treated as an absolute value, particularly since we were only testing the effect in one language group (American English speakers), the possibility that 600 ms marks a point at which social attributions emerge coincides nonetheless with evidence from a meta-analysis of imaging research. That line of work concludes, *inter alia*, that the time from presentation of a picture stimulus to articulation of an appropriate word is 600 ms (Indefrey and Levelt, 2004). Admittedly, such a confrontation naming task and our listeners' rating tasks are very different; however, we submit that the evidence reviewed in Indefrey and Levelt (2004), indicates that 600 ms may be something of a lower bound for tasks that require a single word response *when the thing to be named or the response to be given is not anticipated*. This is, in fact, the general cognitive framework we create by using conversations in which a friendly greeting and quick report of some event is made, without any particular "clue" that a request is coming. Arguably, this simulates the same kind of unanticipated event that a picture-naming task might engender. In other words, our study participants are hearing conversations and making judgments about willingness in the context of unexpected requests.

Because listener-responders are presumably projecting the upcoming end of an interlocutor's turn in preparation to talk, any lapse longer than the average transition time of 250 ms (Stivers *et al.*, 2009) might signal some additional cognitive processing; however, the strongest negative attributions occur after the 600 ms lapses, when the lag is likely no longer related to the unexpected nature of the request. Thus, while a gap around 600 ms is clearly perceptible and viewed as less agreeable than having no gap at all (Roberts *et al.*, 2006), data from the current study show that when the delay extends *beyond* 600 ms, the additional processing may be interpreted as "longer than needed" and presumably (in the present case) related to the respondent's reluctance to agree. In other words, in light of the findings from confrontation naming tasks, the delay in response is perhaps no longer attributable to the unexpectedness of the request (i.e., there is no modeling by listeners of a cognitive "excuse") and speakers are held socially accountable for their delays.

We conclude, therefore, that our findings of social attribution converge in intriguing ways with cognitive processing research, helping to delimit temporal boundaries for further exploration of the relationship between acoustic signals, neural-cognitive processing, and social judgments. As Enfield (2013) argues, the interpretive system is distributed across body, brain, time, and place, a presumption that should motivate further research on the dynamic interplay of social and acoustic signals for the study of attitude and stance. Working to take account of another's possible goals, beliefs, and so on during "implicit metacognition" (Frith, 2012, pp. 2214–2215) our study participants' ratings provide a glimpse, through their increasingly negative judgments, of a moment where social information is built on implicit recognition of the temporal constraints of information processing. Listeners are inherently attentive to temporality as a source of information, and at the discourse level inter-turn silence is a robust signal. It is a dynamic factor whose interpretation will vary depending on interlocutors' activities, identities, goals, and so on, but the door is open for modeling more complex interactions with a range of acoustic, social, and semantic cues that are undoubtedly relevant to signaling stances of interpersonal import.

Acknowledgments

Financial assistance to compensate study participants was provided by the Linguistics Program of Purdue University. An earlier version of this paper was presented in April 2012 by the first author to the Language & Cognition Group of the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. That group helped defray travel costs

for the presentation, and the insight and encouragement arising out of meeting with faculty and staff was invaluable, especially from (in alphabetical order) Joe Blythe, Nick Enfield, Kobin Kendrick, Stephen Levinson, and Francisco Torreira.

References and links

Boersma, P., and Weenink, D. (**2001**). "Praat: A system for doing phonetics by computer," retrieved from University of Amsterdam, Institute of Phonetics Sciences, http://www.fon.hum.uva.nl/praat/ (Last viewed April 23, 2013).

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

Enfield, N. J. (**2013**). *Relationship Thinking: Enchrony, Agency, and Human Sociality* (Oxford University Press, New York) (in press).

Frith, C. D. (**2012**). "The role of metacognition in human social interactions," Philos. Trans. R. Soc. London, Ser. B **367**, 2213–2223.

Hopper, R. (**1992**). *Telephone Conversation* (Indiana University Press, Bloomington).

Indefrey, P., and Levelt, W. J. M. (**2004**). "The spatial and temporal signatures of word production components," Cognition **92**, 101–144.

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

Kendrick, K. H., and Torreira, F. (**2012**). "The timing and construction of preference: A quantitative study," in *Discourse, Communication, and Conversation Conference*, March 22, 2012, Loughborough University, Loughborough, England, UK.

Okada, B. M., Lachs, L., and Boone, B. (**2012**). "Interpreting tone of voice: Musical pitch relationships convey agreement in dyadic conversation," J. Acoust. Soc. Am. **132**, EL208–EL214.

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

Ranganath, R., Jurafsky, D., and McFarland, Daniel, A. (**2013**). "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," Comput. Speech Lang. **27**, 89–115.

Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D., and Aubergé, V. (**2009**). "Multimodal indices to Japanese and French prosodically expressed social affects," Lang. Speech **52**, 223–243.

Roberts, F., Francis, A. L., and Morgan, M. (**2006**). "The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation," Speech Commun. **48**, 1079–1093.

Roberts, F., Margutti, P., and Takano, S. (**2011**). "Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese," Discourse Process. **48**, 331–354.

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

Sacks, H., Schegloff, E. A., and 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 *Everyday Language: Studies in Ethnomethodology*, edited by G. Psathas (Irvington Press, New York), pp. 23–78.

Scherer, K. R. (**2003**). "Vocal communication of emotion: A review of research paradigms," Speech Commun. **40**, 227–256.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (**2011**). "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Commun. **53**, 1062–1087.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K. E., and Levinson, S. C. (**2009**). "Universals and cultural variation in turn-taking in conversation," Proc. Natl. Acad. Sci. U.S.A. **106**, 10587–10592.