Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Our own speech rate influences speech perception

Hans Rutger Bosker [1, 2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands

HansRutger.Bosker@mpi.nl

Abstract

During conversation, spoken utterances occur in rich acoustic contexts, including speech produced by our interlocutor(s) and speech we produced ourselves. Prosodic characteristics of the acoustic context have been known to influence speech perception in a contrastive fashion: for instance, a vowel presented in a fast context is perceived to have a longer duration than the same vowel in a slow context. Given the ubiquity of the sound of our own voice, it may be that our own speech rate - a common source of acoustic context - also influences our perception of the speech of others. Two experiments were designed to test this hypothesis. Experiment 1 replicated earlier contextual rate effects by showing that *hearing* pre-recorded fast or slow context sentences alters the perception of ambiguous Dutch target words. Experiment 2 then extended this finding by showing that *talking* at a fast or slow rate prior to the presentation of the target words also altered the perception of those words. These results suggest that between-talker variation in speech rate production may induce between-talker variation in speech perception, thus potentially explaining why interlocutors tend to converge on speech rate in dialogue settings.

Index Terms: speech rate; rate normalization; self-monitoring; phonetic convergence

1. Introduction

Speech perception typically takes place in a rich acoustic context. Words, uttered by our interlocutor, are heard in the context of earlier speech from that same person (e.g., sentence context), speech we produced ourselves moments earlier (e.g., a question that elicited our interlocutors' utterance), and other non-speech acoustic signals. It has long been recognized that prosodic properties of the acoustic context do not merely serve as background noise, but rather influence subsequent speech perception [1, 2]. For instance, the perception of an ambiguous Dutch vowel midway between short /ɑ/ and long /a:/ may be shifted towards the perception of long /a:/ by presenting it in a context sentence with a fast speech rate [3].

The contrastive influence of contextual speech rate, known as rate normalization, is taken to be a general auditory process [4]. Rate normalization takes place as soon as the target sound is heard [5], may be elicited by non-speech contexts (e.g., tone sequences; [6]), operates independent of cognitive load [7], and interestingly also generalizes across talkers. That is, the perception of talker A is influenced by the speech rate produced by talker B [8, 9].

Given that rate normalization occurs across talker-incongruent speech streams, one may consider yet another source of prosodic context influencing speech perception, namely *the sound of our own voice*. In typical conversations, our own utterances and those of others follow each other in rapid succession. Cross-linguistic research shows that interlocutors universally try to minimize the silence between conversational turns, with a median turn transition duration of approximately 100 ms [10]. As such, the immediate acoustic context of an utterance spoken by our conversational partner includes speech that we produced ourselves some moments earlier, potentially allowing for contextual effects of our own voice on the way we perceive others.

Previous research has tried to find effects of our own habitual speech rate on the way we *evaluate* speech rates produced by other talkers, but the evidence remains tenuous. For instance, Koreman [11] failed to find any effect of listeners' own habitual speech rates on speech rate evaluation. Schwab [12] did show that habitually slow talkers judge speech as faster than listeners with a relatively fast habitual speech rate, but the effect was only observed with slow and neutral rates, not with fast speech. Also, note that both these studies used *explicit* judgments of perceived speech rate, which do not always reflect the acoustic speech rate: acoustic measures of speed of articulation have been found to only explain 53% of the variance of perceived speed judgments [13]. Therefore, the present study targets more local effects of self-produced speech rate on *implicit* rate normalization. That is, does talking at a fast (or slow) speech rate change one's perception of a subsequent utterance, spoken by someone else?

Given the findings on talker-independent rate normalization [8, 9], one may expect a positive answer to this question. However, perception of our own voice differs from perception of other talkers in the fact that self-perception takes place during the execution of a simultaneous task, namely speech production. Neurocognitive studies report differences between perception-during-production and perception-without-production. For instance, activity in the auditory cortex in response to self-produced speech is attenuated relative to hearing tape-recorded speech [speaking-induced suppression; 14]. This attenuation has been attributed to internal forward models that internally simulate the sensory consequences of speech motor actions [15]. Moreover, auditory responses during speech production are not only significantly inhibited, but they have also been found to be slightly delayed [16]. The attenuation and temporal disruption of speech processing during speech production may, in turn, potentially reduce any context effects elicited by our own voice.

 doi: 10.21437/SpeechProsody.2016-47

In fact, if our own speech rate would influence our perception of others, this would introduce considerable variation to the speech perception system. Communication between two interlocutors, with talker A speaking fast and talker B speaking slow, would suffer substantially: speaker A would interpret the slow speech of speaker B relative to his/her own fast speech rate (and vice versa). Speech rate varies considerably both between individuals [17, 18] and within a given speaker; for instance, depending on conversational register, emotions, the length of utterances [18], age [19], etc. Thus, if our own speech rate production would influence our speech perception, it would be a substantial source of variation in speech comprehension.

The present study reports two experiments that targeted effects of preceding slow or fast speech rate on the perception of the Dutch minimal vowel contrast /ɑ/ - /a:/. Experiment 1 aimed to replicate the standard finding of rate normalization. Participants heard manipulated target words, with vowels ambiguous between /ɑ/ and /a:/, embedded in fast or slow pre-recorded context sentences. Fast context sentences were expected to bias perception of the target vowel towards /a:/, and slow context sentences towards /ɑ/.

In Experiment 2, instead of playing pre-recorded context sentences, participants were instructed to produce the context sentences themselves, both at a fast and a slow rate, after which the same manipulated target words from Experiment 1 were played. If self-produced speech serves as acoustic context to the speech of others, the perception of target words may be influenced by the rate at which the participants produced the context sentences themselves. However, as discussed, the effect of self-produced speech rate on the perception of others may also be modulated by speaking-induced suppression of speech perception processes during production.

2. Method

2.1. Experiment 1

2.1.1. Participants

In order to allow for within-subject analyses, the same sample of 45 native Dutch participants with normal hearing was tested in both experiments.

2.1.2. Design and materials

A female native speaker of Dutch was recorded producing the sentence: *Freek ging het hok eerst in en toen weer uit en zei dus het woord...* [target]; "Freek first went into the hut and then out again and therefore said the word... [target]". The sentence did not favor any of the target words semantically and did not contain any /ɑ/ or /a:/ vowels. The sentence ended in one of several monosyllabic target words that either had the short vowel /ɑ/ or the long vowel /a:/. Targets were selected from six minimal word pairs: *zat-zaad* (sat-seed), *Stan-staan* (Stan-stand), *dat-daad* (that-deed), *stad-staat* (city-state), *staf-staaf* (staff-bar), and *zak-zaak* (bag-shop).

From these recordings, context sentences were excised (all speech up to target onset). One token near the speaker's median rate was linearly compressed/expanded into a slow version (ratio = 1.33; total duration 4055 ms) and a fast version (ratio = 0.75; total duration 2512 ms) using PSOLA in Praat [20].

From the same recordings, target words were also excised. One long vowel /a:/ was selected for manipulation. Since the Dutch /ɑ/ - /a:/ vowels are contrastive in both spectral and temporal characteristics, a two-dimensional continuum was created from this one vowel token, comprising 7 duration values and 7 F2 values, all falling within the speaker's natural range. Spectral manipulations were based on Burg's LPC method (implemented in Praat), with the source and filter models estimated automatically from the selected vowel. Filter models were adjusted to have a constant F1 value (739 Hz, ambiguous between /ɑ/ and /a:/) and one of seven desired F2 values (1300 - 1600 Hz in steps of 50 Hz). Source and filter models were then recombined and the new vowels were adjusted to have their original overall amplitude. Based on these spectrally manipulated vowels, duration continua (110 - 170 ms in steps of 10 ms) were created using PSOLA. Finally, the resulting 49 vowel tokens were spliced into the consonantal frames from the six target pairs.

A categorization (2AFC) pretest was run with 26 native Dutch listeners, who categorized the manipulated target words *in isolation* (i.e., without any preceding context) as containing either the short vowel /ɑ/ or the long vowel /a:/. Based on this pretest, three vowel tokens with different F2 values but identical duration (140 ms) were selected for the following experiments, each sampling a different point from the categorization curve: token 1, F2 = 1300 Hz, 27% long vowel categorization; token 2, F2 = 1450 Hz, 48% long vowel categorization; and token 3, F2 = 1550 Hz, 67% long vowel categorization. Finally, all target words were combined with the two (fast and slow) context sentences, adding up to a total of 108 items (2 context rates × 6 target pairs × 3 vowel tokens, all repeated 3 times).

2.1.3. Procedure

Speech stimuli were presented in a fixed random order, with the reversed order presented to half of the participants.

For purposes of comparability across the two experiments, visual displays were identical across both experiments (see Figure 1). Each trial started with a screen showing a (horizontal) hourglass running empty in 5 seconds (from right to left). Above the hourglass, the rate of the context sentence was displayed (*SNEL* "FAST" vs. *TRAAG* "SLOW"). A mark on the hourglass indicated the time point of context sentence onset: early in the case of slow contexts (945 ms after hourglass onset), late in the case of fast contexts (2488 ms after hourglass onset). The hourglass always ran empty at context sentence offset, after which the target word was played. The screen was replaced by two response options after target offset and participants were instructed to indicate what sentence-final target word they had heard: *dat* or *daad*, *zak* or *zaak*, etc. The position of words (left or right) was counter-balanced across participants, who gave their response by pressing "1" for the left word or "0" for the right word. If participants did not respond within 5 seconds, a missing response was recorded and the next trial was presented.

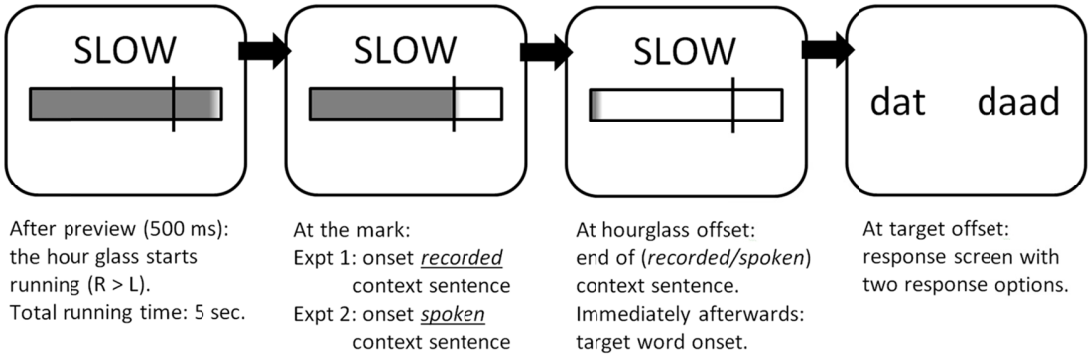


Figure 1: *Schematic diagram of the visual display used in both experiments.*

2.2. Experiment 2

2.2.1. Procedure

After having been familiarized with the speech stimuli, the two speech rates, and the timing of the speech relative to the hourglass in Experiment 1, participants subsequently took part in Experiment 2. This second experiment was identical to Experiment 1, except that participants were now instructed to produce the context sentences themselves. That is, participants imitated the previously heard fast and slow context sentences, crucially, *without* the sentence-final target word. The rate at which the context sentence was to be produced could be gleaned from the rate displayed above the hourglass. Participants were instructed to imitate the rates from Experiment 1 as much as possible and to start speaking at the time point indicated by the mark on the hourglass and to finish when the hourglass ran empty. When the hourglass ran empty, the pre-recorded target words from Experiment 1 were presented. The words of the sentence were not displayed on the screen, but had to be recited from memory. To remind participants, the words of the sentence were displayed on the screen after every sixth trial, but disappeared again for the next trial.

3. Results

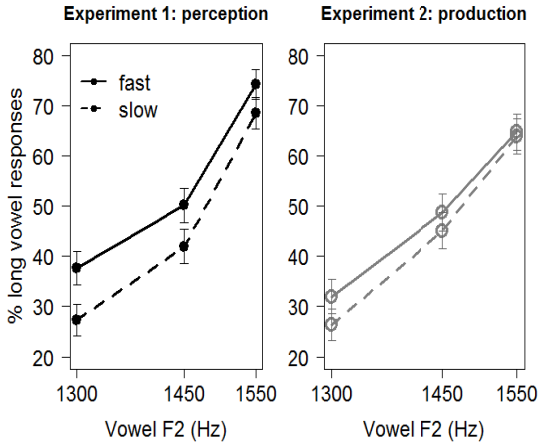
Categorization data, calculated as the proportion of long vowel responses (%long), for each experiment are displayed in Figure 2. The left panel, with the data from Experiment 1, shows that participants reported more long vowels when the target vowel had a higher F2. Moreover, the difference between the two lines suggests that *hearing* a preceding context sentence with a fast speech rate (solid line) biased listeners' perception towards a long vowel. Similarly, the right panel, with the data from Experiment 2, suggests that *speaking* at a fast speech rate also biased listeners' perception towards a long vowel, though the distance between the two lines would seem to be smaller compared to Experiment 1.

A Generalized Linear Mixed Model (GLMM; [21]) with a logistic linking function as implemented in the lme4 library [22] in R [23] tested the binomial responses collected in both experiments for fixed effects of Vowel F2 (continuous predictor, rescaled around the mean), Rate Condition (categorical predictor, intercept is slow), and Experiment (categorical predictor, intercept is Experiment 1), and their interactions, with random effects of Participants and Items (the six different word pairs). Only a by-participant random slope for Rate Condition was included because models with more complex random effects structures failed to converge.

This model revealed significant effects of Vowel F2 (the higher the vowel's F2, the higher the proportion of long vowel responses; $\beta = 1.026$, $z = 18.379$, $p < 0.001$), and Rate Condition (higher proportion of long vowel responses in the fast condition; $\beta = 0.528$, $z = 6.586$, $p < 0.001$). Also, a significant interaction between Rate Condition and Experiment was observed: the effect of Rate Condition was significantly smaller in Experiment 2 ($\beta = -0.363$, $z = -3.399$, $p < 0.001$).

This interaction between Experiment and Rate Condition, however, does not tell us whether the rate effect in Experiment 2 was *eliminated altogether* or whether it was *only reduced*. Therefore, a mathematically equivalent model was built that took Experiment 2 as its intercept. This model revealed that, despite the reduction in effect size, the fast condition was still significantly different from the slow condition in Experiment 2 ($\beta = 0.165$, $z = 1.967$, $p = 0.049$). That is, the Rate Condition of the context sentence affected vowel categorization *in both experiments*, but to a lesser degree in Experiment 2.

Figure 2: *Average categorization data for each experiment (error bars enclose 1.96 × SE on either side, 95% CIs).*



4. Discussion

This study compared the contextual effect of hearing vs. producing a fast or slow speech rate on the perception of a Dutch vowel contrast. Experiment 1 replicated earlier studies on rate normalization showing that *hearing* a fast speech rate changes the perception of a subsequent ambiguous vowel. Experiment 2 extends our understanding of contextual rate effects by showing that *producing* a fast speech rate ourselves changes our perception of vowels produced by someone else.

The effect of self-produced speech rate (Experiment 2) was found to be reduced relative to the effect of perceived speech rate (Experiment 1). This reduction may be explained by two accounts. First, there may have been more variability in the elicited speech rates produced by participants themselves, than in the categorically fast and slow speech materials used in Experiment 1. As such, the acoustic difference between the two rate conditions in Experiment 2 may have been smaller, thus reducing the rate effect. Further analysis of the speech produced by participants in Experiment 2 may reveal whether this account can be corroborated.

Alternatively, the reduction of the rate effect in Experiment 2 may be explained by the different task demands. The rate effect in Experiment 1 was elicited by listening to speech produced by someone else, whereas the rate effect in Experiment 2 was elicited by perception of one's own voice during speech production. Neurocognitive literature suggests that the processes involved in perception-during-production are attenuated relative to those involved in perception-without-production (speaking-induced suppression; [14-16]). Thus, the dual task of perceiving during speech production may have attenuated the perception of the self-produced speech rate, which in turn led to a reduced contextual rate effect. New experiments are currently ongoing to distinguish between different explanations of the smaller contextual rate effect during production.

The ubiquity of the sound of our own voice implies that it forms a great part of the acoustic context in which speech from other speakers occurs. It is therefore not surprising that earlier studies have already targeted influences of listeners' own *habitual* speech rate on *explicit* rate evaluation judgments, but unfortunately with equivocal results [11, 12]. By testing more local effects of self-produced fast and slow context sentences, the present study is the first to reveal direct effects of talking at a fast or slow rate on the way we perceive others. However, since only local effects of self-produced speech rate were tested, the current data do not tell us whether *habitually* slow speakers will perceive the same speech signal differently from *habitually* fast speakers. This remains an open, and intriguing, question for further investigation.

Nevertheless, the finding that talking at a fast pace changes our perception of a subsequent utterance already carries strong implications for our understanding of speech perception and communication in dialogue in general. The extensive variation in speech rate both between and within individuals, combined with the ever-present sound of our own voice, would be expected to induce variation in speech perception, and hence be a source of miscommunication in many dialogue situations. This is where the present study may provide a novel rationale behind the phenomenon of phonetic convergence.

When interlocutors engage in spoken communication, they tend to converge on phonetic/prosodic features of their speech, such as pitch [24], intensity [25], voice onset time [26], and also speech rate [27-30]. Different accounts have been proposed for *how* phonetic convergence arises (self-regulatory convergence dependent on social-motivational factors [31, 32]; brain-to-brain coupling [33]; automatic convergence based on priming [34]), but the *purpose* of convergence has consistently been sought in the social domain. That is, people tend to converge phonetically in order to reduce social distance and facilitate social integration, approval, and conformity [30, 35, 36].

The present study proposes a novel purpose of phonetic convergence, namely to serve speech comprehension. Given the current findings that one's own speech rate influences the perception of the other speaker, communication between speakers with highly divergent speech rates would be predicted to suffer from these cross-talker context effects. Of course, top down information, such as semantic context, may help to avoid misinterpretation of the spoken signal. However, in the absence of such information, comprehension, and hence communication, would be facilitated if interlocutors would try to converge their speech rates, thus minimizing the interference from their own speech rate (in line with findings that phonetic convergence promotes comprehensibility; [37, 38]). Therefore, phonetic convergence on speech rate may not only provide social advantages but may also reduce adverse effects of one's own (divergent) speech rate on the comprehension of the other talker.

5. Acknowledgements

I would like to thank Antje Meyer for useful comments and suggestions. Thanks to Ronald Fischer and Johan Weustink for technical support, to Anne van Hoek for her help in testing participants, and to Annelies van Wijngaarden whose voice was recorded for the speech materials used in this study. This research was supported by a Gravitation grant from the Dutch Government to the Language in Interaction Consortium.

6. References

[1] P. Ladefoged and D. E. Broadbent, "Information conveyed by vowels," *The Journal of the Acoustical Society of America,* vol. 29, pp. 98-104, 1957.

[2] J. L. Miller and A. M. Liberman, "Some effects of later-occurring information on the perception of stop consonant and semivowel," *Perception & Psychophysics,* vol. 25, pp. 457-465, 1979.

[3] H. R. Bosker and E. Reinisch, "Normalization for speechrate in native and nonnative speech.," in *Proceedings of the 18th International Congress of Phonetic Sciences 2015 [ICPhS XVIII], Glasgow*, M. Wolters, J. Livingstone, B. B., R. Smith, M. MacMahon, J. Stuart-Smith*, et al.*, Eds., 2015.

[4] M. J. Sjerps and E. Reinisch, "Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 41, pp. 710-722, 2015.

[5] E. Reinisch and M. J. Sjerps, "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," *Journal of Phonetics,* vol. 41, pp. 101-116, 2013.

[6] T. Wade and L. L. Holt, "Perceptual effects of preceding nonspeech rate on temporal properties of speech categories," *Perception & psychophysics,* vol. 67, pp. 939-950, 2005.
[7] H. R. Bosker, E. Reinisch, and M. J. Sjerps, "Listening under cognitive load makes speech sound fast.," in *Proceedings of the Speech Processing in Realistic Environments [SPIRE] workshop, Groningen*, H. v. d. Heuvel, B. Cranen, and S. Mattys, Eds., 2016, pp. 23-24.
[8] R. S. Newman and J. R. Sawusch, "Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another," *Journal of Phonetics,* vol. 37, pp. 46-65, 2009.
[9] J. R. Sawusch and R. S. Newman, "Perceptual normalization for speaking rate II: Effects of signal discontinuities," *Perception & psychophysics,* vol. 62, pp. 285-300, 2000.
[10] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann*, et al.*, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences,* vol. 106, pp. 10587-10592, 2009.
[11] J. Koreman, "Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech," *The Journal of the Acoustical Society of America,* vol. 119, pp. 582-596, 2006.
[12] S. Schwab, "Relationship between speech rate perceived and produced by the listener," *Phonetica,* vol. 68, pp. 243-255, 2011.
[13] H. R. Bosker, A.-F. Pinget, H. Quené, T. J. M. Sanders, and N. H. De Jong, "What makes speech sound fluent? The contributions of pauses, speed and repairs," *Language Testing,* vol. 30, pp. 157-175, 2013.
[14] J. F. Houde, S. Nagarajan, K. Sekihara, and M. M. Merzenich, "Modulation of the auditory cortex during speech: an MEG study," *Journal of Cognitive Neuroscience,* vol. 14, pp. 1125-1138, 2002.
[15] J. F. Houde and S. S. Nagarajan, "Speech production as state feedback control," *Frontiers in Human Neuroscience,* vol. 5, 2011.
[16] J. Numminen and G. Curio, "Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex," *Neuroscience letters,* vol. 272, pp. 29-32, 1999.
[17] Y.-C. Tsao and G. Weismer, "Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component," *Journal of Speech, Language, and Hearing Research,* vol. 40, pp. 858-866, 1997.
[18] H. Quené, "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo," *The Journal of the Acoustical Society of America,* vol. 123, pp. 1104-1113, 2008.
[19] H. Quené, "Longitudinal trends in speech tempo: The case of Queen Beatrix," *The Journal of the Acoustical Society of America,* vol. 133, pp. EL452-EL457, 2013.
[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer" [computer program], 2012.
[21] H. Quené and H. Van den Bergh, "Examples of mixed-effects modeling with crossed random effects and with binomial data," *Journal of Memory and Language,* vol. 59, pp. 413-425, 2008.
[22] D. Bates, M. Maechler, and B. Bolker, "lme4: Linear mixed-effects models using S4 classes", 2012.
[23] R Development Core Team, "R: A Language and Environment for Statistical Computing", 2012.
[24] J. S. Pardo, "Expressing oneself in conversational interaction," in *Expressing oneself/Expressing one's self: Communication, cognition, language, and identity*, E. Morsella, Ed., New York: Psychology Press, 2010, pp. 183-196.
[25] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability," *Journal of Personality and Social Psychology,* vol. 32, p. 790, 1975.
[26] K. Nielsen, "Specificity and abstractness of VOT imitation," *Journal of Phonetics,* vol. 39, pp. 132-142, 2011.
[27] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction.," in *Proceedings of the 15th International Congress of Phonetic Sciences 2003 [ICPhS XV], Barcelona.* vol. 3, 2003, pp. 833-836.
[28] M. K. Jungers and J. M. Hupp, "Speech priming: Evidence for rate persistence in unscripted speech," *Language and Cognitive Processes,* vol. 24, pp. 611-624, 2009.
[29] I. Finlayson, R. J. Lickley, and M. Corley, "Convergence of speech rate: Interactive alignment beyond representation.," in *Proceedings of the 25th CUNY Conference on Human Sentence Processing*, 2012, p. 24.
[30] H. Giles, N. Coupland, and I. Coupland, *Contexts of accommodation: Developments in applied sociolinguistics*. New York: Cambridge University Press, 1991.
[31] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America,* vol. 119, pp. 2382-2393, 2006.
[32] J. S. Pardo, R. Gibbons, A. Suppes, and R. M. Krauss, "Phonetic convergence in college roommates," *Journal of Phonetics,* vol. 40, pp. 190-197, 2012.
[33] M. Wilson and T. P. Wilson, "An oscillator model of the timing of turn-taking," *Psychonomic Bulletin & Review,* vol. 12, pp. 957-968, 2005.
[34] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences,* vol. 27, pp. 169-189, 2004.
[35] M. Natale, "Social desirability as related to convergence of temporal speech patterns," *Perceptual and Motor Skills,* vol. 40, pp. 827-830, 1975.
[36] J. S. Pardo, I. C. Jay, and R. M. Krauss, "Conversational role influences speech imitation," *Attention, Perception, & Psychophysics,* vol. 72, pp. 2254-2264, 2010.
[37] C. R. Berger and M. E. Roloff, "Social cognition, self-awareness, and interpersonal communication.," in *Progress in communication sciences.* vol. 2, B. Dervin and M. J. Voight, Eds., Norwood, NJ: Ablex, 1980, pp. 1-49.
[38] H. Giles and P. F. Powesland, *Speech style and social evaluation*. New York: Academic Press, 1975.