Converging to the baseline

Corpus evidence for convergence in speech rate to interlocutor's baseline

Uriel Cohen Priva, Lee Edelist, and Emily Gleason

Abstract

Speakers have been shown to alter their speech to resemble that of their conversational partner. Do speakers converge with their interlocutor's baseline, or does convergence stem from conversational properties that similarly affect both parties? Using the Switchboard Corpus, this paper shows evidence for speakers' convergence in speech rate to the other party's baseline, not only to conversation-specific properties. Study 1 shows that the method for calculating speech rate used in this paper is powerful enough to replicate established findings. Study 2 demonstrates that speakers are mostly affected by their own behavior in other contexts, but that they also converge to their interlocutor's baseline, established using the interlocutor's behavior in other contexts. Study 2 also shows that speakers change their speech rate in response to the interlocutor's characteristics: speakers speak more slowly with older speakers regardless of the interlocutor's speech rate, and male speakers speak faster with other male speakers.

I INTRODUCTION

A Overview

Humans have a tendency to mirror one another. A person interacting with someone who shakes their leg or touches their face during conversation is significantly more likely to also take part in these actions (Chartrand and Bargh, 1999), and even newborn infants imitate facial expressions (Meltzoff and Moore, 1983). Speech is no exception to this phenomenon. When speakers are engaged in conversation, various aspects of their speech appear to undergo change and become more similar to those of their interlocutors. The literature refers to this phenomenon by a variety of names (e.g., *convergence*, *accommodation*, *alignment*, *synchrony*, *entrainment*). *Entrainment* and *alignment* typically refer to speech-mirroring as an automatic process (e.g. Pickering and Garrod, 2004). *Accommodation* is associated with more communicational theories (e.g. Communication Accommodation Theory, Giles et al., 1991). This paper uses the more theory-neutral term *convergence* as it does not investigate the mechanisms that lead to mirroring in speech.

Convergence in speech rate is particularly tricky to capture. There are two possible driving factors behind a speaker's observed speech rate similarity with their interlocutor within a conversation. One factor is the interlocutor's baseline speech rate. A speaker may speed up in response to a fast-speaking interlocutor, which critically implies that the interlocutor's speech rate has an influence over the speaker's speech rate. The other factor is the conversation itself: a pleasant conversation and a heated argument are likely to have different properties regardless of speaker-specific aspects. These conversation-level influences may cause both speakers in a conversation to speak quickly because they are speaking on an intense topic, for example. Most previous research does not distinguish between the two (not mutually exclusive) potential factors, limiting what can be inferred from their findings. The influence of conversation-level factors does not necessitate convergence with the other speaker's speech rate, only adaptation to the properties of the conversation. This paper focuses on the direct influence of interlocutor speech rate, and shows that speakers are indeed affected by their interlocutor's baseline speech rate.

B Background

1 Social aspects

There are many factors that have been found to affect degree of convergence. Social attributes such as differing levels of likability and attractiveness of the interlocutor, as well as membership in the speaker's 'in-group' (or lack thereof) have all been shown to affect convergence (Beebe, 1981; Mulac et al., 1988; Purcell, 1984). In particular, group differences in degree and form of convergence have been found between male and female speakers (Hannah and Murachver, 1999; Kendall, 2009; Babel, 2012).[1]

Research has suggested that women generally converge more than men (Bilous and Krauss, 1988; Gallois and Callan, 1988; Willemyns et al., 1997), though such results are often small and complex. Additional effects have been found on the interaction between speaker sex and interlocutor sex. Kendall (2009) found that speech rates were more strongly affected by the interlocutor's sex than by the speaker's sex– both male and female speakers talked in a similar, slow rate when interviewed by a woman, and faster when the interviewer was a man. Other studies showed that although women converge more than men, mixed-sex pairs appear to converge the most (Levitan et al., 2012; Namy et al., 2002). In contrast, Pardo (2006) looked only at same-sex interactions and found that male-male pairs showed the greater degree of convergence.

2 Measuring convergence

We can consider a model in which $A..Z$ are speakers conversing with a number of other speakers. For convenience, we will use S to denote the speaker and I to denote the interlocutor. We will denote the speech of S while speaking with I as S_I, and the speech of I while speaking with S as I_S. The speech of S while speaking with everyone *except* I will be denoted $S_{\neg I}$, and the speech of I while speaking with everyone except S will be denoted $I_{\neg S}$. We will use this notation to discuss the methodologies our study and previous studies use to measure convergence.

Previous work has measured phonetic convergence using a variety of methods. One method of testing convergence is to use third-party similarity judgments (Namy et al., 2002; Goldinger, 1998; Pardo, 2006). Other instrumental methods, such as measurements of speech rate and vowel formants, can be run on repeated words in a lab setting (Babel, 2012), but typically involve a comparison of a speaker's speech in one conversation with their interlocutor's speech in that same conversation, relative to various baselines (comparing S_I with I_S, as in e.g. Street, 1984; Levitan and Hirschberg, 2011; Pardo, 2006; Sanker, 2015). For instance, Levitan and Hirschberg (2011) compare a speaker's speech (S_I) to three reference points: their interlocutor's speech (I_S), the speech of random conversants the speaker did not speak with (various X_Y), and the speaker's speech when speaking with conversants other than the interlocutor ($S_{\neg I}$).

The convergence found by such methods could have been caused by conversational factors, e.g. how well the conversation was going, the topic of conversation, or how well-liked both parties wished to be. Fluent conversation, where the interlocutors "click" and the interaction is going well, is known to elicit speech features associated with excitement, high sociability, and good mood, while awkward, stalled conversations would induce a reduction in those features (Fónagy and Magdics, 1963; Murray and Arnott, 1993; Öster and Risberg, 1986). Participants in conversation may change their speech in the same way to achieve the same social goal, rather than as an effect of convergence. For example, speakers who use faster speech rates are viewed as more competent, whereas more average speech rates are associated with likeability, and speakers may change their rates to fit these impressions (e.g. Smith et al., 1975; Brown, 1980; Putman and Street Jr., 1984). In addition, certain topics might elicit different speech behavior. A person recounting a recent funeral they had attended may have a slower speech rate than while recounting a recent wedding, regardless of interlocutor. Any direct comparison of speakers within a conversation (between S_I and I_S) has the potential to be affected by these conversational factors. To find strict convergence to a partner's speech rate, it is necessary to rule out these additional influences.

Results from other studies would suggest that there are non-conversation-specific effects to be found. Webb (1972) found speech rate convergence in a study which controlled interviewer (interlocutor) behavior by pre-recording interviewer responses and placing the interviewer in a separate room from the interviewee (speaker). This eliminated the possibility of the speaker having some influence on the interviewer's speech rate, and in addition eliminated visual cues, such as facial expression and body language, from having an influence on the speech rate of the speaker. Similarly, an innovative study by Staum Casasanto et al. (2010) controlled interlocutor speech rate by using a virtual speaker in a virtual reality environment. The virtual speaker spoke in either a fast or slow rate while speaking with participants about items in a virtual store using a limited set of responses. They found that only speakers in the fast speech condition increased their rate from baseline, showing convergence to the fast-speaking virtual interlocutor. The designs of both studies do eliminate most conversation-specific effects. However, the interlocutors in both experiments were limited in their responses, and as a result the designs are limited in ecological validity and were only able to capture categorical effects. The highly-controlled designs additionally did not allow for straightforward analysis of the interaction between interlocutor's demographics and convergence.

C New method to calculate convergence

This paper proposes that in order to separate conversational influences from the underlying effect of the interlocutor's speech rate, a speaker's speech in conversation (S_I) should be compared to their interlocutor's baseline speech rate calculated using only speech outside of that conversation ($I_{\neg S}$). If averaged over a number of conversations, this baseline measurement can be an accurate measurement of a particular speaker's specific and individual speech rate preferences. The method described eliminates possible conversational effects, and analyzes only the convergence based on baseline rate. The Switchboard corpus (Godfrey and Holliman, 1997), in which participants took part in multiple conversations, is ideal for these calculations.

Working with the Switchboard corpus has several additional benefits. The Switchboard corpus provides a large collection of naturalistic data with a number of demographic annotations. There are 543 speakers in the corpus (compared to e.g. Babel, 2012, who had 107 speakers, the highest number of participants among the studies mentioned above), and conversations include both same-sex and mixed-sex pairings. The large number of conversations provides the ability to test for more fine-grained gradient effects as well as interactions of sex with convergence and speech rate. In addition, more casual conversations are more likely to avoid effects caused by non-natural lab settings or interview environments (see Willemyns et al., 1997).

Study 1 (section III) replicates known speech rate findings to demonstrate the competency of the proposed method of measuring speech rate, described in section II.B. Study 2 (section IV) measures convergence as suggested above to exclude the influence of conversational factors, and additionally investigates the effect of different demographic variables on convergence.

II STUDIES OVERVIEW: MATERIALS AND METHODS

A Switchboard Corpus

The Switchboard Corpus (Godfrey and Holliman, 1997) contains about 2400 annotated telephone conversations between strangers. Each speaker was paired randomly by computer operator with various other speakers; For each conversation one of 70 possible topics was assigned for discussion between speakers. Each conversation in Switchboard provides two data points: the speech of each of the conversants. For each caller, the corpus provides a number of demographic details.

The manually corrected word duration alignments of the corpus produced at MS State (Harkins et al., 2003) are used. The corpus provides 4876 conversation sides, but only 4850 had information identifying the caller.

B Speech rate

Each word's *expected duration* was taken to be the predicted duration of that word using the median duration of that word across the entire Switchboard corpus, the length of the utterance, and the distance to the end of the utterance (in words) as predictors. Medians were used because the distribution of word durations is not symmetric. The length of the utterance and the distance to the end of the utterance were included because it has been shown that both of these factors can affect rate of speech (Yuan et al., 2006; Quené, 2008; Jacewicz et al., 2010). The resulting measure approximates actual duration, but is less noisy as it is not affected by utterance-specific effects. The mean (245ms for both) and median (202ms for actual, 208ms for expected) are close between the two distributions, but the interquartile range (IQR) is greater for actual duration (190ms) than for expected duration (145ms). Abstracting from noise means that expected duration captures non-contextual effects more easily, e.g. a mixed effects linear regression (using word as a random intercept) that uses log frequency to predict expected duration has greater absolute t values (t=-29.23) than an equivalent regression that predicts actual duration (t=-25.45).

Utterance expected duration was defined as the sum of the expected durations of all words in the utterance, excluding silences, *uh*, *um*, and *oh* (henceforth "filled pauses"). *Utterance duration* was defined as the time from the beginning of the first word in an utterance which was not a silence or filled pause until the end of the last word in that utterance which was not a silence or filled pause, but including intermediate silences and filled pauses. *Pointwise speech rate* at the level of an utterance was defined as the ratio between utterance duration and utterance expected duration. Thus, an utterance whose expected duration is 5 seconds that took 4 seconds to produce would have a pointwise speech rate of 4/5=0.8 (fast), while if the same words took 8 seconds to produce, the pointwise speech rate would be 8/5=1.6 (slow). This methodology follows research in psycholinguistics that compares actual values to expected values (e.g. word duration in Bell et al., 2009), and in particular Cohen Priva (2017), who measured speech rate based on the ratio between actual duration and expected word duration, averaged across a conversation.

It should be noted that utterance-medial silences and filled pauses were used to calculate the actual duration of an utterance, but not the expected duration of the utterance. Doing so collapses two ways in which speech rate can vary: the variation in the duration of the words being used, and the duration and quantity of silences and filled pauses.

All models described below attempt to predict the speaker's *speech rate in a conversation* (S_I). The speaker's speech rate was calculated as the mean of the log pointwise speech rates of all utterances having 4 or more words. Shorter utterances were removed because many of these are backchannels, such as *yeah* and *uhuh*, which may have different behavior. In addition, both the speaker's and the interlocutor's *baseline speech rate* were calculated using the mean speech rate of that caller in other conversations ($S_{\neg I}$ and $I_{\neg S}$, respectively). Not all speakers took part in other conversations, making this calculation impossible for some speakers. Conversations in which either side did not participate in additional conversations were therefore excluded from further analysis, resulting in 4750 conversation sides and 481 speakers (a loss of 2.1% of the data points and 11.4 % of speakers).

C Predictors of speech rate and speech rate convergence

The Switchboard corpus provides several demographic details: sex (M/F), year of birth (translated to age by deducting year of birth from 1991), education (ordinal: 0=less than high school, 1=less than college, 2=college, 3=more than college, with 4 missing values), and broad dialectal regions. Sex and age are studied below to investigate to what extent they (a) affect speakers' speech rates, replicating previous results, and (b) affect rate of convergence.[2]

D Statistical models

All models described below use standardized speech rate as the predicted value in a mixed effects linear regression model. The `lme4` library (Bates et al., 2015) in the R software package (R Core Team, 2016) was used to fit the models and provide t-values. The `lmerTest` package (Kuznetsova et al., 2015), which encapsulates `lme4`, was used to estimate degrees of freedom (Satterthwaite approximation) and calculate p-values. All numerical predictors were standardized as well. All models used the interlocutor, conversation, and topic identity as random intercepts. Study 1 also used the speaker as a random intercept. For binary predictors (speaker and interlocutor sex), we converted the reference level ("FEMALE") to 0, and the non-default value ("MALE") to 1.

We used R's `p.adjust` function to FDR-adjust (Benjamini and Hochberg, 1995) p-values for multiple comparisons in each model separately.

III STUDY 1: REPLICATING ESTABLISHED FINDINGS

A Introduction

Previous studies observed age and sex-based differences in speech rate. In general, younger speakers have been found to have faster rates than older speakers (Duchin and Mysak, 1987; Harnsberger et al., 2008; Horton et al., 2010), and male speakers slightly faster rates than female speakers (Jacewicz et al., 2010; Yuan et al., 2006; Kendall, 2009). Any method that studies speech rate convergence should also be able to accurately measure speech rate, and for this reason this paper seeks to validate our measurement of speech rate by replicating known speech rate findings.

B Methods and materials

The list of conversations described in section II was used. Sex, age, and their interaction were used as fixed effects.

The models described in this paper use a random intercept for conversation. The random intercept captures the possibility of conversation-level deviations from the predictions of the other variables, but crucially allows both conversants to move in the same direction relative to what they would otherwise do: both speak faster or slower beyond what the other predictors predict. The model presented below was significantly better (model comparison $p<10^{-15}$) than a minimally different model that excluded conversation as a random intercept. We considered an alternative in which conversants do not move in the same direction, but toward one another (i.e. they converge) using a conversation random slope for a special dummy variable that was -.5 for one side of the conversation and +.5 for the other side of the conversation (sides were designated by the corpus, we did not attempt to use other possible divisions). That model was not significantly better than a model that did not include the random slope or the random intercept (model comparison p>.9). We therefore use per-conversation random intercepts.

C Results

Older speakers were more likely to have a slower rate of speech (β=0.22, SE=0.054, t=4.137, $p<10^{-4}$, FDR-adjusted $p<10^{-4}$). Male speakers were overall more likely to have a faster rate of speech (β=-0.39, SE=0.076, t=-5.151, $p<10^{-6}$, FDR-adjusted $p<10^{-5}$). Age did not affect male and female speakers differently (β=-0.08, SE=0.076, t=-1.044, unadjusted p=0.297, FDR-adjusted p=0.297). The results are also provided in Table I.

Table I: Study 1 results summary. Estimates, standard errors and t values calculated by the `lme4` package. Degrees of freedom and unadjusted *p*-values calculated by the `lmerTest` packages, using Satterthwaite approximation. FDR-adjusted values (Benjamini and Hochberg, 1995) calculated in R using the `p.adjust` function. Significant effects after FDR-adjustment are in bold.

Variable	Estimate	Std. error	df	t	Unadj *p*	FDR-adj *p*
age	**0.2239**	**0.0541**	**468**	**4.137**	**4.2e-05**	**6.3e-05**
spkr. sex	**-0.3912**	**0.0760**	**472**	**-5.151**	**3.8e-07**	**1.1e-06**
age × spkr. sex	-0.0795	0.0762	470	-1.044	0.297	0.297

Figure 1 visualizes the effect of age on speech rate in the dataset using the average of each speaker's speech rate in that speaker's conversations.

D Discussion

Previous results in the speech rate literature found men to speak slightly faster than women, and younger speakers to speak faster than older speakers. These findings are replicated here using a novel method of calculating speech rate. This serves to validate that this method does accurately capture speech rate and can be used to measure convergence in speech rate, or lack thereof.

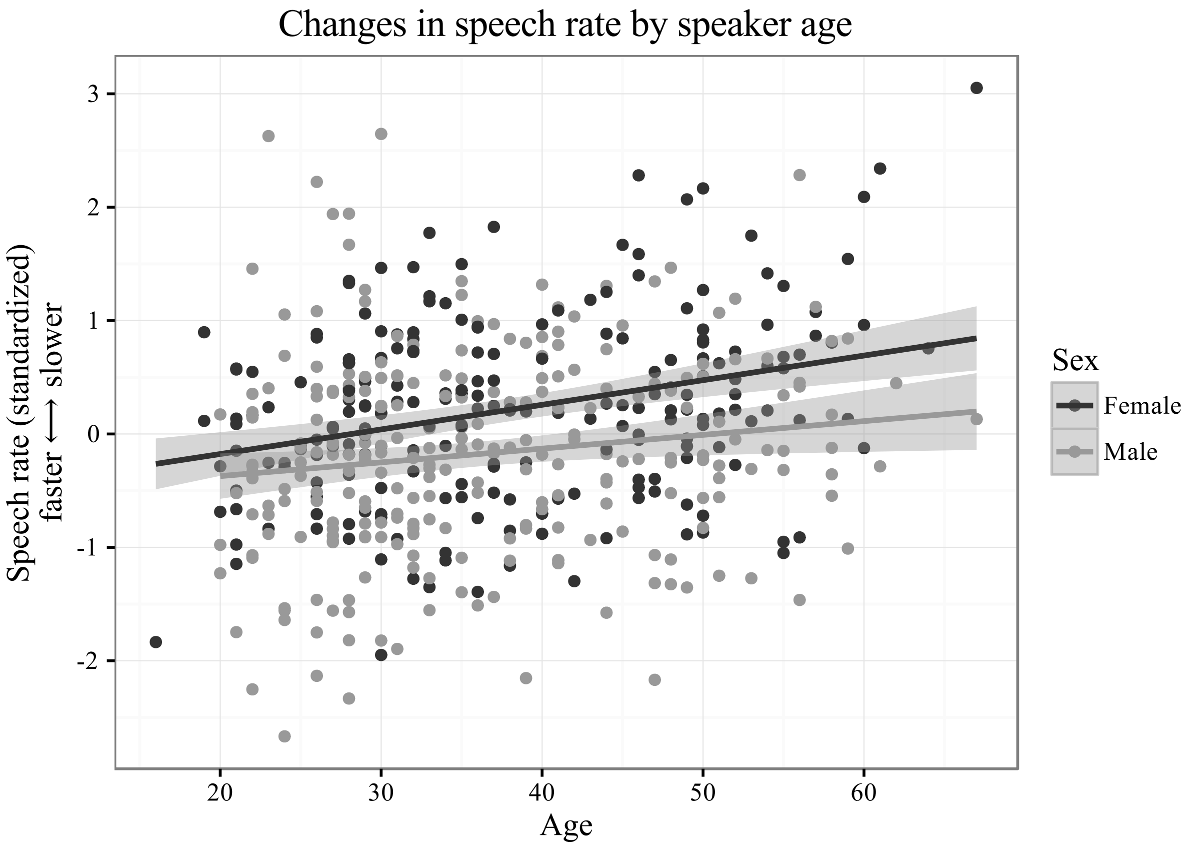

Figure 1: Changes in speech rate by age. Age is measured in years. Speech rate is calculated as mean utterance duration divided by the utterance expected duration, as described in section II.B. Each speaker's speech rate is averaged over the conversations in which that speaker took part. Other controls that are present in Study 1's model are not controlled for in the figure. The lines represent the correlation between age and speech rate, by sex, as calculated by a linear regression.

IV STUDY 2: CONVERGENCE

A Introduction

The method taken in the following study, described in section I.C, checks to what extent speakers converge with their interlocutor's *baseline rate*, rather than conversational pressures common to both parties in the conversation.

In addition to checking whether speakers converge to the other party's baseline rate, we wanted to know whether some speakers converge more or less than other speakers. For instance, do older people converge as much as younger people do? We investigated whether demographic properties affect the degree to which the interlocutor's baseline rate affect the speaker's speech rate.

Lastly, we considered the possibility that modification to speech rate may be categorical, rather than affected by convergence per se. For instance, if speakers speak more slowly to older interlocutors regardless of how fast their interlocutors speak, we can consider this a categorical effect.

B Method and materials

The dataset is the same as the one used in the Study 1, as described in section II.B. First, this study compares two predictors for speech rate: the speaker's baseline speech rate, estimated from their conversation with other interlocutors ($S_{\neg I}$), and the interlocutor's baseline speech rate, estimated from their conversations with others ($I_{\neg S}$). Both estimates are described in section II.B. Both baseline speech rates were added as predictors to the mixed effects linear regression model predicting the speaker's speech rate, as described in section II.D. A significant effect of interlocutor baseline speech rate would indicate that the speaker's speech rate in conversation is affected by their interlocutor's baseline speech rate.

The speaker's baseline speech rate in other conversations is expected to be highly collinear with the intercept fitted by mixed effects models to each speaker. Such high collinearity would keep more complex models from converging. Therefore, we initially tested in a model that contained only these two simple effects whether it was necessary to include speaker as a random effect using model comparison (Kuznetsova et al., 2015). Indeed, the model was not significantly better if speaker identity was included as a random intercept (p>.9). Similarly, a per speaker random slope for interlocutor's baseline speech rate did not improve the model (p>.9). Random intercepts for speakers and per-speaker random slopes for interlocutor's baseline speech rate were therefore removed from further consideration. Conversation as random intercept significantly improved the model ($p<10^{-15}$) and a per-conversation side random slope did not improve the model (p>.9), as in Study 1. Conversation is therefore used as a random intercept in this model as well.

The introduction of baseline speech rates to the model has an implication on the predictive power of demographic variables for predicting speaker's speech rate. Assigning each speaker their own individual baseline speech rate as a predictor assumes that speakers can have any speech rate, regardless of age or sex. Therefore, variance explained in the replication study by speaker's demographic variables would likely instead be attributed to their baseline speech rate, and demographic variables would have little to no effect when speaker's baseline rates are included in the regression. Thus, the speaker's demographic variables are not included in the model used in Study 2.

In contrast, the interlocutor's demographic variables can affect the speaker's speech rate: speakers may change their speech rate in response to the interlocutor's identity. The same holds for the interactions between the interlocutor's demographic variables and the speaker's demographic variables: response to the interlocutor's identity may be conditioned on the speaker's own identity, e.g. men may speak faster, but only with other men. Such effects can be either simple effects or interact with the interlocutor's baseline speech rate. Simple effects would indicate that speakers categorically speak faster or slower when speaking with people with certain demographic background. Interaction between demographic variables (both the speaker's and the interlocutor's) and the interlocutor's speech rate would indicate that demographic variables affect convergence: e.g. perhaps people converge more when speaking with male interlocutors. We therefore included several additional variables:

- The age (standardized) of the interlocutor, as well as its interaction with the (standardized) age of the speaker:
    - Interlocutor age
    - Interlocutor age × speaker age
- The sex of the interlocutor (female as the default), and its interaction with the sex of the speaker:
    - Interlocutor sex
    - Interlocutor sex × speaker sex
- Interactions between the interlocutor's baseline speech rate and all other variables, including speaker's baseline speech rate:
    - Interlocutor baseline × speaker baseline
    - Interlocutor baseline × speaker age
    - Interlocutor baseline × interlocutor age
    - Interlocutor baseline × interlocutor age × speaker age
    - Interlocutor baseline × speaker sex
    - Interlocutor baseline × interlocutor sex
    - Interlocutor baseline × interlocutor sex × speaker sex

C Results and discussion

We list the full results in Table II.

Speaker's baseline speech rate had the most significant effect on their own speech rate in a conversation ($\beta=0.79$, $SE=0.0088$, $t=90.015$, $p<10^{-15}$, FDR-adjusted $p<10^{-15}$). The interlocutor's baseline rate had a smaller yet significant effect on speakers' speech rate ($\beta=0.054$, $SE=0.019$, $t=2.804$, $p<0.01$, FDR-

Table II: Study 2 results summary. Estimates, standard errors and t values calculated by the `lme4` package. Degrees of freedom and unadjusted p-values calculated by the `lmerTest` packages, using Satterthwaite approximation. FDR-adjusted values (Benjamini and Hochberg, 1995) calculated in R using the `p.adjust` function. Significant effects after FDR-adjustment are in bold.

Variable	Estimate	Std. error	df	t	Unadj p	FDR-adj p
spkr. baseline	**0.7940**	**0.009**	**4341**	**90.02**	**<2e-16**	**<2e-16**
intloc. baseline	**0.0540**	**0.019**	**921**	**2.80**	**0.0052**	**0.034**
intloc. age	**0.0249**	**0.010**	**348**	**2.53**	**0.0120**	**0.043**
intloc. sex	0.0099	0.023	618	0.42	0.6726	0.844
spkr. baseline × intloc. baseline	-0.0176	0.010	2623	-1.78	0.0749	0.162
intloc. baseline × spkr. age	0.0025	0.009	4394	0.28	0.7793	0.844
intloc. baseline × intloc. age	-0.0079	0.009	424	-0.86	0.3877	0.630
intloc. age × spkr. age	-0.0230	0.010	2475	-2.32	0.0203	0.053
intloc. baseline × spkr. sex	0.0084	0.025	4470	0.34	0.7368	0.844
intloc. baseline × intloc. sex	-0.0009	0.027	1051	-0.03	0.9738	0.974
intloc. sex × spkr. sex	**-0.0676**	**0.027**	**3772**	**-2.48**	**0.0132**	**0.043**
intloc. baseline × intloc. age × spkr. age	0.0040	0.007	4262	0.54	0.5906	0.844
intloc. baseline × intloc. sex × spkr. sex	-0.0561	0.034	4456	-1.63	0.1041	0.193

adjusted p<0.05). The positive coefficient indicates convergence: when speaking with an interlocutor who spoke slowly or quickly, the speaker's speech rate changed in the same direction.

The interlocutor's baseline rate was a significant predictor of speaker's speech rate in conversation, indicating that speakers did converge to their interlocutor's baseline rate, in support of many previous findings on convergence. These results would suggest that convergence is not solely based on conversational factors, i.e. the flow or topic of the conversation, but also on some base properties of the interlocutor's speech. The effect may be direct or indirect, in that the interlocutors' different speech rates may affect some intermediate factor (e.g. conversation flow) in a predictable way, and thereby affect the speakers' speech rate. The method we take here is not able to separate these (not exclusive) alternatives.

The effect of speaker's baseline rate was an order of magnitude greater than the effect of interlocutor baseline rate. The large difference between the effect of speaker's baseline rate and interlocutor's baseline rate on speaker speech rate suggests that speakers are more consistent than they are convergent, and rely much more on their own baseline. In addition, the high consistency of speaker's speech rate across conversations implies that the baseline rates are fairly reliable estimates for each speaker's behavior in a given conversation.

Interlocutor age had a significant effect on speaker speech rate (β=0.025, SE=0.0099, t=2.527, $p<0.05$, FDR-adjusted $p<0.05$). The positive coefficient of this variable indicates that speakers were categorically slower while speaking with older speakers, regardless of the interlocutor's baseline speech rate. In addition, there was a negative interaction in the unadjusted model between interlocutor age and speaker age, suggesting that possibly older speakers spoke faster to other older speakers, younger speakers spoke faster to other younger speakers, or both, but the effect did not remain significant after p-adjustment (β=-0.023, SE=0.0099, t=-2.323, $p<0.05$, FDR-adjusted $p<0.1$). No other age-based effects or interactions were significant.

The effect of interlocutor sex was not significant, indicating that interlocutor sex did not have an effect on speaker speech rate independently. However, the interaction between speaker sex and interlocutor sex was significant (β=-0.068, SE=0.027, t=-2.478, $p<0.05$, FDR-adjusted $p<0.05$). The reference level for sex is female, so the interaction term indicates the effect of male interlocutors speaking with male speakers. Since the coefficient is negative, it indicates that male speakers spoke faster while speaking with male interlocutors, regardless of the interlocutor's baseline speech rate. No other sex-based effects or interactions were significant.

Interactions with the interlocutor's baseline speech rate would have indicated an effect on convergence. A significant effect for the interaction with the speaker age variable would have indicated that convergence is affected by the age of the speaker, and a significant effect for the interaction with the interlocutor's age would have indicated that speakers converge to varying degrees depending on the age of the interlocutor. Similarly, an interaction with the sex variables would have indicated that speakers of different sexes converge to a different degree, or that speakers converge to a different degree depending on the interlocutor's sex. Finally, convergence with the speaker's baseline speech rate could have indicated that faster speakers converge to a different extent than slower speakers. However, we found no such significant interactions, suggesting that either demographic variables and speaker's baseline speech rate do not have a strong effect on convergence or our measurement is unable to capture such effects. Our model shows that the effect of interlocutor baseline rate on speaker rate is fairly small, and therefore the differences in degree of convergence based on interlocutor baseline rate would be difficult to find.

Given that the conversations in question are between strangers, there is no reason to assume that speakers would have knowledge of their interlocutor's baseline rate. Rather, they would respond to the observed speech rate within the conversation, which the baseline rate estimates. It is likely that the speaker's speech rate would more greatly resemble the *actual* speech rate of their interlocutor, which was observed in conversation, than what is suggested by convergence to baseline rates. Interactions with the actual speech rate (I_S) may have markedly different patterns.

Other studies (e.g. Levitan et al., 2012; Namy et al., 2002) have found that mixed-sex pairs demonstrate the most convergence, while male-male pairs converge the least. This study did not find any differences in convergence between men and women or between mixed and same-sex pairs. As mentioned, the effects are quite small for convergence, and it could be that this measurement is not sensitive enough to capture these differences. Another possibility for these disparities is that the convergence which was found in other studies but not reflected in this study stems mostly from external conversational pressures. These pressures are exactly what this paper's proposed method excludes, and therefore it would be reasonable for other studies which did not exclude these pressures to find increased convergence or increased nuances in their findings compared to the results found here. It is also possible that the difference lies in the type of conversation undergoing analysis. The Switchboard corpus contains casual conversations between strangers over the phone. In other studies, participants were speaking on a very particular subject related to the experiment at hand (i.e. the map task in Pardo, 2006), or in an interview setting (such as in Kendall, 2009). Unlike the telephone conversations represented in the Switchboard corpus, typically in these tasks the participants were physically present, even in cases involving a divider or some other obstruction of visual input between them. In these studies, participants may have behaved differently, or felt different pressures to converge (e.g., different situations may affect speakers' interest in appearing friendly or likeable, see Putman and Street Jr., 1984). In other words, sex differences in convergence may be situation-specific. It would be critical, then, to consider the type of situation at hand in the analysis of any findings.

V GENERAL DISCUSSION

Study 1 used a novel measurement of speech rate calculation to replicate known findings for speech rate: male speakers speak faster (Jacewicz et al., 2010; Yuan et al., 2006; Kendall, 2009), and older speakers speak more slowly (Duchin and Mysak, 1987; Harnsberger et al., 2008; Horton et al., 2010). Study 2 showed (a) that speech rate is largely predicted by a speaker's behavior in other circumstances, (b) that speech rate is affected by the interlocutor's speech rate, such that slower interlocutors lead to slower speech, and faster interlocutors lead to faster speech, and (c) that speaker speech rate is affected by the identity of the interlocutor. Speakers spoke categorically slower to older speakers and male speakers spoke categorically faster while speaking with other male speakers.

In our study, the baseline rate for each speaker is estimated using the mean speech rate of the interlocutor or speaker in all other conversations, where speech rate is calculated using the expected duration of words. It would be possible to extend other methodologies to estimate the baseline rate using other means or to use other measurements of similarity in behavior. For example, a recent study by Solanki et al. (2015) uses techniques designed to recognize attempts to forge another's speech. Their method builds models of speaker and interlocutor behavior and compares speech in conversation to each of these models. Similarity of a speaker's speech to their interlocutor's model can indicate convergence. Preliminary data uses speech taken from the beginning of a conversation (the first task of a series of cooperative tasks) to build a model of each speaker's behavior, and later speech in the conversation as test speech. Given more extensive data, it would be possible to build models of speaker and interlocutor behavior on speech outside of a given conversation (models built on $S_{\neg I}$ and $I_{\neg S}$), which would avoid conversational interference. Another approach could build on the method used in Levitan and Hirschberg (2011), which compares the speech of a speaker in a given conversation with various baselines. One comparison was between the speaker's speech and the interlocutor's speech in the same conversation (S_I and I_S). The method could be extended in a similar way to our own study by using instead the comparison between the speaker's speech and the speech of their interlocutor while speaking in other conversations (S_I and $I_{\neg S}$).

Study 2 also checked whether demographic properties affected convergence and found no significant effects. However, there were effects of interlocutor identity on speaker's speech rate. Speakers spoke categorically slower with older interlocutors, independently from the interlocutor's speech rate, and male speakers spoke categorically faster with other male speakers. Although this was not convergence per se, it is still interesting that changes in speaker speech rate are conditioned on the identity of the interlocutor. Such results mirror a more socially-motivated interpretation of convergence, in which speakers do not only copy the other party's speech rate (cf. Pickering and Garrod, 2004), but rather converge as a result of social interaction.

VI CONCLUSION

Previous studies have shown that speakers converge with their interlocutors, but, with certain notable exceptions (Staum Casasanto et al., 2010; Webb, 1972), did not rule out the possibility that convergence is driven by shared conversational properties, rather than by influence from the interlocutor's speech properties. The present results demonstrate that at least for speech rate, speakers do converge with their interlocutor's baseline behavior in an ecologically valid setting. This provides strong support for previous findings on convergence, especially those on convergence in speech rate (Kendall, 2009; Putman and Street Jr., 1984; Webb, 1972), and suggests that previous findings were not based solely on conversational factors affecting both parties separately. The proposed method shows a particular advantage over methods used in extant research, in that it allows comparisons between speakers' behavior in a given setting with baselines established in other settings. This paper used the method to find convergence to the interlocutor's baseline in speech rate, but it could be straightforwardly extended to other types of convergence such as VOT, pitch, and other spectral properties.

NOTES

[1] For a review of sociological factors shown to affect speech rate see Jacewicz et al. (2009).

[2] Models that conditioned on education did not have a significant effect on speech rate.

REFERENCES

Babel, Molly. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1):177–189. doi: http://dx.doi.org/10.1016/j.wocn.2011.09.001.

Bates, Douglas, Mächler, Martin, Bolker, Ben, and Walker, Steve. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. doi: 10.18637/jss.v067.i01.

Beebe, Leslie M. 1981. Social and situational factors affecting the communicative strategy of dialect code-switching. *International Journal of the Sociology of Language*, 32:139–149.

Bell, Alan, Brenier, Jason, Gregory, Michelle, Girand, Cynthia, and Jurafsky, Daniel. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Benjamini, Yoav and Hochberg, Yosef. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bilous, Frances R and Krauss, Robert M. 1988. Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language & Communication*, 8(3):183–194.

Brown, Bruce L. 1980. Effects of speech rate on personality attributions and competency evaluations. In Giles, H., Robinson, W. P., and Smith, P., editors, *Language: Social psychological perspectives*, pages 293–300. Pergamon Press, Oxford, England.

Chartrand, Tanya L and Bargh, John A. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893.

Cohen Priva, Uriel. 2017. Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition*, 160:27–34. doi: 10.1016/j.cognition.2016.12.002. URL https://urielcpublic.s3.amazonaws.com/Not_so_fast-accepted.pdf.

Duchin, Sandra W. and Mysak, Edward D. 1987. Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders*, 20(3):245–257. doi: 10.1016/0021-9924(87)90022-0.

Fónagy, Ivan and Magdics, Klara. 1963. Emotional patterns in intonation and music. *STUF - Language Typology and Universals*, 16(1-4):293–326. doi: 10.1524/stuf.1963.16.14.293.

Gallois, Cynthia and Callan, Victor J. 1988. Communication accommodation and the prototypical speaker: Predicting evaluations of status and solidarity. *Language & Communication*, 8(3):271–283.

Giles, Howard, Coupland, Nikolas, and Coupland, Justine. 1991. Accommodation theory: Communication, context, and consequence. In Giles, H., Coupland, J., and Coupland, N., editors, *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pages 1–68. Cambridge University Press, New York.

Godfrey, John J. and Holliman, Edward, 1997. Switchboard-1 release 2. Linguistic Data Consortium, Philadelphia.

Goldinger, Stephen D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105(2):251–279. doi: 10.1037/0033-295X.105.2.251.

Hannah, Annette and Murachver, Tamar. 1999. Gender and conversational style as predictors of conversational behavior. *Journal of Language and Social Psychology*, 18(2):153–174. doi: 10.1177/0261927X99018002002.

Harkins, Dan, Feinstein, David, Lindsey, Troy, Martin, Sarah, and Winter, Greg, 2003. Switchboard MS State manually corrected word alignments. https://www.isip.piconepress.com/projects/switchboard/.

Harnsberger, James D., Shrivastav, Rahul, Brown Jr., W. S., Rothman, Howard, and Hollien, Harry.

2008. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, 22(1):58–69. doi: 10.1016/j.jvoice.2006.07.004.

Horton, William S., Spieler, Daniel H., and Shriberg, Elizabeth. 2010. A corpus analysis of patterns of age-related change in conversational speech. *Psychology and Aging*, 25(3):708–713. doi: 10.1037/a0019424.

Jacewicz, Ewa, Fox, Robert A., O'Neill, Caitlin, and Salmons, Joseph. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21:233–256. doi: 10.1017/S0954394509990093.

Jacewicz, Ewa, Fox, Robert Allen, and Wei, Lai. 2010. Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, 128(2):839–850. doi: 10.1121/1.3459842.

Kendall, Tyler S. 2009. *Speech rate, pause, and linguistic variation: An examination through the Sociolinguistic Archive and Analysis Project.* PhD thesis, Duke University.

Kuznetsova, Alexandra, Bruun Brockhoff, Per, and Haubo Bojesen Christensen, Rune. 2015. *lmerTest: Tests in Linear Mixed Effects Models*. URL https://CRAN.R-project.org/package=lmerTest. R package version 2.0-29.

Levitan, Rivka and Hirschberg, Julia Bell. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech 2011*, Brisbane. International Speech Communications Association.

Levitan, Rivka, Gravano, Agustín, Willson, Laura, Benus, Stefan, Hirschberg, Julia, and Nenkova, Ani. 2012. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 11–19, Stroudsburg, PA, USA. Association for Computational Linguistics. ISBN 978-1-937284-20-6.

Meltzoff, Andrew N and Moore, M Keith. 1983. Newborn infants imitate adult facial gestures. *Child Development*, pages 702–709.

Mulac, Anthony, Wiemann, John M., Widenmann, Sally J., and Gibson, Toni W. 1988. Male/female language differences and effects in same-sex and mixed-sex dyads: The gender-linked language effect. *Communication Monographs*, 55(4):315–335. doi: 10.1080/03637758809376175.

Murray, Iain R. and Arnott, John L. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108. doi: http://dx.doi.org/10.1121/1.405558.

Namy, Laura L., Nygaard, Lynne C., and Sauerteig, Denise. 2002. Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, 21(4):422–432. doi: 10.1177/026192702237958.

Öster, A-M. and Risberg, A. 1986. The identification of the mood of a speaker by hearing impaired listeners. *STL-QPSR*, 27(4):79–90.

Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393. doi: 10.1121/1.2178720.

Pickering, Martin J. and Garrod, Simon. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–190. doi: 10.1017/S0140525X04000056.

Purcell, April K. 1984. Code shifting Hawaiian style: children's accommodation along a decreolizing continuum. *International Journal of the Sociology of Language*, 1984(46):71–86.

Putman, William B. and Street Jr., Richard L. 1984. The conception and perception of noncontent speech performance: Implications for speech-accommodation theory. *International Journal of the Sociology of Language*, 1984(46):97 – 114.

Quené, Hugo. 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, 123(2):1104–1113. doi: 10.1121/1.2821762.

R Core Team, . 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sanker, Chelsea. 2015. Comparison of phonetic convergence in multiple measures. In *Cornell Working Papers in Phonetics and Phonology 2015*, pages 60–75.

Smith, Bruce L., Brown, Bruce L., Strong, William J., and Rencher, Alvin C. 1975. Effects of speech rate on personality perception. *Language and Speech*, 18(2):145–152. doi: 10.1177/002383097501800203.

Solanki, Vijay, Vinciarelli, Alessandro, Stuart-Smith, Jane, and Smith, Rachel. 2015. Measuring mimicry in task-oriented conversations: degree of mimicry is related to task difficulty. In *INTERSPEECH 2015*, pages 1815–1819. ISCA.

Staum Casasanto, Laura, Jasmin, Kyle, and Casasanto, Daniel. 2010. Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In Ohlsson, Stellan and Catrambone, Richard, editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 127–132, Portland, Oregon. Cognitive Science Society.

Street, Richard L. 1984. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169. doi: 10.1111/j.1468-2958.1984.tb00043.x.

Webb, James T. 1972. Interview synchrony: An investigation of two speech rate measures in an automated standardized interview. *Studies in dyadic communication*, pages 115–133.

Willemyns, Michael, Gallois, Cynthia, Callan, Victor J., and Pittam, Jeffery. 1997. Accent accommodation in the job interview: Impact of interviewer accent and gender. *Journal of Language and Social Psychology*, 16(1):3–22. doi: 10.1177/0261927X970161001.

Yuan, Jiahong, Liberman, Mark, and Cieri, Christopher. 2006. Towards an integrated understanding of speaking rate in conversation. In *Proceedings of Interspeech*, pages 541–544, Pittsburgh, PA.