Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Early Career Track

Developing an Integrated Model of Speech Entrainment

Rivka Levitan
Department of Computer and Information Science, Brooklyn College, NY, USA
Computer Science Department, Linguistics Department, CUNY Graduate Center, NY, USA
levitan@sci.brooklyn.cuny.edu

Abstract

Entrainment, the phenomenon of conversational partners' speech becoming more similar to each other, is generally accepted to be an important aspect of human-human and human-machine communication. However, there is a gap between accepted psycholinguistic models of entrainment and the body of empirical findings, which includes a large number of unexplained negative results. Existing research does not provide insights specific enough to guide the implementation of entraining spoken dialogue systems or the interpretation of entrainment as a measure of quality. A more integrated model of entrainment is proposed, which looks for consistent explanations of entrainment behavior on specific features and how they interact with speaker, session, and utterance characteristics.

1 Introduction

Entrainment (also termed *accommodation* or *alignment*) is the phenomenon of speakers becoming similar to each other in the course of conversation. Decades of research on entrainment have established that it is prevalent in human-human as well as human-computer interactions, and that it is associated with conversation quality and task success (e.g. [Chartrand and Bargh, 1999; Lee *et al.*, 2010; Levitan *et al.*, 2012; Ireland *et al.*, 2011], inter alia). However, our experience implementing acoustic-prosodic entrainment in a spoken dialogue system, described in Section 2, highlights profound lacunae in the literature.

Empirical studies of entrainment are usually framed in the context of Communication Accommodation Theory (CAT) [Giles *et al.*, 1991], which posits that speakers entrain to or disentrain from their interlocutors in order to decrease or increase social distance. Support for the association between CAT and entrainment comes from studies showing that entrainment is stronger in high-quality conversations or in the speech of individuals with prosocial characteristics [Natale, 1975]. Many studies also cite a cognition-centered "efficiency" theory of entrainment, based on the perception-behavior link [Chartrand and Bargh, 1999] or Interactive Alignment Theory [Pickering and Garrod, 2004]. These theories implicitly postulate that entrainment can be considered a single latent behavior or a structured collection of behaviors.

However, empirical findings on studies with large numbers of entrainment behaviors do not support this conclusion. It is typical to find entrainment on only some entrainment behaviors, with different conversations exhibiting different subsets of entrainment behavior. Similarly, it is typical to find associations between conversation quality and only some measures of entrainment. For example, [Lee *et al.*, 2010] found that entrainment on *pitch* between married couples in therapy was higher in positive interactions, while entrainment on *intensity* was not. Similar examples are common in the literature.

In order to fully understand entrainment, we must understand how personal, interpersonal and session characteristics explain variations in entrainment, how various kinds of entrainment behavior relate to each other, and how all factors affecting entrainment should be weighted when considered in combination. Without such an integrated understanding, it is impossible to create a scientifically-motivated entraining dialogue system, and it is difficult to interpret entrainment as a signal of quality in human-human communication.

Much of this work is motivated by the design of an entraining spoken dialogue system, described in Section 2. Sections 3, 4, and 5 elaborate on some of the issues described in this introduction, and describe some of the progress we have made towards understanding them.

2 Implementing Acoustic-Prosodic Entrainment in a Spoken Dialogue System

Despite strong interest in acoustic-prosodic entrainment, there has been little research on how to implement a spoken dialogue system (SDS) that can entrain on acoustic-prosodic features. Notable *lexically* entraining systems include [Hu *et al.*, 2014] and [Lopes *et al.*, 2013], but since lexical features are discrete, and the technology involved is natural language generation, these approaches are qualitatively different from acoustic-prosodic entrainment, which involves continuous features and text-to-speech (TTS) technologies.

In [Levitan *et al.*, 2016], we presented an architecture for implementing acoustic-prosodic entrainment in a SDS. This architecture is depicted in Figure 1, which shows that an entrainment module can be easily inserted into an existing SDS, and that its processes can complete in parallel with the sys-

Linguistic Level	Measure	Reference
Prosody (pitch, rate, intensity)	local similarity and convergence global similarity and convergence synchrony	[Levitan and Hirschberg, 2011]
Lexical	Perplexity (*PPL*) Kullback-Leibler divergence (*KLD*)	[Gravano *et al.*, 2014]
	High-frequency words (*HFW*)	[Nenkova *et al.*, 2008]

Table 1: Overview of entrainment measures: five per acoustic-prosodic feature, three lexical. (Reproduced from [Weise and Levitan, 2018])

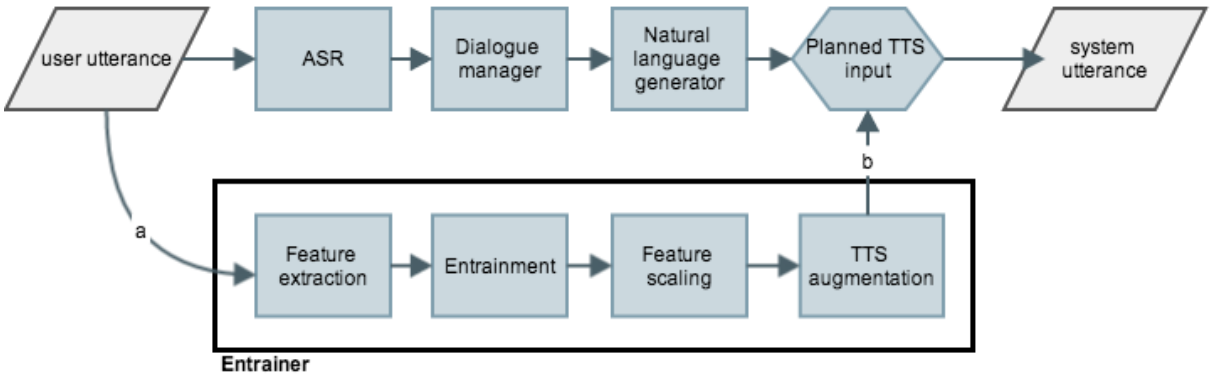

Figure 1: The entrainer integrated in an existing system. (Reproduced from [Levitan *et al.*, 2016])

tem's dialogue management.

Entrainment can be divided into two processes: (1) *Perception* of features in a user's speech, and (2) *production* of those features in the speech output. For (1), we use a Praat [Boersma and Weenink, 2012] script to extract acoustic-prosodic features from an input utterance. For (2), we inject those feature values into SSML markup for a TTS output utterance. In the demo system depicted in Figure 1, these two steps are represented by arcs a and b, respectively.

We tested the accuracy of entrainment using SSML markup for two unit-selection text-to-speech engines, Cepstral (English) and MaryTTS (Spanish), and one parametric TTS implemented on the open-source engine HTS (Slovak). For all three, correlations between the target speech rates and the actual speech rates of the entrained TTS output were strong (over 0.70) and statistically significant. Entrainment on intensity was only tested for the Cepstral TTS; correlations were almost perfect (over 0.90). Feature extraction completes in parallel with Automatic Speech Recognition (ASR) and therefore adds no latency; intermediate calculations take negligible time; and incorporating markup parameters adds no time to speech synthesis.

The difficulty arises in the step we have called "entrainment": when, based on the input utterance's features, we must generate "entrained" output feature values. A naïve approach, exactly matching the input features, has no empirical support - no research has ever suggested that entraining humans exactly match their interlocutors' speech. But none of the entrainment literature provides guidance on how to design a better entrainment algorithm. In general, entrainment studies - including our own - statistically reject the null hypothesis that entrainment does not exist. Specific positive hypotheses are necessary in order to implement humanlike entrainment behavior. This gap in the literature has motivated our recent work on entrainment, described in the rest of this paper.

In [Levitan, 2014] we designed an advice game, called GoFishWithHelpers, to evaluate whether users preferred an avatar that entrained to them - specifically, whether entrainment made them more likely to trust its advice. In [Levitan *et al.*, 2016], we tested GoFishWithHelpers with three entraining systems, in English, Spanish and Slovak. Implementation details differed slightly between the three experiments, as shown in Table 2, and the results (in the rightmost column) were very different. In English, users were *more* likely to trust a male avatar entraining on intensity and speech rate, compared to a similar avatar whose speech features randomly varied within a normal range. In Spanish and Slovak, users were slightly *less* likely to trust a female avatar entraining only on speech rate, compared to a *disentraining* baseline. These differences highlight how crucial it is to choose entrainment features and behaviors correctly.

3 Latent Entrainment Behaviors

Having informally observed differences in the presence of entrainment behaviors (described in Section 1), in [Weise and Levitan, 2018], we explored the hypothesis that these differences can be explained by an underlying structure of latent behavior(s). We considered eighteen entrainment behaviors, enumerated in Table 1. This set of behaviors covers four speech features - three prosodic, one lexical - and various ways of measuring each one, based on the framework described in [Levitan and Hirschberg, 2011].

Language	Avatar gender	Entrainment		Baseline	Entrainment × Advice score
		Features	Method		
English	Male	Intensity Speech rate	Absolute	Random	$+$ ($p < 0.001$)
Spanish	Female	Speech rate	Relative	Disentrain	$-$ ($p < 0.1$)
				Constant	no effect
Slovak	Female	Speech rate	Relative	Disentrain	$-$ ($p < 0.05$)
				Constant	no effect

Table 2: Summary of experiments with entraining system (Reproduced from [Levitan *et al.*, 2016])

The most basic kind of relationship between two entrainment behaviors is correlation: when one is higher than average, the other is also high, and vice versa. We computed Pearson's correlations between each pair of entrainment behaviors and found that there were no correlations between entrainment on different speech features. That is, there is no such thing as a "high-entrainment" conversation. Conversations that were highly entrained on one speech feature would show negligible entrainment on others. They were not even negatively correlated. Similarly, we used a χ^2 analysis to show that no combination of binned "high", "medium", and "low" entrainment on each feature was unusually likely to occur.

We also explored the possibility that entrainment behaviors were related in a nonlinear way that would not be captured by correlations. We represented each session's "entrainment configuration" as a point in a continuous 18-dimensional space, and used k-means clustering to group points in this space. Clusters, if discovered, could be considered "complex" entrainment behaviors composed of lower-level behaviors that are likely to co-occur. We found that the data could not be grouped into any meaningful clusters.

These findings demonstrate that, contrary to our expectations based on the cognitive literature, entrainment cannot be explained as a single construct or latent behavior. Entrainment on individual speech features and at each level of conversation must be considered separately.

4 Entrainment in Meaningful Segments

One way to move forward is to consider the hypothesis that we have been looking at entrainment on too broad a scale. In general, the literature has investigated entrainment at the global level of an entire conversation, or at a local level of discrete points in a conversation. It is possible that this lens is too indiscriminate, and entrainment can only be explained in the context of specific conversational segments.

[Heldner *et al.*, 2010] is an example of a more linguistically meaningful study of entrainment. They investigated pitch entrainment in speech segments classified as *backchannels*, short utterances that signal continued interest and understanding, without attempting to take the floor. With the understanding that backchannels are intended to be unobtrusive, the authors posited that they would be more similar in pitch to their preceding turns than other segments were, and showed that this was indeed the case.

Following their work, in [Levitan *et al.*, 2015] we extended this analysis to multiple acoustic-prosodic features - intensity, pitch, voice quality and speech rate - and a larger set of turn types, whose annotation is defined in [Gravano, 2009]. When simultaneous speech was present and the speaker was successful in taking the floor, the turn is called an *overlap* if the previous utterance was complete (as judged by the annotator) and an *interruption* if it was not. If simultaneous speech was not present, the turn is called a *smooth switch* if the previous utterance was complete and a *pause interruption* if it was not.

Our findings suggested that entrainment was higher in segments which, based on their turn type, were unlikely to begin a new discourse segment. This is consistent with the observation in [Heldner *et al.*, 2010] that entrainment makes segments *unobtrusive*, suggesting a mechanism by which entrainment can organize discourse.

Importantly, these findings were remarkably consistent across the entire set of acoustic-prosodic features we investigated. While there were minor exceptions, and speech rate in particular behaved notably differently than the other features, pitch, intensity, and voice quality features were either all entrained or all un-entrained. This study suggests that while entrainment behaviors are fragmented when all speech is considered together, a more coherent picture may emerge when focus is trained on specific, linguistically meaningful segments - whether the segments are specific turn types, or dialog acts, or bear paralinguistic significance.

Along these lines, [Reichel *et al.*, 2018] focused their entrainment study on dialog acts. They point out that since dialog acts constrain the production of various acoustic parameters, comparisons to a previous turn are often unreasonable. For example, the prosody of an answer *cannot* match the prosody of the preceding question, unless the speaker intends to mock their interlocutor. It is therefore more appropriate to choose utterances of the same category as a baseline for similarity. They analyzed how entrainment varied between dialog acts of high and low cooperativeness, authority, and predictability, demonstrating that speakers use entrainment to structure their joint task - a finding that supports our conclusion in [Levitan *et al.*, 2015] that entrainment plays an organizing role in discourse.

In summary, focusing entrainment research on meaningful segments can follow three paths. Firstly, a study can look for evidence of entrainment in the usual way, but with segment labels as an additional factor. Secondly, the entrainment of a specific segment should be measured with respect to a baseline that is drawn from similar segments. Thirdly, we should explore the possibility that the entrainment observed

in speech features is a "trickle-down effect" of entrainment on higher-level constructs such as dialog acts or paralinguistics.

5 Speech Feature Characteristics

Differences in entrainment behavior may also be explained by the characteristics of individual speech features. For example, Communication Accommodation Theory [Giles *et al.*, 1991] explains entrainment in the context of similarity-attraction theory, suggesting that people entrain to their interlocutors in order to decrease social distance. In that case, it is reasonable to hypothesize that the speech features most associated with a shared identity, such as accent (formants) or lexical choice, would be more likely to be entrained.

Under the theory that entrainment is explained by the perception-behavior link [Chartrand and Bargh, 1999], we can expect features that are more perceptually *salient* to be entrained. We found support for this in our prior work [Levitan *et al.*, 2011; Levitan, 2014], which showed that entrainment was relatively higher on statistical *outliers* for intensity. That is, if a segment's intensity was unusually high or low, the speaker's partner would respond to it with an utterance that was relatively *more* entrained, on average, than the responses to non-outlier utterances.

These hypotheses can be chained together. It is possible that entrainment on *socially meaningful* features is susceptible to interaction characteristics such as liking or power dynamics, while entrainment on other features is more automatic and dependent on salience.

At the most basic level, speech features may be divided between those that are unconstrained and those that are not. This distinction should be able to explain, to some extent, which features are or are not entrained in that utterance. It can also contextualize the use of similarity as a measure of quality. The absence of entrainment on lexical choice, for example, may be revealing in one instance (as when one speaker calls an object a "monkey" while his partner calls it an "ape"). In another instance it may simply reflect the fact that the lexical realization is entirely constrained by the message. [Reichel *et al.*, 2018] make this point in arguing that dialog acts are better units for entrainment than adjacent turns.

6 Speaker Characteristics

Much research has suggested that some of the variation observed in entrainment behavior may be attributable to speaker characteristics. For example, [Natale, 1975] found that entrainment on vocal intensity was higher in speakers with high social desirability, or "propensity to act in a social manner." [Chartrand and Bargh, 1999] found that subjects who scored high for the "perspective-taking" component of empathy entrained more to confederates' gestures. [Cohen Priva and Sanker, 2020] suggests that some people are natural leaders, attracting their interlocutors to entrain to them. [Lehnert-LeHouillier *et al.*, 2020] found that speakers with greater prosodic ability entrained *less* on f0 - an intriguing finding that may relate to our point regarding constraints on speech features (Section 5).

Another speaker characteristic considered central to entrainment is *power* or *dominance*. For example, in a study of Supreme Court oral arguments and discussions among Wikipedia editors [Danescu-Niculescu-Mizil *et al.*, 2012], speakers in a position of dependence converged more on linguistic style towards the individuals who were in a position of power over them.

Gender is also commonly examined as a factor in entrainment analysis. To cite just two studies dealing with task-oriented conversations between strangers, [Pardo, 2006] found that female pairs were less phonetically similar to each other than male pairs were; and [Thomason *et al.*, 2013] found that males entrain more than females on vocal intensity features. In [Levitan *et al.*, 2012], we found that male dyads entrained the least, while mixed-gender dyads entrained the most. However, in a much larger study [Weise *et al.*, 2019], we found that neither gender, native language, nor their combination could explain the variation we observed in the degree and valence of numerous entrainment behaviors.

Since theories of women's language [Lakoff, 1972] relate to social differences and power dynamics between women and men, it is reasonable to expect that gender-based differences in entrainment behavior are likely to be strongly mediated by the social context of the interaction, including (but not limited to) the ages and levels of education of each conversational partner, the degree of familiarity between them, and the topic or task of the conversation. Furthermore, since the theoretical literature casts entrainment as a fundamentally social behavior, social context must be assumed to mediate the influence of *any* personal characteristics on entrainment.

Most discussions connecting entrainment theory to sociolinguistic and psycholinguistic theory have stopped at the point of explaining how gender and other speaker characteristics may influence entrainment as a whole. However, as we have shown (Section 3), entrainment cannot be treated as a single construct: each behavior must be considered individually. Analyses of entrainment and speaker characteristics must go further, to predict - or at least explain post-hoc - variation in how speaker traits affect entrainment on different speech features.

7 Conclusion

This paper discusses what I believe to be the research directions with the most potential to contribute to a model of entrainment that can encompass all the variation observed between measures, features, sessions and speakers. As entrainment research progresses, keeping these factors in mind can help build real understanding out of individual analyses.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Award No. 1845710. Much of this work was done in collaboration with Julia Hirschberg, Agustín Gravano, Štefan Beňuš, and Andreas Weise.

References

[Boersma and Weenink, 2012] Paul Boersma and David Weenink. Praat: doing phonetics by computer [computer program]., 2012. Version 5.3.23, retrieved 21 August 2012 from http://www.praat.org.

[Chartrand and Bargh, 1999] Tanya L. Chartrand and John A. Bargh. The chameleon effect: The Perception-Behavior Link and Social Interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, 1999.

[Cohen Priva and Sanker, 2020] Uriel Cohen Priva and Chelsea Sanker. Natural leaders: Some interlocutors elicit greater convergence across conversations and across characteristics. 2020.

[Danescu-Niculescu-Mizil *et al.*, 2012] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: language effects and power differences in social interaction. In *Proceedings of WWW*, 2012.

[Giles *et al.*, 1991] Howard Giles, Nikolas Coupland, and Justine Coupland. Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*, pages 1–68. 1991.

[Gravano *et al.*, 2014] Agustín Gravano, Štefan Beňuš, Rivka Levitan, and Julia Hirschberg. Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In *Spoken Language Technology (SLT), 2014 IEEE Workshop on*, pages 578–583, 2014.

[Gravano, 2009] Agustín Gravano. *Turn-taking and affirmative cue words in task-oriented dialogue*. PhD thesis, Columbia University, 2009.

[Heldner *et al.*, 2010] Mattias Heldner, Jens Edlund, and Julia Bell Hirschberg. Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech*, 2010.

[Hu *et al.*, 2014] Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Proceedings of 5th International Workshop on Spoken Dialog System*, 2014.

[Ireland *et al.*, 2011] Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science*, 22(1):39–44, 2011.

[Lakoff, 1972] Robin Tolmach Lakoff. *Language and woman's place*. Cambridge Univ Press, 1972.

[Lee *et al.*, 2010] Chi-Chun Lee, Matthew Black, Athanasios Katsamanis, Adam Lammert, Brian Baucom, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth Narayanan. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Proceedings of Interspeech*, 2010.

[Lehnert-LeHouillier *et al.*, 2020] Heike Lehnert-LeHouillier, Susana Terrazas, Steven Sandoval, and Rachel Boren. The relationship between prosodic ability and conversational prosodic entrainment. In *Proc. 10th International Conference on Speech Prosody 2020*, pages 769–773, 2020.

[Levitan and Hirschberg, 2011] Rivka Levitan and Julia Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech 2011*, pages 3081–3084, 2011.

[Levitan *et al.*, 2011] Rivka Levitan, Agustín Gravano, and Julia Hirschberg. Entrainment in Speech Preceding Backchannels. In *ACL HLT*, pages 113–117, 2011.

[Levitan *et al.*, 2012] Rivka Levitan, Laura Willson, Agustín Gravano, Štefan Beňuš, Julia Hirschberg, and Ani Nenkova. Acoustic-Prosodic Entrainment and Social Behavior. In *NAACL HLT*, pages 11–19, 2012.

[Levitan *et al.*, 2015] Rivka Levitan, Stefan Benus, Agustin Gravano, and Julia Hirschberg. Entrainment and turn-taking in human-human dialogue. In *AAAI 2015 Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.

[Levitan *et al.*, 2016] Rivka Levitan, Stefan Benus, Ramiro Gálvez, Agustín Gravano, Florencia Savoretti, Marián Trnka, Andreas Weise, and Julia Hirschberg. Implementing acoustic-prosodic entrainment in a conversational avatar. pages 1166–1170, 09 2016.

[Levitan, 2014] Rivka Levitan. *Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue*. PhD thesis, Columbia University, 2014.

[Lopes *et al.*, 2013] José Lopes, Maxine Eskenazi, and Isabel Trancoso. Automated two-way entrainment to improve spoken dialog system performance. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8372–8376. IEEE, 2013.

[Natale, 1975] Michael Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804, 1975.

[Nenkova *et al.*, 2008] Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High Frequency Word Entrainment in Spoken Dialogue. In *ACL HLT*, pages 169–172, 2008.

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

[Pickering and Garrod, 2004] Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, 27(2):169–190, 2004.

[Reichel *et al.*, 2018] Uwe D. Reichel, Katalin Mády, and Jennifer Cole. Prosodic entrainment in dialog acts. *ArXiv*, abs/1810.12646, 2018.

[Thomason *et al.*, 2013] Jesse Thomason, Huy V Nguyen, and Diane Litman. Prosodic entrainment and tutoring dialogue success. In *Artificial Intelligence in Education*, pages 750–753. Springer, 2013.

[Weise and Levitan, 2018] Andreas Weise and Rivka Levitan. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of NAACL*, 2018.

[Weise *et al.*, 2019] Andreas Weise, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan. Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication*, 115:78–87, 2019.