IDENTIFYING ENTRAINMENT IN TASK-ORIENTED CONVERSATIONS

Run Chen[1,2] *Seokhwan Kim*[2] *Alexandros Papangelis*[2]
Julia Hirschberg[1] *Yang Liu*[2] *Dilek Hakkani-Tür*[2]

[1]Columbia University
[2]Amazon, Alexa

ABSTRACT

Human interlocutors adapt their behavior to each other in a conversation through entrainment. While entrainment has been found in long chit-chat conversations, much less research has been conducted on task-oriented dialogs. In this paper, we investigate short task-oriented Wizard-of-Oz conversations for acoustic-prosodic and lexical entrainment. We conduct significance tests that reveal changes in speech pitch and frequent words as important indicators of entrainment. Our findings will guide user-entraining dialog systems to improve the quality of conversations.

Index Terms— Dialog systems, entrainment

1. INTRODUCTION

Human interlocutors often adapt their behavior to each other in conversations through entrainment, also called accommodation or alignment. People adapt in syntax, word choice, pronunciation, and prosody, as well as in facial expression, posture, and socio-cultural behavior. Studies have found that people who entrain to others are perceived as more socially attractive, more competent and intimate [1, 2]. Entrainment leads subjects to like their conversational partners more and to perceive interactions as more successful [3, 4], and is a good predictor of task success [5]. So producing entrainment in dialog systems is useful in producing more attractive, competent, and intimate conversations that users will enjoy more and consider more successful. To move toward developing entraining dialog systems, we analyze entrainment in human-human conversations collected in a Wizard-of-Oz scenario where two speakers play the user and the agent roles in a task-oriented setup.

Our data is from the DSTC10 Track 2 dataset [6] representing real conversations that a user is expected to have with a task-oriented dialog system. In the context of task-oriented dialog systems, entrainment behaviors can occur in both users and agents. Users may deviate from their typical speech, in sync with what they perceive from the agent and agents may change how they talk to accommodate users: if the user is speaking very slowly, the agent may slow their speech accordingly; if the agent refers to the parking lot as the parking space, the user may also adopt such wording.

As speakers may entrain in both text and speech, we extract acoustic-prosodic features and lexical frequencies to test for proximity, convergence and synchrony aspects of entrainment. We conduct significance tests that reveal changes in speech pitch and frequent words as important indicators of entrainment. We also note the effect of duration and speaker roles on entrainment. We expect these findings will guide us in developing dialog systems to entrain the users.

2. RELATED WORK

Much research on entrainment has been done in a variety of conversational settings, including discussions of married couples about problems in their relationships [7], children adapting their amplitudes to that of an animated character [8], as well as human-computer interactions. Several studies have reported promising results when entrainment between the users' and system's voices took place: [9] reported gains in ASR accuracy when entrainment of speech rate was induced; [10] also found gains in learning when entrainment between humans and a tutoring system occurred in pitch and intensity; and [11] showed that entrainment improved rapport and naturalness when a system shifted the pitch contour of the synthesized speech by the mean pitch of the user. A positive link was found between entrainment and trust for humans using conversational avatars [12] and entraining the system's lexical choices to those of user's increased the dialog success rate [13]. Our study reports evidence of entrainment in a more realistic setup for task-oriented dialog systems, which are typically much shorter than previously studied entrainment datasets. This makes it much more challenging to identify entrainment since the degree of entrainment correlates with the length of conversations [14, 15].

3. DATASET

To study the entrainment behaviors in human-human conversations, we analyzed the DSTC10 Track 2 dataset [6], which

Work done while the first author was an intern at Amazon.

includes about 45 hours of recordings of 917 spoken task-oriented dialogues about touristic information for San Francisco. Each dialogue session was collected by two participants, a user and an agent. While the user-side participant was given a set of specific goals to be achieved in each session, the agent-side participant had access to a database to look up relevant information to provide for particular user requests. Entrainment has been documented in longer conversations such as the Columbia Games corpus, in which sessions between two speakers averaged 45 minutes long [14, 15]. However, our task-oriented conversations are much shorter, with an average of 24.7 turns (min 5 and max 60), and 3 minutes (min 0.45 and max 7.33). Therefore, identifying whether entrainment occurs in much shorter conversations is important to help us incorporate entrainment capabilities into task-oriented human-computer dialogue systems.

4. METHODS

4.1. Entrainment Measures

Following the methods proposed in [14], we evaluate three aspects of entrainment, i) proximity (do speakers in the same conversation sound more similar?), ii) convergence (do speakers entrain more over time?), and iii) synchrony (do speakers' behaviors vary in tandem?) For proximity, we compare the feature differences between partners (participating in the same conversation) with non-partners via a t-test. If a speaker does not entrain, they are expected to sound uniformly talking to any random speaker, whereas when entrainment occurs, the partners within the same conversation adopt each other's text and speech, reducing the feature differences. The partner differences are therefore smaller than the difference between a random non-partner speaker from the dataset.

The turn-level features are extracted and averaged over all turns within a dialog for both the agent and the user as session-level features ϕ_s. A partner difference $\Delta\phi_{\text{partner}}$ is defined as the absolute session-level feature difference between partners in the same conversation. Non-partner differences $\Delta\phi_{\text{nonpartner}}$ are calculated by feature differences between any two speakers that are not in the same conversation.

$$\phi_s = \frac{1}{n}\sum_{i=1}^{n} \phi_{i\text{th-turn}} | \text{ speaker} = s \tag{1}$$

$\Delta\phi_{\text{partner}} = |\phi_s - \phi_{s'}|$ for s and s' in the same dialog

$\Delta\phi_{\text{nonpartner}} = |\phi_s - \phi_{s'}|$ for s and s' not in the same dialog

For convergence, we split each conversation into two halves by the number of turns and compare partner differences between the first and the second halves. If speakers entrain more over time, we expect feature differences to decrease in the second half of the conversation. The half-session-level features $\phi_{k,s}$ are averaged turn-level features in the k-th half session. We calculate half session partner differences $\Delta\phi_k = |\phi_{k,s} - \phi_{k,s'}|$ and juxtapose those from the same conversations in a pair-wise t-test.

We calculate the Pearson's correlation coefficient for the turn-level feature differences between adjacent turns and the rest for levels of synchrony. If speakers entrain over time in synchrony, speakers may adjust their speech and language in accordance with those of their conversational partner. For example, if Speaker A's pitch rises, Speaker B's will too and follow similar patterns in pitch and other features over the course of the conversation.

4.2. Speech-based Entrainment Analysis

The acoustic-prosodic features represent the pitch, energy, voice quality and speaking rate of speakers through 12 features: pitch mean, minimum, maximum, and standard deviation, intensity mean, minimum, maximum, and standard deviation, jitter, shimmer, harmonics-to-noise ratio (HNR), and speaking rate [14, 16]. These features are extracted with praat [17] and parselmouth [18] tools on default parameter settings. The pitch and intensity features are z-score normalized by speaker. The speaking rate is measured by words per second from human transcripts.

4.3. Lexical Entrainment Analysis

Our lexical features include linguistic inquiry and word count (LIWC) [19] category frequencies as well as the 25 most frequent word counts for the corpus (MFC), the agents (MFA) and the users (MFU), accounting for the roles of the speakers. Only 25 frequent words are used here due to the short length of the conversations and thus the sparsity of the text. We use LIWC2015 for the lexical categories covering linguistic dimensions, psychological processes, personal concerns and spoken categories [19]. These features are interpretable and shown to be useful representations for multi-party entrainment [20]. We calculate the percentage of total words in a conversation that match each of the LIWC dictionary categories. Because of the sparsity of lexical features, we only perform analysis of lexical entrainment using session-level proximity and convergence, not turn-level synchrony.

5. RESULTS

Entrainment is found in both the speech and the content of conversations. Despite the brevity of the conversations, speakers assimilate in their pitches, intensity variations and HNR and converge on their intensity variations. They also entrain on the frequency of words that are in many LIWC categories and the most frequent word list.

5.1. Entrainment in Speech and Text

Significant levels of entrainment are observed in the speech data (Table 2). The differences between partners in the same conversation are smaller than those between non-partners

Observation	LIWC Categories
proximity	Total function words, 1st person plural, 3rd person plural, Conjunctions, Comparisons, Quantifiers, Anger, Sadness, Social processes, Family, Friends, Female references, Male references, Cognitive processes, Causation, Tentative, Differentiation, Perceptual processes, Hear, Biological processes, Body, Health, Ingestion, Power, Relativity, Space, Time, Work, Leisure, Home, Money, Religion, Death, Swear words, Netspeak, Assent
no proximity	Total pronouns, Personal pronouns, 1st person singular, 2nd person, Impersonal pronouns, Articles, Prepositions, Auxiliary verbs, Common Adverbs, Negations, Common verbs, Common adjectives, Interrogatives, Numbers, Affective processes, Positive emotion, Negative emotion, Anxiety, Insight, Discrepancy, Certainty, See, Feel, Drives, Affiliation, Achievement, Reward, Risk, Past focus, Present focus, Future focus, Motion, Informal language, Nonfluencies
convergence	1st person singular, 1st person plural, Impersonal pronouns, Articles, Prepositions, Auxiliary verbs, Interrogatives, Family, Friends, Causation, Tentative, Certainty, Differentiation, Perceptual processes, See, Biological processes, Health, Ingestion, Achievement, Power, Past focus, Future focus, Motion, Space, Leisure, Home, Money, Nonfluencies
no convergence	Total function words, Total pronouns, Personal pronouns, 2nd person, Common Adverbs, Negations, Common verbs, Common adjectives, Quantifiers, Affective processes, Positive emotion, Negative emotion, Social processes, Male references, Discrepancy, Feel, Drives, Reward, Risk, Present focus, Relativity, Time, Work, Informal language, Assent

Table 1. *t*-test results for LIWC categories. The table shows the categories with $p < 0.05$. The rows list the LIWC categories that show proximity, no proximty, convergence, and no convergence, respectively.

Feature	Proximity	Convergence
min pitch	-0.03735	1.16309
max pitch	$\mathbf{-8.96402}$**	-1.32656
mean pitch	$\mathbf{-3.83764}$**	-7.21694**
sd pitch	$\mathbf{-3.28754}$**	-2.60738**
min intensity	1.79504*	-1.85953*
max intensity	-0.9323	-4.47257**
mean intensity	0.45932	-7.96386**
sd intensity	$\mathbf{-24.62165}$**	$\mathbf{9.34142}$**
jitter	-0.31599	-7.99437**
shimmer	-0.45399	-2.5374**
HNR	$\mathbf{-26.02003}$**	-2.80482**
speaking rate	-1.00623	-4.47165**

Table 2. *t*-test statistics for speech entrainment. * for $p < 0.1$, ** for $p < 0.05$ and bold for entrainment. Negative for proximity and positive for convergence.

across most speech features (negative t-statistics for proximity). The t-test indicates significant proximity in pitch max, mean, and standard deviation, intensity standard deviation, and HNR. The t-test on the same speech features shows less entrainment observed in the convergence measure than proximity, which may be due to the short length of the conversations in the dataset. The speakers do not have enough time to produce similar significant levels of entrainment as do longer conversations, which typically exhibit higher assimilation. In Table 2 convergence column, a positive t-statistic means that the second half of these conversations has smaller differences, i.e. speakers converge, whereas a negative t-statistic means that they diverge. We only observed convergence in intensity standard deviation at the session level. Most of the other features significantly diverge.

For lexical entrainment, as shown in Table 1, among the LIWC categories, we find proximity in linguistic dimensions (function words, conjunctions), psychological processes (anger, sadness), personal concerns (work, home) and spoken categories (assent). The high level of proximity in the "assent" category also confirms our speculation that cooperative expressions correlate with entrainment. Similarly, we observe lexical convergence in a few LIWC categories such as linguistic dimensions (pronouns, articles) and spoken categories (nonfluencies). The lesser degree of convergence in lexical features aligns with our finding in speech: both are restricted by the short duration of the conversations. The frequent words show high convergence and some proximity (Table 3). Top words such as "that" and "you" appear in all three MFC, MFU and MFA groups, also overlapping with the LIWC categories. Conversational partners assimilate frequencies for assent words such as "okay" and "yeah" but dissimulate those of pronouns and disfluencies, and speakers tend to converge on frequent words rather than diverge.

5.2. Entrainment and Speaker Roles

As the speakers play different roles in the conversations, we explore whether and how their roles affect the degree of entrainment. For agent entrainment, we compare an agent's speech difference from their user partner's and a corpus averaged agent's speech. We found that the agents entrain to the users in intensity standard deviation. In a similar method, we found that the users entrain to the agents in intensity max. In addition to proximity entrainment, we also tested whether an agent's speech varies with their user partner's. We computed the Pearson's correlation coefficient by pairing an agent's turn-level speech with a previous turn uttered by the user (the first column of Table 4). The agents' speech is correlated with their respective users' speech in most acoustic-prosodic features, most notably intensity standard deviation and HNR. Meanwhile, the users also entrain to their agents (the second column of Table 4) and the most correlated features are also intensity standard deviation and HNR.

Observation		25 Most Frequent Words
proximity	MFC:	that, the, for, 's, okay, can, have, yeah, there
	MFA:	that, the, for, 's, okay, yeah, have
	MFU:	the, that, for, can, 's, okay, have, if, of, there
no proximity	MFC:	you, i, and, me, is, a, it, so, one, uhhh, let, to, do, ummm, in, go
	MFA:	is, me, you, one, i, let, so, and, it, go, do, see, ahead, right, four, all, sure, check
	MFU:	you, i, a, uhhh, and, ummm, to, in, do, place, me, 'm, is, great, it
convergence	MFC:	you, i, that, the, and, for, me, is, a, it, so, one, uhhh, 's, let, to, do, can, ummm, in, have, go, yeah, there
	MFA:	that, is, me, you, one, i, let, so, and, it, the, for, 's, go, do, see, ahead, right, yeah, all, sure, have
	MFU:	the, you, i, a, that, uhhh, for, and, ummm, to, in, can, 's, do, place, me, 'm, have, is, if, of, there, it
no convergence	MFU:	great

Table 3. *t*-test results for 25 most frequent words for the corpus (MFC), the agents (MFA), and the users (MFU). The table shows the words with $p < 0.05$. No words found to significantly diverge for MFC and MFA.

Feature	Agent	User	All	Short	Long
min pitch	0.09629**	0.09344**	0.06669**	0.07204**	0.06308**
max pitch	0.21015**	0.1765**	0.18478**	0.22395**	0.15574**
mean pitch	0.01656	0.04185**	0.02708**	0.04359**	0.01674*
sd pitch	0.10931**	0.10655**	0.09609**	0.12639**	0.07436**
min intensity	0.04438**	0.04382**	-0.00024	-0.01105	0.00202
max intensity	0.05345**	0.04854**	0.01153	0.02685**	0.00759
mean intensity	0.00818	0.00712	0.00088	0.00334	-0.00011
sd intensity	**0.55497****	**0.55977****	**0.45446****	**0.45526****	**0.45096****
jitter	0.14807**	0.13655**	0.0574**	0.0677**	0.04929**
shimmer	0.12693**	0.12929**	0.10921**	0.09844**	0.11501**
HNR	**0.41778****	**0.41613****	**0.38746****	**0.3763****	**0.39292****
speaking rate	0.01459	0.02538**	0.00933	0.00409	0.01204

Table 4. Pearson's correlation test for turn-level speech synchrony entrainment. * for $p < 0.1$, ** for $p < 0.05$. The columns show correlation coefficient for agent entraining to user, user entraining to agent, speaker-agnostic synchrony for all turns, short conversations (< 25 turns), and long conversations (≥ 25 turns), respectively. $|r| \geq 0.3$ moderate or strong correlation are in bold.

5.3. Entrainment and Duration

Prior studies report association of entrainment levels with conversation length. To verify that longer conversations indeed lead to stronger entrainment, we split the dataset into two bins based on number of turns: the short conversation set, with 24 turns or fewer (511 conversations), and the long conversation set, with 25 turns or more (406 conversations). Partner differences are significantly smaller for long conversations in session level pitch max and mean, jitter and HNR, confirming that longer conversations indeed show higher levels of speech entrainment. We also found weak correlation between number of turns and partner similarity in features such as pitch, intensity standard deviation, jitter and HNR. For convergence, short conversations converge in the same speech features as all data in Table 2. Long ones converge on a slightly different set of features, converging in min intensity but not on pitch standard deviation, shimmer and HNR; otherwise they are the same as the short conversation set and all data. We compare synchrony entrainment between short and long conversations, finding results comparable to global turn-level synchrony (Table 4). Speech features such as intensity standard deviation and shimmer show moderate positive correlation.

6. CONCLUSION AND FUTURE WORK

Our analysis of entrainment in the DSTC10 dataset demonstrates that entrainment does occur between speakers in task-oriented but shorter human-to-human conversations, which differ from previously studied corpus in style, domain and length. Based on the features of speech and lexical entrainment we have identified, we aim to improve the performance of state-of-the-art dialog system models for similar conversations. For our next step, we will explore other potential factors that may affect the **degree** of entrainment in dialogs. It has been shown that the purpose of the turn also correlates with lexical entrainment levels, and that people tend to entrain more in certain dialog acts, such as conventional-opening and closing [21]. We hope to identify similar results for dialog acts in speech data in our future research.

7. REFERENCES

[1] William B. Putman and Richard L. Street, "The conception and perception of noncontent speech performance: implications for speech-accommodation theory," *Journal of the Sociology of Language*, vol. 1984, no. 46, pp. 97–114, 1984.

[2] Richard Y Bourhis, Howard Giles, and Wallace E Lambert, "Social consequences of accommodating one's style of speech: A cross-national investigation," *International Journal of the Sociology of Language*, vol. 6, no. 5, pp. 5–71, 1975.

[3] Tanya L Chartrand and John A Bargh, "The chameleon effect: the perception–behavior link and social interaction.," *Journal of personality and social psychology*, vol. 76, no. 6, pp. 893, 1999.

[4] Clifford Nass, Youngme Moon, Brian J Fogg, Byron Reeves, and D Christopher Dryer, "Can computer personalities be human personalities?," *International Journal of Human-Computer Studies*, vol. 43, no. 2, pp. 223–239, 1995.

[5] David Reitter and Johanna D Moore, "Predicting success in dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.

[6] Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür, ""how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1147–1154.

[7] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 793–796.

[8] R. Coulston, S. Oviatt, and C. Darves, "Amplitude convergence in children's conversational speech with animated personas," in *Proceedings of ICSLP'02*, 2002.

[9] Linda Bell, Joakim Gustafson, and Mattias Heldner, "Prosodic adaptation in human-computer interaction," in *Proceedings of ICPHS*. Citeseer, 2003, vol. 3, pp. 833–836.

[10] Jesse Thomason, Huy V Nguyen, and Diane Litman, "Prosodic entrainment and tutoring dialogue success," in *International conference on artificial intelligence in education*. Springer, 2013, pp. 750–753.

[11] Nichola Lubold, Heather Pon-Barry, and Erin Walker, "Naturalness and rapport in a pitch adaptive learning companion," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 103–110.

[12] Rivka Levitan, Štefan Beňuš, Ramiro H. Gálvez, Agustín Gravano, Florencia Savoretti, Marian Trnka, Andreas Weise, and Julia Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," in *Interspeech 2016*, 2016, pp. 1166–1170.

[13] José Lopes, Maxine Eskenazi, and Isabel Trancoso, "Automated two-way entrainment to improve spoken dialog system performance," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8372–8376.

[14] Rivka Levitan and Julia Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *INTERSPEECH*, 2011, pp. 3081–3084.

[15] Agustín Gravano, Stefan Benus, Julia Hirschberg, Shira Mitchell, and Ilia Vovsha, "Classification of discourse functions of affirmative words in spoken dialogue," in *Interspeech 2007*, Antwerp, Belgium, 2007.

[16] Rivka Levitan, *Acoustic-prosodic entrainment in human-human and human-computer dialogue*, Ph.D. thesis, Columbia University, Aug. 2014.

[17] Paul Boersma and David Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.2.14, retrieved 24 May 2022 http://www.praat.org/, 1992-2022.

[18] Yannick Jadoul, Bill Thompson, and Bart de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[19] James W Pennebaker, Roger Booth, Ryan L Boyd, and Martha Francis, *Linguistic Inquiry and Word Count: LIWC2015*, Austin, TX, Sept. 2015.

[20] Zahra Rahimi and Diane Litman, "Entrainment2vec: Embedding entrainment for multi-party dialogues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8681–8688, Apr. 2020.

[21] Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David Traum, and Satoshi Nakamura, "Analyzing the effect of entrainment on dialogue acts," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles, Sept. 2016, pp. 310–318, Association for Computational Linguistics.