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, USA [2]Amazon, Alexa, USA

What is Entrainment?

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.

- people who entrain are perceived as more socially attractive, more competent and intimate (Bourhis et al., 1975; Putman & Street, 1984).
- listeners like their entraining conversational partners more and perceive interactions as more successful (Chartrand & Bargh, 1999; Nass et al., 1995)
- entrainment is a good predictor of task success (Reitter & Moore, 2007)

Related Work

Several studies have reported promising results when entrainment between the users' and system's voices took place in conversational systems.

- improved **ASR accuracy** when entrainment of **speech rate** was induced (Bell et al., 2003)
- better learning when entrainment between humans and a tutoring system occurred in **pitch** and **intensity** (Thomason et al., 2013)
- entrainment improved **rapport** and **naturalness** when a system shifted the **pitch** contour of the synthesized speech by the mean pitch of the user (Lubold et al., 2015)
- positive link between entrainment and **trust** for humans using conversational avatars (Levitan et al., 2016)
- entraining the system's **lexical choices** to those of user's increased the **dialog success rate** (Lopes et al., 2013)

Our study reports evidence of entrainment in a more **realistic** situation for **task-oriented dialog systems**, which are typically much **shorter** than previously studied entrainment datasets.

- entrainemnt in acoustic-prosodic features and lexical frequencies
- effect of duration
- effect of speaker roles as system or user

We expect these findings will guide us in developing dialog systems to entrain the users.

Dataset

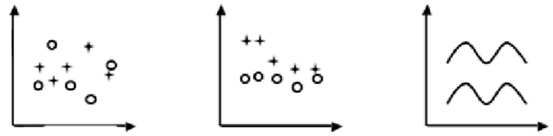
To study the entrainment behaviors in human-human conversations, we analyzed the DSTC10 Track 2 dataset (Kim et al., 2021), which includes about 45 hours of recordings of 917 spoken task-oriented dialogs about touristic information for San Francisco. Each dialog session was collected by two participants, a user and an agent.

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), compared to previously documented entrainment in the Columbia Games corpus (averaged 45 minutes long) (Gravano et al., 2007; Levitan & Hirschberg, 2011).

Speaker	Human Transcript
User	*ummm* i would like to request a place to dine in napa
Agent	sure let me see what there is ok so we do have quite a few options *ummm* is there anything that i can narrow it down for you
User	yeah can you get the cuisine and the price range of of the restaurant
Agent	sure so one of them is called souvla and the cuisine is greek and the price range is moderate is that something that would be interesting to you
User	yes ummm can i have the address and zip code of the restaurant
Agent	sure the address is five three one divisadero street and the zip code is nine four one one seven
User	uhhh can you check whether it's a *good place for groups*
Agent	yeah let me see ok so it says that it is a *good place for groups*
User	awesome thank you so much
Agent	ok have a great time there
User	you too

Table 1. Sample conversation from DSTC10 Track 2 dataset. Textual entrainment is highlighted in italics.

Methods



(a) Proximity (b) Convergence (c) Synchrony

Figure 1. Three views of entrainment (Levitan & Hirschberg, 2011).

Proximity: are speakers in the same conversation more similar, i.e. $\Delta\phi_{\text{partner}} < \Delta\phi_{\text{nonpartner}}$?

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

$\Delta\phi_{\text{partner}} = |\phi_s - \phi_{s'}|$ 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

Convergence: do speakers entrain more over time, i.e. $\Delta\phi_1 > \Delta\phi_2$?

$$\phi_{k,s} = \frac{1}{n}\sum_{i}^{n} \phi_{i\text{th-turn}} | \text{ speaker} = s \;\; \& \;\; (k-1)\cdot\frac{N}{2} < n \leq k\cdot\frac{N}{2}$$

where $k = 1$ or 2 indicates the first or second half of the dialog and N is the total number of turns in a dialog.

$$\Delta\phi_k = |\phi_{k,s} - \phi_{k,s'}| \text{ for } s \text{ and } s' \text{ in the same dialog}$$

Synchrony: do speakers' behaviors vary in tandem, i.e. $\rho(\phi_s, \phi_{s'})$?

Lexical Entrainment Analysis

Linguistic Inquiry and Word Count (LIWC) Category Frequencies

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
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

Table 2. t-test results for LIWC categories (Pennebaker et al., 2015). The table shows the categories with $p < 0.05$. The rows list the LIWC categories that show proximity and convergence, respectively.

Frequent Words Counts

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
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

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 $p < 0.05$.

Speech-based Entrainment Analysis

The acoustic-prosodic features represent the pitch, energy, voice quality and speaking rate of speakers through 12 features (Levitan, 2014; Levitan & Hirschberg, 2011). These features are extracted with Praat (Boersma & Weenink, 1992-2022) and Parselmouth (Jadoul et al., 2018) 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.

Speakers assimilate in their pitches, intensity variations and HNR, and converge on their intensity variations.

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 4. 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.

Speaker Role and Entrainment

- **Proximity**: for agent entrainment, we compare an agent's speech difference from their user partner's and a corpus averaged agent's speech. The agents entrain to the users in intensity standard deviation (sd). Similarly, the users entrain to the agents in intensity max.
- **Synchrony**: the agents' speech is correlated with their respective users' speech in most acoustic-prosodic features, most notably intensity sd and HNR. Meanwhile, the users also entrain to their agents and the most correlated features are also intensity sd and HNR.

Duration and Entrainment

- **Proximity**: better for long conversations in session level pitch max and mean, jitter and HNR. Weak correlation between number of turns and partner similarity in features such as pitch, intensity sd, jitter and HNR.
- **Convergence**: long ones converge in min intensity but not on pitch sd, shimmer and HNR.
- **Synchrony**: positive correlation in intensity sd and shimmer show moderate, similar for any length.

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	$\mathbf{0.55497}^{**}$	$\mathbf{0.55977}^{**}$	$\mathbf{0.45446}^{**}$	$\mathbf{0.45526}^{**}$	$\mathbf{0.45096}^{**}$
jitter	0.14807^{**}	0.13655^{**}	0.0574^{**}	0.0677^{**}	0.04929^{**}
shimmer	0.12693^{**}	0.12929^{**}	0.10921^{**}	0.09844^{**}	0.11501^{**}
HNR	$\mathbf{0.41778}^{**}$	$\mathbf{0.41613}^{**}$	$\mathbf{0.38746}^{**}$	$\mathbf{0.3763}^{**}$	$\mathbf{0.39292}^{**}$
speaking rate	0.01459	0.02538^{**}	0.00933	0.00409	0.01204

Table 5. 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.

Conclusions 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 corpora 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, such as dialog acts.

2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023), Rhodes Island, Greece

runchen@cs.columbia.edu