Global analysis of entrainment in dialogues

Vera Cabarrão[1,2], Isabel Trancoso[1,3], Ana Isabel Mata[2], Helena Moniz[1,2], Fernando Batista[1,4]

[1]L2F, INESC-ID, Lisbon Portugal
[2]Faculdade de Letras da Universidade de Lisboa (FLUL) / Centro de Linguística da Universidade de Lisboa (CLUL), Portugal
[3]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[4]ISCTE-IUL – Instituto Universitário de Lisboa, Lisboa, Portugal

veracabarrao@gmail.com, isabel.trancoso@inesc-id.pt, aim@letras.ulisboa.pt, {Helena.Moniz, Fernando.Batista}@inesc-id.pt

Abstract. This paper performs a global analysis of entrainment between dyads in map-task dialogues in European Portuguese (EP), including 48 dialogues, between 24 speakers. Our main goals focus on the acoustic-prosodic similarities between speakers, namely if there are global entrainment cues displayed in the dialogues, if there are degrees of entrainment manifested in distinct sets of features shared amongst the speakers, if entrainment depends on the role of the speaker as either giver or follower, and also if speakers tend to entrain more with specific pairs regardless of the role. Results show global entrainment in almost all the dyads, but the degrees of entrainment (stronger within the same gender), and the role effects tend to be less striking than the interlocutors' effect. Globally, speakers tend to be more similar to their own speech in other dialogues than to their partners. However, speakers are also more similar to their interlocutors than to speakers with whom they never spoke.

Keywords: entrainment, acoustic-prosodic features, map-task dialogues

1 Introduction

In human-human interactions, interlocutors naturally converge or diverge in their opinions and thoughts. For the dialogue to succeed, it is crucial to understand what is being said and how that content is being expressed. Humans have the ability to do immediate adjustments to their behavior and speech, acting accordingly to the situation [2, 6]. Moreover, several studies have shown that people who adapt to the partner's speech are considered to be more socially attractive and likeable [1], the interactions being much more successful.

The study of speech entrainment, whether between humans or human-computer systems, implies the evaluation of the degree of adaptation one has towards the other.

In Spoken Dialogue Systems, understanding and predicting how to adjust to a human may be more challenging than recognizing the speech signal content.

Entrainment can occur at different levels: acoustic-prosodic [8, 10, 11], phonetic-phonological [15], lexical-syntactic [12, 14, 15], multimodal, via facial expressions and gestures [4], and social [1].

Entrainment has been studied in languages such as English or Mandarin, but in European Portuguese this topic is just starting to be explored. In our study, we adopt an acoustic-prosodic approach to globally identify entrainment in human-human spontaneous speech. We focus on a wide set of features that have already proven to be effective in studying entrainment in American English, namely pitch, energy, speaking rate, and voice quality, and we also analyze duration.

Following the work of Levitan & Hirschberg (2011) [10] and Levitan (2014) [11], we applied the metrics of session-level proximity, meaning we analyze the similarities between speakers per dialogue. Our main goal is to perform a global analysis of entrainment in map-task dialogues, where speakers interact with different partners and also play different roles, giver or follower. This will allow us to understand if speakers entrain differently according to the role they're playing and/or according to the interlocutor.

This paper is organized as follows: Section 2 overviews the work previously reported on this subject. Section 3 describes the data and adopted methodology. Section 4 presents the achieved results, both in terms of global and dyad entrainment. Section 5 presents our conclusions and the future work.

2 Related work

Several studies of people interacting in different communicative situations have emerged, mainly to understand how they adapt to each other to solve problems or specific tasks (e.g., map-task dialogues, card games [8, 10], marital therapy [9], romantic relationships [21]). There is a large amount of literature covering not only linguistic aspects of entrainment, but also task success, social implications of speech adaptation, and its automatic applications.

The seminal work by [11], studying acoustic-prosodic entrainment in multiple dimensions in American English, measured the adaptation of speakers at a global level, the entire session, and at a local level, turn-by-turn, in the Columbia Games Corpus [7]. The authors describe that speakers were globally more similar to their partners than to their non-partners, meaning speakers with whom they were never paired with, in mean and max intensity, max pitch, shimmer, and speaking rate. The authors also found that speakers were more similar to their own speech in different sessions than to their partners in mean pitch, jitter, shimmer, Noise-to-Harmonics-Rate (NHR), and speaking rate.

Building upon the previous study, [10] also found that speakers of Mandarin Chinese were more similar to their partners than to their non-partners in intensity mean, max, pitch max, and speaking rate, differing from American English speakers only in pitch max. The author also describes that speakers paired in different gender groups

(male-male, female-female, and mixed-gender), both in American English and Mandarin Chinese, entrain on intensity mean and max, but differ in all of the other features: mixed-gender pairs present the highest degree of entrainment, being more similar to each other than female-female and male-male pairs.

Pardo [15] describes evidences of phonetic proximity between speakers at the beginning and later in the conversation. In a more recent perceptual study [16], the author found that the gender pairs and the role of the speaker influenced the degree of phonetic convergence during the dialogue. Moreover, the judgments of phonetic convergence were not related to individual acoustic–phonetic attributes, and the situational factors, such as pair gender, speaker role, and imitator role, were the ones with more influence.

While studying affirmative answers in European Portuguese, [3], in the same corpus used in this study, found evidences of pitch concord effects in context-answer pairs with different pragmatic functions. The authors found correlations regarding pitch height between the pairs instruct-agreement and propositional question (yes-no question) - confirm, although expressed in different degrees.

3 Methodology

The corpus used in this paper is the CORAL corpus ([19, 20] (ISLRN 499-311-025-331-2), which comprises 64 dialogues between 32 speakers, amounting to 7 hours orthographically transcribed. The dialogues are produced in map-task format between two speakers, the giver and the follower. The first one has a map with a route drawn and some landmarks and the latter has an incomplete map with different landmarks. The giver's task is to provide the correct directions so that the follower can reconstruct the same route in his/her map.

This work uses a subset of 48 dialogues[1] between 24 speakers (12 male and 12 female divided into 16 male-male pairs, 16 female-female pairs and 16 mixed-gender pairs). The degree of familiarity between interlocutors varied, going from people who never talked with each other to a pair of identical twin sisters (s21 and s24). All speakers play the role of giver and follower twice with different interlocutors. The subset is divided into sentence-like units (SUs), with a total of about 42k words.

In order to measure entrainment, a set of acoustic-prosodic features was extracted for each SU, and their mean values were calculated per speaker in each dialogue. Since our goal is to perform a global analysis of entrainment, this work focuses on the similarities between speakers at the session level and not locally, i.e., between turns.

Two sets of features were used, namely knowledge-based features, i.e., features known in the literature to have impact on the task, and GeMAPS (Geneva Minimalistic Acoustic Parameters for Voice Research and Affective Computing), typically adopted in paralinguistic tasks. The first ones, extracted in the context of [13], are the following: duration of speech with and without internal silences, pitch (f0), and energy normalized maximum, minimum, average, mean, and standard deviation, as well

[1] For the adopted subset, the two recording channels have been post-processed to reduce the interference from the other channel. Using unprocessed data could bias pitch measures.

as pitch and energy slopes. Tempo measures encompass: articulation rate (number of phones or syllables per duration of speech without internal silences); speech rate (number of phones or syllables divided by the duration of speech with internal silences), and phonation ratio (100% times the duration of speech without internal silences divided by the duration of speech including the internal silences). As for GeMAPS [5], this is a set of functionals based on a set of low-level descriptors, totaling 62 acoustic parameters.

$$ENT(s,f) = -|s_f - s_f^i| \tag{1}$$

$$ENTX(s,f) = -\frac{\sum_{x=0}^{n-1} |s_f - s_f^x|}{n} \tag{2}$$

$$ENT_{self}(s,f) = -|s_f - s'_f| \tag{3}$$

Following the work of [10] and [11], we have used equations 1, 2 and 3 to calculate the partner similarity, the non-partner similarity, and the self-similarity, respectively. In the equations, s is a speaker in a session, s_f corresponds to the speaker's mean feature value for that session, ${s^i}_f$ refers to the interlocutor's mean feature for that session, n is the number of non-partners, ${s^x}_f$ is the mean feature for one of those speakers, and ${s'}_f$ is the mean value for feature f of speaker s in another session. Thus, we have used Eq. 1 to measure the level of proximity between interlocutors in the same dialogue by calculating the difference between a speakers' mean for a feature with the same value for his/her partner. Eq. 2 measures the difference between a speakers' mean value for each feature and the ones of the speakers with whom he/she is never paired with. According to [10], such metric establishes a baseline measure of the degree of similarity that one expects to see if there is no entrainment. A final measure of similarity is the comparison between the same speakers' mean values in different dialogues. Given the fact that in CORAL corpus speakers participate in 4 dialogues, 2 as a giver and 2 as a follower, the self similarity or self-entrainment measure was done according to the roles they play, meaning that a speaker was compared to him/herself only when playing the same role. This allows us to verify if speakers are consistent regardless of the role they're playing or if that happens only with a specific role.

Again, in line with the work of [10], all of these measures were compared with a paired t-test to obtain statistically significant differences between them. Moreover, when comparing the means for each feature of the partner, non-partner, and self-similarities, it was possible to see which one has a greater degree of entrainment. For example, if the similarity between partners in a certain feature is greater than between the same speaker in other session, there is entrainment between interlocutors.

4 Results

4.1. Global entrainment

At the session level, there are statistical significant differences ($p<0.05$ represented with **, and $p<0.01$ represented with *) between partners and non-partners in almost all of the features analyzed (Table 1). The comparison of each feature means shows that pitch maxima, minima, average, median, and standard deviation are more similar between interlocutors than between non-partners (see positive t-values in Table 1), clearly showing entrainment between dyads. The same pattern occurs with energy features, except for energy range, standard deviation, and slope. As for speaking rate, even though the differences between partner and non-partner are not significant, partner similarities are greater than non-partner ones expressed in a positive t-value (Table 1). As for voice quality features, namely shimmer and HNR, again speakers are more similar to their interlocutors than to their non-partners. The same happens with jitter, even though this feature does not show statistically significant differences.

When comparing the similarities of a speaker between him/herself in another session and between his/her partner, results show that speakers are more similar with their own speech in all the features, except in energy minima, average, and median. These three features are more similar between partners than between the same speaker in different sessions, an evidence more for entrainment. [11] and [10] also found more similarities between conversational partners in intensity features, namely intensity mean and maxima.

As for the comparison of similarities within roles, results show that speakers have a more consistent style as a giver mainly in tempo measures (positive t-values for articulation rate phone and syllable, 0.785 and 1.200; rate of speech phone and syllable, 1.011 and 1.916; and phonation ratio, 0.089), voice quality features (positive t-values for jitter, 0.572; shimmer, 0.966, and HNR, 1.355), pitch median and slope (t-value of 0.009 and 0.899, respectively), and also energy min, range, and slope (t-values of 0.613, 0.493, and 0.508, respectively). Looking at each speaker individually, results show that almost half of the speakers are more consistent as a giver (N=13) and the other half (N=11) as a follower. These results show trends of more plasticity in speakers adjusting to their partners as followers than as givers.

Given these results, it is possible to conclude that there is global entrainment, but expressed in different degrees: speakers are more similar to their own productions than they are to their interlocutors, but they are also more similar to their interlocutors than they are to speakers with whom they never spoke to.

Features	*t*	*Df*	*Sig.*		*Features*	*t*	*Df*	*Sig.*
duration speech (with internal silences)	-8.195	95	*		duration speech (with internal silences)	-15.001	95	*
duration speech (without internal silences)	-8.645	95	*		duration speech (without internal silences)	-15.204	95	*
pitch max	5.329	95	*		pitch max	-7.415	95	*
pitch min	3.585	95	*		pitch min	-8.686	95	*
pitch range	-2.259	95	**		pitch range	-7.384	95	*
pitch avg	5.592	95	*		pitch avg	-7.705	95	*
pitch med	5.474	95	*		pitch med	-7.788	95	*
pitch stdev	2.627	95	**		pitch stdev	-7.611	95	*
pitch slope	0.213	95			pitch slope	-3.825	95	*
energy max	3.186	95	**		energy max	-3.561	95	*
energy min	6.41	95	*		energy min	0.911	95	
energy range	-0.672	95			energy range	-5.568	95	*
energy avg	10.234	95	*		energy avg	1.013	95	
energy med	9.482	95	*		energy med	0.978	95	
energy stdev	-2.176	95	**		energy stdev	-5.851	95	*
energy slope	-0.676	95			energy slope	-2.99	95	**
articulation rate phone	-1.497	95			articulation rate phone	-6.736	95	*
rate of speech phone	-0.415	95			rate of speech phone	-5.754	95	*
phonation ratio	-2.019	95	**		phonation ratio	-6.75	95	*
articulation rate syl	-0.058	95			articulation rate syl	-4.703	95	*
rate of speech syl	0.12	95			rate of speech syl	-4.808	95	*
jitter amean	1.684	95			jitter amean	-2.675	95	
shimmer amean	2.907	95	**		shimmer amean	-5.474	95	*
HNR amean	3.059	95	**		HNR amean	-7.886	95	*

Table 1. T-tests: partner vs. non-partner differences (left columns), and partner vs. self-differences (right columns)

4.3. Dyad entrainment

Considering the fact that the same speaker interacts with 2 different partners, 2 times as a giver and 2 as a follower, we aim at verifying with whom they entrain more, specifically if they entrain with the same speaker regardless of the role they are playing and in which features.

In order to measure the degree of entrainment between all the pairs, we considered the percentage of similar features each pair shares. If one pair is more similar in a greater number of features, they entrain more than the other pair.

Results show that 14 speakers entrain with the same interlocutor whether they are playing the role of giver or follower (Table 2). Nonetheless, the amount of features is different. Speakers like s1, s2, s3, s5, s9, s19, and s17 entrain in a greater number of features when they are followers and their partner is the giver. The pair s6 (giver)-s1 (follower) entrains in 92% of the features, the highest entrainment found in this data. The remaining 10 speakers entrain only with one partner. In these dyads, speakers are

similar in a smaller number of features (from 50% to 75%), as those who entrain twice with the same partner reveal stronger similarities. Only two speakers entrain equally (50%) in two different dyads, namely the s7-s4/s5 and s11-s21/s9. All of the followers in these pairs (s4, s5, s11, and s9) display more similarities with different partners.

Looking at the type of features where speakers entrain more, results reflect the analysis presented previously (Table 1): partners entrain more on energy (18 dyads), pitch (9 dyads), tempo measures (8 dyads), and voice quality features (12 dyads).

Results also show that pairs within the same gender tend to entrain much more than mixed-gender pairs. There are 10 female speakers that entrain with the same partner in both dialogues they participate, and regardless of the role they're playing. The same only occurs with 4 male speakers. The remaining tends to entrain with speakers of the same gender, but only in one dialogue. Mixed-gender entrainment only occurs with the pairs s7-s5 and s13-s23, the latter being the strongest one (they present similarities in 75% of the features). In the first pair, speaker s7 entrains 50% of the times with both speakers with whom he is paired with. This preliminary gender analysis allows only for an overview of what we can expect from the data. A more fine-grained analysis is needed, specifically a global (session level) and local (turn level) comparison with non-partners playing the same role and with the same gender as the real partners [11].

Entrain both as a giver and as a follower (↔)							**The giver entrains with the follower (→)**					
Speakers/ Gender		% of similar features		Speakers/ Gender		% of similar features	Speakers/ Gender		% of similar features		Speakers/ Gender	
s1	F	71%	↔	s6	F	92%	s4	M	54%	→	s7	M
s2	F	54%	↔	s14	F	67%	s7	M	50%	→	s4/s5	M/F
s3	F	63%	↔	s23	F	75%	s10	M	71%	→	s16	M
s5	F	75%	↔	s8	F	83%	s16	M	63%	→	s15	M
s9	F	54%	↔	s24	F	88%	s15	M	67%	→	s20	M
s19	M	71%	↔	s22	M	88%	s20	M	63%	→	s10	M
s17	M	71%	↔	s18	M	80%	s11	F	50%	→	s21/s9	F/F
							s21	F	54%	→	s24	F
							s11	M	58%	→	s13	M
							s13	M	75%	→	s23	F

Table 2. Partner entrainment per speaker per role

Taking into account that 2 speakers are identical sisters (s21 and s24), we were expecting to find a clear entrainment between them in almost all of the features. In Table 2, we can see that s21, as a giver, has similarities with s24, as a follower, in 54% of the features, as s24 entrains more with s9 twice, as a giver (88%) and as a follower

(54%). These results may be due to the small amount of interactions produced by the sisters. The dialogues where they both interact are the briefest when compared to all the other partners. In the dyad s24-s21, a successful task is achieved with only 36 SUs from the giver and 11 from the follower. In the other dyad, where their roles are reversed, there are 30 and 38 SUs, respectively. These speakers, who already know each other so well, do not need to talk much to complete the task and succeed. This fact points out to a strong entrainment between them. However, due to the almost minimal dialogue we will also perform a turn-by-turn analysis to better understand how they interact and adapt to each other.

As for the dyad entrainment, we can conclude that, despite the role the speakers are playing, they tend to display more sensitivity to some partners. Thus, we can hypothesize that in this data there is a stronger partner effect than a role effect.

In order to verify in detail which speaker adapts the most and which maintains a more consistent personal style, a turn-by-turn analysis (local entrainment) is required, and will be applied in future work, not only to the twin sisters dialogues but to the remaining dyads.

5 Conclusions

This study is our first attempt to describe acoustic-prosodic entrainment in European Portuguese map-task dialogues. Using statistical tests based on proximity metrics at the session level [11], we found evidences of entrainment between partners in pitch, energy, tempo measures, and voice quality features, even though expressed in different degrees. Speakers do not entrain with the same partners and in the same features. We also found that female-female dyads tend to entrain more regardless of the role they are playing, followed by male-male dyads, and, finally, by mixed-gender pairs. These results are not in line with the findings of [11] for American English and Mandarin Chinese, since the author found more entrainment in mixed-gender pairs. Our results also show that speakers are more similar to their interlocutors than to their non-partners, but speakers are also more similar with their own productions in different dialogues than they are to their partners. Despite that, while playing the role of giver, speakers are more consistent, being similar in a greater amount of features to themselves than as a follower, which allows for more adjustment while playing this role.

In a future work, we will explore other metrics at a local level, namely acoustic-prosodic convergence turn-by-turn, and also the progression of entrainment throughout the dialogue (beginning, middle and end). This local entrainment analysis may be also relevant to extend this study to different domains, namely police interrogations, in the scope of the European Project LAW-TRAIN (REF.: H2020-EU.3.7. – 653587).

6 Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references UID/CEC/50021/2013, and UID/LIN/00214/2013,

through the European Project LAW-TRAIN with reference H2020-EU.3.7. – 653587, and under PhD grant SFRH/BD/96492/2013, and Post-doc grant SFRH/PBD/95849/2013.

7 References

1. Beňuš, Š. (2014). Social aspects of entrainment in spoken interaction. Cognitive Computation, 6(4), 802-813.
2. Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482.
3. Cabarrão, V., Mata, A.I., Trancoso, I. (2016) Affirmative constituents in European Portuguese dialogues: prosodic and pragmatic properties. In proceedings of Speech Prosody 2016, Boston.
4. Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. Journal of personality and social psychology, 76(6), 893.
5. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S. & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.
6. Giles, H., Mulac, A., Bradac, J. J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. Annals of the International Communication Association, 10(1), 13-48.
7. Gravano, A. (2009) "Turn-taking and affirmative cue words in task-oriented Dialogue," PhD thesis, Columbia University, 2009
8. Gravano, A., Beňuš, Š., Levitan, R., & Hirschberg, J. (2014, December). Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In Spoken Language Technology Workshop (SLT), 2014 IEEE (pp. 578-583). IEEE.
9. Lee, C. C., Black, M., Katsamanis, A., Lammert, A. C., Baucom, B. R., Christensen, A. & Narayanan, S. S. (2010, September). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In INTERSPEECH (pp. 793-796).
10. Levitan, R., and Hirschberg, J. (2011) "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." Interspeech 2011.
11. Levitan, R. (2014). Acoustic-prosodic entrainment in human-human and human-computer dialogue (Doctoral dissertation, Columbia University).
12. Lopes, J., Eskenazi, M., & Trancoso, I. (2013, May). Automated two-way entrainment to improve spoken dialog system performance. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8372-8376). IEEE.
13. Moniz, H., Batista, F., Mata, A. I., & Trancoso, I. (2014). Speaking style effects in the production of disfluencies. Speech Communication, 65, 20-35.
14. Nenkova, A., Gravano, A., & Hirschberg, J. (2008, June). High frequency word entrainment in spoken dialogue. In Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers (pp. 169-172). Association for Computational Linguistics.
15. Pardo, J. S. (2006). On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America, 119(4), 2382-2393.
16. Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. Attention, Perception, & Psychophysics, 72(8), 2254-2264.

17. Reitter, D., & Moore, J. D. (2007). Predicting success in dialogue. In Proceedings of the 45th Annual Meeting of the Association of Computational
18. Štefan. B. (2012). Social aspects of entrainment in spoken interaction. "Cognitive Computation 6.4 (2014): 802-813.
19. Trancoso, I., do Céu Viana, M., Duarte, I., & Matos, G. (1998). Corpus de diálogo CORAL. in PROPOR'98 - III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, Porto Alegre, Brasil, 1998.
20. Viana, M. C. , Trancoso, I., Duarte, I. , Matos, G., Oliveira, L. C., Campos, H. C. & Correia (1998) "Apresentação do Projecto CORAL-Corpus de Diálogo Etiquetado", in Marrafa, P. & M. A. Mota (orgs) Linguística Computacional: Investigação Fundamental e Aplicações. Lisboa: Edições Colibri/APL, pp. 337-345, 1998.
21. Weidman, S., Breen, M., & Haydon, K. C. (2016). Prosodic Speech Entrainment in Romantic Relationships. In proceedings of Speech Prosody 2016, Boston.