Interspeech 2018
2-6 September 2018, Hyderabad

Acoustic-prosodic entrainment in structural metadata events

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

[1] INESC-ID Lisboa, Laboratório de Sistemas de Língua Falada, Portugal
[2] FLUL - Faculdade de Letras da Universidade de Lisboa; CLUL - Centro de Linguística da Universidade de Lisboa, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[4] Instituto Superior Técnico, Universidade de Lisboa, Portugal

{vera, fmmb, helena.moniz, isabel.trancoso}@inesc-id.pt, aim@letras.ulisboa.pt

Abstract

This paper presents an acoustic-prosodic analysis of entrainment in a Portuguese map-task corpus. Our aim is to analyze how turn-by-turn entrainment varies with distinct structural metadata events: types of sentence-like units (SU) in consecutive turns (e.g. interrogatives followed by declaratives, or both declaratives), and with the presence of discourse markers, affirmative cue words, and disfluencies in the beginning of turns. Entrainment at turn-exchanges may be observed in terms of pitch, energy, duration, and voice quality. Regarding SU types, question-answer turns are the ones with stronger similarity, and declarative-interrogative pairs are the ones where less entrainment occurs, as expected. Moreover, in question-answer pairs, there is also stronger evidence of entrainment with Yes/No and Tag questions than with Wh- questions. In fact, these subtypes are coded in distinctive prosodic ways (moreover, the first subtype has no associated lexical-syntactic cues in Portuguese, only prosodic). As for turn-initial structures, entrainment is stronger when the second turn begins with an affirmative cue word; less strong with ambiguous structures (such as 'OK'), emphatic affirmative answers, and negative answers; and scarce with disfluencies and discourse markers. The different degrees of local entrainment may be related with the informative structure of distinct structural metadata events.

Index Terms: entrainment, acoustic-prosodic features, structural metadata events, dialogues

1. Introduction

Entrainment, also known as accommodation or adaptation between speakers in a conversation, has been described as the ability shared by humans to adjust their speech and behavior to their interlocutors ([1], [2]). This strategy is studied to understand the underlying linguistic, psychological and social mechanisms, as well as to replicate this typically human behavior in automatic systems. It is known that entrainment plays a crucial role in solving specific tasks or making a speaker more likable and attractive to their interlocutor [3], as well as impacting in the success of spoken dialogue systems ([4], [5]).

In an acoustic-prosodic perspective, the topic of our work, entrainment has been largely studied in languages such as English, Mandarin, and even Slovak, but in European Portuguese (EP) this topic is just starting to be explored. Our study aims to analyze local entrainment in spontaneous speech (turn-by-turn), not just to verify if speakers adjust their acoustic-prosodic behavior to their interlocutors in turn-exchanges, but mainly to see if that adjustment is influenced by structural metadata events, namely discourse markers and different types of sentence-like units (SU). More concretely, we aim to verify if this local entrainment is more prone to occur between different types of SU (e.g., question-answer pairs) or when the second turn begins, for example, with a discourse marker (e.g., now; well; so); an affirmative cue word (e.g., yes; exactly; grunts, such as *humhum* and *hum*); or a disfluency (e.g., filled pauses). We also want to study in which acoustic-prosodic features speakers present more similarities.

This paper is organized as follows: Section 2 overviews a selection of literature on this subject; Section 3 describes the corpus used, as well as the set of prosodic features and metrics applied; Section 4 presents our results of local entrainment in structural metadata events. Finally, Section 5 presents our conclusions and future work.

2. State-of-the-art

Our study covers two topics, entrainment and structural metadata events. Although both have been extensively covered separately, we have found no literature on entrainment between different turn types, and very few works on analyzing entrainment with specific structures.

The study of entrainment is not new in the literature (e.g., [6], [7]). In the Accommodation Theory, [7] describe accommodation as a multiply organized and contextually complex set of alternatives that are available to speakers in face-to-face conversations, functioning to achieve solidarity with or dissociation from a conversational partner.

In most recent studies, entrainment is not analyzed per se but in its implications towards a certain goal, whether the success of the task ([8], [9]), social variables ([3], [10]) or power relations ([11]). Entrainment has also been studied under multiple perspectives: acoustic-prosodic ([12], [13], [5]), phonetic-phonological ([14]), lexical-syntactic ([15], [8]), multimodal, via facial expressions and gestures ([16]).

The topic of our paper is acoustic-prosodic entrainment in spontaneous speech. In this line, [5] measured the adaptation of speakers at different levels, all the dialogue (global) and turn-by-turn (local), using the Columbia Games Corpus. Globally, the author found that speakers are more similar to their partners than to their non-partners (i.e., speakers with whom they were never paired with) in mean and max intensity, and speaking rate. Locally, speakers were more similar at adjacent turns than at non-adjacent ones in intensity mean, max, and HNR, even though they do not match in speaking rate. Moreover, the author also found evidences of entrainment on turn-taking cues, namely that a speaker tends to use a cue also used by the interlocutor and that speakers have more cues in common with each

10.21437/Interspeech.2018-2366

other than with random other speakers.

Studying entrainment in the use of 'no' (meaning 'yes' in Slovak) [17], the author found less entrainment than expected in the frequency of 'no' between interlocutors. He also found that speakers tend to entrain (both with 'no' and in the remaining conversation) on intensity and voice quality features.

In EP, only two studies have addressed adaptation between speakers, [18], and [19], both using a subset of the same map-task corpus used in this study. [18] found evidences of prosodic correlations (pitch concord effects) between Yes/No-questions and affirmative answers; and [19] found global entrainment expressed in different degrees, since speakers did not entrained with the same partners and in the same features. Results showed that, despite the role speakers were playing (giver or follower), they tended to display more sensitivity to some partners, which may reveal a stronger partner effect than a role effect.

There is a vast amount of work on identifying structural metadata events, given their relevance to enrich the output of speech recognition systems, namely to recover sentence boundaries, disfluencies and discourse markers [20]. In particular, it has been shown that discourse markers, along with other words that occur mainly turn-initial, are hard to identify, presenting higher error rates when automatically recognized [21].

For EP, [22] performed an automatic classification task to distinguish discourse markers, disfluencies and SU in manually transcribed university lectures and map-task dialogues. Results showed that turn-initially discourse markers can be fairly discriminated from disfluencies and SU, even though their acoustic-prosodic discrimination still poses several challenges, due mostly to the fact that disfluencies and discourse markers share acoustic-prosodic properties.

3. Corpora annotation

The corpus used in this study is the CORAL corpus (ISLRN 499-311-025-331- 2) [23], which comprises 64 dialogues in map-task format between 32 speakers. The dialogues occur between two speakers with different roles (giver, and follower). CORAL is balanced in terms of gender and role played by the speaker (all speakers play both roles twice with different interlocutors). The corpus has 7 hours orthographically transcribed, and totals 61k words. In this work, we use a subset of the corpus, comprising 48 dialogues between 24 speakers. The subset is divided into sentence-like units (SU), with a total of about 42k words.

Table 1 lists the distinct structural metadata events that are studied, showing examples and percentages of occurrences. In terms of SU patterns in consecutive turns, we include declarative (DECL), and interrogative (INT) SU[1], as well as discriminating among three subtypes of interrogatives (Yes/No, Tag, and Wh- questions). The study also covers discourse markers (DMs), affirmative cue words (ACW), ambiguous structures (AMB) – words that can be either a DM or an ACW, and disfluencies (DISF) in the beginning of turns. Moreover, we also analyze other types of structures that are very frequent turn-initially in our data, namely emphatic repetition (EMP), and negative answers (NEG).

3.1. Set of prosodic features and metrics

Our experiments use eGeMAPS [24], a set of 88 acoustic-prosodic features, well-known for their usefulness in a wide range of paralinguistic tasks. To perform a turn-by-turn entrainment analysis, we compared the acoustic-prosodic features between the end of a turn, produced by a speaker, with the beginning of the next one, produced by his/her interlocutor. We applied the metrics defined by [13] and [5], represented in equations 1 and 2. The two metrics are based in Inter-Pausal Units (IPU), pause-free units of speech from a single speaker separated from one another by at least 50ms [25, 5]. When applying these metrics, we had to adjust the unit of analysis to fit EP phonological phenomena, such as truncations of post-tonic material, affrication, or aspiration. Another reason for adjusting the unit of analysis was the delimitation of our target (turn-initial) structures. Instead of selecting the initial and final IPU for each sentence, we selected the initial and final words produced within a 500ms interval. This fixed minimal unit of analysis was empirically tested and proved to be the most fruitful threshold with one or more words per unit of analysis, allowing to extract discourse markers and affirmative cue words, which can be a single word or correspond to a multiword unit (e.g., *então, está bem* / 'so, ok'; *pronto depois* / 'ok then'). Such interval can also be used to facilitate the automatic classification of the target structures and to produce entrainment models for spoken dialogue systems.

$$PartnerDistance = |IPU_t - IPU_p| \quad (1)$$

$$OtherDistance = \frac{\sum |IPU_t - IPU_i|}{10} \quad (2)$$

Several t-tests are then applied, in order to determine: (i) if the similarities are greater between adjacent or non-adjacent turns (Partner distance vs. Other distance); (ii) considering only adjacent turns, if speakers are more similar to each other when the utterance occurs between specific turn types or when a turn begins with a specific structure.

4. Local entrainment results

Evidences of acoustic-prosodic entrainment between speakers per dialogue (globally) were presented by [19], in the same map-task corpus. Now, we aim at verifying if the same speakers also show similarities to each other but at turn exchanges (locally). This study compares entrainment between adjacent turns and non-adjacent ones, revealing that there are substantial statistically significant differences between both groups ($p<0.001$) in 85 out of the 88 of the acoustic-prosodic features analyzed. These results reinforce the ones found for global entrainment, since speakers match their interlocutors at turn-exchanges in pitch, energy, duration, and voice quality features. Globally, speakers matched their partners significantly only in three features: pitch mean rising slope, duration of speech (with and without internal silences), and phonation ratio. These results are not in line with those found by [5] for American English, where speakers match with each other locally in intensity mean, max, and HNR, but not in pitch. Therefore, we can hypothesize that features like energy could be language independent, at least in similar corpora, but not pitch.

4.1. Experiments with different turn types

Considering declarative (DECL) and interrogative (INT) turns, results also show that speakers are more similar between adjacent turns than between non-adjacent ones for the main prosodic parameters: pitch, energy, duration, and voice quality features. However, a Kruskall-Wallis test, comparing only the adjacent

[1] In our corpus, exclamative turns only occur after declarative ones. Therefore, we excluded this SU type from the current analysis.

SU patterns			Turn-initial structures		Examples
DECL-DECL		7%	DM	13%	*agora* 'now'; *bem/bom* 'well'; *portanto* 'ok'; *então* 'so'
INT-DECL	Yes-No (60%)	2%	ACW	44%	*sim* 'yes'; *exacto/exactamente* 'exact/exactly'; *certo* 'certainly'; grunts *(humhum and hum)*; frozen form of the verb *ser* 'to be'
	Tag (15%)				
	Wh- (21%)		EMP	5%	*sim, sim, sim* 'yes, yes, yes'
DECL-INT		2%	AMB	18%	*pronto* 'ok'; ok
INT-INT		0.2%	DISF	13%	filled pauses aa; aam
			NEG	8%	não 'no'; eu não tenho 'I don't have that'

Table 1: *Patterns of turn types and turn-initial structures annotated with the percentage of occurrences.*

	INT-INT	DECL-INT	DECL-DECL
INT-DECL	4/1	**40/15**	**25/5**
DECL-DECL	0/1	**35/10**	
DECL-INT	5/6		

	WH-DECL	TAG-DECL
Yes/No-DECL	6/5	14/10
TAG-DECL	13/7	

Table 2: *Ratio of features where speakers are more similar, per SU patterns.*

	NEG	DISF	DM	AMB	EMP
AFF	19/11	**31/12**	**45/13**	32/26	**20/9**
EMP	16/16	**26/8**	**29/10**	12/22	
AMB	16/10	**22/7**	**49/10**		
DM	**8/30**	**12/30**			
DISF	11/23				

Table 3: *Ratio of features where speakers are more similar, in DECL-DECL turns.*

turns for both turn types, also reveals that there are statistically significant differences between them ($p<0.01$ and $p<0.05$) in the majority of the acoustic-prosodic features. This shows that, even though adjacent turns are always more similar than non-adjacent ones, they also differ according to the turn type, allowing us to hyphotesize that the ending intonation of each turn influences the following one.

In order to verify in which SU types speakers are more similar, we performed a t-test comparing the different patterns between two consecutive turns and reported the amount of features (from a total of 88) where each pair is more similar. Table 2 shows the corresponding results, where each cell presents the ratio of features, with statistically significant differences ($p<0.001$ and $p<0.05$), where speakers are more similar, for each combination of SU patterns. Results show that speakers are more similar between question-answer (INT-DECL) turns than between DECL-DECL sentences (25/5), or between DECL-INT (40/15). In both comparisons, question-answer pairs are more similar in terms of the four main acoustic-prosodic parameters: pitch, energy, duration, and voice quality. Stronger evidences for entrainment are also found between DECL-DECL turns when compared with DECL-INT ones, as speakers show similarities in 35 features in the first pair, opposed to only 10 in the second pair. When question-answer pairs are compared with INT-INT ones, results are less expressive, as fewer features present significant similarities between speakers (4/1). As for the comparison between DECL-DECL vs. INT-INT, and DECL-INT vs. INT-INT, results are very balanced, showing that there is no clear tendency for one pair to be more similar than the other.

To conclude, question-answer turns are the ones with stronger similarities between speakers, and declarative-interrogative pairs are the ones where less entrainment occurs. In our data, we observed that these declarative turns usually correspond to an answer of a previous question or an information about the position in the map followed by a question about the next step to complete the task. This map-task corpus is characterized for its collaborative nature, where speakers interact with the common goal of completing the map as fast as they can. Therefore, it was expected that question-answer pairs were the ones showing more entrainment.

Looking only at question-answer pairs, there are also degrees of entrainment between the different types of interrogatives, even though the differences are not as strong as expected. The comparison between Yes/No and Wh- questions, both followed by a declarative answer, show that both patterns present a statistically significant difference in 11 features: in 6 of them, speakers are more similar when there is a Yes-No question, and in 5 features when there is a Wh-. As for the patterns Yes/No-DECL vs. TAG-DECL, results also show that speakers are more similar between Yes/No questions and the following answer (14/10). The similarities occur in pitch, energy, frequency and spectral parameters. Finally, when comparing TAG-DECL with Wh-DECL, there are stronger evidences for entrainment in the first pattern (13/7). These results may be explained by the fact that these SU are coded in distinctive prosodic ways.

In EP, declarative turns are associated to low/falling nuclear contours (e.g. [26], [27], [28], [29]), and [30] associates the neutral declarative to the contour H+L* L%. As for interrogative turns, Wh- questions are characterized with a descending intonational contour, similarly to declarative sentences ([31, 30]); yes-no questions are characterized in spontaneous speech in EP by [32] with both Low-falling or Low-rising contours; and by [27], with the contour H* HL* H%. In data collected in laboratory, [29] characterized them with the contour H+L* LH%. Moreover, this subtype has no associated lexical-syntactic cues in Portuguese, only prosodic, unlike English where Yes/No questions can be coded with an auxiliary verb and subject inversion. As for Tag questions, [32] associated them to a Low-rising melody. The fact that both Yes/No and Tag questions present high/rising boundary tones, and that declarative sentences tend to present a prenuclear tone H in the first accented syllable may explain why there are evidences for more entrainment between these pairs of SU than with Wh-question-answer pairs.

4.2. Experiments with different turn-initial structures

This analysis was performed only between interrogative turns, namely Yes/No and Tag questions (both showed similar results and were therefore joined as a class to account for more occurrences), followed by a declarative answer, and DECL-DECL

	NEG	DISF	EMP
AFF	13/12	15/6	13/4
EMP	15/15	10/12	
DISF	12/13		

Table 4: *Ratio of features where speakers are more similar, in Tag/Yes-No - DECL turns.*

turns. This selection was due to the small amount of occurrences (less than 20) of our target structures in the beginning of the second turn for the remaining patterns (INT-INT, DECL-INT, and Wh-DECL). As expected, for INT-DECL turns, discourse markers and ambiguous structures also have a small amount of occurrences, 21 and 24, respectively, and are therefore excluded from this analysis. Tables 3 and 4 show the corresponding ratio of features, with statistically significant differences ($p<0.001$, and $p<0.05$) where speakers are more similar.

In question-answer pairs, results show that speakers entrain in more features when the answer is an affirmative cue word rather than an emphatic affirmative answer (13/4), mainly in pitch, jitter, and HNR, or a disfluency (15/6), mainly in voiced quality features; voiced and unvoiced segments length. As for affirmative cue words and negative answers, results are very similar, as speakers show evidences for entrainment in a similar amount of features (13/12).

Regarding DECL-DECL turns, speakers also tend to be more similar to their interlocutors when there is an affirmative cue word than emphatic affirmative answer (more similarities in 20 features, mainly in pitch and energy); discourse markers (45/13), ambiguous structures (32/26), both in pitch, energy, spectral parameters, and voice quality features, and disfluencies (31/12), in energy, voice quality features and voiced/unvoiced segments). Contrarily to question-answer pairs, affirmative cue words and negative answers are not balanced in terms of the amount of features where speakers entrain (19/11). When comparing DMs with all the other structures analyzed, results show that this class is where speakers entrain less. As for disfluencies, there are evidences for more entrainment only when compared with DMs (30/12). Affirmative emphatic, ambiguous structures and negative answers show more entrainment than disfluencies in the majority of the features. Therefore, entrainment is stronger when the second turn begins with an affirmative cue word, both with a declarative and an interrogative context, less strong with ambiguous structures, emphatic affirmative answers, and negative answers; and scarce with disfluencies and discourse markers. These different degrees of local entrainment may be related with the informative structure of these events. In our data, affirmative cue words have multiple pragmatic functions, like expressing feedback or acting as a backchannel. Regardless of their function, they contribute to the fluidity of the dialogue and signal the collaborative nature of the corpus. On the other hand, both discourse markers and disfluencies are generally defined as syntactically detached structures with no propositional content, that share acoustic-prosodic properties according to their pragmatic context: discourse markers that have a function similar to disfluencies, like stalling, may share with them some properties, meaning the plateau contours contrasting with the rises in the following prosodic constituents.

To summarize, in our data, entrainment is influenced by the SU types and by the structures that occur turn-initially: speakers tend to be more similar to their partners at turn-exchanges in question-answer pairs, showing more entrainment in a greater number of features, than with any other of the patterns analyzed; and also entrain more when a turn begins with an affirmative cue word than with a disfluency or a discourse marker.

5. Conclusions

To the best of our knowledge, this paper presents the first local entrainment (turn-by-turn) analysis for EP. We have investigated how distinct structural metadata events, namely types of SU in consecutive turns (e.g. interrogatives followed by declaratives, or both declaratives), and the presence of discourse markers, affirmative cue words, and disfluencies in the beginning of turns, influence the acoustic-prosodic adaptation between speakers.

Our results on local entrainment without considering the sentence types or any kind of specific structures reveals that speakers are more similar between adjacent turns than between non-adjacent ones in the four main acoustic-prosodic parameters: pitch, energy, duration, and voice quality. These results are not in line with a similar analysis performed by [5] for American English, which found that speakers match with each other at turn-exchanges in intensity mean, max, and HNR, but not in pitch. These results may lead us to hypothesize that features like energy could be language independent, at least in similar corpora, but not pitch. The experiments conducted so far show that acoustic-prosodic behavior of local entrainment in EP spans from energy to all the other prosodic parameters.

Considering local entrainment between distinct SU types, question-answer pairs are the ones with stronger similarity in the majority of the pitch, energy, duration, and voice quality features, and declarative-interrogative pairs are the ones where less entrainment occurs. These results were expected given the collaborative nature of the corpus. As for the subtypes of interrogatives in question-answer pairs, with Yes/No and Tag questions there are stronger evidences for entrainment than with Wh- questions. The first two share a high/rising boundary tone, opposed to the low/falling nuclear contour of Wh- questions, a contour similar to neutral declaratives in EP. Moreover, Yes/No questions have no associated lexical-syntactic cues in Portuguese, only prosodic, an evidence more for their contribution for the local entrainment found. It is also worth mentioning that question-answer pairs are the driven force of the dialogic nature of our corpus, very collaborative tasks to be solved together. The fluidity of a dialogue is built upon several strategies and our data shows that the structures evidencing stronger local entrainment are the ones more prone to show collaboration and feedback.

In line with what has been said for SU, the stronger local entrainment occurs with affirmative cue words. This structure is an evidence more of the collaborative effort between the interlocutors to solve the task. On the other hand, disfluencies and discourse markers are the structures showing less degree of entrainment. A possible explanation is the fact that when speakers utter disfluencies and discourse markers they are planning what to say next or stalling. The stalling patterns in EP are plateaus distinguishable from the prosodic patterns of other linguistic structures.

In a future work, we intend to perform a more fine-grained analysis of the different pragmatic functions of discourse markers and affirmative words, to verify how they correlate with entrainment. We also aim at extending this study to other domains.

6. Acknowledgements

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, and under PhD grant SFRH/BD/96492/2013, and Post-doc grant SFRH/PBD/95849/2013.

7. References

[1] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation," *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 22, no. 6, pp. 1482–1496, 1996. [Online]. Available: http://www-psych.stanford.edu/ herb/1990s/Brennan.Clark.96.pdf

[2] H. Giles, A. Mulac, J. J. Bradac, and P. Johnson, "Speech accommodation theory: The first decade and beyond," *Annals of the International Communication Association*, vol. 10, no. 1, pp. 13–48, 1987.

[3] Š. Beňuš, "Social aspects of entrainment in spoken interaction," *Cognitive Computation*, vol. 6, no. 4, pp. 802–813, 2014.

[4] R. Coulston, S. Oviatt, and C. Darves, "Amplitude convergence in children's conversational speech with animated personas," in *Proceedings of the 7th International Conference on Spoken Language Processing*, vol. 4, 2002, pp. 2689–2692.

[5] R. Levitan, "Acoustic-prosodic entrainment in human-human and human-computer dialogue," Ph.D. dissertation, Columbia University, 2014.

[6] H. P. Grice, "Logic and conversation," *1975*, pp. 41–58, 1975.

[7] H. Giles, N. Coupland, and I. Coupland, "1. accommodation theory: Communication, context, and," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

[8] A. Nenkova, A. Gravano, and J. Hirschberg, "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*. Association for Computational Linguistics, 2008, pp. 169–172.

[9] D. Reitter and J. D. Moore, "Alignment and task success in spoken dialogue," *Journal of Memory and Language*, vol. 76, pp. 29–46, 2014.

[10] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "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*. Association for Computational Linguistics, 2012, pp. 11–19.

[11] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg, "Echoes of power: Language effects and power differences in social interaction," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 699–708.

[12] A. Gravano, Š. Beňuš, R. Levitan, and J. Hirschberg, "Three tobi-based measures of prosodic entrainment and their correlations with speaker engagement," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 578–583.

[13] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." in *Interspeech*, 2011, pp. 3081–3084.

[14] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[15] J. D. Lopes, "Lexical entrainment in spoken dialog systems," Ph.D. dissertation, INSTITUTO SUPERIOR TÉCNICO, 2013.

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

[17] Š. Beňuš, "Conversational entrainment in the use of discourse markers," in *Recent Advances of Neural Network Models and Applications*. Springer, 2014, pp. 345–352.

[18] V. Cabarrão, A. I. Mata, and I. Trancoso, "Affirmative constituents in european portuguese dialogues: prosodic and pragmatic properties," in *Proceedings of Speech Prosody*, 2016.

[19] V. Cabarrão, I. Trancoso, A. I. Mata, H. Moniz, and F. Batista, "Global analysis of entrainment in dialogues," in *Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3*. Springer, 2016, pp. 215–223.

[20] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[21] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[22] V. Cabarrão, H. Moniz, F. Batista, J. Ferreira, I. Trancoso, and A. I. Mata, "Cross domain analysis of discourse markers in european portuguese," *Dialogue and Discourse*, accepted for publication.

[23] I. Trancoso, M. do Céu Viana, I. Duarte, and G. Matos, "Corpus de diálogo CORAL," in *PROPOR'98*, Porto Alegre, Brasil, 1998.

[24] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[25] A. Gravano, "Turn-taking and affirmative cue words in task-oriented dialogue," Ph.D. dissertation, Columbia University, 2009.

[26] M. C. Viana, "Para a síntese da entoação do Português," Ph.D. dissertation, University of Lisbon, 1987.

[27] I. Falé, *Fragmentos da prosódia do português europeu: as estruturas coordenadas*, 1995.

[28] M. Cruz-Ferreira, "Intonation in European Portuguese," in *Intonation systems*, D. Hirst and A. Di Cristo, Eds. Cambridge: Cambridge University Press, 1998, pp. 167–178.

[29] S. Frota, *Prosody and Focus in European Portuguese. Phonological Phrasing and Intonation.* New York: Garland Publishing, 2000.

[30] M. C. Viana, S. Frota, I. Falé, I. Mascarenhas, A. I. Mata, H. Moniz, and M. Vigário, "Towards a P_ToBI," in *Workshop of the Transcription of Intonation in the Ibero-Romance Languages, PaPI 2007*, Minho, Portugal, 2007.

[31] M. Cruz-Ferreira, "Intonation in european portuguese," *Intonation systems. A survey of twenty languages*, pp. 167–178, 1998.

[32] A. I. Mata, "Questões de entoação e interrogação no Português. Isso é uma pergunta?" Master's thesis, University of Lisbon, 1990.