Testing the Effects of Acoustic/Prosodic Entrainment on User Behavior at the Dialog-Act Level

Lara Gauder[1], *Marisol Reartes*[1], *Ramiro H. Gálvez*[1,2], *Štefan Beňuš*[3,4], *Agustín Gravano*[1,2]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Instituto de Ciencias de la Computación, CONICET-UBA, Buenos Aires, Argentina
[3] Constantine the Philosopher University in Nitra, Slovakia
[4] Institute of Informatics, Slovak Academy of Sciences, Slovakia

mgauder@dc.uba.ar, mreartes@dc.uba.ar, sbenus@ukf.sk, rgalvez@dc.uba.ar, gravano@dc.uba.ar

Abstract

Entrainment has been documented across several dimensions of human-human dialog. Experimental studies relating dialog success and acoustic/prosodic (a/p) entrainment in spoken dialog systems, point towards a non-neutral effect of a/p entrainment. But results also suggest the presence of positive effects of disentrainment. A plausible driver behind this last result could be that in these experiments both users and avatars were restricted in their use of dialog acts. In particular, systems only produced answers which entrained to the sole dialog act produced by users: requests for advice. Given that there is significant documented correlation between dialog acts and a/p features, it seems reasonable to hypothesize that a/p entrainment may occur at the dialog-act level and that entraining or disentraining across dialog acts may introduce misleading artifacts. This paper presents the design and implementation of an experimental setup which allows to implement entrainment and disentrainment policies at the dialog-act level. It also presents results of a pilot study in Argentine Spanish.

Index Terms: entrainment, dialog acts, spoken dialog systems, wizard-of-oz, competence, trust, Argentine Spanish

1. Introduction

ENTRAINMENT – a phenomenon by which conversational partners become more similar to each other in their behavior [1, 2, 3, 4, 5] – has been documented across several dimensions of human-human dialog [5, 6, 7]. In particular, empirical evidence of entrainment has been found for acoustic-prosodic (a/p) features such as intensity [8, 9, 10], speaking rate [3], and pitch [9, 10]. Moreover, many studies have found links between entrainment and positive social behavior [11, 12, 13, 14, 15, 16, 17, 18].

Most studies relating dialog success and a/p entrainment base their conclusions on the analysis of corpora, and thus only describe correlations of limited applicability. For this reason, recent studies have designed and implemented entraining spoken dialog systems (SDSs) – systems in which the a/p values of the synthesized speech adapt to the way the user speaks – in order to experimentally test the effects of entrainment on several aspects of interaction [19, 20, 21, 22].

Although the direction of these effects is not clear, results point towards a non-neutral effect of a/p entrainment regarding users' behavior. In particular, an intriguing result obtained in a series of experiments reported in [21] indicates that users asked more often for advice from avatars which *disentrained* on speech rate compared to avatars which entrained. DISENTRAINMENT – when speakers adjust *away* from their interlocutors – is a less understood phenomenon; it has also been observed in corpus studies (see, for example, [23, 24]) and has been associated with positive social behavior [25].

A plausible driver behind the results obtained in [21] could be the fact that, in the setup used in that study, both users and avatars were restricted in their use of DIALOG ACTS – i.e., utterance meanings at the level of illocutionary force [26, 27]. In particular, users were only able to request advice from avatars, and avatars were only able to provide advice to users. Given the existence of significant correlations between dialog acts and a/p features [27, 28, 29], it seems reasonable to hypothesize that *acoustic/prosodic entrainment may occur at the dialog-act level* and that entraining or disentraining across dialog acts may introduce misleading artifacts. For example, it may turn out to be inappropriate to extract a/p features from a user's request for advice, and later use them to guide the way an advice is synthesized, as done in [21].

In the present paper we present the design and development of an experimental setup which allows to implement entrainment and disentrainment policies at the dialog-act level. Additionally, we present results of a pilot study testing the proposed setup.

2. Experimental Setup

To test the effects of entrainment at the dialog-act level we created a computer version of the popular children's game GUESS-WHO. In this section we describe in detail our implementation in Argentine Spanish.

2.1. Game rules

In our variant of the GuessWho game, two players (a human subject and a computer avatar) play against each other. Initially, each player has a board of 32 tiles showing the face and the name of cartoon characters (Fig. 1, area 1). They also have an extra, special tile called the TARGET TILE (Fig. 1, area 5). The first player that correctly deduces the character on the opponent's target tile wins the game.

Players take turns to ask yes/no questions about aspects of the cartoon characters on the visible tiles on their board (e.g., *does your character wear glasses?*). The bottom left corner of the screen displays the list of available aspects of the tile characters about which subjects may ask their opponent (Fig. 1, area 6). Upon hearing the answer, the system eliminates candidates by flipping down the tiles which are inconsistent with the new information. This process repeats until one of the players can deduce the character on the opponent's target tile. Subjects can always see the number of remaining tiles on the avatar's

10.21437/SpeechProsody.2018-76

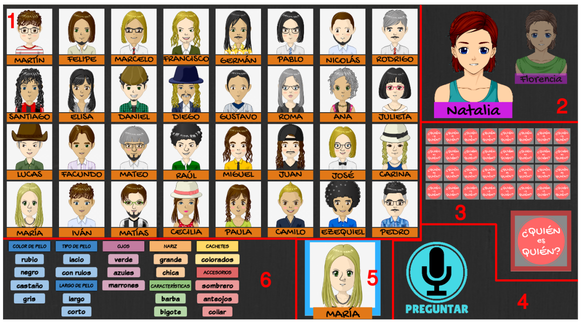

Figure 1: *Screenshot of the main screen of our implementation of GuessWho. Note: see text for the description of the numbered areas.*

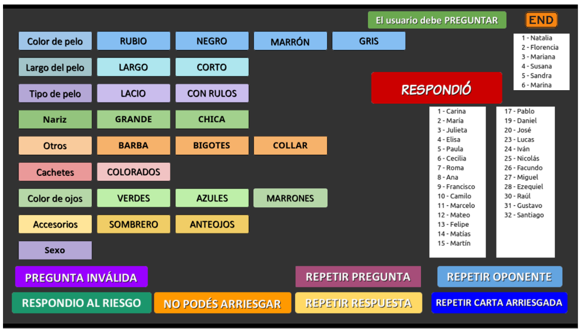

Figure 2: *Screenshot of the Wizard-of-Oz's screen.*

board (Fig. 1, area 3). When a player has only one tile left, the opponent is forced to risk a character.

During the game, subjects wear a Genius HS-400A headset with microphone. The lower right corner of the game screen shows a microphone button (Fig. 1, area 4), which works using the well-known *push-to-talk* metaphor: Subjects are instructed to keep this button pressed while speaking.

In order to have more speech material from the subjects, they are required to ask and answer questions using full sentences. For example, they should answer *no, my character does not wear glasses*, instead of simply answering *no, he doesn't*.

2.2. Session structure

Sessions were recorded, one subject at a time, in a silent laboratory environment. Each session consisted of 16 GuessWho games between a human subject and different computer avatars, divided into five rounds.

Round 1 consists of a single game played against a single avatar. The objective of this round is to familiarize subjects with the game rules and its interface. In this round we measure the subject's base values for the a/p features of interest (see Section 3 below). The avatar's speech is synthesized using the default a/p settings of the TTS engine.

Rounds 2, 3 and 4 share the same structure: the subject plays four games alternating between two avatars, displayed at the top right corner of the screen (Fig. 1, area 2). The avatars' names are Natalia and Florencia in round 2, Susana and Mariana in round 3, and Sandra and Marina in round 4. The two avatars in each round look very similar to one another, except for their hair color and clothes. One minor difference between these rounds is that in round 3, instead of showing cartoon characters, the tiles show geometric figures.

Importantly, subjects are told that each avatar chooses its actions according to two different artificial intelligence algorithms, one believed to be superior to the other. In reality, all avatars use the same algorithm, and we only vary their a/p ENTRAINMENT POLICY. Concretely, in each round one of the avatars entrains on the subject's speech by doing *synchrony*, whereas the other one disentrains by doing *anti-synchrony* [30, 25] (more on this in Section 3). To avoid order effects, we counterbalanced across subjects which avatar starts each round, and which avatar follows each entrainment policy. To avoid gender effects we also balanced female and male subjects across conditions.

In addition, we designed each round so that the subject wins exactly two games (one against each avatar) and loses two games (again, one against each avatar). Specifically, the subject wins, wins, loses, loses in round 2; wins, loses, loses, wins in round 3; and loses, wins, wins, loses in round 4. This design remains constant across all subjects. The game manipulations that we implemented to guarantee these results are explained in Section 2.3.

After the last game in each of rounds 2, 3 and 4, subjects are asked to select an avatar against whom to play in one more game during round 5. This final round consists of three games, but in this case, subjects are paid a small monetary prize for every game they win. Since subjects are aware of these prizes, they have an incentive to choose the avatars they consider to be *less competent*. Which avatars get chosen here, and how these choices relate to their entrainment policies is our main research interest.

2.3. Game manipulation

As explained above, the outcome of each game (whether the subject should win or lose) is defined beforehand. When the subject is supposed to lose a game, the avatar is given a target tile that requires at least six questions to be pinpointed, and the avatar is programmed to pinpoint the subject's tile faster than that. On the other hand, when the subject is supposed to win a game, the avatar's questions are selected so as to flip as few tiles as possible.

As the game develops, two things may happen: 1) If the subject's board has only one tile left, the rules force the avatar to risk a character. In that case, if the subject is supposed to lose, then the avatar risks the correct tile; otherwise, the avatar risks a wrong tile. 2) Analogously, if the avatar's board has one tile left, then the subject must risk a character. If the subject is supposed to win, then whichever character they risk is claimed to be correct (note that the subject has no way of knowing the actual target tile); otherwise, their guess is claimed wrong and the avatar wins.

Finally, the three games in round 5 are not manipulated. The computer avatars choose at random which aspects to ask about in each round.

2.4. Wizard-of-Oz design

In this experiment the subjects' spoken productions were not restricted to a fixed grammar. For this reason, and given that in preliminary tests the transcriptions produced by automatic speech recognition (ASR) systems were not accurate enough, we opted to use a *Wizard-of-Oz* design to guide the system (this design has been used in the context of SDSs successfully in several studies, e.g., [31, 32]).

The wizard's task is to interpret the subject's utterances. For this purpose, we designed a graphical interface (Fig. 2) in which the wizard can press several buttons to indicate the relevant information produced by the subject. Additionally, the wizard may also inform the subject of any particular issues (e.g., releasing the push-to-talk button too early).

3. Acoustic-prosodic entrainment

The computer avatars created for our experiment either entrain or disentrain on the relative level of three a/p features: *speech rate* (measured in syllables per second), *pitch* (measured as F0 mean, in Hz) and *intensity* (measured as mean energy, in dB). We describe next how our avatars adapt at the dialog-act level to users' way of speaking.

3.1. Measuring acoustic/prosodic features values

In order to adapt to the users' speech, we first need to extract the a/p features from their utterances. For speech rate, we first use IBM's *Watson* ASR service[1] to obtain a time-aligned transcription of each utterance, and then estimate its syllable count using the syllabification algorithm presented in [33]. Given that the ASR timestamps are at the word level, we may identify continuous speech segments (i.e., sequences of words with no silence between them) for subsequent computations.

Concretely, suppose we plan to measure a/p feature k on utterance i, which has $j = 1, 2, \ldots, J$ continuous speech segments of duration d_1, d_2, …, d_J. Then for each segment we measure $\phi^k_{i,j}$, the value of a/p feature k on segment j, and then take the value of a/p feature k for the whole utterance i as

$$\phi^k_i = \Big(\sum_{j=1}^{J} d_j * \phi^k_{i,j} \Big) \Big/ \sum_{j=1}^{J} d_j$$

We used Praat [34] to estimate mean pitch[2] and mean intensity[3].

3.2. Speech history

To implement entrainment policies for our computer avatars, we first need to keep track of the evolution of the a/p features on the subject's utterances. In particular, at all times we keep track of $\Phi^k_{h-1,a}$, defined as the mean observed value of a/p feature k over all utterances corresponding to dialog act a produced in round $h-1$ (h being the round played at the moment). Concretely, we calculate:

$$\Phi^k_{h-1,a} = \frac{\sum_{i \in A_{h-1,a}} \phi^k_i}{|A_{h-1,a}|} \tag{1}$$

where $A_{h-1,a}$ is the set of all utterances corresponding to dialog act a produced in round $h-1$. In our game, subjects interact with the avatars mainly by using two dialog acts: *yes/no questions* and *yes/no answers*.[4] Therefore, $a \in \{$*yn-question*, *yn-answer*$\}$.

3.3. Entrainment algorithm

To synthesize the avatars' utterances we use Amazon's Polly Text to Speech (TTS) service.[5] This service allows to modify speech rate and pitch level on a percent basis using SSML tags (e.g., when $+10\%$ is introduced in the pitch tag, the speech is synthesized with a 10% higher pitch level, relative to the system's default value). We modified intensity using the open-source sound processing toolbox SoX.[6] We tested that the desired a/p variations given as input to the TTS system were achieved accurately, for which we used the same procedure described in Section 3.1 on the synthesized speech.

The proposed entrainment/disentrainment policies build upon the ones presented in [23]. Concretely, being i the subject's last utterance with dialog act a during round h, the target variation of a/p feature k for the avatar's following utterance with the same dialog act a is given by

$$\psi^k_{i,a} = \Big(100 * \frac{\phi^k_i - \Phi^k_{h-1,a}}{\Phi^k_{h-1,a}} * \mathit{policy} \Big)\ \% \tag{2}$$

where *policy* equals 1 or -1 if the avatar follows an entrainment or disentrainment policy, respectively. In other words, entraining avatars increase or decrease their a/p features in synchrony with the subject's, along utterances with the same dialog act. Disentraining avatars, on the other hand, do the opposite.

To preserve the naturalness of the synthesized voices, we clip maximum/minimum values of $\psi^k_{i,a}$ (+30%/-30% for speech rate, $+8\%$/-8% for intensity, and $+15\%$/-15% for pitch). These upper and lower bounds were chosen perceptually by the authors. Finally, as all rounds start with subjects questions, ϕ^k_i is undefined for the first answer given by the avatar; in this case we simply use the TTS system's default values.

4. Pilot study

To test our proposed design, we conducted a pilot study. Here we present its main results.

4.1. Subjects

The pilot study took place in August, 2017, in a quiet computer laboratory at the Computer Science Department, University of Buenos Aires. A total of 16 native speakers of Argentine Spanish (8 female, 8 male) between 18 and 65 years old (mean=29.8, sd=12.1) participated voluntarily.[7]

Subjects were told that they would receive a monetary compensation for participating (the local equivalent of roughly five US dollars per hour) plus an additional prize for each game won in the final round (two dollars per won game). The average game lasted 1h 20m.

4.2. Results

The data collected in the pilot study consists of 48 instances of subjects' choices of the avatars against which to compete in the final round (three such choices per subject). In 28 instances (58.3%) subjects selected the *entraining* avatar (first panel in

[1] http://www.ibm.com/watson/services/text-to-speech/

[2] http://fon.hum.uva.nl/praat/manual/Sound__To_Pitch___.html

[3] http://fon.hum.uva.nl/praat/manual/Sound__To_Intensity___.html

[4] Actually, subjects also guess characters and respond to the avatars' guesses at the end of each game, but for simplicity the system does not entrain to any of these contributions.

[5] http://aws.amazon.com/polly/

[6] http://sox.sourceforge.net

[7] This research design was approved by the Ethics Committee of CEMIC (http://www.cemic.edu.ar), under the protocol titled "Relationship between trust and entrainment in speech", PI Agustín Gravano, initially approved on 11/2014 and renewed annually in 12/2015 and 11/2016.

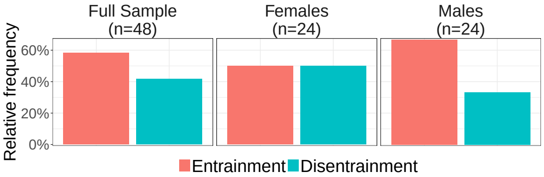

Figure 3: *Percentage of times entrainers and disentrainers avatars were selected.*

Fig. 3). A two-sided exact binomial test assuming a 0.5 probability of choosing the entraining avatar in each instance does not reject at standard significance levels the null hypothesis that the avatars were chosen randomly ($p \approx 0.31$).

When focusing on differential effects across genders (second and third panels in Fig. 3), we find that female subjects selected the entraining avatar in 12 out of 24 occasions (50%), while male subjects did so 16 out of 24 times (66.7%). In either case, an exact binomial test does not reject at standard levels the null hypothesis of choosing at random with probability 0.5, although it approaches significance for males (females: $p = 1$, males: $p \approx 0.15$).

When focusing on differential effects across rounds, we find that the entraining avatar was selected 10 out of 16 times in round 2 (62.5% overall, 62.5% for females, 62.5% for males), 10 out of 16 times in round 3 (62.5%, 50%, 75%), and 8 out of 16 times in round 4 (50%, 37.5%, 62.5%). Note that the observed preference for selecting the entraining avatar tends to be higher for male subjects across all rounds. In no case do exact binomial tests reject the null hypothesis of choosing avatars randomly.

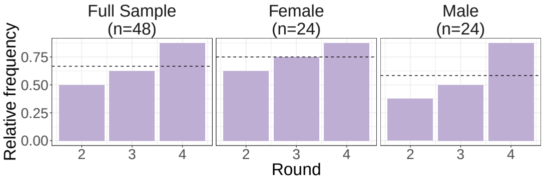

Figure 4: *Percentage of times the heuristic rule was selected. Note: dashed lines represent grand means across rounds.*

Lastly, driven by a subject's comment, we checked if the following heuristic rule was followed consistently: *"Select the avatar I played against in the fourth game if I just defeated her (because she is incompetent); do not select her if she defeated me (because she is skilled)."* Our data suggest that this rule was followed in 32 out of 48 occasions (66.7%, Fig. 4 panel 1). In this case, an exact binomial test rejects that it was followed randomly with an underlying 0.5 probability ($p \approx 0.03$). Moreover, this rule seems to have been followed more closely by female subjects (18/24, 75%, $p \approx 0.02$, Fig. 4 panel 2) than by male subjects (14/24, 58.3%, $p \approx 0.54$, Fig. 4 panel 3). Our data also suggest that this rule was followed more strongly when subjects lost the last game in a round (14/16, 87.5%, $p < 0.01$). However, given that subjects only lost the fourth game in round 4 (see Section 2.2), our design does not allow us to clearly identify whether subjects followed the rule because they had just lost, or simply because by the end of the third round they were tired. These results from rounds 2 and 3 (in which subjects won the last game) point towards some effect relation between tiredness and following the rule. Concretely the rule was followed 8 out of 16 times in round 2 (50%, $p = 1$) and 10 out of 16 times in round 2 (62.5%, $p \approx 0.45$).

4.3. Discussion

In our proposed design and analysis, we take as a measure of perceived competence the fact that a player avoided selecting an avatar as an opponent for the final round. Although not significant in statistical terms, our data suggest a tendency to consider disentraining avatars to be more competent than entraining avatars. This goes in hand with a positive effect of speech-rate disentrainment reported preliminarily from pilot studies by [21]. This effect in our data seems to be higher for male subjects, which also goes in hand with findings reported in previous research [35, 36].

Importantly, an alternative explanation for these results is that entrainment may have a positive effect on likeability, and that subjects may have chosen entraining avatars driven by likeability. Although we are not able to discard this effect as a driver behind the results, it should be noted that the strong effect of the simple heuristic rule described in Section 4.2 – which assumes low recall, the importance of economic incentives and has no relation with avatars' characteristics – suggests that users weighted strongly the opponents' competence level. Future iterations of the game should explore further into the relative importance of likeability in this sense.

5. Conclusions

To test the effects of entrainment at the dialog-act level we created a computer version of the popular children's game Guess-Who. Throughout a session, subjects play a series of Guess-Who games against computer avatars, where some of them follow an intra dialog-act entrainment policy on a/p features while the others follow an intra dialog-act disentrainment policy. We tested the proposed design running a pilot study involving 16 subjects. Results show a tendency to favor choosing entraining avatars as opponents for the final games – suggesting that participants perceived diseantraining avatars as more competent. In line with previous research [21], this places disentrainment as a positive feature regarding perceived avatar's competence. Additionally, effects seem to be stronger in "mixed-gender" games (when male subjects play against avatars using a female voice). Results should be taken as preliminary.

Future work should concentrate on five things. First, making the data acquisition process more scalable. Despite being an open grammar task, the feasibility of using ASR transcriptions instead of the wizard's interpretation should be further checked. Second, the effects of the simple heuristic rule used by subjects should be ameliorated. Our data suggest that one way of doing this could be to make the avatar lose the last game in each round. Third, measures of avatar likeability should be taken, in order to control its effects during data analysis. Fourth, our data suggest stronger effects in mixed-gender games, so the design should also explore the effects of male avatars on users. Fifth, effects of entrainment/disentrainment on different subsets of a/p features should be tested.

6. Acknowledgements

This material is based upon work supported by CONICET, ANPCYT PICT 2014-1561, and the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055.

7. References

[1] R. Levitan, Š. Benuš, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, 2015.

[2] A. Ward and D. Litman, "Measuring convergence and priming in tutorial dialog," University of Pittsburgh, Tech. Rep., 2007.

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

[4] A. Gravano, Š. Benuš, R. Levitan, and J. Hirschberg, "Backward mimicry and forward influence in prosodic contour choice in Standard American English," in *Proceedings of Interspeech*, 2015.

[5] 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, p. 1482, 1996.

[6] 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.

[7] D. Reitter, F. Keller, and J. D. Moore, "A computational cognitive model of syntactic priming," *Cognitive science*, vol. 35, no. 4, pp. 587–637, 2011.

[8] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability," *Journal of Personality and Social Psychology*, vol. 32, no. 5, pp. 790–804, 1975.

[9] S. Gregory, S. Webster, and G. Huang, "Voice pitch and amplitude convergence as a metric of quality in dyadic interviews," *Language & Communication*, vol. 13, no. 3, pp. 195–217, 1993.

[10] A. Ward and D. Litman, "Measuring convergence and priming in tutorial dialog," University of Pittsburgh, Tech. Rep., 2007.

[11] R. Y. Bourhis and H. Giles, "The language of intergroup distinctiveness," *Language, ethnicity and intergroup relations*, vol. 13, p. 119, 1977.

[12] R. L. Street, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[13] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1994, pp. 72–78.

[14] 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, pp. 893–910, 1999.

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

[16] C.-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. G. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proceedings of Interspeech*, 2010.

[17] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker, "Language style matching predicts relationship initiation and stability," *Psychological Science*, vol. 22, no. 1, pp. 39–44, 2011.

[18] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," in *Artificial Intelligence in Education*. Springer, 2013, pp. 750–753.

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

[20] N. Lubold, H. Pon-Barry, and E. Walker, "Naturalness and rapport in a pitch adaptive learning companion," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[21] R. Levitan, Š. Beňuš, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," *Interspeech 2016*, pp. 1166–1170, 2016.

[22] N. Lubold, E. Walker, and H. Pon-Barry, "Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion," in *11^{th} ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 255–262.

[23] R. Levitan, Š. Benuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison," in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, p. 325.

[24] P. G. Healey, M. Purver, and C. Howes, "Divergence in dialogue," *PloS one*, vol. 9, no. 6, p. e98598, 2014.

[25] J. M. Pérez, R. H. Gálvez, and A. Gravano, "Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement," *Interspeech 2016*, pp. 1270–1274, 2016.

[26] J. L. Austin, *How to do things with words*. Oxford university press, 1975.

[27] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[28] V. K. R. Sridhar, S. Narayanan, and S. Bangalore, "Modeling the intonation of discourse segments for improved online dialog act tagging," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 5033–5036.

[29] U. D. Reichel and J. Cole, "Entrainment analysis of categorical intonation representations," in *Proceedings Phonetik und Phonologie*, 2016.

[30] C. D. Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, no. Supplement C, pp. 11 – 34, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639313001386

[31] G. Skantze and A. Hjalmarsson, "Towards incremental speech generation in dialogue systems," in *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, ser. SIGDIAL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1–8. [Online]. Available: http://dl.acm.org/citation.cfm?id=1944506.1944507

[32] M. Ter Maat, K. P. Truong, and D. Heylen, "How turn-taking strategies influence users' impressions of an agent." in *10th International Conference, IVA*, 2010.

[33] Z. Hernández-Figueroa, F. J. Carreras-Riudavets, and G. Rodríguez-Rodríguez, "Automatic syllabification for spanish using lemmatization and derivation to solve the prefix's prominence issue," *Expert Systems with Applications*, vol. 40, no. 17, pp. 7122 – 7131, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741300451X

[34] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[35] 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.

[36] Z. Xia and Q. W. Ma, "Gender and prosodic entrainment in mandarin conversations," in *Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on*. IEEE, 2016, pp. 1–4.