Seeing or Not Seeing Your Conversational Partner: The Influence of Interaction Modality on Prosodic Entrainment

Michelina Savino[1](✉), Loredana Lapertosa[1], and Mario Refice[2]

[1] University of Bari, Bari, Italy
{michelina.savino,loredana.lapertosa}@uniba.it
[2] New York, USA
mario.refice5@gmail.com

Abstract. In speech entrainment research, a less investigated yet crucial aspect for modelling multimodal interactive dialogue systems is the influence of interaction modality, i.e. whether conversational partners who are visible to each other would entrain their speech more with respect to when eye contact is inhibited, or not. In our study, we compared prosodic adaptation behaviour (convergence and synchrony) of the same speaker pairs involved in collaborative game sessions under two conditions: audiovisual vs audio-only interaction. Results provide a complex picture, with a tendency to enhance vocal entrainment when the speech/audio channel is the only one available to conversational partners.

Keywords: Prosodic entrainment · Audiovisual interaction
Audio-only interaction · Italian dialogues

1 Background and Motivation of the Study

A large body of interdisciplinary research has shown that in human-human interaction, participants tend to coordinate their speech along various dimensions [1, 2], typically by making it sounding more similar to that of interlocutors' over the dialogue until they converge (convergence, e.g. [3]), and/or by producing similar speech patterns over time during the interaction (synchrony, [4, 5]). Such an imitation-based behaviour, variously termed as coordination, alignment, adaptation, entrainment or accommodation (all terms will be used interchangeably in this paper), has been basically accounted for as either a totally uncontrolled cognitive device [6], or as a "social device" for modulating social distance among conversational partners (Communication Accommodation Theory [7], henceforth CAT), both being crucial to mutual understanding and effective communication. However, recent outcomes seem pointing to a hybrid model [8], where interpersonal, social, situational, linguistic and cultural factors can strongly condition the automatic alignment process (e.g. [9]). In human-machine interaction, providing interactive dialogue systems with social competence is one of the current challenges in

M. Refice—IEEE Life Senior Member.

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 574–584, 2018.
https://doi.org/10.1007/978-3-319-99579-3_59

the field [10], leading to concentrate most research effort on modelling all the above mentioned factors influencing entrainment in speech communication (see [11] for a review). Among those, one of the less investigated yet particularly crucial for multimodal, interactive dialogue systems or virtual agents modelling [28] is interaction modality, i.e. whether conversational partners who could see each other during the interaction would entrain their speech more with respect to when eye contact is inhibited, or not. Even though some evidence has been provided that visual information enhances speech alignment [12], in a later study by the same authors this result was not confirmed [13]. Moreover, [14] have shown that interlocutors' visibility in prosodic convergence enhancement does not necessarily play a role as a factor on its own, since its influence is conditioned by interpersonal factors (in their study, partner's likeability). Given such a complex picture, this paper aims to contributing to this research issue, focussing on the role of interaction modality on the prosodic manifestation of speech adaptation. Our starting point is a previous study, where we explored prosodic entrainment between pairs of participants involved in a collaborative game task where eye contact was not prevented [15]. In this preliminary investigation, we found different behaviours in terms of manifestation of entrainment across speaker pairs, ranging from those coordinating their speech along a varying number of dimensions and acoustic-prosodic features, to those showing no entrainment at all. Since speakers could see each other during the interaction, for those speaker pairs who did not entrain we hypothesised the use of nonverbal cues (gaze, nodding, etc.) as the preferred strategy for cueing accommodation.

The present study is a follow-up of [15], and here we compare prosodic features produced by the same speaker pairs recorded under two different interaction conditions: Audio-Visual (henceforth AV) modality, implying that speakers could see each other during the game, and Audio-Only (henceforth AO) modality, where partner's visibility was inhibited. The AV dataset is the one already presented and analysed in our previous investigation [15]; in this study, we recorded the same subjects this time interacting without seeing their partner (AO condition) and compared results obtained in the two modalities. We assume that, in principle, both AV and AO interaction modalities presuppose potential advantages and disadvantages in terms of enhancement of prosodic entrainment. In the AV modality, according to [12] visual information should enhance the vocal manifestation of prosodic accommodation; yet on the other hand, the availability of the visual (non-verbal) communication channel could induce dialogue partners to maximise or complement the use of nonverbal cues for manifesting entrainment, at the expense of the speech/vocal cues. In the AO modality, according to [12] the lack of visual information should work against the adaptation process as maximally cued by speech parameters; yet on the other hand, the availability of the speech/auditory channel only could lead to enhancing the vocal manifestation of entrainment as the mostly preferred strategy. In the latter case, we should expect to find speaker pairs who did not exhibit any manifestation of prosodic entrainment in the AV condition (as resulting in [15]), as showing prosodic adaptation in the AO interaction along some dimensions and prosodic parameters.

2 Methodology

2.1 Elicitation of Dialogues

The paradigm adopted for eliciting spoken data is the one developed in the PAGE project [16], and inspired by the “matching task” described in [17]. It basically consists in an adapted version of the old Chinese Tangram Game, as illustrated in our previous study [15]. In our recording sessions, a participant pair in each game round was given a Tangram figure set according to the players’ role in that game round, i.e. Director or Matcher. The Director was provided with a set of four Tangram figures, one of which was marked by an arrow, and the Matcher was given only one of the figures included in the Director’s set. Participants could not see the partner’s figure(s), and goal of the game in each round was to establish whether the figure given to the Matcher was the one marked by the arrow in the Director’s set or not, by exchanging information about figure features. Players were explicitly instructed to come to that decision on the basis of a common agreement. We opted for such a cooperative paradigm basing on the assumption that speakers would more likely to entrain in a collaborative than in a competitive context [18]. A complete Tangram Game session consists in 22 game rounds, with a different Tangram figure set used in each round, and participants alternating their role as Director or Matcher in each round, so that the distribution of role type was balanced between partners in the whole dialogue. Participants in a pair sat at desk in front of each other, each of them wearing head-mounted professional microphones (AKG C520) connected to a Marantz PMD 661 digital recorder. In the AV interaction modality, a cardboard was inserted between the participants’ desks at a suitable height in order to prevent players to see each other’s Tangram figures, yet still preserving eye contact (as reported in [15]). In the AO interaction modality, a high and thick separator was interposed between the participants’ desks during the whole interaction, so that partner’s visibility was inhibited. In order to prevent a possible post-session persistence of entrainment [29] in an immediately subsequent recording, all AO game sessions were recorded about one month later than the AV sessions. Despite such a rather long elapsed time between the two session types, we gave participants two different yet comparable sets of Tangram figures in the AV and AO sessions (subjects were not informed about that). This was decided in order to avoid that participants during the AO sessions could somehow recall the Tangram figures already used in the AV sessions, and consequently come too quickly to the Tangram figure matching solution in each round, leading to shorter and therefore less comparable duration of verbal interaction in the AO with respect to the AV recordings. All sessions resulted as having the same duration, independently from the interaction modality, i.e. approximately 30 min.

2.2 Participants

A total amount of twelve speakers (six pairs) participated in both AV and AO recording sessions. They were young adult females, aged 21–25, and all MA student classmates, i.e. they were familiar with each other as they had met before participating in the experiment. Subjects came from the same geo-linguistic area, i.e. the Bari district in

Apulia, a southeastern region of Italy. Given the interference of interpersonal, social and linguistic factors over the interaction modality in speech adaptation [14], we tried to control as many of them as possible. In fact, age, gender, familiarity, as well as spoken variety are all factors which can influence verbal accommodation (see for example [3, 19–22], among others). Moreover, only female speakers were involved in our experiment, as it has been shown that in shadowing tasks females tend to align more than males [23]. Subjects were all naïve to the research goal of the experiment, and obtained a course credit as reward for participating.

2.3 Data Annotation and Acoustic-Prosodic Measurements

The same methodology adopted in [15] for annotating the AV speech data was extended to the AO dataset recorded in this study. All dialogues were orthographically transcribed, including start/end of each game round, along with the role of each participant (Director or Matcher) in the round. A round is defined as each game dialogue segment starting from when participants receive a set of Tangram figures (according to their role in that round) until they come to the commonly agreed solution as to the Tangram figure matching/not matching in that set. Speech materials produced by all speakers in both interaction modalities were segmented and annotated, by marking intervals corresponding to the following linguistic levels:

1. Game rounds (start-end, each round numbered sequentially).
2. Inter-Pausal Units (IPUs), where an IPU is defined as each speaker's speech bounded by silence longer than 100 ms.
3. Phonological words.
4. Phonological syllables.

Data segmentation and annotation were carried out manually using Praat [24]. In a post-processing step, correctness and consistency of annotations were checked via specifically designed and developed tools, and errors were manually corrected. Prosodic parameters for measuring prosodic entrainment were the same as in [15], namely: articulation rate (number of syllable/sec, excluding pauses), F0 range (F0max–F0min, Hz), F0 level (F0 median, Hz), and intensity (RMS amplitude, dB). Values were all automatically extracted, and acoustic-prosodic measurements obtained by implementing scripts in Praat. The IPU was taken as the speech unit for the measurement of each prosodic feature.

2.4 Similarity Metrics: Convergence and Synchrony at the Dialogue Level

Also in measuring overall speech similarity across the dialogues, we replicated methodology and procedure adopted in [15], and inspired by previous studies on the same topic ([1, 2, 25] among others). According to this approach, similarity metrics includes measuring two underlying basic processes of speech coordination: convergence and synchrony. This type of metrics has been applied at the overall dialogue level, as in [15]. As mentioned in Sect. 1 above, convergence refers to the process by which dialogue partners' speech become more similar over the course of the interaction until they converge, whereas synchrony is defined as occurring when conversational partners' speech patterns become correlated over time. It is worth noting that, according to the literature, convergence and synchrony represent two possible manifestations of entrainment among others [1], and that in speech coordination they do not necessarily co-occur in the same dialogue [1, 2]. Another shared observation is that both convergence and synchrony can be realised on the opposite direction as complementary manifestation of accommodation [7]: according to CAT, intra-speaker coordination can also imply divergence, i.e. speakers can sound more dissimilar over the course of the interaction, in this way marking their social distance (e.g. [26], also reported in [15]). For the same reason, anti-synchrony can also be considered as a possible manifestation of overall speech coordination [2, 27]. Finally, accommodation theory also predicts that speakers can converge by some speech features and diverge by some others in the same interaction [7].

As in [15], for measuring convergence at the overall dialogue level we assumed the speech behaviour as convergent by identifying cases in which speakers mean values were more similar (or, in case of divergence, more distant) to each other later in the dialogue. Accordingly, each game session was divided into two equal-sized windows, each containing the same number of game rounds: the first window corresponding to the interval 1÷11 of game rounds, the second including the sequence 12÷22. Within each of the two windows, we compared (paired t-tests) mean values of speaker1 vs speaker2 for each prosodic parameter. Mean values found as significantly different in the first window but not significantly different in the second window were considered as evidence for convergent entrainment. Mean values found as not significantly different in the first window but significantly different in the second window were taken as evidence of divergence. Cases other from these two (mean values either significantly different or not different in both first and second windows) were considered as providing no evidence for convergence or divergence.

Pearson's correlation was used for measuring synchrony and its complementary dimension (as in [15]). We correlated speaker1 with speaker2 mean values in each game round, over the whole dialogue session. Positive correlation was assumed as evidence of synchrony, and negative correlation as evidence of anti-synchrony.

In determining the possible influence of interaction modality on the manifestation of prosodic adaptation, we assumed the following as evidence of a stronger effect of one modality over the other:

- when entrainment emerges in both modalities, the co-occurring of both convergence and synchrony (including their complementary manifestations) with respect to only one dimension of similarity;
- a comparative larger set of prosodic parameters involved in the vocal manifestation of convergence and synchrony (including complementary dimensions).

3 Results and Discussion

Statistical results on prosodic convergence in all dialogue pairs are shown in Table 1 (AV modality in Table 1a, AO modality in Table 1b). For synchrony, all correlation outcomes are reported in Table 2 (AV modality in Table 2a, AO modality in Table 2b). Results for AV modality (Tables 1a and 2a) are taken from [15] and recalled here for comparison's sake. A summarising view of the distribution of the similarity dimensions (positive and negative convergence and synchrony), and that of the prosodic parameters involved in prosodic entrainment across dyads and interaction modalities – as derived by statistical results in Tables 1 and 2 – is offered in Table 3.

Table 1. a, b Comparison of speaker1 vs speaker2 mean values in the first vs second windows (halves) of each dialogue (two-tailed t-test, t values only when significant (* = p<.05, ** = p<.01, *** = p<.001). Light grey shaded boxes indicate convergence, dark grey shaded ones divergence. Dialogue pairs are identified by speaker's initial name (e.g. CD = participants C and D). Table 1a = Audio-Visual (AV) interaction modality, Table 1b = Audio-Only (AO) interaction modality. Results in Table 1a are taken from [15] and reported here for comparison's sake.

1a

Dialogue pair	**Convergence/Divergence** **Audio-Visual interaction modality sessions** *speaker1-speaker2 mean values comparison, 1 half vs 2 half of dialogue*							
	Art. rate		F0 range		F0 level		Intensity	
	1st half	2nd half	1st half	2nd half	1st half	2nd half	1st half	2nd half
CD	n.s.	n.s	2.18*	n.s.	n.s.	4.18***	2.29*	2.58*
PZ	n.s.	n.s.	n.s.	n.s.	-10.46***	-6.71***	-3.52**	n.s.
RC	n.s.	-2.69*	n.s.	n.s.	n.s.	n.s.	4.88***	4.89***
DS	3.21**	n.s.	2.14*	2.16*	n.s.	n.s.	n.s.	2.16*
PP	n.s.	n.s.	n.s.	n.s.	-8.27***	-4.94***	4.66***	7.10***
BV	-3.73**	-3.97***	-2.33*	-2.34*	-6.42***	-9.35***	6.63***	8.75***

1b

Dialogue pair	**Convergence/Divergence** **Audio-Only interaction modality sessions** *speaker1-speaker2 mean values comparison, 1 half vs 2 half of dialogue*							
	Art. rate		F0 range		F0 level		Intensity	
	1st half	2nd half	1st half	2nd half	1st half	2nd half	1st half	2nd half
CD	n.s.	3.26 **	n.s.	2.21*	n.s.	4.15***	2.60*	4.81***
PZ	n.s.	n.s.	n.s.	n.s.	6.98 ***	8.29 ***	2.23 *	n.s.
RC	-4.89***	n.s.	n.s.	n.s.	-4.02***	n.s.	n.s.	n.s.
DS	-8.13***	-5.60***	-2.75*	n.s.	-2.60*	n.s.	n.s.	n.s.
PP	n.s.	n.s.	2.66*	3.61 **	12.26***	7.82***	n.s.	n.s.
BV	-5.89***	n.s.	n.s.	-4.38***	-17.55***	-11.38***	n.s.	2.76*

Table 2. **a, b** Pearson's correlation (r values) of speaker1 with speaker2 mean values in the overall dialogue session (* = p<.10, ** = p<.05, *** = p<.005). Table 2a = Audio-Visual (AV) interaction modality, Table 2b = Audio-Only (AO) interaction modality. Results in Table 2a are taken from [15] and reported here for comparison's sake.

2a

Dialogue pair	**Synchrony/anti-Synchrony Audio-Visual interaction modality sessions**			
	Art. rate	F0 range	F0 level	Intensity
CD	.034	.185	-.120	-.295
PZ	.465**	-.204	.177	-.053
RC	-.098	-.078	.401*	.047
DS	.523**	.191	-.381*	-.071
PP	-.097	-.217	.452**	.425**
BV	.053	-.346	.048	.219

2b

Dialogue pair	**Synchrony/anti-Synchrony Audio-Only interaction modality sessions**			
	Art. rate	F0 range	F0 level	Intensity
CD	0.123	-0.153	-0.206	-0.505**
PZ	0.048	-0.398**	0.084	-0.617***
RC	0.223	-0.115	-0.006	-0.286
DS	-0.026	-0.025	-0.007	-0.552***
PP	0.117	-0.537**	0.169	-0.064
BV	-0.050	-0.019	0.064	-0.546***

Results for the two dimensions of similarity are in line with findings attested in the research literature on entrainment (as discussed in Sect. 2.4), namely that (a) convergence and synchrony do not necessarily co-occur in the same dialogue: in fact, CD speaker pair show only convergence/divergence in the AV modality, and RC speaker pair only convergence/divergence in the AO session, whereas PP dyad cues speech coordination through synchrony/anti-synchrony only, also irrespective of the interaction modality; (b) convergence and synchrony can be realised on the opposite direction as complementary manifestation of accommodation, and such opposite directions can co-occur in the same interaction. These two types of adaptation behaviour are attested in our data, too. Also, for the negative dimensions a tendency appears to be mostly connected with AO modality, especially anti-synchrony which systematically emerges when partner's visibility is not available (for a distribution of convergence and synchrony directions across dyads and interaction modalities see Table 3, left panel). Besides this aspect, our results do not support the hypothesis that, when entrainment emerges in both conditions, the co-occurring of two instead of just one similarity dimension can be assumed as evidence of prosodic entrainment enhancement in one of the two conditions. An exception is represented by results obtained from CD speaker pair, as they show only convergence/divergence in AV modality, but divergence *and*

Table 3. **Left Panel**: Distribution of convergence and synchrony directions (positive/negative) across speaker pairs and interaction modalities. Legend: conv = convergence; div = divergence; syn = synchrony; a-syn = anti-synchrony. A slash between two items indicates the co-occurring of the same phenomenon in both directions. **Right Panel**: Distribution of prosodic parameters involved in the manifestation of prosodic entrainment across speaker pairs and interaction modalities. Legend for parameters: a = art. rate, r = F0 range, l = F0 level, i = intensity; **blue** = convergence/divergence, **green** = synchrony/anti-synchrony.

Distribution of convergence and synchrony direction across speaker pairs and interaction modality			**Prosodic parameters cueing entrainment across speaker pairs and interaction modalities**		
dyad	*Audio-Visual modality*	*Audio-Only modality*	*dyad*	*Audio-Visual modality*	*Audio-Only modality*
CD	conv/div	div, a-syn	CD	**r**, **l**	**a**, **r**, **l**, **i**
PZ	conv, syn	conv, a-syn	PZ	**i**, **a**	**i**, **r**, **i**
RC	div, syn	conv	RC	**a**, **l**	**a**, **l**
DS	conv/div, syn/a-sy	div, a-syn	DS	**a**, **i**, **a**, **l**	**r**, **l**, **i**
PP	syn	a-syn	PP	**l**, **i**	**r**
BV	-	conv/div, a-syn	BV	-	**a**, **r**, **i**, **i**

anti-synchrony in AO condition (see Table 3, left panel). A clear case of speech adaptation enhanced by lack of interlocutor's visibility is represented by BV speaker pair, who entrained prosodically only during the AO session.

A different picture is obtained when comparing the total number of prosodic features used by speaker pairs across the two interaction conditions, irrespective of the similarity dimensions involved (see Table 3, right panel). Half of the dialogue pairs (CD, PZ, BV) made use of a larger set of prosodic cues in speech adaptation when interacting in AO than in AV session. An extreme case is represented by BV dyad, where speakers do not show any entrainment in the AV game session, but they do manifest vocal entrainment by means of a large number of prosodic cues along multiple dimensions (convergence/divergence, anti-synchrony) when the speech/audio communication channel is the only one available. Of the remaining three dialogues, two (DS and PP) exhibit a rather opposite trend, as they used a larger set of prosodic entrainment cues in AO than in AV modality; whereas RC dyad does not show any influence of interaction modality on the vocal manifestation of entrainment, since speakers made use of the same number of prosodic parameters in both AV and AO interaction conditions.

The picture emerging from our results seems to reflect the complexity in terms of prevailing strategies adopted by conversational partners for speech entrainment depending on interaction modality, as formulated in Sect. 1. The preferred strategy seems to consist in enhancing the vocal manifestation of adaptation when the speech channel is the only one available, and the use of nonverbal cues as possible alternative or complementary coordination strategy is inhibited. However, this is not the only behaviour registered in our data, as we found two cases of a stronger manifestation of vocal entrainment when partners could see each other, instead. This result appears to support evidence found in previous studies as to the positive role of visual information on speech alignment ([12], but see [13]).

As a side observation, our results also indicate that fundamental frequency (F0 level and range) is the most preferred vocal cue for manifesting entrainment across speaker pairs, irrespective of interaction modality. Instead, vocal intensity appears particularly involved when eye contact is inhibited, pointing to a sort of compensatory effect when partner's visibility is inhibited. Articulation rate is the least preferred cue, with no preference in relation to the availability of communication channel modality (audio-visual or audio-only).

4 Conclusions

Results of our study point to a complex account as to the possible influence of interaction modality (in terms of partner's visibility/not visibility) on the enhancement of the vocal manifestation of coordination. A tendency emerges indicating that speakers entrain their speech more when eye contact is inhibited, i.e. when nonverbal cues are not available as alternative or complementary strategy. However, alternative behaviours have been registered in our data, calling for more research on this issue, which would be particularly relevant for modelling socially-competent interactive dialogue systems.

References

1. Levitan, R., Hirschberg, J.: Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Proceedings of Interspeech 2011, Florence, pp. 28–31 (2011)
2. de Looze, C., Rauzy, S.: Measuring speakers' similarity in speech by means of prosodic cues: methods and potential. In: Proceedings of Interspeech 2011, Florence, pp. 1393–1396 (2011)
3. Pardo, J.S.: On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. **119**(4), 2382–2393 (2006)
4. Delaherce, E., Chetouani, M., Mahdhaoul, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: a survey of evaluation methods across disciplines. IEEE Trans. Affect. Comput. **3**(3), 349–365 (2012)
5. Mukherjee, S., D'Ausilio, A., Nguyen, N., Fadiga, L., Badino, L.: The relationship between F0 synchrony and speech convergence in dyadic interaction. In: Proceedings of Interspeech 2017, Stockholm, pp. 2341–2345 (2017)
6. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behav. Brain Sci. **27**, 105–110 (2004)
7. Giles, H., Coupland, N., Coupland, J.: Accommodation theory: communication, context and consequence. In: Giles, H., Coupland, N., Coupland, J. (eds.) Contexts of Accommodation: Developments in Applied Sociolinguistics, pp. 1–68. CUP, Cambridge (1991)

8. Krauss, R.M., Pardo, J.: Is alignment always the result of automatic priming? Behav. Brain Sci. **27**(2), 203–204 (2004)
9. Lewandowski, N.: Automaticity and consciousness in phonetic convergence. In: Proceedings of the International Workshop “The Listening Talker”, Edinburgh, p. 71 (2012)
10. Vinciarelli, A., Pantic, M., Bourland, H.: Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**, 1743–1759 (2009)
11. Benus, S.: Social aspects of entrainment in spoken interaction. Cogn. Comput. **6**, 802–813 (2014)
12. Dias, J.W., Rosenblum, L.D.: Visual influences on interactive speech alignment. Perception **40**, 1457–1466 (2011)
13. Dias, J.W., Rosenblum, L.D.: Visual enhancement of alignment in noisy speech. In: Proceedings of ICA 2013, Montreal, pp. 1–6 (2013). (Proceedings of Meetings of Acoustics, vol. 19, p. 060139. JASA)
14. Schweitzer, K., Walsh, M., Schweitzer, A.: To see or not to see: interlocutor visibility and likeability influence convergence in intonation. In: Proceedings of Interspeech 2017, Stockholm, pp. 919–923 (2017)
15. Savino, M., Lapertosa, L., Caffò, A., Refice, M.: Measuring prosodic entrainment in italian collaborative game-based dialogues. In: Ronzhin, A., Potapova, R., Németh, G. (eds.) SPECOM 2016. LNCS (LNAI), vol. 9811, pp. 476–483. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43958-7_57
16. PAGE (Prosodic And Gestural Entrainment in conversational interactions across diverse languages) project funded by the Volkswagen Foundation. http://page.home.amu.edu.pl/
17. Wilkes-Gibbs, D., Clark, H.H.: Coordinating beliefs in conversation. J. Mem. Lang. **31**, 183–194 (1992)
18. Manson, J.H., Bryant, G.A., Gervais, M.M., Kline, M.A.: Convergence of speech rate in conversation predicts cooperation. Evol. Hum. Behav. **34**, 419–426 (2013)
19. Kim, M., Horton, W.S., Bradlow, A.R.: Phonetic convergence in spontaneous conversation as a function of interlocutor language distance. J. Lab. Phonol. **2**, 125–156 (2011)
20. Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., Nenkova, A.: Acoustic-prosodic entrainment and social behaviour. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, pp. 11–19 (2012)
21. Pardo, J., Gibbons, R., Suppes, A., Krauss, R.: Phonetic convergence in college roommates. J. Phon. **40**, 190–197 (2012)
22. Babel, M.: Evidence for phonetic and social selectivity in spontaneous phonetic imitation. J. Phon. **40**, 177–189 (2012)
23. Namy, L.L., Nygaard, L.C., Sauerteig, D.: Gender differences in vocal accommodation: the role of perception. J. Lang. Soc. Psychol. **21**(4), 422–432 (2002)
24. Boersma, P.: Praat, a system for doing phonetics by computer. Glot Int. **5**(9/10), 131–151 (2001)
25. Eldlund, J., Heldner, M., Hirschberg, J.: Pause and gap length in face-to-face interaction. In: Proceedings of Interspeech 2009, Brighton, pp. 2779–2782 (2009)
26. Schweitzer, A., Lewandowski, N.: Convergence of articulation rate in spontaneous speech. In: Proceedings of Interspeech 2013, Lyon, pp. 525–529 (2013)

27. Dale, R., Fusaroli, R., Håkonsson, D.D., Healey, P., Mønster, D., McGraw, J., Mitkidikis, P., Tylen, K.: Beyond synchrony: complementarity and asynchrony in joint action. In: Proceedings of the Annual Meeting of the Cognitive Science Society, Austin, TX, vol. 35, pp. 79–80 (2013)
28. Watanabe, T.: Human-entrained embodied interaction and communication technology for advanced media society. In: Proceedings of the 16th IEEE International Conference on Robot & Human Interactive Communication, Jieu, pp. 31–36 (2007)
29. Yoonjeong, L., Gordon Danner, S., Parrell, B., Sungbok, L., Goldstein, L., Byrd, D.: Prosodic convergence during and after a cooperative maze task. In: LabPhon 2015 Conference, Abstract 245, Cornell (2015)