ELSEVIER
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP) and of the International Speech Communication Association (ISCA)

Entrainment profiles: Comparison by gender, role, and feature set

Uwe D. Reichel, Štefan Beňuš, Katalin Mády

PII: S0167-6393(17)30351-5
DOI: 10.1016/j.specom.2018.04.009
Reference: SPECOM 2560

To appear in: *Speech Communication*

Received date: 15 September 2017
Revised date: 16 March 2018
Accepted date: 25 April 2018

Entrainment profiles: Comparison by gender, role, and feature set

Uwe D. Reichel[a], Štefan Beňuš[b,c], Katalin Mády[d]

[a] University of Munich, Germany
[b] Constantine the Philosopher University, Nitra
[c] II SAS Bratislava, Slovakia
[d] Hungarian Academy of Sciences, Budapest, Hungary

Abstract

We examine prosodic entrainment in cooperative game dialogs for new feature sets describing register, pitch accent shape, and rhythmic aspects of utterances. For these as well as for established features we present entrainment profiles to detect within- and across-dialog entrainment by the speakers' gender and role in the game. It turned out, that feature sets undergo entrainment in different quantitative and qualitative ways, which can partly be attributed to their different functions. Furthermore, interactions between speaker gender and role (describer vs. follower) suggest gender-dependent strategies in cooperative solution-oriented interactions: female describers entrain most, male describers least. Our data suggests a slight advantage of the latter strategy on task success.

Keywords: entrainment, prosody, profile, gender, social role, dialog

1. Introduction

In spoken conversations, multiple aspects of interlocutors' utterances and their speaking behavior tend to become more similar to each other. This phenomenon is called entrainment in the computer science literature and is also commonly referred to as alignment, accommodation, audience design, mimicry, priming, or other in psychology, sociology and other disciplines. There are several well established and relatively non-controversial aspects of entrainment. First, entrainment affects not only speech but also other modalities such as gaze, facial expression, mannerisms, or posture [1]. In this paper, we concentrate only on entrainment in the speech modality as entrainment in most studies was observed in spoken interactions (and possibly even without visual contact between interlocutors [2]), which points to speech as playing an important and natural role for entrainment also in other modalities.

Second, entrainment affects both linguistic and para-linguistics domains of speaking. On the linguistic level entrainment affects amongst others the choice of words [3, 4, 5] or syntactic constructions [6, 7, 8]. While the text/transcript discrete data are predominantly used for analyzing the linguistic aspects, the continuous acoustic-prosodic features extracted directly from the speech signal have been commonly used to explore entrainment in the para-linguistic domain (speech rate, intensity, pitch, voice quality [9, 10, 11, 12, 13]). A notable exception is the study analyzing entrainment in terms of linguistically meaningful aspects of intonational contours via discrete ToBI labeling [14].

Third, speech entrainment tends to correlate with positive perception of the interlocutor and/or interactions in which entrainment took place. Entrainment has been shown to increase the success of conversation in terms of low inter-turn latencies and a reduced number of interruptions [12, 3] as well as with objective task success measures [15], and people are generally perceived as more socially attractive and likable, more competent and intimate if they entrain to their interlocutors (reviews in [16] and [17]). More recently, entrainment was also found to play an important role in the perception of social attractiveness and likability [18, 19] This extends also to some aspects of human-machine spoken interactions in which bi-directional entrainment between humans and machines improved the effectiveness and user's experience of the interactions (review in [17]) and several approaches are proposed for endowing synthesizers with speech entrainment capabilities [20, 21, 22].

However, recent research also suggests that the link between speech entrainment and aspects characterizing spoken interaction is more complex. First, as also pointed out by an anonymous reviewer, a causal link between entrainment and task success has not been clearly established and the observed positive correlations may stem from a stronger social relationship reflected by greater collaboration, engagement, and/or entrainment. Moreover, several studies also suggest that both entrainment and disentrainment co-occur integrally in conversations and that positive aspects perceived in the interactions may be linked to their combination [23, 24, 25]. This complexity is further corroborated by studies showing that convergence and synchrony in pitch features have complex and complementary relationships with the speakers impression of their interlocutor's visual attractiveness and likability [26, 19].

In addition, there are some other aspects of entrainment that are still not well understood. The first general issue of contention in cognitive science and psychology is the degree of control a speaker has over entrainment to her interlocutor. Despite differences, two influential approaches to entrainment ([27, 28] and [1]) suggest that entrainment is in general an automatic priming-type mechanism rooted in the perception-production link in which the activation of the linguistic representations or other behavior from the interlocutor increases the likelihood of producing such representations/behavior by the speaker. On the other hand, the Communication Accommodation Theory (CAT) [29] maintains that speakers use entrainment or dis-entrainment in order to attenuate (or accentuate) social differences and thus actively negotiate social distance in spoken interactions. Several studies propose a hybrid approach in which the link between processes of perception and production is not automatic, but can be mediated by pragmatic goals or social factors [30, 31].

A more specific issue, that is directly relevant to the first one, involves the role of gender and power relations of the interlocutors. Entrainment turns out to be stronger in case of mutual positive attitude of the interlocutors, than in case of negative attitude [32], which is in line with the predictions of theoretical models such as the CAT [29]. The CAT also predicts a dependence of entrainment on dominance relations. In case of a misbalanced power of two interlocutors the one with the lower status (or authority, dominance) will entrain more to the one with the higher status [33]. Empirical evidence for this claim has been found amongst others for talkshow data [9], the judicial domain [4], or in task-oriented dialogues [34], where hierarchies turned out to be well reflected in the amount of entrainment. Combining this with the male-dominance hypothesis [35], we may hypothesize that female speakers generally entrain more than males. In addition to this sociological reasoning, greater entrainment of females compared to males might be hypothesized based on the above mentioned link between entrainment and the perception-production loop: females might be capable to entrain more, since they are more sensitive to fine phonetic detail than males [1].

Support has been found for both the male-dominance hypothesis in terms of higher frequencies of interruptions and ego first-person singular pronouns [35], and for the higher phonetic sensitivity of female speakers [36].

However, the picture of gender-related entrainment differences is much less clear than to be expected based on the literature. Some studies explore only mixed-gender dyads [26], others [37, 38] revealed complex patterns of gender-related entrainment in same- and mixed-gender dyads that are furthermore feature- and language-dependent. Similarly, the interplay between gender and the conversation role on entrainment is not clear. [8] for example analyzed data from multi-party picture-describing task in which the degree of syntactic entrainment of the participants in a current picture-description was affected by the speaker's role in the previous description (addressee or side-participant) but not by the addressee's role. However, the gender of the participants is not specified in this study and these conversational roles do not yield straightforwardly to power differences.

Another specific issue involves the the type of features commonly used in entrainment research. Since the linguistic features require transcripts and (shallow) parsing or expensive annotation of the data (e.g. ToBI labeling), studies exploring entrainment based only on the signal focused on coarse acoustic-prosodic (a/p) features. This makes sense also for applied research since the upshot of understanding speech entrainment in human-human spoken interactions is in designing interactive spoken dialogue systems with online entrainment capabilities so that human-machine spoken dialogue systems in the future are more effective and more positively perceived by humans. The coarse a/p features are easily extractable from the signal and can be in turn easily adaptable in speech synthesis for entrainment purposes. However, the speech signal may also contain automatically extractable information about higher-level features that are intermediate between para-linguistic and linguistic and include, for example, features characterizing the shape of intonational contours in relevant speech intervals. Analyzing the relevance of such features for entrainment, and their relationship to the traditional a/p features will fill the current gap in our understanding of speech entrainment.

Goals of the current study. We will address the two specific issues mentioned above by disentangling the gender and communicative role in analyzing how they participate on entrainment. That is, we will not predefine male and female authority, as a special case of 'power', in terms of the male-dominance hypothesis, but assign it to the speaker's role in a cooperative game. Technically, in order to examine entrainment selectively by speaker role and gender we propose an asymmetric turn pairing procedure that yields separate entrainment values for each speaker. We also will address a potential impact of entrainment behavior by role and gender on task success. Furthermore, we will extend the prosodic feature pool to be investigated. All pitch examinations cited above were restricted to rather coarse acoustic measures such as the mean or maximum value of the fundamental frequency (f0) [11, 12], its variance [9] and the distance between raw f0 contours [13]. We will add features derived from a parametric superpositional intonation stylization, that allow for the comparison of more complex pitch patterns in different prosodic domains. These contextualized features furthermore allow for a positional examination of entrainment, that is, whether more entrainment occurs in the beginning or the end of a turn.

Finally, although we introduce some new a/p features and factors (role/gender) in exploring entrainment, we strive to make our results comparable to the existing literature by basing our quantification of entrainment on the notions of synchrony and proximity. In this we follow previous studies [39, 11, 24] that explored the signal-based continuous features. Another line of alignment/

accommodation research bases their analyses on discrete data and are largely dependent on quantifying the ratios or probabilities of exact repetitions of certain text-based linguistic structures or lexical items [40, 41, 42]. We leave the comparison of the signal-based and text-based operationalizations of entrainment for future research.

After the presentation of our data and the extracted prosodic features (sections 2 and 3) we will introduce profiles of several operationalizations of entrainment (section 4). The observations obtained from these profiles will be tested and discussed in sections 5 and 6.

2. Data

2.1. Corpus

The Slovak Games Corpus (SK-games) was used; e.g. [43]. The corpus was recorded with slight modifications following the Object games of the Columbia Games Corpus [44, 45]. Briefly, pairs of subjects were seated in a quiet room opposite each other but without any visual contact and used the mouse to move images on the screens from their initial positions to the target positions. One of the subjects saw the target position on her screen (the Describer) and guided the other player (the Follower) to place the image into that position. The players were awarded points based on a pixel-match between the target position on the Describer's screen and the placement on the Follower's screen. In each session the subjects placed 14 images and they regularly switched roles of the Describer and Follower. This design resulted in natural task-oriented collaborative dialogues. The material comprises 9 sessions of approximately 6 hours of dialogs by 11 speakers (5 female, 6 male; 5 mixed gender, 2 female-female, and 2 male-male dialogs).

2.2. Preprocessing

Alignment. The manually derived text transcription within the semi-automatically determined inter-pausal units (IPUs, threshold of 100ms) was automatically aligned to the signal on the sound and word levels using the SPHINX toolkit adjusted for Slovak [46]. This forced alignment occasionally produced short silent periods within the originally determined IPUs and the entire alignment was manually corrected by the second author.

F0 and energy. F0 was extracted by autocorrelation (PRAAT 5.3.16 [47], sample rate 100 Hz). Voiceless utterance parts and f0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky-Golay filtering [48] using third order polynomials in 5 sample windows and transformed to semitones relative to a base value. This base value was set to the f0 median below the 5th percentile of an utterance and serves to normalize f0 with respect to its overall level.

Energy in terms of root mean squared deviation was calculated with the same sample rate as f0 in Hamming windows of 50 ms length.

Prosodic structure. The dialogs were segmented into turns and interpausal units. The latter we employed as a coarse approximation of prosodic phrases given that speech pauses are among the most salient phrase boundary cues [49]. By this simplifying assumption we use the terms “interpausal unit” and “prosodic phrase” interchangeably in the following. Automatic syllable nucleus assignment follows the procedure introduced in [50] to a large extent. An analysis window w_a and a reference window w_r with the same time midpoint were moved along the band-pass filtered signal in 50ms steps. Filtering was carried out by a 5th order Butterworth filter with the cutoff frequencies 200 and 4000Hz. For a syllable nucleus assignment the energy in the relevant frequency range r is required to be higher in w_a than in w_r by a factor v, and additionally had to surpass a threshold x relative to the maximum energy RMS_{max} of the utterance, i.e. $\text{RMS}(w_a) > \text{RMS}(w_r) \cdot v \wedge \text{RMS}(w_a) > \text{RMS}_{max} \cdot x$. Based on the tuning results in [51] the parameters were set to the following values: $w_a = 0.05s$, $w_r = 0.11s$, $v = 1.1$, $x = 0.1$.

Pitch accents were detected automatically by means of a bootstrapped nearest centroid classifier as described in detail in [51]. Based on pitch accent-related features derived for each word-initial syllable introduced in section 3.3 as well as by vowel length z-scores, two centroids for accented and non-accented syllable were bootstrapped based on two simplifying assumptions: (1) all words longer than a threshold t_a are likely to be content words that contain a high amount of information and are thus taken as class 1 (accented) representatives, and (2) all words shorter than a threshold t_{na} are likely to be function words with a low amount of lexical information and are thus taken as class 0 (no accent) representatives. t_a and t_{na} were set to 0.6s and 0.15s, respectively. For words fulfilling criterion (1) the first syllable (Slovak has fixed word-initial stress) was added to the class 1 cluster. For words fulfilling criterion (2) all syllables were added to the class 0 cluster.

From this initial clustering feature weights were calculated from the mean cluster silhouette derived separately for each feature. The weights thus reflect how well a feature separates the seed clusters.

After this cluster initialization the remaining word-initial syllables are assigned to the classes 0 or 1 in a single pass the following way: for each feature vector i its weighted Euclidean distances $d_{i,0}$ and $d_{i,1}$ to the class 0 and class 1 centroids are calculated, and the quotient of both distances $q_i = \frac{d_{i,0}}{d_{i,1}}$ is recorded. All items with a q_i above a defined percentile p are assigned to class 1, and the items below to class 0. By choosing a percentile threshold well above 50 the skewed distribution of class 0 and class 1 cases for both boundaries and accents can be tackled, i.e. more items receive class 0 than class 1. The percentile threshold p was set as in [51] to 82.

In [51] this procedure yielded F1 scores up to 0.63 on spontaneous speech data, which clearly indicates moderate precision and recall values for pitch accent detection. However, the choice of the feature sets for accent detection ensures that syllables with salient pitch and energy movements are identified for further analyses.

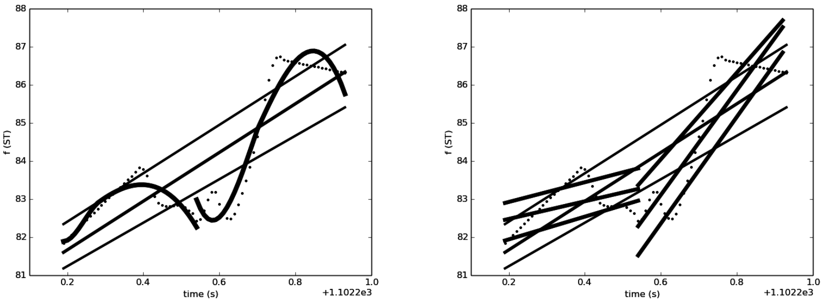

Figure 1: Superpositional f0 stylization within the CoPaSul framework. On the interpausal unit (IPU) level a base, mid- and topline (solid) are fitted to the f0 contour (dotted) for register stylization. Level is represented by the midline, range by a regression line fitted to the pointwise distance between base and topline. On the local pitch event level comprising accents and boundary tones the f0 shape is represented by a third-order polynomial (left). It's Gestalt properties, i.e. its register deviation from the phrase-level register is quantified by generating a local register representation the same way as for the phrase level (right) and by calculating the root mean squared deviations between the midlines and the range regression lines.

3. Prosodic features

Next to general f0 and energy features we derived register and local pitch event related features from the contour-based, parametric, and superpositional CoPaSul stylization framework [52] representing f0 as a superposition of a global register and a local pitch accent component. This stylization is presented in Figure 1. Furthermore rhythmic features were extracted as described below. All features introduced here as well as the automatic extraction of prosodic structure can be carried out by means of the open source CoPaSul prosody analyses software [53, 54].

All features are listed in Table 1 together with the feature set name they belong to and a short description. A more detailed description is given in the subsequent sections.

Feature set	Feature	Description
gnl_en	max	energy maximum in turn
gnl_en	med	energy median in turn
gnl_en	sd	energy standard deviation in turn
gnl_f0	max	f0 maximum in turn
gnl_f0	med	f0 median in turn
gnl_f0	sd	f0 standard deviation in turn
phrase	rng.c0.F/L	f0 range intercept of first/last phrase
phrase	rng.c1.F/L	f0 range slope of first/last phrase
phrase	lev.c0.F/L	f0 level intercept of first/last phrase
phrase	lev.c1.F/L	f0 level slope of first/last phrase
acc	c0-3.F/L	polynomial coef of the first/last pitch accent
acc	rng.c0.F/L	f0 range intercept of first/last pitch accent
acc	rng.c1.F/L	f0 range slope of first/last pitch accent
acc	lev.c0.F/L	f0 level intercept of first/last pitch accent
acc	lev.c1.F/L	f0 level slope of first/last pitch accent
acc	gst.lev.F/L	level deviation of first/last pitch accent
acc	gst.rng.F/L	range deviation of first/last pitch accent
rhy_en	syl.rate	mean syllable rate
rhy_en	syl.prop	syllable influence on energy contour
rhy_f0	syl.prop	syllable influence on f0 contour

Table 1: Description of prosodic features grouped by feature sets. "first/last" refers to the position of the prosodic event within a turn.

3.1. General f0 and energy features

For the feature sets *gnl_f0* and *gnl_en* within each turn we calculated the median, the maximum, and the standard deviation of the f0 and the energy contour, respectively.

3.2. Prosodic phrase characteristics

The *phrase* feature set describes f0 register characteristics. According to [55] f0 register in the prosodic phrase domain can be represented in terms of the f0 range between high and low pitch targets, and the f0 mean level within this span. To capture both register aspects, level and range, within each prosodic phrase we fitted a base-, a mid, and a topline by means of linear regressions as shown in Figure 1. This line fitting procedure works as follows: A window of length 50 ms is shifted along the f0 contour with a step size of 10 ms. Within each window the f0 median is calculated (1) of the values below the 10th percentile for the baseline, (2) of the values above the 90th percentile for the topline, and (3) of all values for the midline. This gives three sequences of medians, one each for the base-, the mid-, and the topline, respectively. These lines are subsequently derived by linear regressions, time has been normalized to the range from 0 to 1. As described in further detail in [56] this stylization is less affected by local events as pitch accents and boundary tones and does not need to rely on error-prone detection of local maxima and minima. Based on this stylization the midline is taken as a representation of pitch level. For pitch range we fitted a further regression line through the pointwise distances between the topline and the baseline. A negative slope thus indicates convergence of top- and baseline, whereas a positive slope indicates divergence.

From this register level and range representation we extracted for the first and for the (occasionally identical) last prosodic phrase in a turn the following features: intercept and slope of the midline, and intercept and slope of the range regression line. That gives eight features subsumed to the *phrase* feature set.

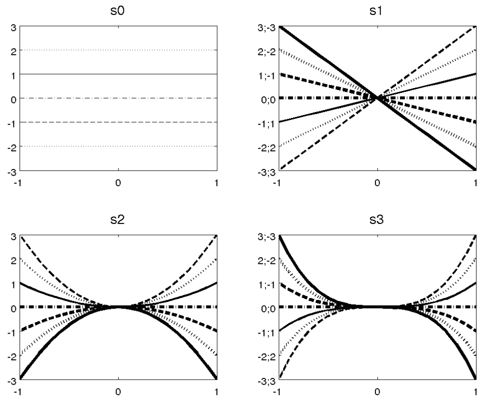

Figure 2: Influence of each coefficient of the third order polynomial $\sum_{i=0}^{3} s_i \cdot t^i$ on the contour shape. All other coefficients set to 0. For compactness purpose on the y-axis both function and coefficient values are shown if they differ.

3.3. Pitch accent characteristics

After subtracting the midline derived on the phrase level as described in section 3.2 we fitted third-order polynomials to the residual f0 contour around the syllable nuclei associated with the first and the last local pitch event (accent or boundary tone) in a turn. The stylization window of length 300 ms was placed symmetrically on the syllable nucleus, and time t was normalized to the range from -1 to 1. This window length of approximately 1.5 syllables was chosen to capture the f0 contour on the accented syllable in some local context.

As can be seen in Figure 2 the coefficients represent different aspects of local f0 shapes. Given the polynomial $\sum_{i=0}^{3} s_i \cdot t^i$, s_0 is related to the local f0 level relative to the register midline. s_1 and s_3 are related to the local f0 trend (rising or falling) and to peak alignment. s_2 determines the peak curvature (convex or concave) and its acuity.

Next to the polynomial coefficients we measured local register values by reapplying the stylization introduced in section 3.2 within the analysis window around the pitch accent.

Finally, pitch accent Gestalt was measured in terms of local register deviation from the corresponding stretch of global register. This was simply done by calculating the RMSD between the pitch accent midline and the corresponding part of the phrase midline. For the accent and phrase range regression lines we did the same.

From these stylizations the feature set *acc* emerges for the first and for the last local pitch event in a turn. It contains (1) the polynomial coefficients describing the local f0 shape, (2) the intercept and slope coefficients for the mid- and the range regression line describing the local register, and (3) the local level and range deviation from the underlying phrase in terms of the RMSD between the accent- and phrase-level regression lines.

3.4. Rhythm features

In our approach, rhythm within a turn is represented in terms of syllable rate (number of detected syllable nuclei per second) and the influence of the syllable level of the prosodic hierarchy on the energy and f0 contours. To quantify the syllabic influence on any of these contours we performed a discrete cosine transform (DCT) on this contour as in [57]. We then calculated the syllable influence w as the relative weight of the coefficients around the syllable rate r ($+/-1$ Hz to account for syllable rate fluctuations) within all coefficients below 10 Hz as follows:

$$w = \frac{\sum_{c:r-1\leq f(c)\leq r+1\text{Hz}} |c|}{\sum_{c:f(c)\leq 10\text{Hz}} |c|}$$

The higher w the higher thus the influence of the syllable rate on the contour. This procedure which is shown in Figure 3 was first used to quantify the impact of hand stroke rate on the energy contour in counting out rhymes [58]. The upper cutoff of 10 Hz goes back to the reasoning that contour modulations above 10 Hz do not occur due to macroprosodic events as accents or syllables, but amongst others due to microprosodic effects.

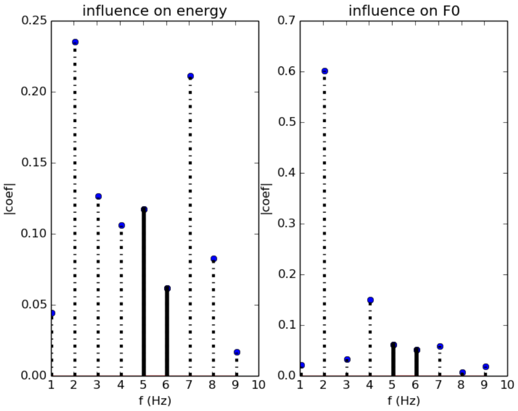

Figure 3: Rhythm features: Quantifying the influence of syllable rate on the energy and f0 contour. For this purpose a discrete cosine transform (DCT) is applied to the contour. The absolute amplitudes of the coefficients around the syllable rate are summed and divided by the summed absolute amplitudes of all coefficients below 10 Hz. This gives the proportional influence of the syllable on the contour. In the shown case the syllable rate of 4.5 Hz has a relatively high impact on the energy contour but not on the f0 contour. For both contours a high influence in the 2 Hz region related to pitch accents can be observed.

4. Entrainment profiles

For all feature sets described in the previous section we generated entrainment profiles that document in how far speakers entrain with respect to these features and depending on the speaker's gender and role in the dialog.[1]

Entrainment generally is expressed in low feature distances relative to a reference. We address two types of feature distance, one related to proximity, the other to synchrony. Additionally, we examine entrainment on a local and a global level based on an asymmetric pairing of turns to tease apart the impact by speaker genders and roles. We describe these operationalizations of entrainment in two subsections 4.1 and 4.2 below and then proceed to describing the profiles themselves as a means to visualize the data and generate hypotheses on entrainment for further statistical testing.

4.1. Proximity- vs. synchrony-related distance

As pointed out in [39, 11, 24] accommodation can be expressed, among others, in terms of proximity (or similarity), convergence and synchrony. Convergence and proximity are linked in a way that the former describes an increase of the latter and thus a decrease in distance over time, which is visualized in Figure 4. Convergence- and proximity-related distance is trivially represented by the absolute distance of the feature value pair, the lower the distance, the higher the proximity. In the following we restrict the analysis to proximity, thus we are measuring pointwise distances of single turn pairs without their time course. Synchrony means that feature values move in parallel. [24] proposes to calculate correlations over a sequence of turn pairs. Here as for proximity we choose a more straight-forward approach operating on a single turn-pair only. We simply subtract the respective speakers' mean values from the feature values before calculating the absolute distance. Synchrony-related distance is thus low, if the speakers realize a feature either both above or below their respective means. By that we derive for each feature and each turn pair one proximity-

[1] The usage of such profiles was inspired by the speaker profile study of [59].

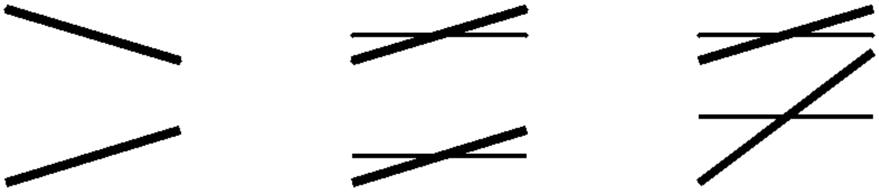

Figure 4: Convergence (left) vs. synchrony (mid) vs. convergence+synchrony (right) of some feature. Convergence describes an increase in proximity, which is given by the absolute distance of the feature values. For synchrony the feature values are centered on the speaker-dependent mean value before calculating their absolute distance.

and one synchrony-related distance value. It is likely that some of the examined features preferably undergo one entrainment type only. Pitch accent shape coefficients for example cannot simply be shifted in parallel by the interlocutors due to non-linearities in f0 contour continua as found e.g. by [60], so that for these parameters entrainment is expected rather to happen not in terms of synchrony but of proximity.

4.2. Directed local vs. global entrainment

Turn pairing was carried out on two levels to account for local and for global entrainment. Local entrainment refers to a greater similarity in adjacent compared to non-adjacent turns in the same dialog. Global entrainment refers to an overall greater similarity within a speaker pair (or dialog) than across dyads (dialogs) [11]. For **local entrainment** we compared the feature distances between adjacent and non-adjacent turns within the same dialog and task. For the *adjacent* sample we paired each turn with the one directly preceding it in the dialog. For the *non-adjacent* sample for each turn a non-adjacent turn was drawn randomly (if available) from the preceding part of the dialog within the same task among those turns that fulfill the constraint of a minimum inter-onset interval of 15 seconds. For **global entrainment** feature distances were compared between turn pairs in the same dialog and the same number of turn pairs across dialogs. For the *same dialog* sample we paired each turn with a randomly drawn turn from the preceding part of the same dialog and task. For the *different dialog* sample we randomly paired turns of unrelated speaker pairs, i.e. speaker pairs not engaged in any common game conversation.

Our sample generation approach differs from previous approaches as in [11] in several respects: first, we apply a directed pairing of turns to the left dialog context only. This enables us to compare entrainment behavior asymmetrically across speaker genders and roles since for the statistical analyses described below we relate the obtained distance values not to both speakers but to the second one only, i.e. for each turn pair we examine how similar the second speaker gets to the first, and not vice versa.

Second, for global entrainment we are not comparing mean feature values calculated for each speaker as [11], but analogously as for local entrainment we work on the raw turn pair data. This ensures comparable sample sizes in local and global entrainment examination making the results less dependent on the number of speakers in the corpus, especially if this number is low. And again this approach allows for asymmetric examination of gender and role influence also on global entrainment. As opposed to mean value comparison that yields one distance value for both speakers in a dialog, directed turn comparison assigns a distinct value to each interlocutor.

Third, our approach differs with respect to IPU pairing. *Adjacency* refers to the turn level and not to the compared events themselves. For each turn pair we compared separately their initial and their final phrase and accent characteristics, which implies that also for adjacent turns the compared events generally are not adjacent. This approach is motivated first by our goal to compare entrainment effects in dependence of the position within a turn. Furthermore, it serves to reduce value range differences across different positions within an utterance. These differences are amongst others caused by declination and locally restricted event functions such as pitch accents vs. boundary tones. By our positional restriction we obtain distance values with a less obscured link to entrainment.

4.3. Profiles

For each feature set we generated for proximity and for synchrony each an entrainment profile in the following way: the features of the respective set are plotted on the y axis, and their mean distance values on the x axis. Mean distance values are separately calculated for adjacent turns (a), non-adjacent turns in the same dialog (na) and turn pairs across dialogs (u). The latter two define the references for local and global entrainment, respectively. The adjacent turn distances are further split by speaker role and gender, to visualize the impact of speaker type on entrainment. Distance and type specification always refers to the responding speaker, i.e. the speaker uttering the later turn in the turn pair. For visual inspection a local entrainment tendency is indicated by a-lines left of the na-reference line. Global entrainment is reflected in a-lines and the na-line left of the u-reference line. For both the local and global domain the opposite order indicates a disentrainment tendency. Figure 5 shows mean proximity distance values for the feature sets *phrase* and *acc*. By visual inspection female speakers (solid lines), especially the followers (thick solid) show smaller distances in adjacent turns than in non-adjacent or unrelated ones for most features indicating entrainment. Male speaker profiles (dashed lines), especially the describer ones (thick dashed), in contrast are right of the reference lines indicating higher distance values and thus a disentrainment tendency.

4.4. Descriptive observation

By visual inspection of such entrainment profiles in Figures 5 and 6 the following observations can be made:

- There is a role-gender interaction; female describers (thick solid) generally entrain most, male describers (thick dashed) entrain least.

- The zigzag lines for entraining speakers in Figure 5 for set *acc* indicate that more entrainment takes place in turn-final than in turn-initial position (**_L* and **_F* features, respectively).

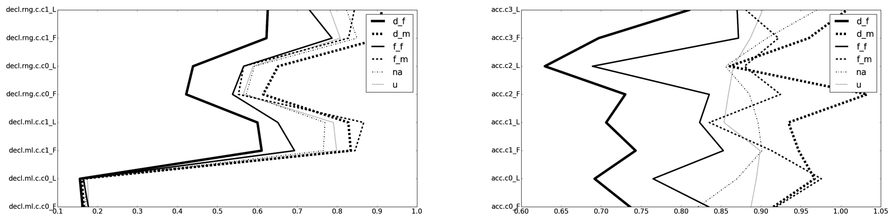

Figure 5: Entrainment profiles for features from the sets *phrase* (left) and *acc* (right). The y-axis gives the features described in table 1, the x-axis gives their mean proximity distances. For each speaker type *role_gender* defined by role (describer *d* or follower *f*) and gender (female *f* or male *m*) a profile graph relates each feature to its mean proximity distance in adjacent turns. Describers *d_** profiles are given in thick lines, follower *f_** profiles in thin lines. Solid indicates female **_f*, dashed male **_m*. Two reference profiles are given for non-adjacent turns in the same dialog (*na*, dash-dotted) and for unrelated turns in different dialogs (*u*, dotted).

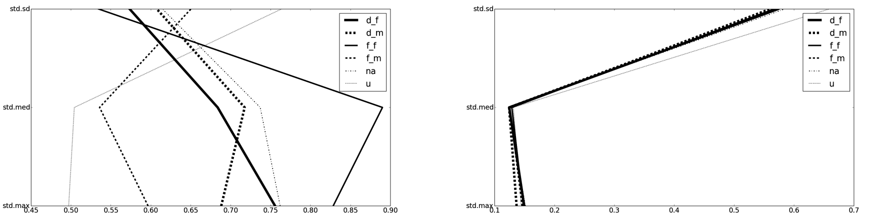

Figure 6: Entrainment profiles for the feature set *gnl_f0* for proximity (left) and synchrony (right). For this feature set these two entrainment measures behave very differently. For details please see the caption of Figure 5.

- The feature sets undergo entrainment to a different degree. While *gnl_en* features do not entrain at all, *acc* features show entrainment for certain speaker types.
- The profile pair in Figure 6 suggests that there is a bias of some feature sets towards proximity or synchrony.

These descriptive observations obtained from the visualization of entrainment profiles serve as hypotheses for further statistic examinations that are described in the following section.

5. Harvesting and condensation of entrainment data

To cope with the complexity of our data – 37 acoustic features times 2 entrainment domains times 2 distance measures times each 2 roles and genders – we employed a two-step approach consisting of data harvesting and condensation. By harvesting we collect the entrainment behavior of all speaker types for all prosodic features. Subsequent condensation serves to structure the data in terms of probabilistic relations between entrainment on one hand, and feature sets, speaker types, and segment positions on the other hand.

5.1. Harvesting

5.1.1. Methods

We used linear mixed-effect models for each prosodic feature based on the lmer() function in the *lme4* package in the statistics software R [61]. The dependent variable *dist* refers to proximity and synchrony each in global and local entrainment turn pairs. Thus for each prosodic feature, 4 distance values are tested. The fixed effects are *pairing, role*, and *gender.* For local entrainment *pairing* stands for *adjacent vs. non-adjacent* turn. For global entrainment it stands for *same vs. different* dialog. *role* and *gender* refer to the replying speaker and define his/her role in the play (*describer* or *follower*) and the gender (*female* vs. *male*). The identities of the initiating and the replying speaker are considered to be random factors for which a random intercept model was calculated. Significant interactions ($p < 0.05$) of the fixed effects calculated by the Anova() function of the *car* package in R [62] were subsequently examined by re-applying the tests on corresponding subsets. To account for the large number of tests, p-values were corrected for false discovery rate [63].

5.1.2. Results

From these tests we derived two tables 2 and 3 for global and local entrainment, respectively.

In Tables 2 and 3 the columns *prox* and *sync* contain all speaker types for which the linear mixed-effect models introduced in the previous section revealed entrainment for a certain feature and distance measure ($\alpha = 0.05$, p-values corrected for false discovery rate). Speaker types are composed of the speaker's role (describer *d* vs. follower *f*), and gender (female *f* vs. male *m*). For local entrainment this means, that the distance of a feature is significantly smaller in neighboring turns opposed to non-neighboring turns. For global entrainment it indicates, that the distance is significantly smaller within a dialog than across dialogs. The *–prox* and *–sync* columns show all disentraining speaker types for a feature and a distance measure, that is, for adjacent or within-dialog turn pairs the distance turned out to be significantly higher than for non-adjacent/cross-dialog turn pairs.

5.2. Condensation

5.2.1. Method

From the tables obtained by harvesting we infer conditional entrainment probabilities separately for proximity and synchrony for feature sets, position within a turn, and speaker type as exemplified for the feature set *gnl_f0* and proximity. In Table 3 in one out of three cases (row 6 out of 4–6) column *prox* reports entrainment evidence, which is defined by the observation that at least one of the speaker types (x_x and d_f in row 6) shows entrainment. Thus the conditional proximity entrainment probability for feature set *gnl_f0* amounts

	Features		Entrainment		Disentrainment	
	set	name	prox	sync	–prox	–sync
1	gnl_en	max	–	–	x_x,x_m,d_m,f_m	x_x,x_m,d_m,f_m
2	gnl_en	med	–	–	–	–
3	gnl_en	sd	–	x_f	x_m,d_m,f_m	x_m,d_m,f_m
4	gnl_f0	max	–	f_f	x_x,x_f,x_m,d_x, d_f,d_m,f_x,f_f,f_m	d_f
5	gnl_f0	med	f_m	x_x	x_x,x_f,x_m,d_f, d_m,f_f	–
6	gnl_f0	sd	x_x,x_f,x_m,d_f, d_m,f_f,f_m	x_x,f_f	–	–
7	phrase	lev.c0.F	–	x_x,d_m f_x	x_x,x_f,x_m,d_x,f_x	–
8	phrase	lev.c0.L	f_m	x_x	x_x,x_f,x_m,d_x, d_f,d_m,f_x,f_f	–
9	phrase	lev.c1.F	x_x	x_x	–	–
10	phrase	lev.c1.L	x_x	–	–	–
11	phrase	rng.c0.F	x_x,x_f,d_f	d_f	–	d_m
12	phrase	rng.c0.L	x_x,x_f,d_f,f_f	d_f,f_f	–	d_m
13	phrase	rng.c1.F	d_f	d_f	x_m,d_m,f_m	x_m,d_m,f_m
14	phrase	rng.c1.L	d_f	d_f	x_m,d_m	d_m
15	acc	c0.F	x_x	–	–	–
16	acc	c0.L	x_x	x_x,x_f	–	–
17	acc	c1.F	–	–	x_m	x_m
18	acc	c1.L	–	–	–	–
19	acc	c2.F	d_f	d_f	x_m	x_m,d_m
20	acc	c2.L	x_x,d_f	x_x,d_f	–	–
21	acc	c3.F	d_f	d_f	d_m,f_m	d_m,f_m
22	acc	c3.L	–	–	–	–
23	acc	lev.c0.F	f_m	x_x	x_x,x_f,x_m,d_f,d_m, f_f	–
24	acc	lev.c0.L	f_m	x_x	x_x,x_f,x_m,d_m,f_f	–
25	acc	lev.c1.F	–	–	x_m,d_m	x_m,d_m
26	acc	lev.c1.L	x_x	x_x	–	–
27	acc	rng.c0.F	x_x,x_f,d_f	–	x_m,f_m	x_m
28	acc	rng.c0.L	x_x,d_f	–	–	–
29	acc	rng.c1.F	d_f	d_f	x_m,d_m	x_m,d_m
30	acc	rng.c1.L	–	–	x_m,d_m,f_m	x_m,d_m,f_m
31	acc	gst.lev.rms.F	x_f	d_f	x_m	x_m,d_m
32	acc	gst.lev.rms.L	x_x,x_f,d_x	–	–	f_m
33	acc	gst.rng.rms.F	x_f	–	–	f_m
34	acc	gst.rng.rms.L	x_x,x_f	–	–	–
35	rhy_en	syl.prop	–	–	–	–
36	rhy_en	syl.rate	–	–	–	f_f
37	rhy_f0	syl.prop	x_f	–	–	–

Table 2: Global entrainment and disentrainment by feature and speaker type for proximity *prox* and synchrony *sync*. Speaker type is encoded as *role_gender*; role: describer *d* vs. follower *f*; gender: female *f* vs. male *m*; *x* denotes *not specified*. To give an example how to read this table: line 14 refers to the feature *rng.c1.L* of the *phrase* set, i.e. the range slope of the turn-final phrase. For this feature female describers *d_f* entrain with respect to both proximity and synchrony. Proximity disentrainment is observed for male speakers *x_m* which turned out to be significant due to the disentraining behavior of male describers *d_m*.

	Features		Entrainment		Disentrainment	
	set	name	prox	sync	–prox	–sync
1	gnl_en	max	–	–	x_x,d_f,d_m,f_f	x_x,d_f,d_m,f_f
2	gnl_en	med	–	–	x_x	x_x
3	gnl_en	sd	–	–	x_x,d_x,d_f,d_m, f_x,f_f,f_m	x_x,d_x,d_f,d_m, f_x,f_f,f_m
4	gnl_f0	max	–	d_x,d_f	–	–
5	gnl_f0	med	–	d_x	–	–
6	gnl_f0	sd	x_x,d_f	d_f,f_m	–	–
7	phrase	lev.c0.F	f_f	–	–	–
8	phrase	lev.c0.L	x_f	–	–	–
9	phrase	lev.c1.F	d_x	d_x	–	–
10	phrase	lev.c1.L	d_x	d_x	–	–
11	phrase	rng.c0.F	–	–	–	–
12	phrase	rng.c0.L	–	d_f	–	–
13	phrase	rng.c1.F	–	x_x,x_m,f_m	–	–
14	phrase	rng.c1.L	–	–	–	–
15	acc	c0.F	–	–	–	–
16	acc	c0.L	f_x	–	–	–
17	acc	c1.F	–	–	–	–
18	acc	c1.L	–	–	–	–
19	acc	c2.F	–	–	–	–
20	acc	c2.L	x_x,f_f	x_x,f_f	–	–
21	acc	c3.F	–	–	–	–
22	acc	c3.L	x_x	x_x	–	–
23	acc	lev.c0.F	–	–	–	–
24	acc	lev.c0.L	–	–	–	–
25	acc	lev.c1.F	–	–	–	–
26	acc	lev.c1.L	–	–	–	–
27	acc	rng.c0.F	–	–	–	–
28	acc	rng.c0.L	–	–	–	–
29	acc	rng.c1.F	–	–	–	–
30	acc	rng.c1.L	–	–	–	–
31	acc	gst.lev.rms.F	–	–	–	–
32	acc	gst.lev.rms.L	–	–	–	–
33	acc	gst.rng.rms.F	–	–	x_x	–
34	acc	gst.rng.rms.L	–	–	–	–
35	rhy_en	syl.prop	x_x,d_f,d_m	–	–	–
36	rhy_en	syl.rate	x_x,d_f,f_m	x_x,d_m	–	–
37	rhy_f0	syl.prop	–	–	–	–

Table 3: Local entrainment and disentrainment by feature and speaker type for proximity *prox* and synchrony *sync*. Speaker type is encoded as *role_gender*; role: describer *d* vs. follower *f*; gender: female *f* vs. male *m*; *x* denotes *not specified*. To give an example how to read this table: line 16 refers to the feature *c0.L* of the *acc* feature set, i.e. the coefficient c_0 of the polynomial stylization of the turn final local pitch event. For this feature all followers *f_x* entrain with respect to proximity.

to $\frac{1}{3}$. For synchrony entrainment occurs for all features, thus conditional synchrony entrainment is 1. Analogously, given no disentrainment evidence, the disentrainment probability is 0 both for proximity and synchrony.

5.2.2. Results

Tables 2 and 3 show which speaker types entrain or disentrain in terms of proximity or synchrony for each feature. The features are further categorized into feature sets. In Table 2 the global entrainment data is collected, in Table 3 the local one. Position of the compared segments within the turns is indicated in the column *feat* by the final capital letters F and L (for first and last segment, respectively). This categorization only applies to the feature sets *phrase* and *acc*. The conditional probabilities for feature sets, position in the turn, and speaker types which were derived from the tables as described in section 5.1.2 are visualized by stacked barplots in Figures 7, 8, and 9.

From the barplots we infer the following observations that will be discussed in section 6:

1. The feature sets *gnl_f0*, *phrase* and *acc* show strong entrainment tendencies, so that especially the newly introduced features of *phrase* and *acc* are worth to be looked at more closely. In contrast, feature set *gnl_en* is very much biased towards disentrainment.
2. The new feature sets undergo local and global entrainment to different proportions. While *phrase* and *acc* undergo more global entrainment, the opposite is to be observed for *rhy_en*.
3. Some feature sets such as *acc* tend to show proximity whereas other feature sets as *gnl_f0* tend to show synchronization.
4. Entrainment takes place in turn-final position more than in turn-initial position.
5. Entrainment is highly speaker-type dependent, more precisely there is an interaction between role and gender. Female describers entrain most, male describers entrain least, female and male followers entrain to approximately the same extent.

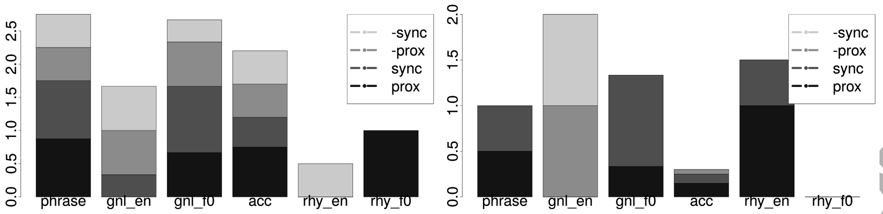

Figure 7: Conditional global (left) and local (right) entrainment and disentrainment probabilities for each feature set derived from Tables 2 and 3. Disentrainment for proximity and synchrony is denoted by *–prox* and *–sync*, respectively. Each partition in the stacks denotes a probability with values between 0 and 1.

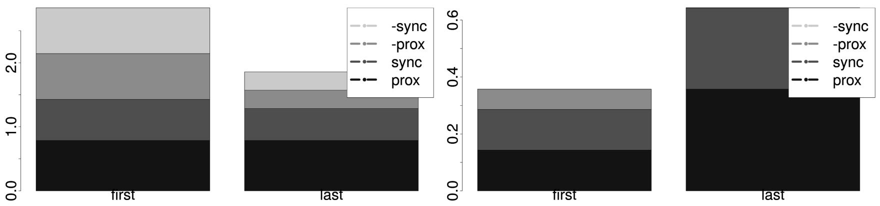

Figure 8: Conditional global (left) and local (right) entrainment and disentrainment probabilities for the first and last position within turns derived from Tables 2 and 3. Disentrainment for proximity and synchrony is denoted by *–prox* and *–sync*, respectively. Each partition in the stacks denotes a probability with values between 0 and 1.

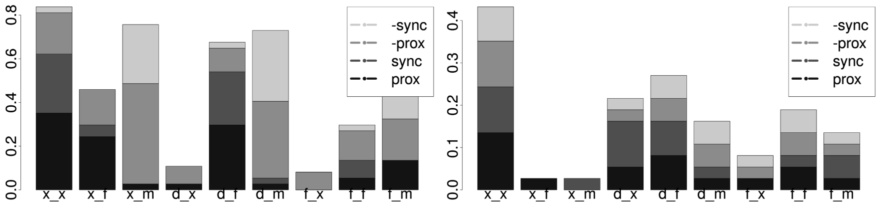

Figure 9: Conditional global (left) and local (right) entrainment and disentrainment probabilities derived from Tables 2 and 3 for each speaker type defined by *role_gender*; role: describer *d* vs. follower *f*; gender: female *f* vs. male *m*; *x* denotes *not specified*. Disentrainment for proximity and synchrony is denoted by *–prox* and *–sync*, respectively. Each partition in the stacks denotes a probability with values between 0 and 1.

5.2.3. Task success

Next to the entrainment plots we recorded several task success measures in Figure 10 for all gender/role combinations in order to examine whether the speaker-type related entrainment behavior has an impact on task success. *Score* measures the distance between the reference and the game outcome of the target object location as described in section 2. *Duration* gives the time it took to solve the task, *efficiency* is *score* divided by *duration*, and *smooth* stands for the proportion of smooth turn transitions in the entire dialog. A transition was defined to be smooth, if it falls in the interval between -0.5 and 0.5 seconds. Overlap values below $-0.5s$ and delays above $0.5s$ indicate interruptions and vacillations, respectively. These values were selected for two reasons. First, turns with minor overlaps or delays within one or two syllables are commonly perceived as 'smooth' in high-involvement interactions ([64]). Moreover, we also examined latencies in an almost identical corpus of collaborative games in English ([45]) with hand-annotations of turn-types such as smooth switch, overlap, interruption, or pause interruption. We found that interruptions were more likely than plain overlaps for overlaps greater than 350ms, and that pause interruptions (signaling non-smooth hesitations from the current speaker) were also more likely than smooth switches with more than 500ms latency.[2]

We tested differences for each of the four success measures by linear mixed-effect models with task success as the dependent variable, the describer and follower gender as the independent variable. The Ids of both speakers were taken as random effects, for which a random intercept model was calculated.

Only for *score* we found a significant difference for which only the describer's gender is responsible ($t = 1.696$, $p = 0.0220$; follower's gender: $t = 1.016$, $p = 0.1362$; interaction: $t = 0.012$, $p = 0.9903$). Pairs with male describers tend to achieve higher scores. For none of the other success variables any significant relationship has been found, neither to the describer's or follower's gender, nor

[2] We thank A. Gravano for providing us with mean latencies for turn types in Columbia Games Corpus.

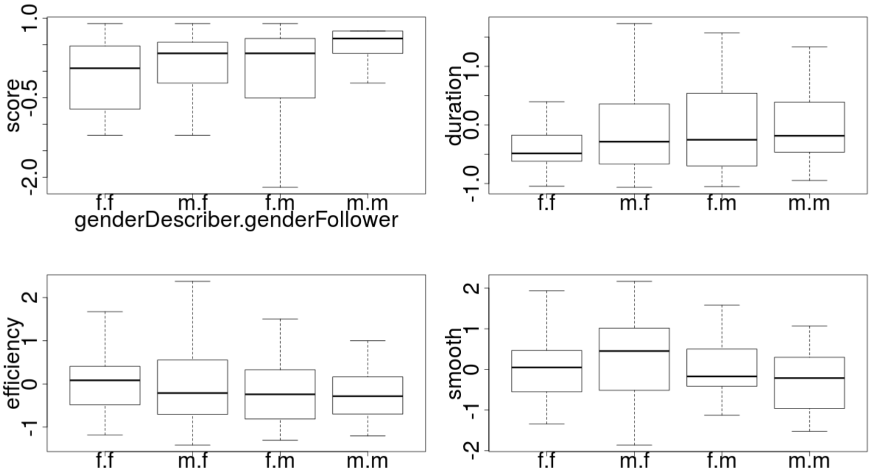

Figure 10: Task success measures (z-transformed) for all gender pairings in the role of describers and followers.

to their interaction ($t < 1.02$, $p > 0.135$).

6. Discussion

In the following we will elaborate on the different behavior of feature sets with respect to proximity and synchrony, as well as with respect to global and local entrainment (see observed tendencies 1–3 in section 5.2.2). Then we discuss reasons for the observation that entrainment predominantly occurs in turn-final position (tendency 4). Finally, gender-role interactions (tendency 5) in entrainment will be explained by different gender-related strategies in solution-oriented collaborative interactions.

6.1. Feature set

What are the general tendencies of set-related entrainment?. As can be seen in Figure 7 next to the well-examined set *gnl_f0* also the new sets *phrase* and *acc* not yet examined in previous studies show clear entrainment tendencies showing that speakers do not only accommodate in terms of coarse but also more fine-grained local prosodic characteristics.

Clearly, *gnl_en* shows disentrainment, especially for local comparisons. This might be due to the cooperative dialog situation in combination with the potential, especially of energy-related entrainment, to hinder cooperation, e.g. in cases when both speakers start to raise their voice as in competitive turn taking situations. From the more general perspective of Causal Attribution Theory [65] one interprets other people's behavior with respect to their intentions and motivations. [66] argue in the Conversation Accommodation Theory (CAT) framework, that proximity might also be considered as negative if the supposed intent has negative connotations. Such a negatively received accommodation occurs for example in *patronizing communication* [29, 67], which can manifest itself in mimicking dialectal features [29] and in a slow and less complex speaking style of young adults when talking to older adults based on negative age-related stereotypes [67]. Applied to our data, a joint increase in energy might be considered as a negative accommodation which is mutually interpreted as confrontational, so that speakers rather diverge on this feature.

This finding also extends previous observations regarding (dis)entrainment in this corpus. [68] and [69] analyzing local and global entrainment in the data in terms of proximity, convergence, and synchrony and using different methodological approaches than this paper, found tendencies for entrainment in intensity that were, nevertheless, stronger than for other features. On the one hand, this supports the analysis of intensity as a feature with low-functional load and thus relatively free to participate in negotiating social relations during the dialogue. On the other hand, the diverging tendencies in the current and previous results suggest the complex nature of entrainment in speech and the possibility that the entrainment potential of certain features within a dataset might be sensitive to different operationalizations of entrainment in terms of synchrony, convergence, local and global domains, or units of analysis.

Do sets differ with respect to global vs. local entrainment?. **More global en-**

trainment indicates that overall speakers accommodate, but local linguistic (e.g. sentence type, dialog act, information status) variation inhibits local accommodation. This can be observed for the feature sets *phrase* and *acc* that clearly are affected by such linguistic parameters. Analogously, disentrainment for such features systematically occurs only on the global level.

rhy_en in contrast can undergo much **more local entrainment** since such rhythm features are much less constrained by linguistic context than *phrase* and *acc*.

Proximity vs synchrony. Feature sets show tendencies to undergo entrainment **either in terms of proximity or of synchrony**. Features defining f0 shape, mainly contained in feature set *acc*, show similarity to a higher extent than synchrony. In contrast overall f0 median, maximum, and standard deviation features from set *gnl_f0* rather synchronize than become similar (e.g. both speakers deviate in the same direction from their mean instead of getting closer). This implies that a mixed-gender conversation does not lead to a mutually approaching f0 mean, i.e. that the female speaker lowers her pitch, while the male speaker raises it. Rather the speakers accommodate in such a way that they both use a high or low register relative to their personal reference, thus they synchronize. Furthermore (as mentioned above), synchrony does not disentangle entrainment from competition as in competitive turn taking situations in which both speakers might signal their interest to keep/get the turn by a relatively high f0 register [70]. For f0 shapes in contrast, synchrony is much less likely, since speakers cannot simply shift different f0 contours in parallel due to non-linearities (e.g. early vs late peak [60]). Rather they accommodate to more similar f0 shapes.

For the feature set *phrase* **both synchrony and proximity apply** to the same extent as is visualized in the right part of Figure 4. This indicates, that the features are varied in parallel but not to the same degree, i.e. one speaker additionally becomes similar towards the other. This asymmetric behavior can be observed predominantly for describers (cf Tables 2 and 3, columns *prox, sync*, rows 7–14) and among them rather for females (cf Table 2, rows 7–14).

6.2. Segment position

Differentiating between turn-initial and turn-final position reveals an imbalance in local but not in global entrainment (cf Figure 8). Local entrainment is more likely to occur in turn-final position. Generally speaking, these different amounts of local and global as well as of turn-initial and -final entrainment support the notion of hybrid causes for accommodation as proposed by [30, 31], cf section 1. Next to automatic priming mechanisms applying throughout the entire turn it seems that in turn-final position pragmatic goals are an additional trigger for entrainment. Turn-finally local pitch events have a higher likelihood to carry dialog structuring functions: while turn-initial pitch events are mostly pitch accents, turn-final events often refer to boundary tones indicating amongst others utterance finality or continuation. Thus in spoken dialogs they serve as turn-taking and backchanneling-inviting cues. For both entrainment has been reported in previous studies by [71] and [72], respectively. Further evidence for entrainment in discourse markers has been found by [73]. Thus, one possible explanation for the higher amount of turn-final as opposed to turn-initial entrainment is the voluntary dialog structuring influence which adds on to automatic entrainment especially at the end of turns.

6.3. Speaker type

Figures 9 shows, that describers d_* entrain more than followers f_* and among describers, it's the female speakers d_f who entrain. For females *_f, describers entrain more than followers, for males *_m it's the opposite.

Globally, disentrainment is to a higher extent found among male speakers x_m, above all among the male describers d_m.

Given these findings one can again conclude that entrainment cannot exhaustively be explained biologically emerging from the perception behavior-link [1], since this explanation does not account for the role-related variation of female and male speakers.

Neither can one conclude that entrainment is a straightforward function of dominance as predicted by the CAT [33] in that sense that the less dominant interlocutor entrains more. Female and male speakers behave differently in their roles of describers and followers, describers being equipped with higher authority than followers due to their lead in knowledge. While men behave in line with the CAT predictions, i.e. highly disentrain in a high authority position, female speakers do the opposite.

One motivation for the female behavior might emerge from the cooperative setting of the game. In this context females might rather use entrainment to increase communication efficiency instead of marking authority. However, as shown in Figure 10 for almost none of the success measures a significant difference between male and female describers has been observed. Only for the *score* variable we found a significant advantage for male describers. Thus, even if the female strategy was to increase communication efficiency, it was not necessarily successful.

Given the cooperative setting of the game, and the finding that male speakers did not perform worse in solving the task than female speakers, it can be concluded that entrainment is used differently across gender in cooperative solution-oriented interactions. Male speakers in the role of describers tend to mark hierarchy by disentrainment, which can be as, or even more, beneficial for task success as the female strategy of common ground creation by entrainment. The amount of entrainment for female and male followers is about the same. Thus male followers entrain more maybe to signal that they accept the describer's authority, and female followers entrain less, since it is less their but rather the describer's responsibility to establish a common ground.

7. Conclusion

In this paper we set to provide a novel approach to analyzing speech entrainment in collaborative dialogues. We focused on disentangling the role of gender and communicative role of the speakers by directed turn pairing and used novel features for characterizing prosody, an extended set of analysis units (turn-initial and final IPUs), and a modified formalization of global and local proximity and synchrony between interlocutors. The results showed that speech entrainment is a highly multi-faceted phenomenon as different groups of features show different entrainment and disentrainment behavior in the local/global and synchrony/proximity domains. Furthermore, entrainment predominantly occurs in the turn-final position, which supports a hybrid account stating both automatic and voluntary triggers. Finally, the observed gender-role interactions might be linked to different strategies in solution-oriented collaborative interactions for males and females.

8. Acknowledgments

The work of the first author is financed by a grant of the Alexander von Humboldt-foundation. This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055 to the second author.

References

[1] T. Chartrand, J. Bargh, The chameleon effect: The perception-behavior link and social interaction, Journal of Personality and Social Psychology 76 (6) (1999) 893–910.

[2] S. K., S. M.V., C. Fowler, Mutual interpersonal postural constraints are involved in cooperative conversation, Journal of Experimental Psychology: Human Perception & Performance 29 (2003) 326–332.

[3] A. Nenkova, A. Gravano, J. Hirschberg, High frequency word entrainment in spoken dialogue, in: Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, Columbus, Ohio, 2008, pp. 169–172.

[4] D. Danescu-Niculescu-Mizil, L. Lee, B. Pang, J. Kleinberg, Echoes of power: Language effects and power differences in social interaction, in:

Proc. 21st international conference on World Wide Web, Lyon, France, 2012, pp. 699–708.

[5] S. Brennan, H. Clark, Conceptual pacts and lexical choice in conversation, J Exp Psychol Learn Mem Cogn 22 (6) (1996) 1482–93.

[6] A. Cleland, M. Pickering, The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure, Journal of Memory and Language 49 (2003) 214–230.

[7] S. Gries, Syntactic priming: A corpus-based approach, Journal of Psycholinguistic Research.

[8] H. Branigan, M. Pickering, J. McLean, A. Cleland, Participant role and syntactic alignment in dialogue, Cognition 104 (2007) 163–197.

[9] S. Gregory, S. Webster, A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions, J. Pers. Soc. Psychol. 70 (1996) 1231–1240.

[10] S. Gregory, K. Dagan, S. Webster, Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality, J. Nonverbal Behavior 21 (1997) 23–43.

[11] R. Levitan, J. Hirschberg, Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions, in: Proc. Interspeech, Florence, Italy, 2011, pp. 3081–3084.

[12] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, A. Nenkova, Acoustic-prosodic entrainment and social behavior, in: NAACL HLT '12 Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada, 2012, pp. 11–19.

[13] M. Babel, D. Bulatov, The role of fundamental frequency in phonetic accommodation, Language and Speech 55 (2012) 231–248.

[14] A. Gravano, v. Beňuš, R. Levitan, J. Hirschberg, Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement, in: Proc. IEEE Spoken Language Technology Workshop, South Lake Tahoe, NV, 2014, pp. 578–582.

[15] D. Reitter, J. Moore, Alignment and task success in spoken dialogue, Journal of Memory and Language 76 (2014) 29–46.

[16] J. Hirschberg, Speaking more like you: Entrainment in conversational speech., in: Proc. Interspeech, Florence, Italy, 2011, pp. 27–31.

[17] Š. Beňuš, Social aspects of entrainment in spoken interaction, Cognitive Computation 6 (4).

[18] K. Schweitzer, M. Walsh, A. Schweitzer, To see or not to see: Interlocutor visibility and likeability influence convergence in intonation, in: Proc. Interspeech, Stockholm, Sweden, 2017, pp. 919–923.

[19] J. Michalsky, H. Schoormann, Pitch convergence as an effect of perceived attractiveness and likability, in: Proc. Interspeech, Stockholm, Sweden, 2017, pp. 2253–2256.

[20] R. Levitan, Š. Beňuš, R. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, J. Hirschberg, Implementing acoustic-prosodic entrainment in a conversational avatar, in: Proc. Interspeech, San Francisco, California, 2016, pp. 1166–1170.

[21] N. Lubold, H. Pon-Barry, E. Walker, Naturalness and rapport in a pitch adaptive learning companion, in: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, Arizona, 2015, pp. 103–110.

[22] E. Raveh, I. Gessinger, S. Le Maguer, B. Möbius, I. Steiner, Investigate phonetic convergence in a shadowing experiment with synthetic stimuli, in: J. Trouvain, I. Steiner, B. Möbius (Eds.), Elektronische Sprachverarbeitung

2017, Vol. 86 of Studientexte zur Sprachkommunikation, TUDpress, Dresden, Germany, 2017, pp. 254–261.

[23] J. Perez, R. Galvez, A. Gravano, Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement, in: Proc. of Interspeech, San Francisco, 2016, pp. 1270–1274.

[24] C. De Looze, S. Scherer, B. Vaughan, N. Campbell, Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction, Speech Communication 58 (2014) 11–34.

[25] P. Healey, M. Purver, C. Howes, Divergence in dialogue, PLoS ONE 9 (6) (2014) 6p.

[26] J. Michalsky, H. Schoormann, O. Niebuhr, Turn transitions as salient places for social signals – Local prosodic entrainment as a cue to perceived attractiveness and likability, in: Proc. Phonetics&Phonology in German-speaking countries, Berlin, Germany, 2017, pp. 125–128.

[27] M. Pickering, S. Garrod, Toward a mechanistic psychology of dialogue, Behavioral and Brain Sciences.

[28] M. J. Pickering, S. Garrod, An integrated theory of language production and comprehension, Behavioral and Brain Sciences 36 (4) (2013) 329–347.

[29] H. Giles, N. Coupland, Language: Contexts and Consequences, Brooks/Cole, Pacific Grove, CA, 1991.

[30] T. Kraljic, S. Brennan, A. Samuel, Accommodating variation: Dialects, idiolects, and speech processing, Cognition 107 (1) (2008) 54–81.

[31] A. Schweitzer, N. Lewandowski, Social factors in convergence of F1 and F2 in spontaneous speech, in: Proc. 10th International Seminar on Speech Production, Cologne, 2014, pp. 391–394.

[32] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, S. Narayanan, Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples, in: Proc. Interspeech, Makuhari, Chiba, Japan, 2010, pp. 793–796.

[33] H. Giles, T. Ogay, Communication accommodation theory, in: B. Whaley, W. Samter (Eds.), Explaining Communication: Contemporary Theories and Exemplars, Lawrence Erlbaum, Mahwah, NJ, 2007, pp. 293–310.

[34] Š. Beňuš, A. Gravano, J. Hirschberg, Pragmatic aspects of temporal accommodation in turn-taking, Journal of Pragmatics 43 (12) (2011) 3001–3027.

[35] D. Zimmermann, C. West, Sex roles, interruptions, and silences in conversation, in: B. Thorne, N. Henley (Eds.), Language and Sex: Difference and Dominance, Newbury House, Rowley, M.A., 1975, pp. 105–129.

[36] L. Namy, L. Nygaard, D. Sauerteig, Gender differences in vocal accommodation: The role of perception, Journal of Personality and Social Psychology 21 (4) (2002) 422–432.

[37] F. Bilous, R. Krauss, Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads, Language & Communication 8 (3/4) (1988) 183–194.

[38] Z. Xia, R. Levitan, J. Hirschberg, Prosodic entrainment in mandarin and english: A cross-linguistic comparison, in: Proc. Speech Prosody, Dublin, Ireland, 2014, pp. 65–69.

[39] J. Edlund, M. Heldner, J. Hirschberg, Pause and gap length in face-to-face face interaction, in: Proc. Interspeech, Brighton, England, 2009, pp. 2779–2782.

[40] R. Fusaroli, B. Bahrami, K. Olsen, A. Roepstorff, G. Rees, C. Frith, K. Tylen, Coming to terms quantifying the benefits of linguistic coordination, Psychological Science 23 (8) (2012) 931–939.

[41] Y. Xu, D. Reitter, An evaluation and comparison of linguistic alignment measures, in: Proc. CMCL, Denver, Colorado, 2015, pp. 58–67.

[42] S. Jones, R. Cotterill, N. Dewdney, K. Muir, A. Joinson, Finding Zelig in text: A measure for normalising linguistic accommodation, in: Proc. COLING, Dublin, Ireland, 2014, pp. 455–465.

[43] Š. Beňuš, The prosody of backchannels in Slovak, in: Proc. Speech Prosody, 2016, pp. 75–79.

[44] G. Agustín, Turn-taking and affirmative cue words in task-oriented dialogue, Ph.D. thesis, Columbia University, NY (2009).

[45] G. Agustín, J. Hirschberg, Turn-taking cues in task-oriented dialogue, Comp. Speech and Language 25 (3) (2011) 601–634.

[46] S. Darjaa, M. Cerňak, M. Trnka, M. Rusko, R. Sabo, Effective triphone mapping for acoustic modeling in speech recognition, in: Proc. Interspeech, 2011, pp. 1717–1720.

[47] P. Boersma, D. Weenink, PRAAT, a system for doing phonetics by computer, Tech. rep., Institute of Phonetic Sciences of the University of Amsterdam, 132–182 (1999).

[48] A. Savitzky, M. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Analytical Chemistry 36 (8) (1964) 1627–1639.

[49] M. Swerts, R. Geluykens, Prosody as a marker of information flow in spoken discourse, Language Speech 37 (1994) 21–43.

[50] H. Pfitzinger, S. Burger, S. Heid, Syllable Detection in Read and Spontaneous Speech, in: Proc. ICSLP, Vol. 2, Philadelphia, 1996, pp. 1261–1264.

[51] U. Reichel, Unsupervised extraction of prosodic structure, in: J. Trouvain, I. Steiner, B. Möbius (Eds.), Elektronische Sprachverarbeitung 2017, Vol. 86 of Studientexte zur Sprachkommunikation, TUDpress, Dresden, Germany, 2017, pp. 262–269.

[52] U. Reichel, Linking bottom-up intonation stylization to discourse structure, Computer, Speech, and Language 28 (2014) 1340–1365.

[53] U. Reichel, CoPaSul Manual – Contour-based parametric and superpositional intonation stylization, RIL, MTA, Budapest, Hungary, https://arxiv.org/abs/1612.04765 (2016).

[54] U. Reichel, Copasul software, GitHub Repository, `https://github.com/reichelu/copasul` (September 14th 2017).

[55] T. Rietveld, P. Vermillion, Cues for Perceived Pitch Register, Phonetica 60 (2003) 261–272.

[56] U. Reichel, K. Mády, Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for hungarian, in: Proc. Interspeech 2014, Singapore, 2014, pp. 111–115.

[57] C. Heinrich, F. Schiel, The influence of alcoholic intoxication on the short-time energy function of speech, J. Acoust. Soc. Am. 135 (5) (2014) 2942–2951. `doi:10.1121/1.4870705`.

[58] S. Fuchs, U. Reichel, On the relation between pointing gestures and speech production in german counting out rhymes: Evidence from motion capture data and speech acoustics, in: Proc. P&P, Munich, Germany, 2016, pp. 51–54.

[59] O. Niebuhr, J. Voše, A. Brem, What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice, Computers in Human Behavior 64 (2016) 366–382.

[60] K. Kohler, Categorical pitch perception, in: Proc. ICPhS, Tallinn, 1987, pp. 331–333.

[61] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67 (1) (2015) 1–48. `doi:10.18637/jss.v067.i01`.

[62] J. Fox, S. Weisberg, An R Companion to Applied Regression, 2nd Edition, Sage, Thousand Oaks CA, 2011.
URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`

[63] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Annals of Statistics 29 (2001) 1165–1188.

[64] D. Tannen, Gender and discourse, Oxford University Press, Oxford, 1994.

[65] F. Heider, The psychology of interpersonal relations, Wiley, New York, 1958.

[66] H. Giles, P. Smith, Accommodation theory: Optimal levels of convergence, in: H. Giles, R. St. Clair (Eds.), Language and Social Psychology, Basil Blackwell, Baltimore, 1979, pp. 45–65.

[67] J. Soliz, H. Giles, Relational and identity processes in communication: A contextual and meta-analytical review of Communication Accommodation Theory, in: E. Cohen (Ed.), Communication Yearbook, Vol. 38, Routledge, 2014, pp. 107–144.

[68] Š. Beňuš, R. Levitan, J. Hirschberg, A. Gravano, S. Darjaa, Entrainment in Slovak collaborative dialogues, in: Proc. 5th IEEE Conference on Cognitive Infocommunications, Vietri sul Mare, Italy, 2014, pp. 309–313.

[69] R. Levitan, Š. Beňuš, A. Gravano, J. Hirschberg, Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison, in: Proc. 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czech Republic, 2015, pp. 325–334.

[70] P. French, J. Local, Turn-competitive incomings, J. Pragmatics 7 (1983) 701–715.

[71] A. Gravano, J. Hirschberg, Conversational entrainment in the use of discourse markers, in: Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium, Springer, 2015, pp. 345–352.

[72] R. Levitan, A. Gravano, J. Hirschberg, Entrainment in speech preceding backchannels, in: Proc. 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 2011, pp. 113–117.

[73] Š. Beňuš, Conversational Entrainment in the Use of Discourse Markers, in: S. Bassis, A. Esposito, F. Morabito (Eds.), Recent Advances of Neural Networks Models and Applications, Smart inovations, systems, and technologies, Vol. 26, Springer, 2014, pp. 345–352.