Original Paper

Phonetica
DOI: 10.1159/000353473

Received: December 14, 2012
Accepted: June 5, 2013

On How Accent Distribution Can Signal Speaker Adaptation

Constantijn Kaland Emiel Krahmer Marc Swerts

Tilburg Centre for Communication and Cognition (TiCC), Tilburg, The Netherlands

Abstract

Some dialogues are perceived as running more smoothly than others. To some extent that impression could be related to how well speakers adapt their prosody to each other. Adaptation in prosody can be signaled by the use of pitch accents that indicate how utterances are structurally related to those of the interlocutor (prosodic function) or by copying the interlocutor's prosodic features (prosodic form). The same acoustic features, such as pitch, are involved in both ways of adaptation. Further, function and form may require a different prosody for successful adaptation in certain discourse contexts. In this study we investigate to what extent interlocutors are perceived as good adapters, depending on whether the prosody of both speakers is functionally coherent or similar in form. This is done in two perception tests using prosodically manipulated dialogues. Results show that coherent functional prosody can be a cue for speaker adaptation and that this cue is more powerful than similarity in prosodic form.

1 Introduction

Speakers in a dialogue adapt to each other. In particular, speaker adaptation is often analyzed regarding the extent to which interlocutors copy certain verbal or nonverbal features of their conversation partners. This kind of adaptation behavior has been modeled in different, but related accounts, most notably in terms of alignment [Pickering and Garrod, 2004], convergence [Pardo, 2006] or mimicry [Bargh and Chartrand, 1999]. As we discuss in this introductory section, there is a large body of work to support the claim that people indeed adapt to each other by copying each other's conversational characteristics, which also holds – more specifically – for the way speakers copy formal aspects of each other's prosody. In addition, while adaptation has mostly been studied in terms of such copying behavior, we will argue that interlocutors may also adapt to each other in a different way, namely in signaling that their utterances

Some results of experiment I were presented in similar fashion at the Speech Prosody 2012 Conference [paper PS2B-9, see Kaland et al., 2012].

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0031–8388/13/0694–0■■■
$38.00/0
Accessible online at:
www.karger.com/pho

C.C.L. Kaland
Tilburg Centre for Communication and Cognition (TiCC)
PO Box 90153
NL–5000 LE Tilburg (The Netherlands)
E-Mail c.c.l.kaland@uvt.nl

are meaningfully related to the utterances of their conversation partners. As we will see, the latter type of adaptation can be reflected in the functional use of prosody, in particular in how speakers distribute pitch accents in their utterances. Interestingly, the formal and functional forms of prosodic adaptation can be in conflict with each other, which begs the question whether one is more important than the other. To answer this general question, we will use a perceptual approach to investigate whether different kinds of prosodic adaptation have a different impact on the way listeners perceive the smoothness of an interaction between two interlocutors.

Adaptation may appear from the fact that the form of speakers' utterances is determined by that of the utterances they received from their conversation partners. This happens at different linguistic levels and is referred to as alignment [Garrod and Pickering, 2004; Pickering and Garrod, 2004]. For example, at the syntactic level it has been shown that speakers repeat prepositions from their conversation partner when answering a question [Levelt and Kelter, 1982]. In their experiment shopkeepers were asked the question 'What time do you close?' or 'At what time do you close?'. They frequently responded 'Five o'clock' to the former and 'At five o'clock' to the latter question. Furthermore, Brennan and Clark [1996] showed that conversation partners in a referential task often formed an agreement on how to refer to an object, i.e. a *conceptual pact*, and ended up using the same lexical description when repeatedly referring to a concept. For example, speakers may agree to refer to a tangram figure that vaguely resembled a skater by 'the ice-skater'. These studies reveal that by taking over lexical features, speakers show that they share information with their interlocutor [common ground; Clark, 1996]. It has also been shown that people in interaction are likely to take over nonverbal aspects. For example, in a conversation speakers may copy each other's body postures [Bargh and Chartrand, 1999], such as having the arms crossed. The process in which people in interaction exhibit postural similarity is called *mimicry*. According to some models, mimicry happens automatically and is of great importance in that people ground themselves in their social environment by taking over the interactional behavior that surrounds them [Bargh and Chartrand, 1999]. In this way mimicry supports the social bonding process between people. The mechanism behind alignment in dialogue is assumed to require a tight link between perception and behavior [Dijksterhuis and Bargh, 2001; Garrod and Pickering, 2004]. In other words, speakers' behavior is a product of the behavior they perceive. The studies just discussed show that adaptation by means of copying certain behavioral features usually concerns the formal characteristics of that behavior. That is, taking over linguistic features from an interlocutor rather affects *how* (form) speakers say what they mean rather than *what* (function) they say. For example, 'Five o'clock' and 'At five o'clock' are semantically equivalent in the contexts sketched above. However, at the level of prosody the distinction between form and function may be more vague as they both manifest themselves in the same acoustic features, such as pitch, duration and intensity. The aim of the present study is to explore to what extent speaker adaptation at the level of prosody is related to function, form or a combination of both.

There is evidence that speakers copy the prosodic form of each other's utterances. For example, dialogue interlocutors are likely to converge on a global pitch range [Gregory, 1986; Couper-Kuhlen, 1996]. That is, speakers' fundamental frequency tends to become similar within certain frequency bands once they engage in a dialogue. Also speech intensity [Natale, 1975] and speech rate [Giles et al., 1991; Szczepek Reed, 2010] are likely to converge to similar levels once speakers interact in

a dialogue. These processes are attested for both young and old age groups [Coulston et al., 2002; Nilsenová et al., 2009]. It is debatable whether prosodic adaption processes like the ones just mentioned are fully automatic. As shown by Pardo [2006], the extent to which interlocutors sound similar on repeated words is also largely determined by the gender- and the task-specific role of the speaker. That is, in a map task male speakers converged more than female speakers and instruction givers converged more than instruction receivers [Pardo, 2006]. In addition, it has been shown that speakers are likely to take over prosodic features that are somehow related to the type of dialogue act in a conversation. For example, Okada et al. [2012] investigated dialogues in which speakers described maps that either matched or mismatched the map of their interlocutor in a map task. In the former type of dialogues interlocutors exhibited agreement, whereas in the latter type of dialogue interlocutors exhibited disagreement. Okada et al. [2012] found that the pitch intervals (rounded off to the nearest semitone) of successive utterances had the same key in agreement dialogues and a different key in disagreement dialogues. Further, Nilsenová et al. [2009] tested whether speakers are more likely to take over prosody when it was related to the semantics (i.e. function) of the utterance compared to when it was not. Participants in their study interacted with a manipulated computer voice. The manipulated prosodic features were boundary tones, which mark phrase beginnings and endings, and pitch range, which has no direct linguistic function. Results showed that speakers prefer to take over boundary tones over pitch range [Nilsenová et al., 2009].

Speakers may also use prosodic features like pitch, duration and intensity in a functional way to adapt to their interlocutor. Most clearly, they can use pitch accents to highlight particular words that are important to the discourse [Pierrehumbert and Hirschberg, 1990]. In most Germanic languages, pitch-accented words are phonetically realized by a certain pitch contour, generally accompanied by an increase in duration and intensity, which lends prominence to the accented word [Pierrehumbert and Hirschberg, 1990]. In Dutch, as well as many other languages, speakers use such accents to make new or contrastive information more prominent, whereas given information is typically deaccented. In interaction the appropriate use of accentuation has to do with the fact that such accents signal coherence between successive utterances by different speakers. That is, by accenting new information and by deaccenting given information speakers show that they account for the information they share with their partners.

At first sight it appears to be the case that speakers often copy the accentuation pattern of their interlocutor when interacting in dialogue. This becomes especially clear when considering contrastive intonation. For example, a first speaker may ask 'Would you like to have a green or a red apple?' with a pitch accent on *green* and on *red* to signal the distinguishing feature (color) between the proposed apples. A second speaker may respond 'I would like to have a red apple' with a pitch accent on *red*. In the latter utterance the second speaker copies the contrastive intonation to acknowledge that he or she is aware of the color contrast as well. In that respect, this example is analogous to the one discussed above where speakers tend to use a preposition in a response when that preposition was also present in the syntactic structure of the conversation partner's question. However, there are other dialogue contexts in which it is not appropriate to copy the interlocutor's accentuation pattern. For example, speakers may start a conversation with their partner by introducing the phrase: 'I have a green apple'. When this utterance represents entirely new information, speakers will typically produce that

sentence with a main accent on the word *apple*. Suppose that their addressee continues the interaction by stating: 'And I have a red apple'. A felicitous use of accents would lead to a realization in which only the word *red* is accented in the final noun phrase (henceforth: NP) of that sentence, to signal the difference in color between the piece of fruit in that sentence and the one mentioned in the preceding utterance. Thus, in this type of context it may be appropriate for speakers to use an accentuation pattern that has a different form compared to the one of the other speaker, as a single accent on the noun *apple* in the second sentence would signal that the second speaker does not fully take into account what has been said in the preceding sentence. Pechmann [1984, 1989] showed that speakers indeed systematically use a contrastive intonation pattern to describe an object that is minimally distinguishable from another object.

With regard to perception, it has been shown that listeners update the information status of both members of a contrast set (i.e. in the example *red* and *apple* as contrastive and given information, respectively) when hearing a contrastive intonation [Braun and Tagliapietra, 2010]. Thus, like copying certain prosodic features, accentuation is a cue for the adaptation process of speakers, as it signals coherence between utterances of both interlocutors. In general, there is much work in which it was shown that accent distributions have an impact on the way speech is perceived. For example, the quality of the spoken output of a speech synthesis system is related to the extent to which the accent distribution in the generated utterances is appropriate [Quené and Kager, 1989]. Similarly, Nooteboom and Kruyt [1987] found that sequences of utterances are perceived as more natural if the distributions of the accents in these consecutive utterances correctly match given/new distinctions. Terken and Nooteboom [1987] showed that inappropriately accented information confuses listeners, resulting in longer comprehension times. Finally, Swerts and Marsi [2012] found that the perceived difference between 'good' and 'bad' newsreaders is related to how these speakers distributed accents in their spoken texts. However, such studies on the perception of accent distributions have mostly been based on analyses of single-speaker productions (monologues), so that it remains to be seen how accent distributions across turns from different interlocutors are perceived. More specifically, the current study focuses on how pitch accents that cue differences in information status in sequences of turns from different speakers may be indicative of how well these speakers adapt to each other.

In sum, when it comes to prosody speakers have different ways to adapt to each other. On the one hand speakers may achieve similarity at the level of prosodic form, by copying each other's acoustic characteristics. On the other hand speakers may achieve coherence at the level of prosodic function, by means of pitch accents to link their utterances semantically to what has been uttered by their conversation partner. Although form and function may apply to different levels of prosody they are both signaled by the same acoustic features. Research has shown that speakers start to sound similar to each other in the course of their interaction, but it is not clear – as we argued above – whether this easily generalizes to accent distributions. Speakers may not simply copy each other's accent patterns because that would interfere with informational constraints, like the existence of contrasts across speaking turns. Much previous research investigated speaker adaptation processes from a production side [i.e. Natale, 1975; Couper-Kuhlen, 1996; Pardo, 2006; Nilsenová et al., 2009]. In the current article, we aim to investigate how listeners perceive dialogue fragments that contain different accent distributions across speaking turns. Our perceptual approach is in line with that of earlier studies, even when these were not concerned with prosody. For

example, it has been shown that the level of adaptation between conversation partners can have an effect on the subjective impression of that interaction [e.g. Chartrand and Bargh, 1999]. As speaker adaptation is commonly defined with respect to copying certain formal linguistic features [Pickering and Garrod, 2004], we first explore whether functional prosody by means of accentuation is a perceptual cue for speaker adaptation. This is tested in experiment I. Thereafter, in experiment II we test the role of formal prosody in addition to functional prosody as perceptual cues for speaker adaptation. We hypothesize that functional prosody, due to its systematic use [Pechmann, 1984, 1989] and its close relation to semantics [Nilsenová et al., 2009] is a cue for speaker adaptation and that it is stronger compared to similarity in prosodic form. Both experiments I and II are perception tasks eliciting listeners' judgments about the adaptation process between speakers who use contrastive intonation in prosodically manipulated dialogues. Listeners hear the dialogues passively, as overhearers [Schober and Clark, 1989].

2 Experiment I: Prosodic Function

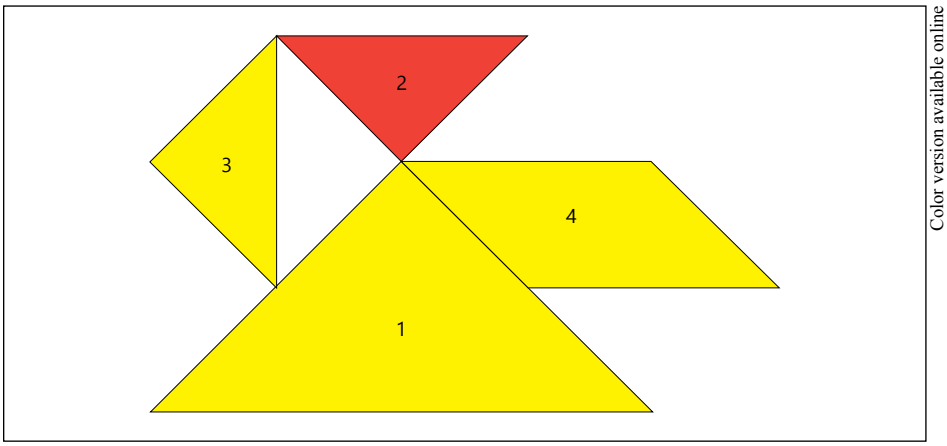
Experiment I elicits judgments about the speaker adaptation process by presenting prosodically manipulated dialogues to listeners. Dialogues consist of utterances that are elicited in a data collection task. Elicited utterances have a natural contrastive intonation. Prosodic manipulation is obtained by the order in which utterances are put together in a dialogue.

2.1 Data Collection

We used a specific semispontaneous procedure to obtain naturally sounding utterance stimuli for our perception test. That is, to elicit utterances with a contrastive intonation a cooperative object naming task was carried out. The participants were instructed to construct a specific shape by using geometrical figures in different forms (triangle, square, parallelogram) and colors (red, blue, yellow). The order in which the geometrical figures should be put together was manipulated such that two successive figures were minimally distinguishable. For example, a blue square may be followed by a blue triangle (shape contrast) or a yellow square (color contrast). In this way shape contrasts elicited accented nouns and color contrasts elicited accented adjectives. Twelve participants carried out the object naming task together with an experimenter (6 males, 6 females, mean age = 27.8 years, age range: 20–33). All participants were native speakers of Dutch. The experimenter and the participant alternately put a geometrical figure in place. Participants were instructed to describe their act using the same matrix sentence for each description: *Ik leg de [blauwe driehoek] hier* (I put the [blue triangle] here). In this way the grammatical position of the NP referring to the geometrical figure was kept constant. Furthermore, each utterance ended with the word *hier* (here) so that boundary tones did not occur within the target NP. A computer screen showed how the target shape consisted of the geometrical figures. In addition, geometrical figures were numbered according to the order in which they had to be constructed (fig. 1). In total 40 shapes were constructed such that each description of a geometrical figure occurred repeatedly in each accent condition. All utterances were recorded as wave files using a Marantz PMD-600 solid state recorder.

2.2 Acoustic Analysis

A total of 288 out of 480 NPs were selected from the 12 speakers that were recorded during the data collection. To meet the requirements of the perception experiment 6 speaker pairs consisting of 1 male and 1 female speaker were formed (see section 2.3). NPs produced by the experimenter were omitted. On the basis of auditory judgments NPs with the clearest accent on either the adjective or the noun were selected. A software program [Audacity Team, 2006] was used to extract the NPs from the wave-file recordings. As a sanity check, we investigated the effect of accent on the NPs by carrying



Color version available online

Fig. 1. Example shape consisting of geometrical figures.

Table 1. Mean pitch maxima (ERB) and standard deviations (in parentheses) as a function of accent and word type (n = 288)

Word type	Accented	Deaccented
Adjective	5.21 (0.19)	4.40 (0.27)
Noun	5.32 (0.36)	4.66 (0.18)

out an acoustic analysis. That is, for each NP the pitch maximum of the adjective and the noun was measured automatically with a script using Praat [Boersma and Weenink, 2011]. Pitch was measured in ERB [Glasberg and Moore, 1990] for two reasons. First, ERB closely resembles the perception of prominence due to accentuation. Second, its logarithmic scale abstracts largely over gender differences. A repeated measures analysis of variance (RM-ANOVA) was carried out with maximum pitch as dependent variable and with accent (two levels: accented, deaccented) and word type (two levels: adjective, noun) as independent variables.

Speakers produce both accented adjectives and accented nouns with a higher pitch than their deaccented counterparts: $F(1, 5) = 65.55$, $p < 0.001$, $\eta_p^2 = 0.93$ (table 1). Although nouns exhibit slightly higher pitch maxima than adjectives, there is no significant effect of word type nor an interaction effect of accent and word type. Further, it is known that in Dutch the perception of contrastive intonation depends on both the prominence of the adjective and the noun [Krahmer and Swerts, 2001]. Therefore, a pitch maxima difference measure is computed by subtracting the maximum pitch value of the deaccented word from the maximum pitch value of the accented word. The difference measure covers the values of both adjectives and nouns. Thereafter, values are computed per speaker pair to investigate whether they differ individually. For example, the value of pair AB is the average pitch maxima difference of speaker A and of speaker B (table 2). In this way we obtain one value per speaker pair that resembles the clarity of their contrastive intonation. Results indicate that speakers from different pairs differ in the degree to which they use pitch to mark the difference between accented and deaccented words. Table 2 shows that pitch maxima differences are the smallest for pair EF and the largest for pair GH. An RM-ANOVA with the difference measure as dependent variable and with speaker pair as independent variable shows an effect of speaker pair: $F(5, 55) = 3.56$, $p < 0.007$, $\eta_p^2 = 0.25$. Pairwise comparisons (after Bonferroni correction) show significant differences between the pairs EF-GH [mean (SE) = 0.62 (0.08), 95% CI = (0.32, 0.91), $p < 0.001$] and GH-IJ [mean (SE) = 0.54 (0.08), 95% CI = (0.25, 0.84), $p < 0.001$].

Table 2. Mean pitch maxima differences and standard deviations (SD) per speaker and per speaker pair

Speaker	Mean (SD)	Speaker pair	Mean (SD)
A	0.83 (1.01)	AB	0.76 (0.49)
B	0.69 (0.28)		
C	0.63 (0.33)	CD	0.80 (0.38)
D	0.97 (0.59)		
E	0.47 (0.18)	EF	0.45 (0.26)
F	0.45 (0.42)		
G	0.85 (0.67)	GH	1.07 (0.25)
H	1.29 (0.48)		
I	0.28 (0.42)	IJ	0.53 (0.22)
J	0.78 (0.26)		
K	1.44 (1.40)	KL	0.85 (0.76)
L	0.25 (0.35)		

Table 3. Examples of functionally coherent (+) and incoherent (–) dialogues per accentuation pattern (accented words in small capitals)

Function	Pattern	NP1	NP2	NP3	NP4
+	AANN	RED triangle	BLUE triangle	yellow SQUARE	yellow TRIANGLE
	NNAA	yellow SQUARE	yellow TRIANGLE	RED triangle	YELLOW triangle
–	AANN	BLUE square	YELLOW triangle	red TRIANGLE	yellow TRIANGLE
	NNAA	red TRIANGLE	blue TRIANGLE	YELLOW square	YELLOW triangle

2.3 Stimuli

Short dialogues were constructed consisting of four NPs (table 3). NPs were taken alternately from 1 male and 1 female speaker (2 NPs each) to balance out possible effects of participants' preferences for male or female voices. Background noise was added to the dialogue to mask edges at turn shifts, which favors the perception of the dialogue as a whole. Dialogues were presented as fragments cut from the middle of an entire dialogue. Therefore, fragments of NPs (approximately one syllable) were added to dialogue edges of which the volume faded in or faded out, respectively. NPs were ordered in such a way that accents were either used coherently or incoherently throughout the dialogue. For functionally coherent dialogues, the contrastive word in the NP was accented and the given word in the NP was deaccented. Contrastiveness and givenness were defined with respect to the previous utterance (table 3). That is, in *blue triangle* after *red triangle* the adjective *blue* would be contrastive and the noun *triangle* would be given. For functionally incoherent dialogues the contrastive word in the NP was deaccented and the given word was accented. The accentuation pattern was balanced over the dialogues to compensate for habituation effects. That is, in one half of the dialogues the first two NPs of a dialogue had an accented adjective whereas the last two NPs had an accented noun (AANN) and vice versa in the second half of the dialogues (NNAA). The duration of one dialogue was approximately 22 s. In total 72 dialogues were constructed using 6 speaker pairs. Per speaker pair 6 coherent and 6 incoherent dialogues were constructed.

2.4 Procedure

Coherent dialogues were coupled with incoherent dialogues in fixed pairs (6 in total). A web-based perception task presented all dialogue pairs per speaker pair (36 pairs in total). Each pair of dialogues originated from the same speaker pair. In order to reduce effects of other sound sources,

Table 4. Means and standard deviations (SD) of proportions of choices for functionally coherent dialogues and χ^2 tests per speaker pair

Speaker pair	Mean (SD)	χ^2 (d.f. = 1, n = 120)
AB	0.76 (0.22)	32.03 ($p < 0.001$)
CD	0.72 (0.28)	22.53 ($p < 0.001$)
EF	0.63 (0.26)	7.5 ($p < 0.01$)
GH	0.82 (0.16)	48.13 ($p < 0.001$)
IJ	0.78 (0.22)	36.30 ($p < 0.001$)
KL	0.63 (0.29)	7.5 ($p < 0.01$)

participants were instructed to do the experiment in a quiet room or to wear headphones. The entire experiment lasted about 20 min. Dialogue pairs occurred in a different random order for each participant. Each dialogue pair was presented on an html-page designed using WWStim [Veenker, 2003]. Each pair of dialogues was represented by two buttons which participants could click to hear them. Participants were told that speakers in the dialogue constructed a tangram figure together while describing the geometrical figures they used. Instructions to participants noted that some speaker pairs collaborated better than others. The participants' task was to choose the dialogue in which speakers account for each other's utterances best. For each stimulus participants were shown the question *In welke dialoog houden sprekers het beste rekening met elkaar?* (In which dialogue do speakers account for each other best?). They were instructed to pay close attention to intonation. Participants were allowed to listen to the dialogues as often as they liked. Once they had made their choice this was registered digitally and could not be altered. The data was analyzed in terms of proportions where 1 corresponds to a choice for a functionally coherent dialogue and 0 corresponds to a choice for a functionally incoherent dialogue.

2.5 Participants

The perception experiment was done by 20 naïve participants (9 males, 11 females, mean age = 27.5 years, age range: 20–35). None of the participants of experiment I took part in the data collection task. All participants were native speakers of Dutch without hearing problems and had no phonetic or linguistic training.

2.6 Statistical Analysis

To investigate participants' preference for coherent or incoherent dialogues a χ^2 test is performed on the proportion of choices (overall and per speaker pair). Furthermore, we test whether participants' choices depend on the speaker pair they heard. Therefore, an RM-ANOVA and pairwise comparisons with mean proportion of choices as dependent variable and speaker pair (six levels: pairs AB to KL) as independent variable is carried out.

2.7 Results

Participants have a preference for indicating the functionally coherent dialogue as the dialogue in which speakers account for each other's utterances best. That is, the coherent dialogues are chosen in 71.94% of the cases. This rate is statistically above the chance level of 50%: χ^2 (1, n = 720) = 138.69, $p < 0.001$, for all speaker pairs (table 4). Results further indicate that the factor speaker pair has an effect on participants' choices: F (5, 95) = 4.78, $p < 0.01$, $\eta_p^2 = 0.20$ (table 4). Pairwise comparisons (after Bonferroni correction) show that the pairs EF-GH and GH-KL differ significantly: mean (SE) = 0.19 (0.05), 95% CI = (0.02, 0.37), $p < 0.05$. Other pairs do not differ significantly. This shows that to some extent participants' choices depend on which speaker pairs they had heard.

Table 5. Mean pitch maxima (ERB) and standard deviations as a function of accent and word type (n = 96)

Word type	Accented	Deaccented
Adjective	5.27 (0.38)	4.09 (0.78)
Noun	5.01 (1.00)	4.70 (0.40)

2.8 Conclusion

Experiment I shows that listeners can use the functional coherence of contrastive intonation patterns at a dialogue level to judge the interlocutors' adaptation process. That is, our judges appear to assess dialogue partners in a conversation as being more cooperative if the accent distributions match the given-new distinctions across speaking turns. Thus, functional prosody is not only a cue within discourse to signal its structure. It also exhibits information about the discourse process itself as a collaborative interaction between speakers.

Interestingly, there is an effect of speaker pair in that listeners judge some pairs as adapting better than other pairs. Inspection of both tables 2 and 4 suggests that this side effect is not simply a listeners' preference for certain speakers. Rather, speaker pairs that show large acoustic differences between accented and deaccented words seem to be perceived as better adapters compared to pairs showing small acoustic differences (see pair EF and GH). Thus, the effect of speaker pair seems to be grounded in the acoustic characteristics of accentuation. Such a conclusion favors the view that functional prosody plays an active role in the perception of speaker adaptation. The question remains, however, how this role relates to prosodic form, as both function and form are expressed in the same acoustic features. In particular, it remains to be investigated to what extent speaker adaptation is signaled by functional or formal prosody. This will be tested in experiment II.

3 Experiment II: Prosodic Function versus Prosodic Form

3.1 Data Collection

NPs used in experiment II were collected in the same data collection procedure as described for experiment I (section 2.1). Selected NPs for experiment II partly overlap those of experiment I.

3.2 Acoustic Analysis

A total of 96 out of 480 NPs were selected from the 12 speakers (6 pairs) that were recorded during the data collection (see section 2.1) to meet the requirements of the perception experiment (see section 3.3). Acoustic analysis was carried out as described for experiment I (section 2.2).

Speakers produce both accented adjectives and accented nouns with a higher pitch than their deaccented counterparts: $F(1, 5) = 61.03$, $p < 0.001$, $\eta_p^2 = 0.92$ (table 5). Although nouns exhibit slightly higher pitch maxima than adjectives, there is no significant effect of word type nor an interaction effect of accent and word type.

3.3 Design

We constructed short dialogues that consist of four NPs collected during the object naming task. The construction of dialogues is similar with respect to balancing male and female speakers, the addition of background noise and volume fading at dialogue edges, as described

Table 6. Dialogues used in experiment II as a function of prosodic function and prosodic form

Function	Form	NP1	NP2	NP3	NP4
+	+	BLUE triangle	YELLOW triangle	RED triangle	BLUE triangle
+	–	BLUE square	blue TRIANGLE	YELLOW triangle	yellow SQUARE
–	+	YELLOW square	YELLOW triangle	BLUE square	BLUE triangle
–	–	blue TRIANGLE	YELLOW square	red TRIANGLE	BLUE square

Accented words in small capitals; + = coherent/similar, – = incoherent/dissimilar.

in section 2.3. However, the constructed dialogues were not identical to the ones of experiment I as in those dialogues only one variable (functional coherence) was manipulated. In the current design two variables were manipulated: *functional coherence* (two levels: coherent, incoherent) and *formal similarity* (two levels: similar, dissimilar; see table 6). For functionally coherent dialogues the accentuation pattern is coherent with respect to the content of what the previous speaker said. Functionally incoherent dialogues follow the opposite pattern. This manipulation of functional coherence is identical to experiment I. For formally similar dialogues speakers are perceived as copying the accentuation pattern of the previous speaker. That is, when the previous speaker accentuates the adjective and deaccentuates the noun, the next speaker does so too, and similarly for cases where only the noun is accented. In formally dissimilar dialogues speakers accentuate the word that the previous speaker deaccentuated and deaccentuate the word that the previous speaker accentuated (table 6). The four types of dialogues were constructed for each speaker pair, resulting in a total of 24 dialogues.

3.4 Materials

The perception task was again web-based and designed using WWStim [Veenker, 2003]. Stimuli were presented online as html pages. One stimulus was considered as an html page displaying two buttons with which two different dialogues of the same speaker pair could be played (fig. 2). Participants made a forced choice between the two dialogues. Dialogue pairs were constructed in such a way that the four dialogue types were uniquely combined (6 per speaker pair). This setup had a total of 36 stimuli with all individual dialogues having an equal chance to be chosen.

3.5 Procedure

The experimental procedure is identical to the one described for experiment I (section 2.4).

3.6 Participants

A total of 20 naïve participants completed the experiment (5 men, 15 women, mean age = 26.6 years, age range: 21–39). They were all native speakers of Dutch without hearing problems. None of the participants acted in the data collection task or in experiment I.

3.7 Statistical Analysis

For each dialogue a score was computed as a proportion for which 1 corresponds to a chosen dialogue and 0 to an unchosen dialogue; χ^2 tests are performed on the proportion data for the four dialogue types, both overall and per speaker pair. Further, χ^2 tests are performed on the proportion data of the variables *functional coherence* and *formal similarity* individually to explore participants' preferences in more detail.

3.8 Results

A χ^2 test on the proportion data of all four dialogue types shows a significant effect [χ^2 (3, n = 720) = 34.68, $p < 0.001$] indicating that participants' preferences clearly differ from a random distribution (25% preference for all dialogue types). Table 7 shows

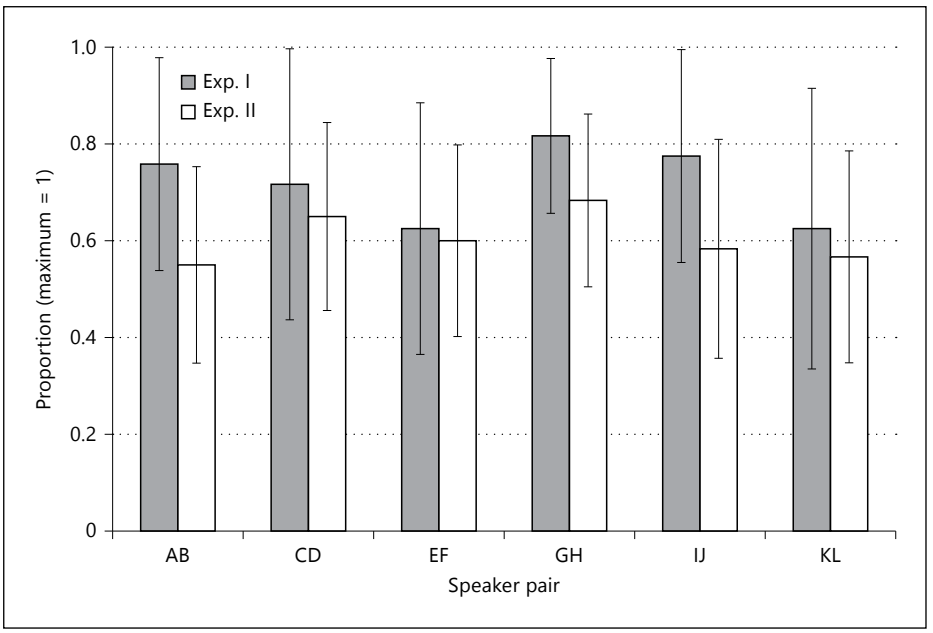


Fig. 2. Proportion of choices for functionally coherent dialogues in experiments I and II. Note that values of experiment I are identical to table 4. For experiment II values in table 7 are added up (see section 3.8).

Table 7. Proportion of chosen dialogues and standard deviations (in parentheses) crossed for prosodic function and prosodic form and χ^2 tests overall and per speaker pair

Speaker pair	Function: +		Function: –		χ^2 (d.f. = 3)
	form: +	form: –	form: +	form: –	
AB (n = 120)	0.24 (0.43)	0.31 (0.46)	0.27 (0.44)	0.18 (0.39)	3.93 (NS)
CD (n = 120)	0.29 (0.46)	0.36 (0.48)	0.20 (0.40)	0.15 (0.36)	12.47 ($p < 0.01$)
EF (n = 120)	0.32 (0.47)	0.28 (0.45)	0.24 (0.43)	0.16 (0.37)	6.73 ($p = 0.08$)
GH (n = 120)	0.35 (0.48)	0.33 (0.47)	0.17 (0.37)	0.15 (0.36)	16.27 ($p < 0.01$)
IJ (n = 120)	0.24 (0.43)	0.34 (0.48)	0.19 (0.40)	0.23 (0.42)	6.00 (NS)
KL (n = 120)	0.31 (0.46)	0.26 (0.44)	0.23 (0.42)	0.21 (0.41)	2.80 (NS)
Total (n = 720)	0.29 (0.46)	0.31 (0.46)	0.22 (0.41)	0.18 (0.38)	34.68 ($p < 0.001$)

+ = Coherent/similar, – = incoherent/dissimilar.

that there are differences per speaker pair in that for some speaker pairs listener's judgments are randomly distributed (AB, IJ, KL) whereas for other speaker pairs this is not the case (CD, EF and GH). The data in table 7 further suggest that listeners are especially sensitive to functional coherence. Indeed, if we do focused analyses, it turns out that participants choose the functionally coherent dialogues more often than the functionally incoherent dialogues: χ^2 (1, n = 720) = 32.09, $p < 0.001$. To compare experiment I and experiment II, figure 2 shows the proportions of choices for functionally coherent dialogues. Note that values of experiment I are identical to table 4. Values of experiment II are taken from table 7 by taking the sum of values of choices for functionally coherent dialogues which abstracts over formal similarity. Therefore, the computation for speaker pair AB is 0.24 + 0.31 = 0.55. The results of experiments I and II are comparable, with similar patterns across speaker pairs (fig. 2). As for formal similarity, participants choose the dialogues with formally similar prosody equally often as dialogues with formally dissimilar prosody: χ^2 (1, n = 720) = 0.14, p = 0.71.

3.9 Conclusion

Experiment II shows that speakers are perceived as better adapters when they use prosody in a functionally coherent way. Prosodic form, i.e., whether or not speakers literally copy the accent distribution of their partners, does not have an effect on the subjective impression of the interaction. This result is in line with experiment I, in which it is shown that accentuation is a cue for the adaptation process between interlocutors. Experiment II extends this finding by showing that functional prosody between speakers' successive utterances is a stronger cue for the adaptation process than mere prosodic form.

4 Discussion

The present study shows that accentuation patterns that are functionally coherent with respect to the previous speaker's utterance can act as a cue for the adaptation process between interlocutors. That is, when speakers use their accents to mark contrasts with information of their partner's preceding turn, then this leads to an overall impression of the dialogue being coherent, when compared to cases in which the accent distributions are inconsistent with such contrasts. Moreover, the current results indicate that functional prosody, as far as accent distributions are concerned, is a stronger cue for speaker adaptation than formal prosody. Such an outcome is in accordance with studies indicating that speaker adaptation in prosody is more likely when the prosody does relate to the semantics of an utterance [Nilsenová et al., 2009; Okada et al., 2012] compared to when it does not. That is, when compared to formal prosody, functional prosody is a stronger indication that dialogue partners are well adapted to each other.

Both experiments I and II indicate that listeners' adaptation judgments differ per speaker pair. Although listeners were instructed to judge adaptation on the basis of intonation, we cannot rule out the possibility that certain preferences for speaker pairs did play a role in the judgments. Those preferences may relate to the sound of a certain voice or a combination of voices in particular pairs. For example, in the current approach we did not control for pitch range differences between speakers, a feature that often converges to similar levels when speakers interact [Couper-Kuhlen, 1996]. Therefore, it may have been the case that certain speaker pairs matched acoustically

better than others, which could have caused differences in how participants judged the adaptation process. However, such a conclusion cannot be drawn from the current results. Our findings do suggest that speaker pair differences found in experiment I are grounded in the acoustic realization of pitch accents. This favors the view that the prosodic form did play a minor role as a cue for speaker adaptation. A closer investigation of individual differences is left for future research.

The current findings confirm previous work on the use of accents in discourse. That is, functionally coherent accentuation provides cues for how successive utterances are related [Pechmann, 1984, 1989; Pierrehumbert and Hirschberg, 1990]. The structure that can be derived from accents thus is a useful tool for both speakers and listeners to build new information on top of what has been said before. The current study seems to suggest that the clearer the accents are used to signal discourse structure the better the interlocutors are judged to account for each other. Thus, pitch accents signal both structure and speakers' interactional behavior in discourse.

Before concluding, three points about the current methodology should be noted explicitly. First, listeners in experiment II did not show a preference for formally similar dialogues compared to the formally dissimilar dialogues. This does not allow us to conclude that prosodic form is not a cue for speaker adaptation at all. Rather, the outcome may be related to how the variables' function and form were manipulated. That is, for both variables the location of the pitch accent was the cue on which listeners should have based their judgments. Typically, the location of pitch accents signals the information status of words rather than whether 2 speakers sound similar. In other words, adaptation at the level of prosodic form may be less bound to a specific location in an utterance [i.e. Natale, 1975; Gregory, 1986; Giles et al., 1991; Couper-Kuhlen, 1996; Szczepek Reed, 2010] than adaptation at the level of prosodic function. The choice for a manipulation of the pitch accent's location in the current approach may have restrained an effect of prosodic form.

Second, listeners in the current perception experiments are not addressees. In the perception experiment listeners hear the dialogue passively. That is, they are overhearers [Schober and Clark, 1989] and do not actively build common ground with an interlocutor. With respect to this distinction one could think of a variant of the perception experiment in which participants are both interlocutor and judge. For example, they act in a dialogue with a partner that either responds with a coherent or incoherent intonation pattern. Participants may then be expected to judge the adaption process in a more radical way due to their active role in dialogue. This paradigm is carried out in a follow-up study [Kaland et al., in preparation].

Third, it has to be noted that in the current setup dialogues are reduced to combinations of short utterances (NPs). In spontaneous speech there may be more complex discourse structures and therefore other prosodic cues to partner adaptation. One such cue could be the use of boundary tones. It is known that those tones mark whether the speaker has more to say, asks for response or has finished the utterance. Nilsenová et al. [2009] showed that speakers are likely to adapt to their interlocutor by copying their boundary tone in the course of the interaction between them. With respect to utterance finality, however, copying the interlocutor's boundary tone may not be appropriate in the middle of a dialogue. Future work could investigate, along the lines of experiment II, to what extent speaker adaptation in boundary tones is signaled by its functional or formal use. We conjecture that, as shown in the current study for accentuation, the function of prosody is a stronger cue for the interlocutor's adaptation than the form of prosody.

To conclude, this study has shown that adaptation is signaled merely by speakers' coherence in prosodic function and not by speakers' similarity in prosodic form. This nuances a common definition of prosodic adaptation that relies on the speakers' tendency to copy their interlocutors' prosodic form. A stronger form of adaptation may be found in the way speakers use prosody to signal coherence between their own and their interlocutor's utterances in a functional way.

References

Audacity Team: Audacity; version 1.2.6. Retrieved from http://audacity.sourceforge.net (2006).

Bargh, J.A.; Chartrand, T.L.: The unbearable automaticity of being. Am. Psychol. *54:* 462–479 (1999).

Boersma, P.; Weenink, D.: Praat: doing phonetics by computer; version 5.2.19. Retrieved from http://www.praat.org (2011).

Braun, B.; Tagliapietra, L.: The role of contrastive intonation contours in the retrieval of contextual alternatives. Lang. cogn. Processes *25:* 1024–1043 (2010).

Brennan, S.E.; Clark, H.H.: Conceptual pacts and lexical choice in conversation. J. exp. Psychol. Learn. Memory Cogn. *22:* 1482–1493 (1996).

Chartrand, T.L.; Bargh, J.A.: The chameleon effect: the perception-behavior link and social interaction. J. Personality soc. Psychol. *76:* 893–910 (1999).

Clark, H.H.: Using language (Cambridge University Press, Cambridge 1996).

Coulston, R.; Oviatt, S.; Darves, C.: Amplitude convergence in children's conversational speech with animated personas. Proc. Int. Conf. Spoken Lang. Processing, Denver 2002, pp. 2689–2692.

Couper-Kuhlen, E.: The prosody of repetition: on quoting and mimicry; in Couper-Kuhlen, Selting, Prosody in conversation: interactional studies, pp. 366–405 (Cambridge University Press, Cambridge 1996).

Dijksterhuis, A.; Bargh, J.A.: The perception-behavior expressway: automatic effects of social perception on social behavior; in Zanna, Adv. exp. soc. Psychol., vol. 33, pp. 1–40 (Academic Press, San Diego 2001).

Garrod, S.; Pickering, M.J.: Why is conversation so easy? Trends cogn. Sci. *8:* 8–11 (2004).

Giles, H.; Coupland, N.; Coupland, J.: Accommodation theory: communication, context and consequences; in Giles, Coupland, Coupland, Contexts of accommodation: developments in applied linguistics, pp. 1–68 (Cambridge University Press, Cambridge 1991).

Glasberg, B.R.; Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data. Hear. Res. *47:* 103–138 (1990).

Gregory, S.W.: Social psychological implications of voice frequency correlations: analyzing conversation partner adaptation by computer. Soc. Psychol. Q. *39:* 237–246 (1986).

Kaland, C.C.L.; Krahmer, E.J.; Swerts, M.G.J.: Accentuation as a cue for speaker adaptation; in Ma, Ding, Hirst, Proc. 6th Int. Conf. on Speech Prosody 2012, vol. I, pp. 290–294 (Tongji University Press, Shanghai 2012).

Kaland, C.C.L.; Krahmer, E.J.; Swerts, M.G.J.: Adapting to inappropriate prosody (Tilburg University, in preparation).

Krahmer, E.J.; Swerts, M.G.J.: On the alleged existence of contrastive accents. Speech Commun. *34:* 391–405 (2001).

Levelt, W.J.M.; Kelter, S.: Surface form and memory in question answering. Cogn. Psychol. *14:* 78–106 (1982).

Natale, M.: Converge of mean vocal intensity in dyadic communication as a function of social desirability. J. Personality soc. Psychol. *32:* 790–804 (1975).

Nilsenová, M.; Swerts, M.G.J.; Houtepen, V.; Dittrich, H.: Pitch adaptation in different age groups: boundary tones versus global pitch. Proc. Interspeech 2009, Brighton 2009, pp. 1015–1018.

Nooteboom, S.G.; Kruyt, J.G.: Accents, focus distribution, and the perceived distribution of given and new information. J. acoust. Soc. Am. *82:* 1512–1524 (1987).

Okada, B.M.; Lachs, L.; Boone, B.: Interpreting tone of voice: musical pitch relationships convey agreement in dyadic conversation. J. acoust. Soc. Am. *132:* 208–214 (2012).

Pardo, J.S.: On phonetic convergence during conversational interaction. J. acoust. Soc. Am. *119:* 2382–2393 (2006).

Pechmann, T.: Überspezifizierung und Betonung in referentieller Kommunikation (Universität Mannheim, Mannheim 1984).

Pechmann, T.: Incremental speech production and referential overspecification. Linguistics *27:* 89–110 (1989).

Pickering, M.J.; Garrod, S.: Toward a mechanistic psychology of dialogue. Behav. Brain Sci. *27:* 169–226 (2004).

Pierrehumbert, J.; Hirschberg, J.: The meaning of intonational contours in the interpretation of discourse; in Cohen, Morgan, Pollack, Intentions in communication, pp. 271–312 (MIT Press, Cambridge 1990).

Quené, H.; Kager, R.: Automatic accentuation and prosodic phrasing for Dutch text-to-speech conversion. Proc. Eur. Conf. on Speech Commun. Technol. (Eurospeech), Edinburgh 1989, pp. 214–217.

Schober, M.F.; Clark, H.H.: Understanding by addressees and overhearers. Cogn. Psychol. *21:* 211–232 (1989).

Swerts, M.; Marsi, E.: Prosodic evaluation of accent distributions in spoken news bulletins of Flemish newsreaders. J. acoust. Soc. Am. *132:* 2616–2624 (2012).
Szczepek Reed, B.: Speech rhythm across turn transitions in cross-cultural talk-in-interaction. J. Pragmatics *42:* 1037–1059 (2010).
Terken, J.; Noteboom, S.G.: Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. Lang. cogn. Processes *2:* 145–163 (1987).
Veenker, T.J.G.: WWStim: a CGI script for presenting web-based questionnaires and experiments; version 1.4.4 (Utrecht University, 2003). Retrieved from http://www.let.uu.nl/~Theo.Veenker/personal/projects/wwstim/doc/en/.