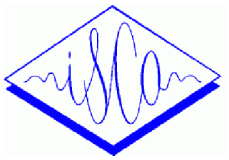

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Contrastive intonation in autism: The effect of speaker- and listener-perspective

Constantijn Kaland, Emiel Krahmer, Marc Swerts

Tilburg centre for Communication and Cognition (TiCC), Tilburg University, The Netherlands

{c.c.l.kaland, e.j.krahmer, m.g.j.swerts}@uvt.nl

Abstract

To indicate that a referent is minimally distinguishable from a previously mentioned antecedent speakers can use contrastive intonation. Commonly, the antecedent is shared with the listener. However, in natural discourse interlocutors may not share all information. In a previous study we found that typically developing speakers can account for such perspective differences when producing contrastive intonation. It is known that in autism the ability to account for another's mental state is impaired and prosody is atypical. In the current study we investigate to what extent speakers with an autism spectrum disorder account for their listeners when producing contrastive intonation. Results show that typical and autistic speakers produce contrastive intonation similarly although both groups differ with respect to other aspects of prosody.

Index Terms: prosody, contrastive intonation, autism, prominence

10.21437/Interspeech.2012-308

1. Introduction

Speakers use prosody to mark information status in discourse. For example, to indicate that certain information contrasts with what is previously mentioned a specific intonation pattern may be used. Consider the utterance "last week I saw a yellow bird, but yesterday I saw a red bird". Typically, speakers accentuate *red* and deaccentuate *bird* to indicate that the colour distinguishes the two birds that were seen [1], [2]. Contrastive intonation can be used with respect to information that is previously mentioned and shared with the listener (i.e. the antecedent). In a previous study we investigated to what extent typically developing speakers account for the listener when they produce contrastive intonation [3]. Results show that speakers produce accented words less prominently and deaccented words more prominently when the antecedent is not shared with the listener than when the antecedent is shared with the listener. On the one hand this shows that speakers account for listeners that have a different perspective. On the other hand, speakers still produce some form of contrastive marking that only makes sense from their own perspective. So, typically developing speakers producing contrastive intonation take into account perspectives of both interlocutors. This ability to account for another perspective is claimed to be impaired in speakers with an autism spectrum disorder. Therefore, the present research investigates to what extent autistic speakers account for the information they share with a listener when they produce contrastive intonation.

Autism is often characterized as having difficulties reflecting on the contents of one's own and other's mind [4]. The ability to do so is referred to as having a Theory of Mind (ToM). An impaired ToM can result communicative problems. Studies often point to prosody as a key feature of communicative difficulties in speakers with autism [5]. Those difficulties relate to both production and perception. As for production, prosody is assigned a variety of impressionistic and contradictory classifications, such as 'monotonous', 'exaggerated' and 'bizarre' [5], [6]. Studies that acoustically analyse speech in autism show that those impressions relate to general 'paralinguistic' aspects of prosody [7], such as pitch range [8]. That is, speakers with autism often use a wider pitch range than typical speakers. A few studies look at the relation between ToM and prosody directly by carrying out perception tasks. In those tasks participants have to recognize an emotional or mental state on the basis of vocal cues. Results show that autistic listeners perform worse than typicals, although differences are sometimes small (cf. [9] and [10]). So far, research lacks a direct investigation of how the production of prosody relates to ToM.

Studies that investigate the production of contrastive intonation find that this pattern is particularly problematic for speakers with an autism spectrum disorder. Several studies find that accents are placed on more than one syllable [11] or on inappropriate words [12], [13]. An interesting finding is done by [6] who investigated both the perception and production of contrastive intonation. They show that autistic children have difficulties interpreting contrastive intonation when the adjective is accented. Nevertheless, those children have a tendency to produce an accent on the adjective even when inappropriate.

To sum up, aspects related to prosodic form are found to be atypical in autism. This atypicality can be ascribed to ToM, although evidence from production is lacking. Further, the use of contrastive intonation, a functional aspect of prosody, is problematic in autism. The present research focuses explicitly on the production of contrastive intonation in autism to investigate to what extent speaker- and listener-factors are taken into account. Presumably, autistic speakers have difficulties accounting for their listener and produce contrastive intonation with respect to their own perspective. Furthermore, we expect autistic speakers to make more accent placement errors and to exhibit atypical prosody in general. We investigate contrastive intonation as a functional aspect of prosody in a production experiment and by means of prominence analysis. We investigate speech dynamicity and speaker commitment as formal aspects of prosody in a perception experiment. Results of typically developing speakers collected in [3] serve as control.

2. Method

Utterances with a contrastive intonation are elicited in a production experiment. A perception experiment is carried out to collect judgements about formal aspects of prosody.

2.1. Production experiment

2.1.1. Participants

20 different participants act as speaker in the production experiment (6 women, 14 men, M_{age} = 28.9 years, age range: 18-51 years). They are all native speakers of Dutch with high

functioning autism (HFA), diagnosed between November 2005 and October 2011. All participants are diagnosed by either a psychiatrist or psychologist on the basis of DSM-IV [14] as having Asperger Syndrome (1 woman, 6 men) or Pervasive Developmental Disorder – Not Otherwise Specified (PDD-NOS; 5 woman, 8 men). They are given a small present for their effort.

2.1.2. Design and procedure

The experimental design and procedure is identical to the one in [3]. To elicit references to contrastive information, participants act as speakers in a referential communication task in which they instruct two different listeners to put figures on a bingo card. The order of instructions is manipulated so that successive instructions refer to figures that can be distinguished by just their colour or just their shape (test stimuli) or by both their colour and their shape (fillers). A test stimulus concerns the latter of two successive instructions, as the present study investigates contrastive intonation with respect to the previous utterance. Two successive instructions are either uttered to the same listener or to different listeners (*listener*: same, different). The setup ensures that only successive instructions to the same listener make sense in terms of contrastive intonation, not successive instructions to different listeners. That is, speakers are told that when addressing one listener, the other listener hears music via a headphone so that the instruction cannot be heard. In reality, listeners are confederates and hear all instructions (see below). Contrastive information in the test stimuli concerns either the colour or shape of the target figure, thus the focused word is either the adjective or noun (*focus*: adjective, noun).

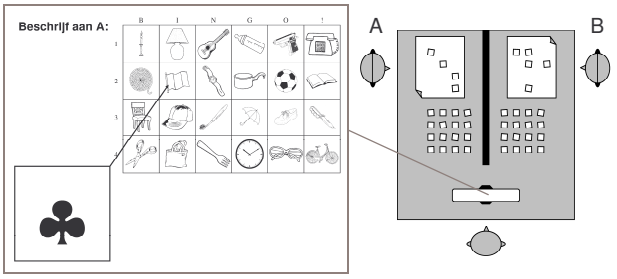

Figure 1: *Left: example of the speaker's screen, showing in Dutch "Beschrijf aan A" (describe to A), the target figure (bottom left) and A's bingo card. An instruction could be: "put the red clover on the flag". Right: birdseye view of the setup showing the speaker facing the screen (bottom) and the listeners facing their bingo cards and figures (top).*

The communication task is played as a bingo game with the speaker as the game leader and listeners as players. Each listener has a different bingo card displaying 24 common objects (e.g. fruit, tools, means of transport). Bingo cards are 6 x 4 grids with rows numbered from 1 to 4 and columns marked by each character of "bingo!" (Figure 1). In addition, listeners each have a set of paper card figures; a drop, clover, canoe or triangle (in Dutch *druppel*, *klaver*, *kano* and *driehoek* respectively) coloured red, yellow, green or blue (in Dutch *rood, geel, groen* and *blauw* respectively). Different rounds are played, which begin by the speaker's announcement of which row or column has to be covered by target figures (for example a figure on each cell of row 2). The listener who achieves the right pattern first shouts "bingo!", upon which that listener receives a point and the round ends. The speaker has to keep the scores. The first instruction of each new round is a filler to account for speakers' pitch reset upon switching discourse contexts. The stimulus order occurs in two randomizations; each of which is presented to 10 participants. Speakers utter 48 instructions in total (equally spread over listeners, crossed for the factors listener and focus). The speaker is seated at one end of a table and listeners, who cannot see each other but who are both visible to the speaker, at the other end (Figure 1). Before the game begins speakers receive instructions and play a training round. Listeners wear open-ear headphones to facilitate the speaker's illusion that the listener who is not addressed hears music. After each experiment speakers are asked whether they indeed believe that listeners heard music and not the instruction (all responded affirmative).

Speakers (not listeners) see a screen displaying the target figure and the bingo card of the listener to be addressed (Figure 1). The screen's lay-out indicates when speakers have to switch between listeners. That is, for listener A the target figure is displayed on the screen's left side and for listener B the target figure is displayed on the right side. In accordance, speakers have to look past the left side of the screen when addressing listener A and past the right side of the screen when addressing listener B (Figure 1). Additionally, speakers are told that the software responsible for the instruction slides on the screen also switches music between listeners. Speakers' speech is digitally recorded by a headset microphone and saved as wave-file.

2.1.3. Prosodic analysis

We extract noun phrases (NPs) referring to target figures in the test stimuli (n = 480) from the wave-file recordings using Praat [15]. They are analysed for prominence by perception ratings. For this, a web-based task [16] presents the NPs to three intonation experts. They rate the strength of the accent on a three point scale (0: no accent, 1: weak accent, 2: strong accent). Adjectives are rated in the first part of the task, nouns are rated in the second part. Experts hear the full NPs in both parts. In this way we obtain a prominence judgement per word in the NP. The presentation order of NPs is randomized so that experts are blind for condition. To abstract over the experts' ratings, the prominence scores per word are added up so that they range from 0 to 6 (0 when all experts rate the accent as absent, 6 when all experts rate the accent as strong). Pearson's correlation coefficients as computed for the adjective and noun ratings indicate that the experts' ratings are consistent [r(478) range = .59 - .69, $p < .001$].

A contrastively focused word in Dutch obtains prominence by both accentuation and deaccentuation [2]. To account for this a difference score is computed. That is, the prominence score of the unfocused word is subtracted from the prominence score of the focused word. In this way, positive difference scores indicate that the focused word is more prominent than the unfocused word and negative scores indicate that the unfocused word is more prominent than the focused word.

2.1.4. Statistical analysis

Repeated measures analysis of variance (RM-ANOVA) is performed on prominence difference scores collected in [3] and the production experiment as dependent variables with listener (2 levels: same, different) and focus (2 levels: adjective, noun) as within-subject factors and with development (2 levels: typical and HFA) as between subject factor.

2.1.5. Results

Data of both typical and HFA speakers show that addressing the same listener (M = 2.52) results in significantly larger prominence difference scores than addressing a different listener (M = 1.64): [$F(1,38) = 25.72$, $p < .001$, $\eta_p^2 = .40$]. As for focus, speakers produce larger differences when the focused word is the adjective (M = 3.31) than when it is the noun (M = .84): [$F(1,38) = 26.52$, $p < .001$, $\eta_p^2 = .41$]. The factor development shows a trend [$F(1,38) = 2.96$, $p = .093$] in that typical speakers produce larger (M = 2.42) differences between focused and unfocused words than HFA speakers (M = 1.74).

Table 1. *Mean prominence scores and standard deviation for adjective, noun and their difference as a function of development, listener and focus. Data typicals obtained in [3].*

Development	Listener	Focus	Prominence score *M* (*SD*)		
			Adjective	Noun	Difference
Typical	Same	Adjective	5.21 (*1.00*)	1.32 (*1.44*)	3.89 (*2.23*)
		Noun	2.42 (*1.96*)	4.30 (*1.83*)	1.88 (*3.63*)
	Different	Adjective	4.79 (*1.43*)	1.67 (*1.63*)	3.13 (*2.84*)
		Noun	2.94 (*1.84*)	3.71 (*1.92*)	0.76 (*3.57*)
HFA	Same	Adjective	5.03 (*1.42*)	1.76 (*1.40*)	3.27 (*2.61*)
		Noun	2.96 (*2.06*)	3.98 (*1.78*)	1.03 (*3.62*)
	Different	Adjective	4.86 (*1.48*)	1.90 (*1.49*)	2.96 (*2.67*)
		Noun	3.66 (*1.98*)	3.36 (*1.90*)	-0.30 (*3.55*)

Results show that typical and HFA speakers produce contrastive intonation in a similar way. That is, they both account for the listener by producing contrastive intonation in an attenuated way when the antecedent is not shared with the listener. Further, they both account for their own perspective as well, as focus remains marked when addressing a different listener. One exception in HFA speakers concerns the adjective when addressing a different listener. Here, speakers produce the adjective slightly more prominent than the noun, although the noun is focused. In general, there is no evidence that HFA speakers produce contrastive intonation with respect to their own perspective nor that they make accent placement errors. Those findings are counter to what has been suggested in the literature concerning functional aspects of prosody in autism. However, there is other evidence suggesting that differences between typical and HFA speakers are related to formal aspects of prosody. To investigate this issue a perception experiment is carried out.

2.2. Perception experiment

2.2.1. Participants

30 different participants do the perception experiment (22 women, 8 men, M_{age} = 22.6 years, age range: 18-60 years). They are all native Dutch speaking students of Tilburg University without hearing problems participating for course credit. None of them participated in the production experiment or in [3].

2.2.2. Design and procedure

The perception experiment elicits judgements about two general aspects of prosody: speech dynamicity and speaker commitment. Dynamicity is taken as an aspect that directly relates to the speech sound, whereas commitment relates to the speaker and indirectly to the speech sound. The participant's task is to judge these aspects on a five point scale; ranging from 'monotonous' (1) to 'dynamic' (5) and from 'not committed' (1) to 'very committed' (5). Participants are given no definition of the concepts dynamicity or commitment. Stimuli consists of NPs collected in the production experiment and in [3]. That is, four NPs are taken from each typical and HFA speaker such that each condition is represented (*listener* crossed with *focus*). In total 160 NPs are presented to the participants.

To avoid surrounding noise, participants do the perception experiment in a sound booth and wear headphones. Before the start of the actual experiment participants can adjust the audio volume, receive instructions and have to complete an example stimulus. NPs are presented on html-pages designed using WWStim [16]. NPs are presented in a random order that is different for each participant. During the experiment each NP can be played as often as needed. The judgement can be altered before proceeding to the next stimulus. Participants cannot alter choices made previously. The task lasts about 25 minutes. Results are collected on a web server.

2.2.3. Statistical analysis

An RM-ANOVA is performed on judgement scores of dynamicity and commitment as dependent variables with development (2 levels: typical and HFA), listener (2 levels: same, different) and focus (2 levels: adjective, noun) as within-subject factors.

2.2.4. Results

Table 2. *Mean speech dynamicity, speaker commitment and standerd deviation as a function of speaker development.*

Development	Dynamicity	Commitment
Typical	3.22 (*1.11*)	2.95 (*1.17*)
HFA	3.14 (*1.12*)	2.85 (*1.15*)

Listeners perceive the speech of typical speakers as more dynamic (M = 3.22) than the speech of HFA speakers (M = 3.14): [$F(1,29) = 7.34$, $p < .05$, $\eta_p^2 = .20$], see Table 2. The dynamicity scores show no effects of listener or focus. Further, listeners judge typical speakers as more committed (M = 2.95) than HFA speakers (M = 2.85): [$F(1,29) = 6.77$, $p < .05$, $\eta_p^2 = .19$]. As for listener, the commitment scores show no significant effect. Focus does show an effect, in that speakers producing focused adjectives are perceived as more committed (M = 2.93) than speakers producing focused nouns (M = 2.86): [$F(1,29) = 4.57$, $p < .05$, $\eta_p^2 = .14$]. Moreover, focus interacts with development in that the production of focused adjectives is perceived as much more committed than the production of focused nouns for typical speakers than for HFA speakers: [$F(1,29) = 7.94$, $p < .01$, $\eta_p^2 = .22$]. These effects suggest that accentuation affects how committed a speaker is perceived.

Analysis of both the speech dynamicity and speaker commitment scores shows that these judgements largely correlate: [$r_{typical}(78) = .95$, $p < .001$] and [$r_{HFA}(78) = .93$, $p < .001$]. This indicates that when speech is more dynamic listeners judge the speakers to be more committed.

3. Conclusions

3.1. Prominence

The production experiment shows that HFA speakers incorporate both speaker- and listener-factors when producing contrastive

intonation. That is, speakers produce contrastive intonation differently when the antecedent is shared with their listener than when the antecedent is not shared with their listener. As for speaker-factors, HFA speakers produce some form of focus marking when addressing a different listener. When HFA speakers would fully account for their listeners, no focus marking would be expected when they address a different listener. Concerning listener-factors, results shows that HFA speakers produce contrastive intonation more clearly when addressing the same listener. Thus, HFA speakers account for a different listener by producing an attenuated contrastive intonation pattern for them.

Comparing the typical speakers [3] with the HFA speakers provides no clear differences. That is, both development groups produce contrastive intonation less clearly when addressing a different listener. So both groups show evidence for taking into account both speaker- and listener-perspectives. Only subtle differences are found in that HFA speakers have a larger tendency to produce the adjective more prominently when addressing a different listener. Typical speakers in the same condition do not show this tendency. More data is needed to asses whether this difference is meaningful.

3.2. Speech dynamicity and speaker commitment

Both the speech dynamicity and commitment are judged higher for typical than for HFA speakers. The effects indicate that the prosody of both development groups indeed differs. With respect to dynamicity, we can conclude that HFA speakers sound slightly more monotonous than typical speakers. This confirms observations by [17], but not those of [18]. The different methodology in this study may explain this conclusion. As for speaker commitment, HFA speakers are found to be slightly less committed than typical speakers. Our results suggest that speaker commitment relates to focus marking in that larger prominence differences found for adjectives (cf. nouns) result in the perception of a more committed speaker.

3.3. General discussion

The present results show that typical and HFA speakers produce contrastive intonation by taking into account both their own and their listener's perspective. Expected differences between both development groups are not found. Only subtle differences show that HFA speakers atypically produce the adjective with more prominence when it is unfocused. Such a finding confirms, among others, [6]. This deviation can however not be explained by a strict version of an impaired ToM [4]. This theory would predict HFA speakers to show identical behaviour no matter whether the listener is the same or different. Our results show that HFA speakers only make errors when addressing a different listener. When addressing the same listener HFA speakers are similar to typical speakers.

Differences between the development groups are found for speech dynamicity and speaker commitment. Those features could also explain why HFA speakers have a tendency to produce contrastive intonation overall less clearly, as measured by prominence difference scores. Although this difference statistically is a trend, it can be the result of a monotonously sounding speaker. That is, a flat intonation predicts smaller differences between focused and unfocused words.

Contrastive intonation, in sum, is produced by taking into account both speaker and listener factors by typical and HFA speakers. However, subtle differences found can be explained on the basis of speaker development. With respect to those differences, the present study shows that it is useful to distinguish aspects of prosody that relate to function (focus marking) and those that relate to form (dynamicity, commitment). The difference between typical and HFA speakers in this study seems to be grounded in the latter rather than in the former aspects of prosody.

4. Acknowledgements

We thank Sonny de Nijs of Zintri Zorggroep for recruitment of participants and assistance during the production experiment, Marieke Hoetjes for help with the prominence ratings and three anonymous reviewers for their comments.

5. References

[1] Pechmann, T., "Überspezifizierung und Betonung in referentieller Kommunikation", Dissertation, Universität Mannheim, 1984.

[2] Krahmer, E., and Swerts, M., "On the alleged existence of contrastive accents. Speech Communication", 34:391-405, 2001.

[3] Kaland, C., Krahmer, E., and Swerts, M., "Contrastive intonation: Speaker- or listener-driven?", Paper presented at the 17th International Congress of Phonetic Sciences, Hong Kong, 2011.

[4] Baron-Cohen, S., "Theory of mind and autism: a review", International Review of Mental Retardation, 23:169-184, 2001.

[5] McCann, J., and Peppe, S., "Prosody in Autism Spectrum Disorders: A Critical Review", International J. of Language and Communication Disorders, 38(4):325-350, 2003.

[6] Peppé, S., McCann, J., Gibbon, F., O'Hare, A., and Rutherford, M., "Receptive and expressive prosodic ability in children with high-functioning autism", J. of Speech, Language, and Hearing Res., 50:1015–1028, 2007.

[7] Ladd, D. R., "Intonational phonology", [2nd ed.], Cambridge; New York: Cambridge University Press, 2008.

[8] Nadig, A., and Shaw, H., "Acoustic and Perceptual Measurement of Expressive Prosody in High-Functioning Autism: Increased Pitch Range and What it Means to Listeners", J. of Autism and Developmental Disorders, 42:499-511, 2011.

[9] Rutherford, M. D., Baron-Cohen, S., and Wheelwright, S., "Reading the Mind in the Voice: A Study with Normal Adults and Adults with Asperger Syndrome and High Functioning Autism", J. of Autism and Developmental Disorders, 32:189-194, 2002.

[10] Chevallier, C., Noveck, I., Happé, F., and Wilson, D., "What's in a voice? Prosody as a test case for the Theory of Mind account of autism", Neuropsychologia, 49(3):507-517, 2011.

[11] Baltaxe, C., "Use of contrastive stress in normal, aphasic, and autistic children", J. of Speech and Hearing Res., 27:97–105, 1984.

[12] McCaleb, P., and Prizant, B. M., "Encoding of new versus old information by autistic children", J. of Speech and Hearing Disorders, 50:230-240, 1985.

[13] Fine, J., Bartolucci, G., Ginsberg, G., and Szatmari, P., "The use of intonation to communicate in pervasive developmental disorders", J. of Child Psychology and Psychiatry, 32:771–782, 1991.

[14] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders", 4th ed., Washington: Author, 2000.

[15] Boersma, P., and Weenink, D., "Praat: doing phonetics by computer", v5.2.19, 2011.

[16] Veenker, T. J. G., "WWStim: A CGI script for presenting web-based questionnaires and experiments", v1.4.4, Utrecht, 2003.

[17] Von Benda, U., "Zur Auditiven Beurteilung der Intonation Autistischer Kinder: ein Hörexperiment", In D.-W. Allhoff [Ed.], Mündliche Kommunikation, Störungen und Therapie (Vol. 10). Frankfurt am Main: Scriptor, 1983.

[18] Simmons, J. Q., and Baltaxe, C. (1975). "Language patterns of adolescent autistics", J. of Autism and Developmental Disorders, 5:333-351, 1975.