INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Similar prosodic structure perceived differently in German and English

Heather Kember[1,2], Ann-Kathrin Grohe[3], Katharina Zahner[4], Bettina Braun[4], Andrea Weber[3], and Anne Cutler[1,2]

[1]The MARCS Institute, Western Sydney University, Australia;
[2]ARC Centre of Excellence for the Dynamics of Language, Australia;
[3]University of Tübingen, Germany;
[4]Konstanz University, Germany

h.kember@westernsydney.edu.au

Abstract

English and German have similar prosody, but their speakers realize some pitch falls (not rises) in subtly different ways. We here test for asymmetry in perception. An ABX discrimination task requiring F0 slope or duration judgements on isolated vowels revealed no cross-language difference in duration or F0 fall discrimination, but discrimination of rises (realized similarly in each language) was less accurate for English than for German listeners. This unexpected finding may reflect greater sensitivity to rising patterns by German listeners, or reduced sensitivity by English listeners as a result of extensive exposure to phrase-final rises ("uptalk") in their language.

Index Terms: prosody, speech perception, German, English

1. Introduction

English and German are closely related languages, and are similar in their phonology. This is especially true of prosodic structure [1]. At the word level, both languages have lexical stress, a preference for stress on word-initial syllables, and a strong tendency for vowels in unstressed syllables to be reduced. At the sentence level, both use pitch accents to indicate the relative importance of words within an utterance, with different pitch accent types being selected, depending on, for example, information structure. Statements and yes/no questions are similarly distinguished in the two languages; the statement *That was an elephant /Das war ein Elefant* would be uttered with a falling pitch movement, while a yes/no question such as *That was an elephant?/Das war ein Elefant?* would be realized in each case with a rising intonational contour [2, 3].

Despite the broad similarities, however, subtle differences have been discovered in the way prosodic elements such as pitch accents and final rises and falls are realized acoustically in English versus German. Speakers have different means of coping with situations that make the matching of prosodic to segmental structure non-trivial. For instance, if the utterance contains limited sonorant material (because consonants are unvoiced and vowels are short, for instance), then it can be hard to realize a given intonational contour on a designated syllable. Two strategies that can be used here are *truncation* and *compression*. Truncation involves the pitch slope staying the same but stopping when voiced material runs out, whereas compression implements a faster pitch change so that the full accent is realized, but over a shorter time span. English and German speakers make different choices in this situation.

Grabe [4] asked speakers of each language to produce comparable surnames in questions versus statements, i.e., in contexts requiring respectively rising and falling contours. The consonants in these surnames were all voiceless, and the vowels were chosen to provide progressively less voiced material (*Schiefer*, *Schief*, *Schiff* in German; *Sheafer*, *Sheaf*, and *Shift* in English). What she found was that for English, in both falling and rising contexts, rate of fundamental frequency (F0) change increased significantly with decreasing syllable duration. English speakers were thus compressing pitch movements when the segmental material was shorter. The two pitch accent types showed no significant differences in realization; English speakers used compression for both rises and falls. German speakers, however, showed this pattern of compression only in the case of rises, and tended to truncate falls. Thus pitch accent realization differed across the two languages, in that German speakers truncated falls but compressed rises, whereas English speakers compressed both falls and rises. (Grabe viewed this as a phonetic implementation effect rather than any difference in the underlying meaning associated with a fall or rise.)

Not only have language-specific differences been found in how truncation and compression are used, but there can also be differences within the same language, across varieties. The English stimuli used in [4] produced evidence of different strategies when they were spoken by users of four separate UK dialects; Cambridge English and Newcastle English were consistent with the results described above for Standard Southern British, but Belfast and Leeds English were not [5]. Alignment differences in acoustic realization of pitch accents have also been reported across dialects of German [6].

In the present study we consider pitch accent truncation and compression, and ask whether difference in how nuclear tunes are actually realized has consequences for native speakers' perception of rises and falls. If we base our expectations solely on [4], we would predict that German and English speakers will differ in perceptual discrimination of falls, as this is where production differences are seen. In that case we expect no cross-language differences in the discrimination of rises, given that in both languages the same compression strategy is applied in rise production.

On the other hand, perception will be affected by prosodic patterning across the whole language, including prosodic effects not in fact mandated by underlying phonology. Many varieties of English show use of rising terminal contours even in statements ("uptalk"; [7]), and this intonational pattern is certainly found in the language of our participant population, Australian English [8]. Such patterns have not been reported in Southern German varieties. Thus responses to rising accents in our study might also reflect this cross-language difference in the overall frequency of rising contours in speech.

http://dx.doi.org/10.21437/Interspeech.2017-544

2. Method

2.1. Participants

Twenty-four native speakers of German (M_{age} = 24.22 years, SD = 3.23, 18 females) and 24 native speakers of Australian English (M_{age} = 27.96 years, SD = 11.09, 18 females) participated. All German speakers grew up with Southern German dialects and were recruited in Tübingen and Konstanz in southern Germany. They were paid a small sum for their participation. English speakers were recruited in Sydney and reimbursed with either course credit or a small payment.

2.2. Stimuli

From tokens of each of the surnames *Schief* (German) and *Sheaf* (English) as used in [4], we extracted a single [i:] token to manipulate for our stimuli. The tokens were recorded by one female native speaker of Australian English and one female native speaker of Southern German, with similar f0. The English-speaking participants listened to the Australian English speaker stimuli and the German-speaking participants listened to the German speaker stimuli. Apart from the speaker difference, the stimuli were manipulated in the exact same way for both languages using Praat [9], to produce stimuli modeled on data of [4].

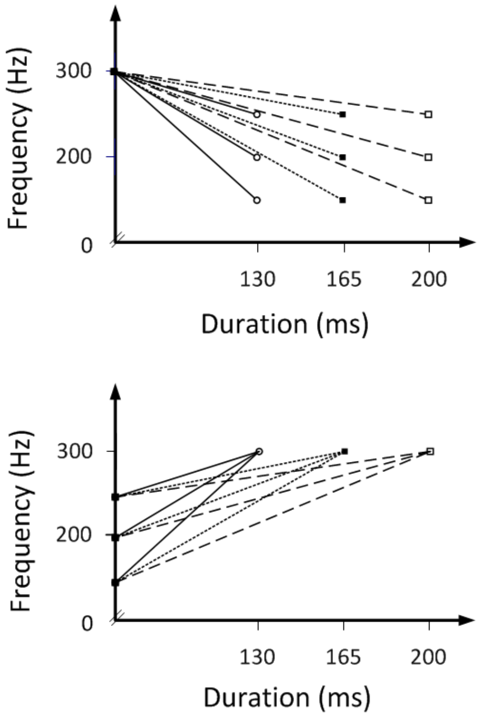

Figure 1: *Each of the steps on the perception continuum, with in the top panel all 9 falls (3 fall slopes x 3 durations) and in the bottom panel all 9 rises in (3 rise slopes x 3 durations).*

We created a three-step continuum of rises and falls (three rises and three falls), with three duration values for each rise and fall, thus giving nine realizations of each accent type. The values we chose were based on the values reported in [4] for the continuum of three surnames. Falling contours started at 300Hz and ended at either 150, 200, or 250Hz (f0 range of 12, 7 and 3st, respectively). For the rising contours it was the reverse, such that all contours ended at 300Hz and started at either 150, 200, or 250Hz. For each pitch contour, duration was in three steps: 130, 165, and 200ms. The total set thus consisted of 18 stimuli (3 durations x 3 slopes x 2 pitch contours). This is displayed in Figure 1.

2.3. Procedure

The existing equipment in each of the three testing locations differed in some minor respects; the experimental procedure however was identical across sites. Participants sat in front of a laptop computer and a button box; they wore headphones fitted to be comfortable; and sound volume was adjusted such that each participant judged the stimuli to be easily audible.

We used an ABX paradigm that was presented using Presentation software [10]. and paired stimuli such that pairs differed only on one dimension, for example two falling contours with the duration held constant, or two falls with the same starting and ending frequencies, but differing durations. With three values for each dimension, participants were thus making judgements about stimuli that were either adjacent or two steps removed on a continuum. Falling contours were only paired with other falling contours, and rising contours with other rising contours. This resulted in 72 stimuli pairings in total. For each combination of A and B, two trials were created: one in which X was identical to A, and one in which X was identical to B. Two versions of the experiment were created, with half of the total experimental trials in each version. Participants were randomly assigned to one of the experimental versions, and presentation of each version was balanced across language groups. Presentation order of items within each version was randomized.

Participants were informed that they would hear three successive sounds, with the third one being identical to either the first or the second sound. They were instructed to press the corresponding button on the button box to indicate whether they thought the third sound was identical to either the first or the second. Each trial began by displaying “new trial” on screen for one second, then a fixation cross for 500ms. Participants then heard the three sounds, with a one-second inter-sound interval. Instructions then appeared on screen to press the left button to indicate the first sound, or the right button to indicate the second sound. No time limit was given to respond. Participants were given three practice trials to ensure they were comfortable with the method prior to starting the experimental trials. The entire study (including a speech production task completed after the perceptual study, of which the results are not reported here) took approximately one hour.

3. Results

Results were analysed using binomial mixed effects regression models in R [11]. For each model we created a three-level independent variable, enabling a comparison of participants’ ability to perceive the difference between stimulus 1 and 2, stimulus 1 and 3, and stimulus 2 and 3. For the models of falling and rising contours, stimulus 1 had the smallest pitch excursion, stimulus 2 had the intermediate pitch excursion, and stimulus 3 had the largest pitch excursion. For the model of duration contrasts, stimulus 1, 2, and 3, refers to the short, medium, and long stimuli respectively.

3.1. Perception of falling contours

For the initial model, we entered language background (2 levels: English, German), stimulus pairing (3 levels), and duration (3 levels: short, medium, and long) as fixed factors, with participant entered as a random factor. There was no main effect of language background, or of duration, but there was a significant effect of stimulus pair, such that all participants were more accurate at differentiating between the falls that were further apart on the continuum (fall 1 and 3) compared to an adjacent pair fall (fall 1 and 2; β = 1.29, SE = .25, z=5.19, p<.001). There was no difference in discrimination ability between the two adjacent pair falling contours, β = .32, SE = .19, z = 1.67, p=.09. This is illustrated in Figure 2. By adding in interaction terms, we created two further models. Neither interaction (between duration and stimulus pairing, or between language background and stimulus pairing) was significant.

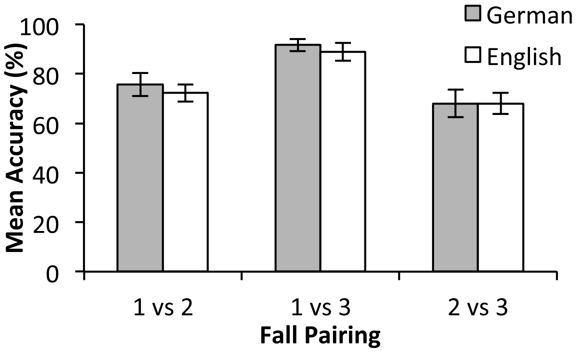

Figure 2: *Mean accuracy for discriminating between pairs of falling contours, by language group. Error bars represent standard error of the mean.*

3.2. Perception of rising contours

As above, the initial model included language background (2 levels: English, German), stimulus pairing (3 levels), and duration (3 levels: short, medium, and long) as fixed factors, with participant entered as a random factor. In contrast to the perception of falls, there was here a significant main effect of language background, with English speakers less accurate than German speakers, β = .86, SE = .23, z=3.67, p<.001.

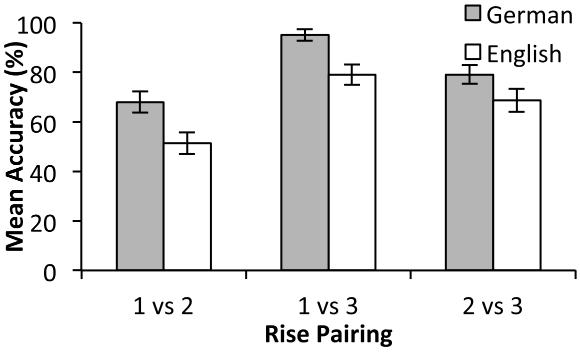

Figure 3: *Mean accuracy for discriminating between pairs of rising contours, by language group. Error bars represent standard error of the mean.*

There was again a significant effect of stimulus pairing, with stimuli 1 and 3 (more distant) more accurately differentiated than stimuli 1 and 2 (adjacent), β = 1.67, SE = .22, z=7.43, p<.001. Participants were here also significantly more accurate at differentiating stimuli 2 and 3 versus 1 and 2, β = .73, SE = .19, z = 3.80, p<.001. Again, differentiation was also more accurate when stimulus pairs were longer versus shorter, β = .64, SE = .21, z = 3.06, p<.001, again in contrast to the findings for falls. As before, we also entered interaction terms into the model. Neither the interaction between stimulus pairing and duration, nor the interaction between language background and stimulus pairing proved significant. Figure 3 displays mean accuracy for each of the rise stimulus pairings.

3.3. Perception of duration contrasts

As above, the initial model contained main effects of language background and stimulus pairing, but this time the third variable was f0-excursion (3 levels: 150Hz, 200Hz, 250Hz). Participant was entered as a random factor. There was no main effect of language background, β = .06, SE = .17, z = .39, p = .69, nor of slope, β = .02, SE = .14, z = 1.4, p=.89. There was a significant effect of stimulus pairing such that participants more accurately distinguished short and long compared to short and medium, β = 1.38, SE = .15, z = 2.04, p<.001, which were in turn more accurately distinguished than medium and long, β = .25, SE = .12, z = 2.04, p=.04. When the interaction between stimulus pairing and slope was entered into the model it failed to converge. We then entered the interaction between language background and stimulus pair into the model, however neither of these interaction effects was significant.

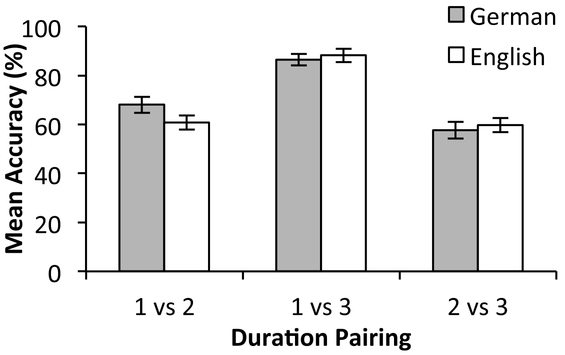

Figure 4: *Mean accuracy for discriminating between pairs of duration contrasts, by language background. Error bars represent standard error of the mean.*

4. Discussion

Our results showed that German and English listeners were equally accurate at discriminating between pairs of falling contours and duration contrasts. In general, both groups were more accurate when distinguishing between more distant pairs in comparison to pairs that were adjacent on a continuum. There was however, a cross-language difference in ability to differentiate pairs of rising contours. German listeners were significantly more accurate at discriminating between rising contour pairs compared to Australian English listeners.

In the production data of Grabe [4], English and German speakers differed in the use of truncation versus compression when realizing pitch accents with limited scope for voicing: both German and English speakers compressed falling contours (i.e. realized the full contour over the shortened syllable), but German speakers did not use this strategy when realizing rising contours. Instead they used truncation, i.e., the same slope but “cut off”. These data could suggest that our perception data might have shown cross-language differences in discrimination of falling contours, but equivalent performance for rising contours. However, this was not the pattern revealed in our results. Our German and English listeners discriminated falling contours and duration contrasts equally accurately, but rising contours, that are realized similarly in production in each language, were discriminated less accurately by English than by German listeners.

Several alternative explanations of this unexpected result present themselves, and each of them points the way to further empirical tests. Firstly, German- and English-speaking listeners may actually differ in their sensitivity to rises and falls due to attending to different aspects of pitch contours in natural discourse. The intonation of English and German tends to be analyzed in different terms [2,3]. While English analyses focus on the f0-movement after the tone that is associated with the stressed syllable (e.g., H*L, L*H), many German analyses focus on the f0-movement leading up to the starred tone (e.g., L+H*, H+L*). Recent perception data from German showed that listeners are sensitive to the manipulation of such onglide movements in rising prenuclear and nuclear accents [12, 13]. While no data from English are as yet available to be compared with this finding, it is possible that German listeners redeployed their sensitivity for rising onglides to the discrimination of the current stimulus pairs. In order to tease apart whether listeners are attending more to onglides and offglides versus to patterns of truncation and compression, one approach could be to replicate the present design with stimuli more closely resembling pitch accents as they are realized in natural conversation. Secondly, the ability to discriminate rising patterns could be affected by more global prosodic patterns in the listeners’ everyday experience. Consider the rising terminal contour (“uptalk”) known to be widely used in Australian English [7, 8]; this high rising terminal contour has no phonological meaning attached to it, but instead may serve pragmatic functions (such as holding the conversational turn). Thus it may in effect conflict with meanings associated with falls and rises. Australian English listeners might then have a reduced sensitivity to rising contours simply because they hear so many exemplars in everyday discourse, with no phonological meaning ever attached. One way to test this hypothesis, in turn, would be to replicate the study with a group of English speakers from a variety where uptalk is not present.

On the other hand, note also that Truckenbrodt [14] has argued that rising accents (albeit in non-terminal - in fact, prenuclear) position are particularly common in Southern German. If this were to be postulated as a source of the better performance of our German listener group in discriminating rises, it would obviously amount to a counter-argument to the potential role of uptalk as a source of our English listener group’s worse performance with the same task. Clearly, both cannot simultaneously be true!

Finally, although there is neither in our present results nor in the literature any evidence that production findings such as in [4] directly predict influences on perception, such a relationship would be potentially important and should certainly merit further exploration. In this context, we cannot rule out the possibility that the listeners in our study would as speakers use different truncation and compression strategies than those reported in the literature, e.g. in [4], simply because they speak Australian English and Southern German (whereas the productions measured by Grabe came from speakers of Northern German and Standard Southern British English). Recall that Grabe et al. [5] showed that varieties within British English differed in their use of truncation versus compression strategies in final pitch accent realization. In any event, replications of the study in [4] using the two speaker groups whose perception we have tested here would be an appropriate further test of this possibility.

5. Conclusion

In the present study we have demonstrated that even for closely related languages such as English and German, language background can modulate listeners' sensitivity to variation in pitch contours. Such variation is known to occur across less closely related pairs of language (for instance, Rathcke [15] compared perception of phrase-final pitch rises versus falls in German and Russian, and showed differences across her two listener groups as to which cues were most critical for making this distinction). Prior to the present study, however, no such comparisons had to our knowledge been undertaken for English versus German, or for other Germanic language pairs, or for other linguistic close relatives. A close linguistic relationship is no bar to the development of subtle variation in perceptual sensitivity!

6. Acknowledgements

Funding was provided by the ARC Centre of Excellence for the Dynamics of Languages (project ID: CE140100041) and by an Australia-Germany Joint Research Cooperation Scheme award to A. Cutler and A. Weber (for SpeechNet BaWü).

7. References

[1] H. van der Hulst, *Word prosodic systems in the languages of Europe*. Berlin: Walter de Gruyter, 1999.

[2] C. Féry, *German intonational patterns*. Tübingen: Walter de Gruyter, 1993.

[3] C. Gussenhoven, *The phonology of tone and intonation.* Cambridge: Cambridge University Press, 2004

[4] E. Grabe, "Pitch accent realization in English and German," *Journal of Phonetics*, vol. 26, no. 2, pp. 129-143, 1998.

[5] E. Grabe, et al., "Pitch accent realization in four varieties of British English," *Journal of Phonetics,* vol. 28, no. 2, pp. 161-185, 2000.

[6] M. Atterer, and D.R. Ladd, "On the phonetics and phonology of "segmental anchoring of F0: evidence from German," *Journal of Phonetics,* vol. 32, no. 2, pp. 177-197, 2004

[7] P. Warren, *Uptalk: The phenomenon of rising intonation*. Cambridge: Cambridge University Press, 2016.

[8] Fletcher, J., et al., "Intonational rises and dialog acts in the Australian English map task," *Language and Speech*, vol. 45, no. 3, pp. 229-253, 2002.

[9] P. Boersma, and D. Weenink. *Praat: doing phonetics by computer*. www.praat.org 2014 [cited 2014 2 January]; 5.3.62

[10] Version 18.0, Neurobehavioral Systems, Inc.: Berkeley, CA.

[11] R Core Team, R: A language and environment for statistical computing (Version 3.0. 2). *R Foundation for Statistical Computing,* Vienna, Austria, 2014.

[12] Ritter, S. and M. Grice, "The role of tonal onglides in German nuclear pitch accents," *Language and Speech*, vol. 58, no. 1, pp. 114-128, 2015.

[13] Braun, B., "Phonetics and phonology of thematic contrast in German," *Language and Speech,* vol. 49, no. 4, pp. 451-493, 2006.

[14] H. Truckenbrodt, "Upstep on edge tones and on nuclear accents," *Tones and tunes,* vol. 2, pp. 349-386, 2007.

[15] T. Rathcke, "On the neutralizing status of truncation in intonation: A perception study of boundary tones in German and Russian," *Journal of Phonetics,* vol. 41, no 3, pp. 172-185, 2013.