Special Issue: Mechanisms of regulation in speech, eds. Mücke, Hermes & Cho

Integrating the discreteness and continuity of intonational categories

Martine Grice *, Simon Ritter, Henrik Niemann, Timo B. Roettger

University of Cologne, Germany

1. Introduction

In Autosegmental–Metrical phonology (Ladd, 2008; Pierrehumbert, 1980), intonation contours are decomposed into sequences of postlexical tones associated with prosodic and metrical structures. These tones can be cumulative or delimitative in nature, corresponding to descriptive classes such as pitch accents and boundary tones, respectively (e.g. Beckman & Venditti, 2011). Here we are concerned with pitch accents, tones that co-occur – either singly or in combination – with metrical strong positions in the phrase, and that commonly lend prominence to the words they are on. In languages such as German and English, the inventory of pitch accents ranges from four to six, depending on the model (e.g. compare for German: Grice, Baumann, & Benzmüller, 2005; Peters, 2014, and Kügler et al., 2015; and for English: Beckman, Hirschberg, & Shattuck-Hufnagel, 2005; Grabe, Post, & Nolan, 2001, and Dilley & Heffner, 2013). It is true for all of these models, however, that whilst establishing the presence or absence of a pitch accent is relatively straightforward, distinguishing between different types of pitch accent is not.

In the above intonation models, what is taken to be a pitch accent category is characterised by a bundle of phonetic properties (Ladd, 2008) that serve a specific communicative function in a given context. To establish an inventory of pitch accents thus involves the analysis of a number of continuous phenomena into discrete categories, which are taken to be phonological representations. Transcribers are trained to pay attention to those phonetic properties that are deemed to be the main characteristics of each accent. This is also the case for the German GToBI transcription system used in this study (Grice et al., 2005, see also Grice & Baumann, 2002; Grice & Baumann, 2016). Thus, transcribers attend to gradual dimensions of the speech signal, and are trained to draw a line between pitch accent categories based on multiple gradual acoustic and auditory dimensions.

These dimensions typically involve f0 peak alignment, which is taken to be a major cue to pitch accent type in a number of languages (see Arvaniti, 2011; D’Imperio, 2011, and Prieto, 2011, for extensive overviews). Early work on English by Pierrehumbert and Steele (1989) showed that a distinction could be made between L+H* and H*, simply by adjusting the timing of the f0 peak (in a rising-falling-rising contour from medial to late) in relation to the segments. Dilley and Heffner (2013) showed that a distinction between H+L* and H* could

* Corresponding author at: University of Cologne, IFL-Phonetik, Herbert-Lewin-Str. 6, 50931 Koeln, Germany. Fax: +49 681 302 4694. E-mail address: martine.grice@uni-koeln.de (M. Grice).

Please cite this article in press as: Grice, M., et al. Integrating the discreteness and continuity of intonational categories. Journal of Phonetics (2017). http://dx.doi.org/10.1016/j.wocn.2017.03.003
also be made by adjusting the position of the f0 peak (from early to medial). Such a distinction between medial and early peaks was also found in German by Kohler (1987). Subsequent work has shown that the alignment of an f0 peak can interact with the shape of the f0 curve (Barnes, Veilleux, Brugos, & Shattuck-Hufnagel, 2012; Knight, 2008), and can be subject to considerable speaker specific variation (Niebuhr, D’Imperio, Gili Fivela, & Cangemi, 2011). Moreover, it can be strongly affected by the segmental makeup of the words (Barnes et al., 2012).

Target height, i.e. the height of the f0 target corresponding to the starred tone, is also a dimension that plays a role in pitch accent categorisation, although to a lesser extent than alignment (Ladd & Morton, 1997 for English, Kügler & Gollrad, 2015 for German). Target height is of particular importance when comparing H* to L+H*. The perceived height of an f0 target can also be affected by alignment, such that peak delay has been treated as a substitute parameter for a raised peak (Gussenhoven, 2004).

A further dimension that has been shown to play a role in the categorisation of pitch accents is the tonal onglide (Ritter & Grice, 2015). This is the f0 movement leading towards the f0 target on the accented syllable and thus takes into account not only the f0 on the accented syllable but also on the syllable before the accented syllable. This part of the f0 contour is generally accounted for phonologically in terms of a leading tone, such as the L in L+H*, or the first H in H+H*.

In German, as in English, different pitch accents typically express information structure (the division of sentences into focus and background) and information status (the degree of givenness of a referent in the current discourse): A pitch accent with a late (and high) f0 peak and a rising onglide (L+H*) generally signals contrastive focus and one with a medial peak and shallow rising onglide (H*) new information, while a pitch accent with an early f0 peak and a falling onglide (H+H*) tends to signal broad focus, or given information (cf. Baumann, Becker, Grice, & Mücke, 2007; Féry & Kügler, 2008; Grice et al., 2005; Kohler, 1991; Kohler, 2005; Ritter & Grice, 2015). A similar form-function mapping has been argued for English too (Pierrehumbert & Hirschberg, 1990). However, studies looking at speaker specific behaviour in detail have found that these statements do not hold for all speakers of a language, or even for one speaker all of the time (Cangemi, Krüger, & Grice, 2015; Grice, Baumann, & Jagdfeld, 2009; Peppé, Maxim, & Wells, 2000). As Cruttenden (1986) pointed out, there are usually alternatives to the more common, preferred intonation patterns, which at the time of writing his book were usually neglected. More recent work takes an approach that provides information as to the distribution of alternative realisations of a given function (Grabe, 2004 and Yoon, 2010 for English, Baumann, 2006; Grice et al., 2009, and Baumann, Röhr, & Grice, 2015 for German, see also Savino & Grice, 2011 and Cangemi & Grice, 2016 for Italian).

This paper is concerned with a corpus of read speech produced by five speakers in German. The corpus was designed to investigate the distinction between different types of focus, in terms of both supralaryngeal articulation and intonation. We shall first discuss the corpus with regard to categorical annotations of pitch accent types. It has already been shown for this corpus that there is no consistent relation between pitch accent categories and pragmatic function at this level (Mücke & Grice, 2014). In this paper we examine the relation between the phonetic parameters characterising pitch accent categories assigned by the transcribers, and how they are distributed across the different functions and speakers.

The main reason for exploring this particular corpus is that the productions of all five speakers have been subjected to perceptual testing in which it was shown that all five speakers are able to convey their pragmatic intentions to listeners in a comparable way (Cangemi et al., 2015; Krüger, 2009). Taking all listeners together, the intended focus structure was correctly identified on average between 64% and 67% of cases. This indicates that the different speakers were comparable in their adequacy to carry out the task, thus motivating further exploration of this dataset. We discuss the perception results in detail in Section 5.

In this exploration we do not a priori attempt to reject theoretically driven null hypotheses, but rather aim to analyse the data post hoc and to generate hypotheses for further study. In line with standards of reproducible research (Peng, 2011), the obtained data table and the R scripts produced for data processing and plotting are made available on GitHub: https://github.com/troettge/grice-et-al-2017-german-focus.

The goal of this paper is to undergo a detailed analysis of the f0 contour and how aspects of this contour (peak alignment, target height and tonal onglide) relate to the focus structures expressed and perceived. More specifically, we show that there is no one-to-one-mapping between the pitch accent type and the expressed focus structure. Even though speakers differ in their choice of a pitch accent type for an intended focus type, the intended focus type is almost equally well perceived. In other words, listeners perceive the focus type not exclusively on the basis of the given accent type but are sensitive to more fine-grained details such as the peak alignment, target height and onglide. These continuous measures show the same pattern across all speakers and can be linked to the perception of a focus type.

The paper is structured as follows: Section 2 presents the methods including the speech material and the analyses focusing on the quantitative evaluation of the pitch accent types produced by the speakers. While Section 3 examines the distribution of discrete pitch accent types as transcribed by the labellers, Section 4 examines the produced intonational patterns in terms of three continuous parameters (F0 peak alignment, tonal onglide, target height) and relates these to articulatory parameters. Section 5 takes a closer look at individual differences in the mapping between intended focus structure and perceptual recovery of these functions, showing that pitch accent type cannot solely account for the perceptual results. Section 6 explores the implications for intonation theory and calls for an integrated analysis of the discrete phonological pitch accents and the modulation of continuous phonetic parameters that characterise them.

2. Method

2.1. Speakers and recordings

In this paper we closely examine a dataset produced by five native speakers of Standard German (3 female, 2 male, mean
age = 28, age range = 22–37). Speakers F1, F3, and M2 were from the Franconian area, and F2 and M1 from the Western Low German area. Acoustic and articulatory recordings were made at the phonetics laboratory at the University of Cologne, employing a Digital Audio Tape recorder (TASCAM DA-P1) and a condenser microphone (AKG C420 headset) at a sampling rate of 44.1 kHz (16 bit), along with an Articulograph (AG100). Mücke and Grice (2014) reported on the articulatory data, touching only briefly on the pitch accent analysis, whereas this paper focuses on the intonation, both in terms of pitch accent types as well as three continuous parameters contributing towards the categorical analysis (peak alignment, target height and tonal onglide) and how they relate to the focus structures expressed.

2.2. Speech material

The speech material was designed such that the three target words (nonce names) Bieber [biːbɐ], Bahber [baːbɐ] and Bohber [boːbɐ] appeared in three different focus structures: either in broad, narrow or contrastive focus. All target words were disyllabic trochaic words. The different focus structures were elicited using question–answer pairs in which the answer was the target sentence. Examples for the context and elicited focus types are shown in Table 1.

Our dataset consisted of 310 tokens (3 target words × 3 focus structures × 7 repetitions × 5 speakers = 315 tokens; 5 tokens were excluded due to recording or processing issues). In contrast to the data in Mücke and Grice (2014) we excluded target words produced as part of the background (where the nuclear accent was at the beginning of the phrase, on Melanie), as none of them were produced with a pitch accent, therefore being no peak or target to measure.

2.3. Procedure

Subjects were presented with a question both auditorily and orthographically on a computer screen. Questions were presented via the computer’s built-in loudspeakers and were spoken by a professional male speaker.1 Subsequently, subjects were presented with the answer orthographically and instructed to read it aloud in a contextually appropriate manner. Each recording block was preceded by a training block of five question–answer pairs. The question–answer pairs in the test materials were pseudo-randomised, making sure that no two consecutive target sentences contained the same test word.

2.4. Analysis

Two trained transcribers independently placed GToBI labels using their auditory impressions and the f0 contour in PRAAT (Boersma & Weenink, 2016), using the GToBI training materials for orientation (Grice & Baumann, 2002; Grice et al., 2005, website for GToBI: http://www.gtobi.uni-koeln.de). These training materials make transcribers aware of the specific characteristics of each pitch accent type, as shown in Table 2.

In addition to the GToBI standard labels, individual targets for bitonal pitch accents were labelled separately (e.g. H+H* has a separate label for each H). Transcribers agreed in 80% of the cases (or 84% of the cases if the focus type “background” was included). We tested the interrater reliability by calculating Cohen’s kappa (Cohen, 1960) for pitch accent categorisation (H+H*, H*, L+H*) using the R package irr (Gamer, Lemon, Fellows, & Singh, 2012), revealing substantial agreement between transcribers (0.712). Apart from cases of uncertainty, where one labeller did not place a label, disagreement involved H* being confused with other accents: either with L+H* (49% of disagreements) or with H+H* (42% of disagreements). The cases where they did not agree were subject to further scrutiny, leading to a consensus transcription, which was used for further analysis.

Table 1

<table>
<thead>
<tr>
<th>Questions</th>
<th>Answers</th>
<th>test word /baːbɐ/ in:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was gibs Newes?</td>
<td>Melanie will Doktor Bahber treffen</td>
<td>broad focus</td>
</tr>
<tr>
<td>2. Wen will Melanie treffen?</td>
<td>Melanie will Doktor Bahber treffen</td>
<td>narrow focus</td>
</tr>
<tr>
<td>3. Will Melanie Dr. Werner treffen?</td>
<td>Melanie will Doktor Bahber treffen</td>
<td>contrastive focus</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Accent</th>
<th>Schematised f0</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>H+H*</td>
<td><img src="image" alt="Graph" /></td>
<td>• Fall from previous syllable onto accented syllable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Mid target on accented syllable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Accented syllable sounds mid</td>
</tr>
<tr>
<td>H*</td>
<td><img src="image" alt="Graph" /></td>
<td>• High accent with a shallow rise (sagging interpolation)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Medial f0 peak</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Accented syllable sounds high (not as high as L+H*)</td>
</tr>
<tr>
<td>L+H*</td>
<td><img src="image" alt="Graph" /></td>
<td>• Low tone before accented syllable and rise to accented syllable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The extent of the rise is great</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Late f0 peak</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Accented syllable sounds high</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Grice, M., et al. Integrating the discreteness and continuity of intonational categories. Journal of Phonetics (2017), http://dx.doi.org/10.1016/j.wocn.2017.03.003
For the quantitative evaluation of the contours, we calculated three phonetic parameters explicitly used for the categorisation of pitch accents (see Table 2): the alignment of the f0 peak, the tonal onglide (cf. Ritter & Grice, 2015), and the height of the f0 target on the accented syllable. These are explained in detail below and schematically depicted in Fig. 1.

**F0 peak alignment**

The peak alignment measure indicates the temporal alignment of the f0 peak in relation to the onset of the nuclear accented vowel (in milliseconds). The f0 peak can occur either within or after the accented syllable, as in the case of H* and L+H* accents (Fig. 1a), or before the accented syllable, as in the case of H+!H* (Fig. 1b). Positive alignment values correspond to f0 peaks occurring within or after the accented syllable and negative values correspond to f0 peaks occurring before the accented syllable.

**Tonal onglide**

The tonal onglide measure indicates whether the f0 is rising or falling towards the accented syllable. It is calculated by relating the f0 value in the nuclear accented syllable (at the location where the GToBI label for the starred tone was placed) to a reference point 30 ms before the start of the syllable (in semitones, cf. Ritter & Grice, 2015). The starred tone label was placed on the f0 peak in the case of rising pitch accents (H* and L+H*) and on a change of gradient in the contour or in the middle of the syllable if the f0 trajectory had no inflection point (Fig. 1b). A positive tonal onglide value indicates a rise up to the peak while a negative value indicates a fall to the tonal target on the accented syllable.

The reference point 30 ms before the accented syllable was chosen in order to obtain a point in the low central vowel of the last syllable of "Doktor" ['dɔktɐ], i.e. the syllable preceding the accented syllable of the target word. This point within the centralised vowel [ɐ] provided a consistent f0 measurement and a good approximation to the position of the acoustic realisation of a potential leading tone, if there was one.

**Target height**

The target height measure is operationalised as the difference in semitones between the tonal target on the accented syllable, corresponding to the starred tone, and a following low reference point towards the end of the phrase, corresponding to a low boundary tone. In cases of a low boundary tone, the sentence-final low point is generally assumed to be stable for a speaker across different contexts (cf. Pierrehumbert, 1980, and Liberman and Pierrehumbert, 1984 for English, and Féry & Kügler, 2008; Kügler & Féry, 2016, and Grabe,
1998 for German). Since transcribers, however, faced problems when determining the position at which to place the low boundary tone label, due to microprosody and creaky voice, we automatically labelled a constant reference point 210 ms after the end of the target word. See Appendix 1 for information as to the constancy in F0 values of this reference point for each speaker.

3. Results in terms of pitch accent types

Although the results reported by Mücke and Grice (2014) concentrated on supralaryngeal kinematics, they also looked at transcribed pitch accent placement and type. In all three focus conditions (broad, narrow, contrastive) there was always a pitch accent but no one-to-one mapping between the focus type and the transcribed accent type. There were, however, probabilistic preferences.

Fig. 2 shows the distributions for each focus condition pooled over speakers. Here the most common pitch accent type in the broad focus condition was the early peak pitch accent, H+!H* (66% of cases), exemplified in Fig. 3a. This accent type has an “early peak” (H) before the target word, and a downstepped target (!H*) on the accented syllable. The most common pitch accent in the contrastive focus condition was L+H* (79%), exemplified in Fig. 3c, with a peak late in the accented syllable.

Although there is a predominance of H+!H* for broad focus and L+H* for contrastive focus, the percentages displayed in Fig. 2 also reveal that both of these focus conditions had alternative realisations. These were often H*, see the example in Fig. 3b (middle panel), although there were a few cases of L+H* in the broad focus condition. The narrow focus condition had a slightly more even distribution of accent types: 53% for L+H*, 35% for H* and 12% for H+!H*. These results show that there was no condition with an exclusive pitch accent realisation, nor one particular pitch accent used exclusively in one of the conditions. This supports the observation made in Grice et al. (2009) that there is considerable variation within and across speakers with regard to accent choice when expressing pragmatic functions of this type. The only form-function generalisation we can make for this dataset is that H+IH* never occurs in the contrastive condition (see also Kügler & Gollrad, 2015 for similar results).

Looking at individual speakers’ productions (see Fig. 4), it becomes clear that not all speakers make use of each of the three pitch accent types, accounting for some of the variation in the distributions. One speaker (F3) almost exclusively uses only one category of pitch accent (H*) across all three focus types while two speakers (F1, M1) hardly use this pitch accent at all.

All of the distributions relating pitch accent type to focus condition rely on a discrete decision made by the transcribers to categorise what they observe as one pitch accent type over another. In what follows, we take a step back from the discrete analysis into pitch accent categories and examine the global contours and the continuous parameters that are taken to correspond to the characteristics of each pitch accent.

4. Results in terms of continuous parameters

4.1. Global contours

Looking at the superimposed f0 contours in Fig. 5, the three focus conditions show a great deal of speaker-specific variation. Some speakers make a clear distinction between at least two conditions. Take for example speaker M1: The global
Fig. 3. Representative waveforms and f0 trajectories for three pitch accent types (a) H+H*, top panel; (b) H*, middle panel; (c) L+H*, bottom panel, all produced by speaker F2.
contours he produces for signalling broad focus (a hat pattern) are clearly distinct from those signalling narrow or contrastive focus. This difference corresponds to the distribution in nuclear pitch accents as analysed by the transcribers (shown in Fig. 4), where for this speaker, broad focus was produced consistently with H+!H* and narrow and contrastive focus with L+H*.

Speakers F1 and F2 appear to distinguish between broad and contrastive focus (a hat pattern for the former, a low prenuclear stretch in the latter) but produce more variation in the contours for narrow focus than in the other conditions. On the other hand, speakers F3 and M2 use what appears to be the same global contour to mark all three focus conditions (broad, narrow and contrastive).

In the next section we explore whether measuring selected parameters can reveal quantifiable differences that might be missed by the categorical pitch accent analysis and which can only be impressionistically observed in global contours.

4.2 Continuous f0 parameters

A closer look at the three parameters, f0 peak alignment, tonal onglide and the f0 target height on the accented word explores how individual speakers modify each parameter, and the relation of this modification to the transcribed categories. In the following we will descriptively explore the three continuous parameters using violin plots, which, based on kernel density estimations, show the probability density of data points at different values in a similar way to a histogram. The estimated likelihood of a value is proportional to its “bulginess”. In addition, black dots and dashed lines indicate mean values of distributions for better comparison across focus types within and across speakers. Due to the multimodal nature of distributions, the mean values provided should, however, be treated with caution. Here, they merely serve the purpose of visually guiding the reader to notice differences in central tendencies.

**F0 peak alignment**

Fig. 6 displays the results for peak alignment, i.e. the alignment of the f0 peak relative to the vowel onset.2

We can observe that speakers F1, F2, and M1 make a clear distinction in the alignment of the peak, with broad focus almost exclusively having peaks well before the beginning of the accented syllable (negative values, corresponding to H+!H*), and narrow and contrastive focus almost exclusively having peaks late in this syllable (positive values corresponding to H* and L+H*). For these speakers, values for narrow focus fall between these two but are highly variable, both within and across speakers, indicated by the large spread of the narrow focus distribution. F1 and M1 do not appear to differentiate clearly between narrow and contrastive focus, but each speaker’s distributions, albeit overlapping heavily, are different across the two conditions. Both speakers exhibit later peak alignment in contrastive focus compared to narrow focus. F1 and F2 clearly show multimodal distributions of alignment values in the narrow focus condition, with some values negative and others positive, reflecting the variation in the assigned pitch accent transcriptions shown in Fig. 4. For both speakers

---

2 Alignment is presented here in absolute terms. Relative alignment is presented in Appendix 2.
pitch accents in the narrow focus condition consisted of both early peaks (H+H*) and medial and/or late peaks (H*/L+H*).

For speakers F3 and M2, the difference in alignment is subtler (barely detectable for M2). Speaker F3’s contrastive focus peaks are somewhat earlier than the average peak alignment values for the other speakers in this condition, contributing towards the transcription of H* rather than L+H*. Across the three conditions, there are only subtle differences in alignment, with broad focus showing the largest spread in distribution, with peaks aligned with either the accented vowel or before it. Speaker M2 shows similar distributions across focus types, however, the weight of data points slightly shifts, with narrow focus further to the right than broad focus, and contrastive focus slightly more to the right than narrow focus. Comparing the sequence of dashed lines indicating the means of the distributions, all speakers show the same relative mapping of peak alignment and focus types.

Tonal onglide

Similar to peak alignment, the onglide results in Fig. 7 show a clear differentiation between broad focus and the other two conditions for the same three speakers (F1, F2 and M1). The broad focus condition exhibits negative values reflecting the falling nature of the early peak accent. While F1 shows a large spread of distributions, F2 and M1 show a clearer separation at least for broad and contrastive focus. F2, again, shows a clearly separated distribution for narrow focus as compared to contrastive focus, while the distributions for F1 and M1 in these two conditions overlap heavily. Regardless of this overlap, the weighting of distributions nevertheless consistently tends towards larger onglide values for contrastive focus, resulting in overall higher mean values.

The accents of speakers F3 and M2 almost exclusively have rising onglides, but they exhibit a subtler differentiation within the rising accents that mirror the global trends: greater rising onglides in narrow focus than in broad focus and even greater rises for contrastive focus. M2 shows substantial overlap across the three focus conditions. However, distribution weights clearly follow the same global trend. Again the subtle differences in the productions of these two speakers mirror neatly the rather categorical patterns found for the other three speakers.

Fig. 5. Time-normalised, superimposed f0 contours for each intended focus type (from left to right: broad focus, narrow focus and contrastive focus, displayed for each speaker separately.
Comparing the sequence of dashed lines indicating the means of the distributions, it is clear that all speakers show the same relative mapping of onglide values and focus types. As with the alignment distributions, the broad focus line is more to the left than the narrow focus line, which, in turn, is more to the left than the contrastive focus line.

**Target height**

Finally, looking at the f0 height of the target on the accented syllable (see Fig. 8), again the same pattern emerges. Distributions for this measurement overlap heavily as the spread is large. Regardless of whether a falling or a rising accent is used, f0 is higher in narrow focus than in broad focus and even higher in contrastive focus. This is true for speakers using contours that were transcribed as categorically different pitch accents, such as (F1, F2 and M1) as well as for speakers using contours categorised as having the same pitch accent types (F3 and M2). Again, speakers F1, F2 and M1 show distributions with no or minimal overlap for broad and contrastive focus. For two of the speakers (F1 and F2), narrow focus is in between, overlapping with both of the other conditions, whereas for M1 narrow focus and contrastive focus are highly overlapping. For the subtler speakers, F3 and M2, all three focus types overlap substantially. However, again, the weight of distributions shifts systematically from broad through narrow to contrastive focus. Except for narrow vs. contrastive focus of speaker M1, the mean differences of target height values go in the same direction for all five speakers. The broad focus line is more to the left than the narrow focus line, which, again, is more to the left than the contrastive focus line.

It is interesting to note that transcribers disagreed most when labelling speaker M2, especially the broad focus and narrow focus conditions, but not in the contrastive focus condition in which the L+H* label was transcribed by both labellers almost all of the time. Recall that in the contrastive condition this speaker produced slightly higher targets with a greater onglide (despite there being barely any difference in alignment), possibly making the realisations easier to categorise.

### 4.3. Discussion of f0 parameters and their relation to articulation

We have shown that speakers vary greatly in the extent to which they use intonational parameters to distinguish between the different focus conditions. For instance, some speakers...
make a clear difference between broad and narrow focus or broad and contrastive focus, whereas others make rather subtle differences. The question arises as to whether the intonationally more subtle speakers make more extensive use of other parameters instead. To address this question we examined individual productions in terms of the two most informative kinematic parameters found in Mücke and Grice (2014), i.e. the duration and displacement of the opening gesture. Specifically, the duration of the opening gesture refers to the time between the maximum lip closure in the word-initial [b] and the maximum lip opening during the following vowel in the target word ([baː], [boː], [biː]). The displacement refers to the interlip distance, i.e. the spatial distance between lower and upper lips. More specifically, the displacement of the opening gesture refers to the lip distance difference between the maximum closure and the maximum opening corresponding to the production of the word-initial [b] and the following vowel, respectively.

For comparison with the intonation results, violin plots are presented in Fig. 9 (duration of opening gesture) and Fig. 10 (displacement of opening gesture). Looking at these plots, it is evident that there is considerable overlap across the three focus types for all five speakers for both articulatory parameters. Mücke and Grice (2014) paid particular attention to the differences in these parameters when comparing broad and contrastive focus. They found that articulatory expansion (longer and larger lip opening gestures) not only depended on whether there was an accent or not, but was also affected by other factors, such as degree of prominence (as a function of the focus structure). These differences can be observed when comparing the means (dots and dashed lines) and distributions for broad and contrastive focus in the violin plots, the latter always tending towards being longer (greater duration of the opening gesture) and larger (greater displacement of the opening gesture).

Even when comparing broad and narrow focus, we can observe the same trend across all speakers. That is, all speakers tend to produce longer and larger opening gestures in narrow focus than in broad focus. Likewise, all speakers tend to produce longer and larger opening gestures in contrastive focus as compared to narrow focus. Overall, speakers use
Looking at individual speakers, F2 makes a clear distinction between the three focus types in the intonation, whereas speaker M2 makes very subtle adjustments (see Figs. 6–8). However, when looking at the articulation, both speakers have a comparable separation of the three focus types. Moreover, F3, who also made fairly subtle adjustments in the intonational parameters, tends to distinguish between the focus types in the duration of the opening gesture, although to a smaller extent in displacement. M1, who did not clearly separate narrow and contrastive focus intonationally (especially in terms of alignment and target height, Figs. 6 and 8 respectively), tended to differentiate these in terms of displacement. Thus, the distributions presented in this section suggest that the high variability in the realisation of intonational parameters in this dataset is neither mirrored nor compensated for by articulatory parameters.

In the following section we return to the pitch accent distributions for each speaker and relate them to how listeners perceived intended focus conditions. It will become apparent that pitch accent type alone cannot account for the perceptual results, and that thus some of the continuous parameters examined above play a role in the signalling of these pragmatic functions.

5. Perception of focus type and pitch accent categories

Recall that it has been reported for this dataset that all five speakers are able to convey their pragmatic intentions to listeners in a comparable way. Averaging over all listeners, the intended focus structure was correctly identified to a similar extent for each speaker. However, beyond overall identification performances, the following paragraphs will provide more detailed information as to the pragmatic meanings confused.

Fig. 8. Violin plots of target height values (in semitones) as a function of focus type and speaker. Mean values are indicated by black dots and dashed lines for comparison across focus types. Target height is operationalized in relation to a low reference point for each speaker and condition separately.
Fig. 11 shows matrices depicting intended focus type by speaker on the y-axis plotted against either the transcribed pitch accent category (left column) or the perceived focus type by all listeners from the perception experiment (right column). The darker the shade, the greater the match between intended focus type and pitch accent, or intended focus type and perceived focus type respectively. Pitting pitch accent distributions against the perceptual confusion of categories allows us to evaluate how informative the transcribed pitch accent type really is for listeners’ assessment of communicative intentions.

In an ideal situation in which pitch accent and focus type would map on to each other in a one-to-one fashion, we would expect a black diagonal from top left to bottom right. As is described in Section 3, however, pitch accent distribution is highly speaker specific. This is particularly apparent for speaker F3, who almost exclusively produces H’ accents for all focus types, or speaker M1, who produced H+H’ accents for broad focus and L+H’ accents for both narrow and contrastive focus.

Despite the high degree of variability in the mapping between intended focus type and pitch accent type, the perceptual confusions look strikingly similar. As might be expected, there is frequent confusion between broad and narrow focus and between narrow and contrastive focus, whereas broad and contrastive focus are seldom confused with each other. The mismatch between the pitch accent types used and the perceptual results is particularly evident when looking at speaker F3 where confusion patterns are comparable to those for other speakers, despite the fact that this speaker only produces one pitch accent pattern for all intended focus types.

We conclude that, in our sample, discrete pitch accent categories are only a poor approximation for how listeners comprehend intended pragmatic meanings. The confusion matrices suggest that irrespective of the assigned pitch accent type, listeners show similar patterns of confusion, an observation that is very much in line with the idea that listeners are sensitive to continuous acoustic parameters in recognising speakers’ intentions.

However, recall that speakers made similar use of the continuous parameters investigated, although they differed substantially with regard to the extent to which they modulated these parameters: Some speakers had distinct distributions...
while others had highly overlapping distributions. In light of the perceptual confusion matrices, it appears that listeners are able to tune into the speaker-specific mapping of continuous phonetic parameters and intended focus types.

6. General discussion

To summarise the results, looking at the distributions of pitch accent types as categorised by the transcribers, it appears that certain focus types are typically expressed by certain pitch accents. However, there is variation, both within and across speakers, with no one-to-one mapping between focus structure and pitch accent type.

Looking at the means of the phonetic parameters analysed, all five speakers have strikingly similar tendencies, although for some speakers the differences between realisations for each focus type are greater than for others. For some speakers, these parameters are large enough (either singly or in combination) to correspond to a discrete shift in the choice by the transcribers of one category over another; in other cases, the change appears to stay within the limits of one category. In these cases, different focus types are expressed by instances of the same pitch accent type. Most importantly, regardless of the mapping onto pitch accent categories, all speakers show the same relative pattern: All speakers exhibit a later peak alignment, a greater tonal onglide and a higher target for contrastive focus than for narrow focus, and likewise for narrow focus in relation to broad focus, regardless of whether they use different pitch accent types or not.

A purely categorical (pitch accent-based) account would miss the continuous differences across and within speakers. Crucially, such an approach would also miss the similarities in the expression of the different focus conditions. By looking at the parameters behind the phonological categories and by keeping in mind which categories they are involved in distinguishing, we can see that all speakers are modulating the continuous parameters in the same way, even if they do so to different extents. We can thus conclude that all five speakers essentially use the same relative system.

Our data indicate that it is not possible to provide a clear-cut categorisation of the intonation contour as relating to a particular function. Nevertheless, aggregating continuous
phenomena into descriptive bins, such as the GToBI pitch accents, allows us to examine distributions indicating relative frequencies of such categories. Here we see that not only are there alternative realisations but also general preferences for certain contours to express a given function. By looking at continuous parameters, we can supplement this information, showing that some speakers manipulate parameters in such a subtle way that they are not reflected in the categorisation of the analyst (the transcriber). The preferred categories for each function (the ones that are most frequent across speakers)

Fig. 11. Matrices for production (left) and perception (right). Production matrices show proportions of coded pitch accent by intended focus; perception matrices show rated focus type by intended focus type.
involve parameters that are at the extreme end of the scale for that category.

The results discussed above, are difficult to reconcile with traditional linguistic descriptions that focus entirely on abstract phonological categories, not least because these categories are by definition devoid of gradient phonetic information (Féry, 1993; Grabe, 1998; Grice et al., 2005; Uhmann, 1991). This has both practical and conceptual implications for phonological accounts of intonation.

In cases of well-described languages such as German, an intonational system, i.e., an inventory of contrastive intonational events represented as symbols (H+!H* vs. L+H*), has already been established. These symbols are used by transcribers to annotate corpora such as the present one. Even though transcribers are guided by a number of different parameters, they usually only transcribe the result of their categorisation. If only these rigid symbolic representations are considered when analysing and comparing intonational phenomena, we falsely assume that certain speakers do not express certain pragmatic functions intonationally. As described above, some speakers used the same pitch accent type to express all focus conditions. This conclusion, however, misses important generalisations that can be observed through inspection of continuous parameter values.

In cases of language documentation, in which a system has to be established bottom-up, these generalisations can easily be missed and might even lead to an ill-defined inventory of descriptive categories. As the present study has shown, looking at both a categorical assessment in terms of distributions of symbolic representations, as well as continuous modulations of certain acoustic parameters within these categories, allows a more general assessment of an intonation system. Thus, the transcription of pitch accents would benefit from a combined approach, complementing the abstract symbolic representations with the quantification of continuous parameters that define these symbolic representations. The extraction of a number of parameters can be carried out automatically if each target point is transcribed as a separate label (e.g. one for L and one for H* in L+H*, a practice that has been suggested for the new consensus German transcription system (Kügler et al., 2015)). This is in line with recent arguments made by Cole and Shattuck-Hufnagel (2016) in favour of a combined approach to the transcription of prosody in general.

Fig. A1. F0 values at the reference point across focus types for each speaker separately.
Apart from the practical implications for language documentation, these findings have important consequences for our conceptualisation of the knowledge a speaker has to have to use intonation. Generally, this issue lies at the threshold of the long-debated relation between phonology and phonetics. Traditionally, it has been assumed that grammatical computation is based on discrete categories (e.g. H\* vs. L+H\*). These categories are manifested in the continuous substance of the speaking event, the language’s phonetics (e.g. the actual alignment of the f0 peak). Because of this disparity, phonology and phonetics are assumed to be fundamentally separate but related through a process of translation – or mapping – from discrete symbols to continuous properties in the phonetic signal. This is the view behind most prominent work on language in particular, and, in fact, on cognitive science in general (Fodor, 1975; Fodor & Pylyshyn, 1981; Harnad, 1990; Haugeland, 1985; Newell & Simon, 1976).

Our results provide evidence from intonation in favour of a view that maintains that what we refer to as “phonology” and “phonetics” are two sides of the same coin, best understood as a single system. Such a system could be described using a formal language able to express both discrete and continuous phenomena, in which the key constructs are not symbol strings (representations) and algorithms for their manipulation (discrete computation), but rather laws stated in the form of differential equations. This approach has been successfully used to account for various aspects of cognition (Haken, 1977; Kelso, 1995; Port & Van Gelder, 1995). In phonology this approach was first used within the framework of Articulatory Phonology to represent coordination patterns within the speech apparatus (among others: Browman & Goldstein, 1986; Fowler, Rubin, Remez, & Turvey, 1980; Saltzman & Munhall, 1989). Gafos and Benus (2006) have applied a similar approach to capture quantitative and qualitative aspects of final devoicing in German and vowel harmony in Hungarian.

Further research is needed before we can model quantitative and qualitative aspects of intonation and its interaction with grammatical and non-grammatical aspects of language. Although this is beyond the scope of the present paper, the data presented provide an indication that such a dynamical approach might adequately account for the patterns of intonation we have described so far.

Fig. A2. Alignment in relation to the duration of the syllable.
7. Conclusion

This paper has explored the mapping of pragmatic functions onto intonation in terms of the discrete choice of pitch accent type made by the transcriber as well as continuous phonetic parameters that contribute towards the pitch accent categorisation. Although distributions indicating relative frequencies of intonational categories (accent types) expressing each function can be shown to provide a first approximation of the variation found, an important generalisation is missed, namely that there is a parallel between quantitative and qualitative effects: Some speakers’ productions lead to a categorical distinction (reflected in the analysis as a different pitch accent category), others being more subtle (resulting in no difference in the assigned category). What is particularly striking is that regardless of the mapping onto proposed categories, all speakers show the same relative pattern.

Focusing on the similarities (rather than the differences in intonational categories used) allows us to provide a plausible account for the fact that all five speakers were perceived at a similar level of accuracy: It appears that listeners are able to “tune in” to the productions of each speaker. This would simply calibrating their perception in terms of the degree to which each parameter is modulated, rather than learning new parameters for each speaker.

Moreover, these results indicate the need for a model of intonation that treats continuous modulation of individual parameters as contributing towards the discrete interpretation in terms of intonational categories.

Acknowledgements

This project was funded by the projects DASS (Degrees of Activation and Focus-Background Structure in Spontaneous Speech) and IndiBe (Individual Behaviour in the Encoding and Decoding of prominence) awarded by the German Research Foundation (GR 1610/5-2 and SFB 1252/1) to Martine Grice.

References


Please cite this article in press as: Grice, M., et al. Integrating the discreteness and continuity of intonational categories. Journal of Phonetics (2017), http://dx.doi.org/10.1016/j.wocn.2017.03.003