Dynamic speech adaptation to unreliable cues during intonational processing

Timo B. Roettger (timo.roettger@northwestern.edu)
[1]Northwestern University, Department of Linguistics
2016 Sheridan Rd Evanston, IL 60208 USA
[2]University of Cologne, I*f*L Phonetics
Herbert-Lewin-Strasse 6, 50931 Köln, Germany

Michael Franke (mchfranke@gmail.com)
Universität Tübingen, Seminar für Sprachwissenschaften
Wilhelmstr. 19–23, 72074 Tübingen, Germany

Abstract

Human behavior is often remarkably flexible, showing the ability to quickly adapt to the statistical peculiarities of a particular local context. When it comes to language, previous work has shown that listeners' anticipatory interpretations of intonational cues are adapted dynamically when cues are observed to be stochastically unreliable. This paper reports novel empirical data from manual response dynamics (mouse-tracking) on how listeners adapt their predictive interpretation when some intonational cues are occasionally unreliable while others are consistently reliable. A model of rational belief dynamics predicts that listeners adapt differently to different unreliable intonational cues, as a function of their initial evidential strength. These predictions are borne out by our data.

Keywords: intonation; mouse-tracking; prosody; rational predictive processing; speech adaptation

Introduction

Variable environments require the ability to quickly adapt expectations and behavior. Language is no exception. Indeed, language users have been shown repeatedly to adapt readily to their immediate local context in syntax (e.g. Fine & Jaeger, 2013; Jaeger & Snider, 2013), pragmatics (e.g. Grodner & Sedivy, 2011; Yildirim, Degen, Tanenhaus, & Jaeger, 2016), and speech (e.g. Norris, McQueen, & Cutler, 2003; Kleinschmidt & Jaeger, 2015). While previous work looking at speech adaptation has mainly focused on local phenomena within small temporal windows (e.g. adaptation to segments), there is only little work on adaptation of speech patterns across larger domains. Here, we focus on listeners' ability to adapt to the selective and partial reliability of intonational cues, and ask whether observed adaptations are consistent with a model of rational belief dynamics.

Intonation plays an integral role in comprehending spoken language. In English or German, for instance, the position and form of a pitch accent can signal a referent as discourse-new or contrastive (Ladd, 2008). Deviating from a traditional categorical view (e.g. Pierrehumbert & Hirschberg, 1990), recent work identifies intonational form-function mappings as highly variable and probabilistic (e.g. Grice, Ritter, Niemann, & Roettger, 2017; Roettger, 2017). Comprehenders can nevertheless rapidly process intonational cues to anticipate a likely speaker-intended referent even before encountering disambiguating lexical material (e.g. Dahan, Tanenhaus, & Chambers, 2002; Weber, Braun, & Crocker, 2006; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014a).

Listeners also adapt their anticipatory cue interpretation based on experimental pre-exposure to either reliable or unreliable input (Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014b). Unfortunately, pre-exposure manipulation of cue reliability does not allow inferences about the temporal dynamics of listener adaptation during exposure. Roettger and Franke (under review) therefore investigated the development of listener interpretation behavior over the course of the experiment when listeners are exposed to either occasionally unreliable or, in a different group, consistently reliable intonational cues. For reliable input, listeners quickly learned to predictively exploit the *absence* of an early pitch accent, while exploiting the *presence* of this cue right from the start (see Materials section for details). Unreliable exposure occasionally featured unnatural uses of all relevant intonational cues. This inhibited anticipatory interpretations mainly for the *presence* of an early pitch accent, but did not strongly affect the condition with an absent pitch accent.

Roettger and Franke (under review) argue that these results are compatible with the assumption that comprehenders expect reliable intonational information initially and gradually adapt these expectations rationally under reliable or unreliable input. A Bayesian model of the *evidential strength* of intonational cues (to be introduced presently) predicts that the stronger (weaker) a cue, the more (less) it will be affected by learning that it is unreliable, and the less (more) it will be affected by learning that it is reliable. This model also predicts that listeners should adapt differently to scenarios where only one cue is learned to be unreliable, while the other is reliable. We here report on an experiment, extending that of Roettger and Franke (under review), designed to test these predictions.

Rational Predictive Processing

Bayesian comprehenders derive their rational predictive interpretation from differences in the likelihood with which they expect speakers to produce particular intonational contours to signal a certain discourse status of a referent. By Bayes rule, the *posterior odds* in favor of referent r_1 over r_2 after observing a (possibly partial) utterance u are calculated as the product of the *likelihood ratio* (how likely a speaker produces u

for r_i) and the *prior odds* (how likely a speaker refers to r_i):

$$\underbrace{\frac{P(r_1 \mid u)}{P(r_2 \mid u)}}_{\text{posterior odds}} = \underbrace{\frac{P(u \mid r_1)}{P(u \mid r_2)}}_{\text{likelihood ratio}} \underbrace{\frac{P(r_1)}{P(r_2)}}_{\text{prior odds}} \quad (1)$$

If utterance u with its specific intonational contour is more likely to be produced for r_1 than for r_2, an observation of u shifts the listener's beliefs towards r_1 and away from r_2. Observing u would therefore be *observational evidence* in favor of r_1 relative to r_2 (Jaynes, 2003). The likelihood ratio therefore quantifies the *evidential strength* of a cue u.

A direct experimental measure of comprehenders' dynamically evolving posterior odds between two candidate interpretations can be obtained from mouse-movements in a forced-choice decision task. Roettger and Stoeber (2017) and Roettger and Franke (under review) show that listeners integrate intonational information early on and move their mouse towards a likely target referent even before they have processed disambiguating lexical information. This is in line with numerous experiments demonstrating that the continuous uptake of sensory input and dynamic competition between simultaneously active representations is reflected in subjects' hand or finger movements (e.g. Magnuson, 2005; Spivey, Grosjean, & Knoblich, 2005; Freeman & Ambady, 2010) and falls in line with recent papers using mouse tracking to investigate the processing of pragmatic inferences (e.g. Tomlinson, Gotzner, & Bott, 2017).

Concrete model predictions for a mouse-tracking experiment in which some intonational cues are unreliable while others are reliable are spelled out below, after the design and materials have been introduced in more detail.

Experiment

The following experiment was preregistered on the 27th of November 2017, prior to data collection. The preregistration file can be retrieved with all materials, data, and analysis scripts from `https://osf.io/49q2r/`.

Participants and Procedure

Sixty native German speakers participated, all with self-reported normal or corrected-to-normal vision and normal hearing (21 male, 39 female, mean age = 24.4 (SD = 3.4)).

Subjects were seated in front of a Mac mini 2.5 GHz Intel Core i5. They controlled the experiment via a Logitech B100 corded USB Mouse. Cursor acceleration was linearized and cursor speed was slowed down (to 1400 sensitivity) using the CursorSense© application (version 1.32). Slowing down the cursor ensured that motor behavior was recorded in a smooth trajectory as the acoustic signal unfolded.

Subjects learned about a 'wuggy', a fantasy creature which picks up objects. There were 12 objects to pick up (bee, chicken, diaper, fork, marble, pants, pear, rose, saw, scale, vase, violin), all with German grammatical gender feminine.

Each trial exposed subjects to a context screen, shown for 2500 ms and providing a specific discourse context. Participants heard either a *topic question* like (1), which introduced a referent as discourse-given, or the *neutral question* (2):

(1) Hat der Wuggy dann die Geige aufgesammelt?
Did the wuggy then pick up the violin?

(2) Was ist passiert? What happened?

Next, participants saw a response screen with 2 response alternatives, each depicting one object in the upper left and right corner, respectively. After 1000 ms a yellow circle appeared at the bottom center of the screen. A click on it initiated playback of an audio recording of a statement specifying which object was picked up, e.g. (3) or (4).

(3) Der Wuggy hat dann die Geige aufgesammelt.
Then the wuggy has picked up the violin.

(4) Der Wuggy hat dann die Birne aufgesammelt.
Then the wuggy has picked up the pear.

Subjects were instructed to move their mouse upwards immediately after clicking the initiation button (see Spivey et al., 2005) and to choose their response as quickly as possible. After each response selection, the screen was blank for a 1000 ms inter-stimulus interval. Subjects familiarized themselves with the paradigm during 16 initial practice trials.

Material

Statements were acoustically manipulated to exhibit three different intonation contours. Depending on the preceding context question (1) or (2), statements in (3) and (4) are prototypically realized with different intonation contours (e.g. Grice et al., 2017). After a neutral question (2), both subject and object are discourse-new which can be prosodically encoded by specific pitch accents on both constituents (often referred to as *broad focus*). A common contour in these cases is a rising accent on the subject, followed by a high stretch of f_0 and a high or falling accent on the object. After a polar topic question (1), the utterance in (3), which affirmatively picks up the *given referent*, can prosodically emphasize that the proposition in question is true, for example, by *verum focus*, which manifests itself here in the form of a high rising accent on the German auxiliary *hat* (Engl. *has*). Finally and as opposed to the latter, the answer in (4) corrects the topic question (1). It affirmatively mentions a *contrastive referent*, which is typically realized by *contrastive focus*, an intonation contour with a high rising accent on *Birne* (*pear*). Figure 1 illustrates the three contours schematically. Statements for each experimental item (n = 12) came with these three intonation contours (Broad, Verum, and Contrast), resulting in 36 different target sentences overall.

Visual stimuli were taken from the BOSS corpus (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010). There were two sets of acoustic stimuli: questions providing a discourse context presented on the context screen and statements triggering participants' responses on the response screen, with one question and one statement corresponding to each object.

Acoustic stimuli were recorded by a trained phonetician in a sound-attenuated booth with a headset microphone (AKG

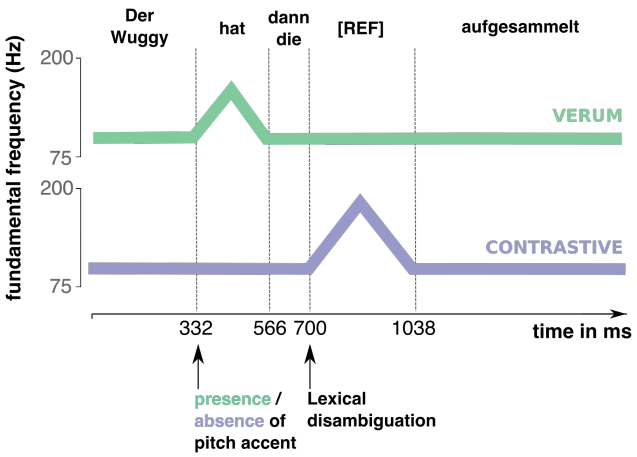

Figure 1: f_0 contours and average temporal landmarks for the resynthesis of Verum and Contrastive focus.

C420) using 48 kHz/16-bit sampling. To ensure that the three different contexts exhibit the same temporal characteristics for each sentence (i.e. the lexical information become available at the same time across focus conditions), sentences were manipulated and resynthesized using Praat (Boersma & Weenink, 2016). The resulting stimuli differed only in the pitch contour and accompanied intensity envelope. The preregistration report `https://osf.io/49q2r/` contains additional information about the resynthesis process.

Design

There were two experimental groups. The unreliable verum (UV) group was exposed to consistently natural contrastive focus contours but occasionally encountered unreliable verum focus contours; reversely for the unreliable contrast (UC) group. 'Unreliable' use of intonation is defined as follows. In the context of question (1), the speaker would use statement (3) realized with a pitch accent on the object as if to indicate a contrastive referent and statement (4) realized with a pitch accent on the auxiliary as if to indicate a given referent, creating a mismatch between information status, pitch accent position and disambiguating lexical information. Occasional exposure to unreliable cues undermines the possibility to confidently predict the likely speaker-intended referent earlier than after lexical disambiguation.

Subjects are exposed to 12 blocks of 8 stimuli each. In the UV group, each block contained 2 reliable contrastive focus statements, 2 reliable verum focus statements, 1 unreliable verum focus statement, and 3 broad focus statements. In the UC group, each block contains 2 reliable verum focus statements, 2 reliable contrastive focus statements, 1 unreliable contrastive focus statement, and 3 broad focus statements.

Analysis

The screen coordinates of the computer mouse were sampled at 100 Hz using the mousetrap plugin (Kieslich & Henninger, 2017) implemented in the open source experimental software OpenSesame (Mathôt, Schreij, & Theeuwes, 2012). Trajectories were processed with the package `mousetrap` (Kieslich & Henninger, 2017) using `R` (R Core Team, 2017).

There were a total of 84 target trials per participant. We only analysed target trials with reliable mappings between discourse context and intonation. For each trial, we compute the *turn towards the target* (TTT) as the latest point in time at which the trajectory did not head towards the target (where "heading towards the target" is operationalized by approximating the first derivative to the x- and y-coordinates of a trajectory; see function `get_TTT_derivative()` in the analysis scripts at `http://osf.io/49q2r`).

We fitted Bayesian hierarchical linear models which predict TTT measurements as a function of FOCUS, GROUP and BLOCK and their three-way interaction, using the Stan modelling language (Carpenter et al., 2016) and the package `brms` (Bürkner, 2016). The models included maximal random-effect structures, allowing the predictors and their interactions to vary by subjects (FOCUS - BLOCK) and experimental items (FOCUS - BLOCK- GROUP). We used weakly informative Gaussian priors centered around zero with $\sigma = 100$ for all population-level regression coefficients (Gelman, 2006), as well as standard priors of the `brms` package for all other parameters. Four sampling chains with 4000 iterations each were run for each model, with a warm-up period of 2000 iterations, ensuring convergence. We report, for relevant predictor levels and difference between predictor levels, 95% credible intervals (CIs) and the posterior probability that a respective posterior distribution β is smaller than zero $P(\beta < 0)$. We judge there to be evidence for an effect if zero is (by a reasonably clear margin) not included in the CI and $P(\beta < 0)$ is close to zero or one.

Model predictions

Our link hypothesis is that the TTT measure is a strictly decreasing function of posterior odds, as defined in Eq. (1). In the experimental context, where any object appeared equally likely as given/contrastive referent, it is reasonable to assume that prior odds are roughly 1. Consequently, we consider a mapping from likelihood ratios (i.e., evidential strength) to TTT, which ideally should have a finite lower bound to which it converges from above as evidential strength grows to infinity. One natural choice is an exponential decay function: $\text{TTT} \sim \exp(1 - \text{evidential strength})$.

To model belief dynamics, we assume that listeners increment non-normalized scores, which might, for simplicity, represent numbers of recent remembered instances where utterance u was used to refer to referent r. For example, the speaker's propensity to choose verum (V) or contrast (C) focus for either given (r_g) or competitor referent (r_c) could be derived from scores like in the adjacent table. This yields conditional production probabilities after normalization, like so: $P(V \mid r_g) = \frac{35}{35+15} = 0.7$. In each experimental trial, listeners observe both utterance and referent (by final lexical disam-

	V	C
r_g	35	15
r_c	5	45

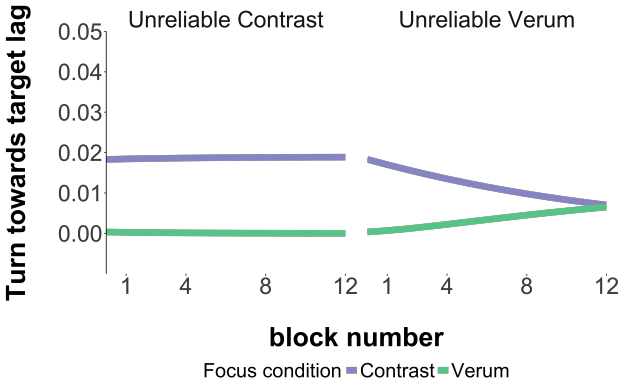

Figure 2: Model predictions for main text example.

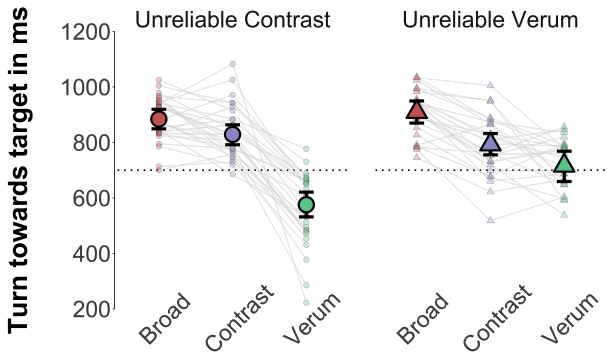

Figure 3: Estimates and CIs for the TTT measurement. Semi-transparent points are average values for each subject. Solid grey lines group individual subject's values across conditions. The dotted line indicates average acoustic onset of referent.

parameter	mean	95% CI	$P(\beta < 0)$
Contrast (UC) - Broad (UC)	−56	(-88;-26)	1
Contrast (UC) - Verum (UC)	252	(206;297)	0
Broad (UC) - Verum (UC)	309	(265;356)	0
Contrast (UV) - Broad (UV)	−117	(-151;-81)	1
Contrast (UV) - Verum (UV)	77	(23;136)	0
Broad (UV) - Verum (UV)	194	(139;254)	0
Contrast (UC) - Contrast (UV)	36	(-16;87)	0.09
Broad (UC) - Broad (UV)	−25	(-75;22)	0.84
Verum (UC) - Verum (UV)	−140	(-207;-73)	1
Slope Contrast (UC)	2	(-5;8)	0.3
Slope Broad (UC)	−2	(-8;4)	0.68
Slope Verum (UC)	1	(-7;9)	0.4
Slope Contrast (UV)	−2	(-9;5)	0.72
Slope Broad (UV)	3	(-4;10)	0.21
Slope Verum (UV)	7	(-2;16)	0.08

Table 1: Posterior estimates of differences between conditions (rows 1-9) and posterior estimates of the effect of experimental block for each condition (rows 10-15).

biguation), and so increment the relevant score by 1.

Experiment initially, listeners will have *default expectations* about speakers' form-function mappings. For proficient speakers of German, these should satisfy some natural constraints. We assume that (i) the *base rate* of C is higher than that of V, $P(C) > P(V)$ where $P(u) = \sum_{r_i} P(u \mid r_i)\ P(r_i)$, as verum focus is infrequent in German; (ii) verum focus is more natural for a given than for a competitor referent, $P(V \mid r_g) > P(V \mid r_c)$; (iii) for realizing reference to the competitor, contrastive focus is more natural, $P(C \mid r_c) > P(V \mid r_c)$.

These assumptions and constraints define an infinite class of models. Nonetheless, we can derive general qualitative predictions. Given the assumed base rate difference, V has higher evidential strength than C. For the example above, we have $\frac{P(V|r_g)}{P(V|r_c)} = 7$, but $\frac{P(C|r_c)}{P(C|r_g)} = 3$. Most importantly, we predict that the higher (lower) the evidential strength of a cue, the more (less) it will be affected by learning that it is unreliable, and the less (more) it will be affected by learning that it is reliable. An example, based on the numbers from the table above for the exact belief updates induced by the present experiment is in Figure 2. Concretely, in the UC condition we predict no noteworthy additional inhibition of contrast focus (intuitively: because it is a weak cue from the start, i.e. the absence of a pitch accent is not an informative predictor of the discourse status of an upcoming referent, because its presence is not very expected by low base rate) and consider even a small facilitation possible for some model parameter values (because participants still see more reliable contrast focus uses than unreliable ones even in the UC condition). Neither would we expect a noteworthy additional facilitation of verum focus exploitation (because it is a reliable cue from the start). In the UV condition we predict both a noteworthy inhibition of verum focus exploitation (because a high-fidelity cue's reliability is now undermined), as well as a learning effect in the exploitation of contrast focus (because learning raises the initially low evidential value). These model-derived predictions were preregistered at `https://osf.io/49q2r/`.

Results and Discussion

Following pre-registered protocol, the whole data set of a participant was excluded whenever he/she (a) exhibited initiation times above 350 ms in more than 15% of the trials, (b) exhibited more than 10% errors, or (c) exhibited movement behavior violating instructions in more than 15% of the trials. We excluded one subject for each exclusion criterion. We further had to exclude two subjects due to experimental malfunctions. Trials with initiation times greater than 350 ms (1.3%) and incorrect responses (0.3%) were discarded on a trial-by-trial basis. Additionally, trials that exhibited movement behavior violating instructions were discarded, too (1%).

Figure 3 displays the mean and CIs of the posterior distribution (conditioned on the middle of the experiment, i.e. scaled block number = 0). There is substantial evidence that the three different focus conditions elicit different TTT patterns, with Broad being the slowest (UC: β = 884, CI = (850;920); UV: β = 909, CI = (869;949)) followed by Contrast (UC: β = 828, CI = (791;863); UV: β = 793, CI = (756;833)) and Verum (UC: β = 576, CI = (532;621); UV: β = 716, CI = (660;769)). (Posterior differences between conditions are summarized in Table 1.)

These patterns are in line with Roettger and Franke (under review). The acoustically early cue associated with verum fo-

cus allows listeners to infer the intended referent long before the lexical material becomes available. Beyond that, listeners also use the absence of this cue (no accent on the auxiliary) to anticipate the contrastive interpretation. This inference does not happen as fast as in the verum focus condition but earlier than lexical disambiguation (Broad > Contrastive > Verum).

Looking across groups, neither Contrast nor Broad show clear indications of an impact of the group manipulation (Contrast: β = 36, CI = (-16;87), $P(\beta < 0)$ = 0.09; Broad: β = -25, CI = (-75;22), $P(\beta < 0)$ = 0.84). Verum, however, is clearly slower in the UV group (β = -140, CI = (-207;-73), $P(\beta < 0)$ = 1). These results are compatible with our predictions. We predicted a difference mainly in the Verum condition, where TTT measures should be slower. Since we only predicted no facilitation for the Contrast condition in the UC group, these results are fully compatible with model-derived predictions.

Figure 4 displays how these temporal effects change over the experiment. In comparison to the patterns described by Roettger and Franke (under review), the present effects do not change much across the experiment. There is not sufficient evidence that the development of participants' anticipatory behaviour over the course of the experiment (slope of the lines) is different from zero (= a flat line) (see Table 1), although our posterior belief in the predicted positive slope for the Verum condition in the UV group is about 0.92.

Despite the absence of conclusive evidence for dynamic changes of TTT measures throughout the experiment, there are suggestive patterns comparing the start and end of the experiment. In the UC group, Contrast is initially similarly slow as Broad (CI intervals overlap, see Figure 4). Broad seems to become slower and Contrast faster, leading to a substantial differences between these categories by the end of the experiment. Thus, listeners seem to learn to exploit the absence of a pitch accent on the auxiliary as a predictive cue to an upcoming contrastive referent. Learning happens despite occasional unreliable form-function mappings in the Verum condition.

Contrary to this, at the beginning of the experiment, the Verum condition in the UV group starts with a temporal advantage over Contrast. However, throughout the experiment, Verum appears to become slower approaching the temporal performance of Contrast by the end of the experiment (CI intervals are heavily overlapping). Listeners appear to selectively unlearn the expected speaker production probabilities for verum focus, while learning to predictively exploit the form function mapping in the Contrast condition.

General Discussion

This study replicates earlier findings that listeners rapidly exploit intonational cues to predict speaker intentions (Dahan et al., 2002; Weber et al., 2006; Kurumada et al., 2014a). Hearing an early pitch accent (or its absence), listeners' manual response dynamics indicate an early bias towards one interpretation over another (Roettger & Stoeber, 2017). Our results further replicate and expand findings by Roettger and Franke

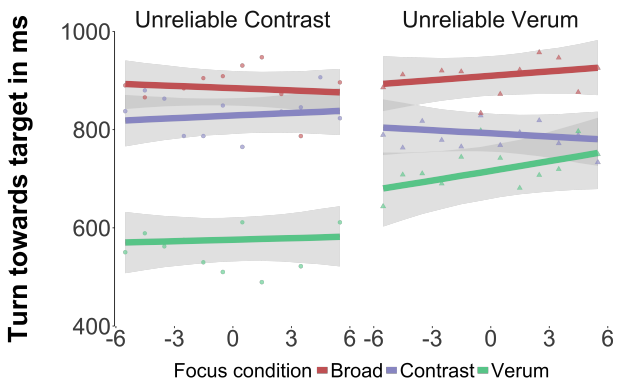

Figure 4: Estimated TTT values (lines) as a function of block number (scaled), dependent on focus condition and listener group. Shaded ribbons are 95% CIs. Semi-transparent points correspond to average values for each block.

(under review) showing that intonational cue exploitation depends on the estimated reliability of form-function mappings. If listeners learn that a cue is uninformative, they appear to weigh down the informational value of that cue (c.f. Kurumada et al., 2014b). This selective adaptation further shows a clear tendency to change dynamically throughout exposure. A Bayesian model of predictive cue integration and belief dynamics, paired with an exponential link function from posterior odds to the TTT measure, predicts interesting asymmetries in listeners' responses and their temporal development, which are supported by the data.

These results also point to interesting follow-up research. Infinitely different numerical models are compatible with the naturalness constraints on speaker production we postulated here. We have focused on assessing general qualitative predictions only. The question arises whether a quantitative fit, using model parameter estimation based on the data, is possible. Doing so will likely also highlight aspects in our data that the present model does *not* seem to capture. Figure 4 suggests that already in the first block there is a large effect of unreliability on the processing of verum focus. Our model does not predict this (Figure 2). It is conceivable if not likely that listeners have a more elaborate belief update process than modelled here. Already after the first example of an unreliable use of what is normally a high-fidelity cue, listeners might be immediately alerted. This, for instance, could lead them to immediately adjust their readiness to deviate from their default beliefs. Plasticity of listener beliefs is represented as the sums over rows in our tables of non-normalized weights: the higher the sum, the less swiftly beliefs adapt. Our model predictions were derived based on fixed plasticity rates for which the model gives non-trivial predictions, but it is worthwhile for future work to explore, both empirically and in modelling, the possibility that listeners also quickly adjust their beliefs about optimal plasticity based on the relative surprisal of observed speaker utterances.

Despite these open issues, the present study contributes to our understanding of how listeners deal with ubiquitous un-

certainty in processing intonation; how listeners infer speaker intentions based on bottom-up acoustic cues and probabilistic expectations about likely intonational contours; and whether listeners' flexible adaptation behavior is compatible with rational belief dynamics.

Acknowledgments

Timo Roettger was supported by the "Zukunftskonzept" of the University of Cologne as part of the Excellence Initiative.

References

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer. [computer program]. version 6.0.17.*

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, *5*(5), e10773.

Bürkner, P.-C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*, 1–37.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, *47*(2), 292–314.

Fine, A. B., & Jaeger, F. T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*(3), 578–591.

Freeman, J. B., & Ambady, N. (2010). Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, *1*(3), 515–534.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, *64*, 90-107.

Grodner, D., & Sedivy, J. C. (2011). 10 the effect of speaker-specific information on pragmatic inferences. *The processing and acquisition of reference*, 239.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, 1–16.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, *122*(2), 148.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. (2014b). Rapid adaptation in online pragmatic interpretation of contrastive prosody. In *Proceedings of the cognitive science society* (Vol. 36).

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014a). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*(2), 335–342.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Magnuson, J. S. (2005). Moving hand reveals dynamics of thought. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 9995–9996.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, *44*(2), 314–324.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*(2), 204–238.

Pierrehumbert, J., & Hirschberg, J. B. (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in communication*, 271–311.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Roettger, T. B. (2017). *Tonal placement in Tashlhiyt: How an intonation system accommodates to adverse phonological environments* (Vol. 3). Language Science Press.

Roettger, T. B., & Franke, M. (under review). *Task-oriented adaptation in intonation-based intention recognition.* `https://osf.io/dnbuk`. Open Science Framework.

Roettger, T. B., & Stoeber, M. (2017). Manual response dynamics reflect rapid integration of intonational information during reference resolution. In G. Gunzelmannn, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of CogSci 39* (pp. 3010–3015). Cognitive Science Society.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

Tomlinson, J., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of memory and language*, *69*(1), 18–35.

Tomlinson, J., Gotzner, N., & Bott, L. (2017). Intonation and pragmatic enrichment: How intonation constrains ad hoc scalar inferences. *Language and Speech*, *60*(2), 200–223.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, *49*(3), 367–392.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of memory and language*, *87*, 128–143.