CAN SEGMENTAL DETAIL INFLUENCE PROSODIC ANALYSIS? THE CASE OF CONTRACTIONS IN ENGLISH

Holger Mitterer[1,3], Sahyang Kim[2], Taheong Cho[3]

[1]University of Malta, [2]Hongik University, [3]Hanyang Institute for Phonetics & Cognitive Sciences of Language, Hanyang University
holger.mitterer@um.edu.mt, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

Prosodic structure is phonetically reflected not only on suprasegmental dimensions (e.g., f0, duration, and amplitude) but also on segmental dimensions (e.g., coarticulation, segmental strengthening or reduction). This leads to the question whether the listener makes use of the prosodic-structurally conditioned segmental detail to compute the prosodic structure of a given utterance. We generated small dialogues for an online mouse-tracking task to test whether participants could use focus-related prosodic information (pitch-accented or not) and segmental realization of *have* (full or contracted) in predicting an upcoming target referent in a sentence. Results showed that a pitch-accented full form of 'have' did facilitate the listener's prediction of the upcoming referent. But results did not show such an effect of segmental realization (full or contracted) in the absence of pitch accent on 'have', suggesting that segmental information may not always be exploited by listeners when strong suprasegmental cues are available associated with contrastive focus.

Keywords: prosody, perception, prosody-syntax interface, online mouse-tracking

1. INTRODUCTION

Prosodic structure influences not only phonetic realization of suprasegmental features (such as f0, duration and amplitude), but segmental realization as reflected, for example, in articulatory strengthening versus weakening, localized hyper- or hypo-articulation and coarticulatory variation [1]-[3]. Cho et al. [1], for instance, found less coarticulatory nasalization when the word was carrying a pitch accent (see also [4]–[6]). Similarly, Cho & Keating [7] reported that alveolar stops are articulatorily strengthened, leading to longer VOTs for unvoiced stops, in domain-initial position or in a pitch-accented condition. These studies indicate that segmental realization may carry prosodic weight possibly to serve as a cue to prosodic structure.

In speech comprehension, prosodic information is known to influence listeners' parsing of syntactic structure. For instance, Steinhauer et al. [8] showed that prosodic breaks can easily overrule the otherwise well supported minimal-attachment heuristic in sentence processing, while others have shown that pitch accents or their absence can lead listeners to expect new or given information [9]–[11]. In these studies, the focus was on the suprasegmental aspects that signal prosodic structure. For instance, Roettger & Franke [11] used small dialogues and the German *verum* (focus-related) accent (see below) to show that listeners use this accent to predict how a sentence will continue.

(1a) *Hat der Wuggy dann die Geige aufgesammelt?*
Has the wuggy then the violin pick-up?
Did the wuggy then pick up the violin?
(1b) *Der Wuggy HAT dann die Geige aufgesammelt.*
the wuggy has then the violin picked-up.
The wuggy then picked up the violin.

As indicated by uppercase HAT (Engl., 'to have') in (1b), an affirmative response to a yes/no question in German can be marked by a pitch accent on the conjugated verb [12]. Roettger & Franke [11] used a mouse tracking task and listeners had to move the cursor to the object that the "wuggy" had picked up. When there was a *verum* accent on the auxiliary, listeners turned the mouse towards the target even before that word was heard.

In this and other similar eye-tracking studies [9], [10], the prosodic information used was encoded in variation of f0 and duration. This gives rise to the question whether similar results can be obtained if prosodic information is encoded in segmental detail. The processing of prosodic and segmental information is often conceptualized as being independent, so that a prosodic analysis (carried out by a so-called 'Prosody Analyzer') occurs in parallel with a segmental analysis (see [14] for related discussion), sometimes even conceptualized as being divided between the two hemispheres [15]. A strong version of this approach may therefore postulate that segmental information would not influence sentence processing in a similar way as is modulated by prosodic information embedded in suprasegmental features. Alternatively, however, there is some evidence that allows us to predict similar effects of segmental detail on sentence processing. Scott & Cutler [16] showed that listeners can use flapping (or

its absence) to decide whether a noun should be attached to an ongoing phrase or not ([*The day we met*] *Anne* … vs. [*The day we met Anne*]…). More recently, Mitterer et al. [17] showed a similar effect in Maltese using the glottal stop, which is a phoneme in Maltese. These two studies, however, made use of offline measures. They cannot therefore be seen as direct evidence for the listener's use of segmental process on the fly at the early processing stage but may be attributed to a late integration of prosodic and segmental information that is also implied in the Prosody Analyzer account [13].

Two mouse-tracking experiments were carried out in this study to explore how segmental information of 'have' in English may be used online by listeners in predicting an upcoming referent in sentence. We used a similar mouse-tracking set up following [11], but we had to implement various adaptations for the online setting (e.g., where the mouse cursor speed and location cannot be influenced by a script[i]). Moreover, we added one more condition in which the prosodic *weight* on the auxiliary (i.e., to what extent the production of the auxiliary 'have' can carry information of prosodic structure) was exclusively coded in segmental terms with no further bottom-up support of suprasegmental features for an assumed prosodic structure. In an example below, (2a) is the question and (2b-d) are three possible answers.

(2a) The aliens haven't shot the robot, right?
(2b) Well, they HAVE shot the robot. (*full, pitch accent*)
(2c) Well, they've shot the violin. (*reduced*)
(2d) Well, they have shot the violin (*full, no pitch accent*)

While participants heard a dialogue with one of the three answers (2b-d), they were presented with pictures of two objects (e.g., a violin and a robot) located left and right at the top of the screen. As in a simple computer game, the objects were falling down from the top with some left-right wriggling, and there was a picture of a spaceship which can shoot the falling objects. Participants could move the spaceship left and right using the mouse and were instructed to press a mouse button to fire a shot at a falling object in accordance as indicated in the answer sentence.

For (2b), the pitch-accent on *HAVE* should be interpreted as being contrastive to the *haven't* in the question, which is likely to lead listeners to predict that the upcoming object noun phrase would be the same as the one given in the question, thus moving the spaceship more quickly towards the given object. On the other hand, the absence of pitch accent as in (2c-d) is likely to indicate that an upcoming object is different from what has already been given in (2a), thus leading listeners to move towards the new object. One could, however, further expect the segmental difference between the reduced form (2c) and the full form (2d). Given that the full form is consistent with the accent-induced segmental realization, one could hypothesize that it carries more prosodic weight towards focus-induced prominence than the reduced form, thus rendering listeners *less* likely to expect a new object in the upcoming part of the sentence.

We tested these predictions in two experiments using different control conditions. Exp. 1 used a neutral (or broad focus) question (*What has happened?*) as a control condition as well as the question which may induce a narrow focus in the answer as in (2a). In the narrow focus condition, the presence or absence of focus on '*have*' in the answer sentence should be matched with the pragmatically appropriate referent (i.e., a given referent for (2b), and a new reference for (2c,d)). This, however, has a potential drawback. While results of Exp. 1 may reflect the processing of prosody, they may also be influenced by the learning of contingencies in the experiment ('narrow' question + *HAVE* → given). Therefore, Exp. 2 used the narrow questions only, and the referent in the answer sentences were not always pragmatically matched with the question—i.e., both the 'new' and 'given' referents were targets on the screen regardless of the information structure of the dialogue. This incongruent mapping between prosody and information structure may lead participants to ignore the phonetic details of the answer sentences to be used for computing prosodic structure. Given no contingencies within the experiment that participants could learn to exploit, however, if participants showed any preference for a given object, it could be interpreted as being due to the processing of the prosodic information. That is, the two designs have complementary weaknesses, but together, should paint a more complete picture.

2. GENERAL METHOD

Native speakers of American English were recruited via the *prolific.co* platform (24 for Exp 1 and 36 for Exp 2). We constrained participants to be in the age range 18-40, living in the USA, having English as their first language and not have reported any language-related disorder in their sign-up to *prolific.co*. Participants could only participate in one of the experiments. We used a larger sample in Exp 2 since it contained fewer trials.

We recorded two male American English speakers; one speaker provided the questions and the other the replies in the three versions as indicated in (2). Questions were edited to shorten the pause before the *right?*, since a long pause might be irritating in a repetitive experiment with many trials. For the answers, we cross-spliced the different parts of the sentence, so that participants could not learn that a given pronunciation of *Well* would lead to preference

of a particular object. Splicing points were after *Well* and after *shot*. For each combination of the target and the condition, all experimental materials were cross-spliced so that four different versions were generated with a randomly selected precursor from the same condition, but always for a different target. The stimulus was presented in a random order. We used copyright-free images for the 16 target objects in a 200x200 pixels frame, which were used as the targets to be shot by the spaceship in the game.

Participants were instructed on the structure of the task and conducted 16 practice trials to familiarize themselves with the 16 objects. They then continued to complete 144 (Exp. 1, nine blocks with 16 objects) or 96 (Exp. 2, six blocks with 16 objects) experimental trials. In Exp. 1, each target was presented two times in each of the three experimental conditions with the 'narrow' question (***HAVE*** with pitch accent; ***have*** *or* ***'ve*** with no pitch accent) and three times in the control condition (with the question *What has happened?*) with each version of the answer used once. In Exp. 2, each target was presented six times, once in each cell of the design by which the target type ('given' or 'new') was crossed with the three versions of the answer.

Similar to eye-tracking data, we investigated how close to the target the mouse position was in the horizontal dimension. For each trial, the screen coordinates were re-scaled to the interval [-1,1], with zero being the middle of the screen. The target side was mapped onto the positive side. Co-ordinates above .6 on either side were set to .6, so that the mouse was positioned to shoot the target or the competitor in a corridor within which the target wriggled left and right. Since we were interested in predictive processing, we analysed the mean mouse position in a time window from -200ms to 200ms around the target onset, a range which cannot be influenced yet by the pronunciation of the target word itself. The data were analysed with linear mixed-effects models (LME) that used the maximally converging random-effects structure and contrast coded predictors. Degrees of freedom were estimated using the lmerTest [17] package in R.

3. RESULTS

3.1 Experiment 1

Fig. 1 shows the results of Exp. 1. Panel A shows that participants move towards an expected target (i.e., towards a 'given' target in the pitch-accented HAVE condition or a 'new' target in the unaccented condition) well before the onset in all conditions if the question is narrow. But they wait on the target onset in the control 'broad' condition (*What has happened*?) in which the mouse moving towards the target starts around 200ms after the target onset. Panel B shows that this pattern arises only after the first block.

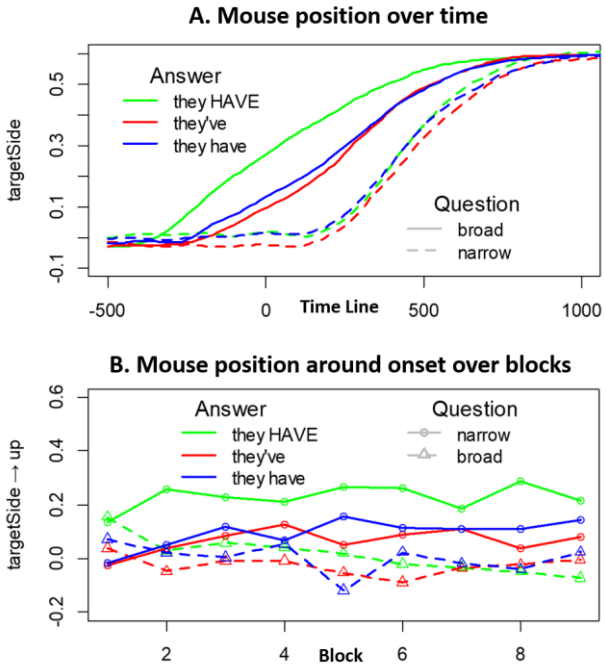

Fig. 1: Results from Experiment 1. Panel A shows the mouse position relative to the target over time; Panel B shows the average position in a -200 – 200ms time window around target onset over the course of the experiment.

The LME used contrast-coded predictors for Question (-0.5 = broad, 0.5 = narrow), Block (ranging from -4 till 4) and Answer with two contrasts, one for the presence of accent (*HAVE* vs. others) and one for the use of contraction (*they've* vs. *they have*). The analysis revealed that participants were significantly closer to the target with the 'narrow' question (b=0.137, $t(23.61)=5.53$ p<.001) and this effect became larger over blocks (b = 0.021, $t(28) = 4.14$, p<.001) and was larger for the answer with a pitch accent than not (b = 0.135, $t(24)=3.181$, p=.004). Participants also moved the mouse closer to the target when there was an accent in the answer (b=0.094, $t(22) = 4.554$. p<.001), which again increased over blocks (b=0.01, $t(2994)=1.994$, p=0.046). However, there were no significant differences between the two no-pitch-accent conditions that differed in segmental detail (full vs. reduced/contracted).

3.2 Experiment 2

In Exp. 2, mouse movements were analysed based on whether the target object was 'given' or 'new' (regardless of whether or not it was supported by pitch accent) and the form of the answer with the same three levels as in Exp. 1. Fig. 2 shows the mouse position relative to the target side. Panel A shows that participants move the mouse towards the 'given' target when supported by pitch accent ('HAVE'). Likewise, when the target is a 'new' object

(mismatched with pitch accent), they still move the mouse towards the 'given' object (the competitor) as can be predicted by the pitch-accented 'HAVE.' Additionally, as the solid lines in Fig. 2 indicate, at the beginning participants tend to move the mouse closer to the 'given' target in all conditions.

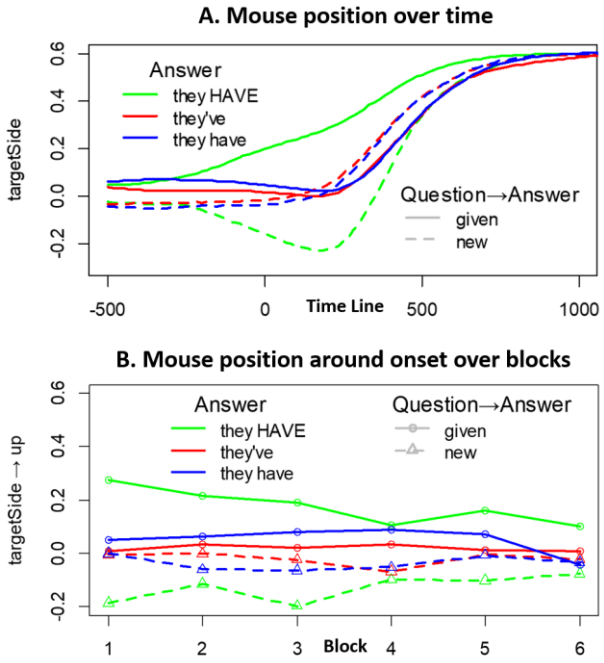

Fig. 2: Results from Experiment 2. Panel A shows the mouse position relative to the target over time, while Panel B shows the average position in a time window from -200 to 200ms around target onset over the six blocks.

The statistical analysis generally bore out these observations. There was a significant effect of Object Type ('given' vs. 'new', b = 0.132, t(33) = 6.00, p < .001) which became smaller over blocks (b = -0.019, t(35) = 2.45, p= .019). An interaction of Pitch Accent with Object Type (given/new) (b=0.221, t(35) = 5.34, p<.001) reflects the preference for the 'given' object only in the pitch-accented condition. This interaction is further supported by an interaction with block (b = -0.04, t(2894)=-3.12, p=.002), again reflecting that the preference for the given object in the pitch-accented condition gets smaller over the course of the experiment. It appears that listeners learned to overcome the mismatched (inappropriate) prosody and the expected referent (i.e., 'HAVE' followed by a 'new' target or a contracted form followed by a 'given' target). Importantly, as Fig. 2B shows, there was still a clear preference for the 'given' object in the 'HAVE' condition even in the last block (solid green line), indicating that the listeners predict the upcoming referent based on the prosody information. As in Exp. 1, there were no clear differences between the two conditions differing by segmental detail.

4. DISCUSSION

Our results first demonstrate that our online mouse tracking method gave rise to anticipatory response patterns with mouse movements towards the target well before target onset. In Exp. 1, participants performed as expected with earlier moves towards a 'given' object when 'have' was pitch-accented, but towards a 'new' object when 'have' was unaccented. Our results also showed that participants adapted to the prosodic contingencies as they were able to predict the target better in later blocks during the experiment, even in conditions with a-priori uninformative prosodic cues. In Exp. 2, when prosody and the expected referent were mixed (matched or mismatched), we found a robust and consistent predictive effect of the presence of pitch accent prosody (but not its absence) on 'have', showing a strong bias towards a 'given' object in all conditions despite the fact that a half of the stimulus sentences could contain a prosodic incongruence (e.g., a pitch accented 'HAVE' followed by a 'new' object). The fact that only the effect of the presence of an accent clearly survives further suggests the necessity to distinguish adaptation to prosody from adaptation to the experimental situation.

Crucially, however, we failed to find any evidence that listeners can use prosodically-conditioned segmental detail (reduced vs. full forms of *have*) as a cue to prosodic structure in the same way that they use suprasegmental cues that are embedded in our pitch-accented condition. In fact, suprasegmental cues remained important even when they were uninformative within the experimental context (as in Exp. 2). This primacy of suprasegmental prosodic cues in sentence processing may help understand why listeners failed to show a segmental effect. Given that the prosodic information was so salient in the pitch-accented 'HAVE' condition in the current experimental settings, the lack of pitch accent could have led the listeners to treat two unaccented variants with the equal prosodic weight without paying attention to the segmental detail.

In conclusion, our results confirm that sentence processing is modulated by prosodic structure that is computed primarily by suprasegmental cues driven by narrow focus. The results also indicate that segmental information alone may not provide robust perceptual bottom-up support to prosodic structure in the absence of suprasegmental cues, making it difficult to inform how segmental and prosodic information may be integrated in the time course of speech perception. It remains to be seen whether and how segmental information may be exploited by listeners in an experimental setting when other prosodic information is weaker.

5. ACKNOWLEDEMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5C2A02086884).

6. REFERENCES

[1] T. Cho, D. Kim, and S. Kim, 'Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English', *J. Phon.*, vol. 64, pp. 71–89, Sep. 2017, doi: 10.1016/j.wocn.2016.12.003.

[2] P. Keating, T. Cho, C. Fougeron and C-S. Hsu, 'Domain-initial strengthening in four languages', In J. Local, R. Ogden & R. Temple (Eds.), *Phonetic interpretation: Papers in Laboratory Phonology VI*, (pp.143-161) et al., 2003.

[3] K. de Jong, 'The suprasegmental articulation of prominence in English: Linguistic stress as localized hyperarticulation. *J. of the Acous. Soc. of Am.*, 97, 491-504, 1995.

[4] J. Jang, S. Kim, and T. Cho, 'Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences', *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. EL33–EL39, Jul. 2018, doi: 10.1121/1.5044641.

[5] H. Li, S. Kim, and T. Cho, 'Prosodic structurally conditioned variation of coarticulatory vowel nasalization in Mandarin Chinese: Its language specificity and cross-linguistic generalizability', *J. Acoust. Soc. Am.*, vol. 148, no. 3, pp. EL240–EL246, Sep. 2020, doi: 10.1121/10.0001743.

[6] J. Jang, S. Kim, and T. Cho, 'Prosodic Structural Effects on Non-Contrastive Coarticulatory Vowel Nasalization in L2 English by Korean Learners', *Lang. Speech*, p. 00238309221108657, Jul. 2022, doi: 10.1177/00238309221108657.

[7] T. Cho, and P. Keating, Effects of initial position versus prominence in English, J. of Phon, 466-485, 2009.

[8] K. Steinhauer, K. Alter, and A. D. Friederici, 'Brain potentials indicate immediate use of prosodic cues in natural speech processing', *Nat. Neurosci.*, vol. 2, no. 2, pp. 191–196, 1999, doi: 10.1038/5757.

[9] D. Dahan, M. K. Tanenhaus, and C. G. Chambers, 'Accent and reference resolution in spoken-language comprehension', *J. Mem. Lang.*, vol. 47, no. 2, pp. 292–314, Aug. 2002, doi: 10.1016/S0749-596X(02)00001-3.

[10] A. Weber, B. Braun, and M. W. Crocker, 'Finding referents in time: Eye-tracking evidence for the role of contrastive accents', *Lang. Speech*, vol. 49, no. 3, pp. 367–392, Sep. 2006, doi: 10.1177/00238309060490030301.

[11] T. B. Roettger and M. Franke, 'Evidential Strength of Intonational Cues and Rational Adaptation to (Un-)Reliable Intonation', *Cogn. Sci.*, vol. 43, no. 7, p. e12745, 2019, doi: 10.1111/cogs.12745.

[12] G. Turco, B. Braun, and C. Dimroth, 'When contrasting polarity, the Dutch use particles, Germans intonation', *J. Pragmat.*, vol. 62, pp. 94–106, Feb. 2014, doi: 10.1016/j.pragma.2013.09.020.

[13] K. B. Shatzman and J. M. McQueen, 'Prosodic knowledge affects the recognition of newly-acquired words', *Psychol. Sci.*, vol. 17, pp. 372–377, 2006, doi: 10.1111/j.1467-9280.2006.01714.x.

[14] J. Steffman, S. Kim, and S-A. Jun, 'Prosodic phrasing mediates listeners' perception of temporal cues: Evidence from the Korean Accentual Phrase', *J. Phon*, vol. 94, pp.1-13, 2022.

[15] G. Hickok and D. Poeppel, 'The cortical organization of speech processing', *Nat. Rev. Neurosci.*, vol. 8, no. 5, pp. 393–402, May 2007, doi: 10.1038/nrn2113.

[16] D. R. Scott and A. Cutler, 'Segmental phonology and the perception of syntactic structure', *J. Verbal Learn. Verbal Behav.*, vol. 23, no. 4, pp. 450–466, Aug. 1984, doi: 10.1016/S0022-5371(84)90291-3.

[17] H. Mitterer, S. Kim, and T. Cho, 'The Role of Segmental Information in Syntactic Processing Through the Syntax–Prosody Interface', *Lang. Speech*, vol. 64, no. 4, pp. 962–979, Dec. 2021, doi: 10.1177/0023830920974401.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, 'Package "lmerTest"', *R Package Version*, vol. 2, no. 0, 2015.

[19] M. S. Slim and R. J. Hartsuiker, 'Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js', *Behav. Res. Methods*, Nov. 2022, doi: 10.3758/s13428-022-01989-z.

[20] A. P. Salverda, D. Kleinschmidt, and M. K. Tanenhaus, 'Immediate effects of anticipatory coarticulation in spoken-word recognition', *J. Mem. Lang.*, vol. 71, no. 1, pp. 145–163, Feb. 2014, doi: 10.1016/j.jml.2013.11.002.

[21] F. Eisner and J. M. McQueen, 'Perceptual learning in speech: Stability over time', *J. Acoust. Soc. Am.*, vol. 119, no. 4, p. 1950, 2006, doi: 10.1121/1.2178721.

[i] If mouse control was possible in JavaScript, websites could force the mouse cursor to hoover over ads and lower the speed with which the ad could be left.