The role of talker-specific prosody in predictive speech perception

Giulio G.A. Severijnen (Donders Centre for Cognition, MPI for Psycholinguistics), Hans Rutger Bosker (MPI for Psycholinguistics), Vitória Piai (Donders Centre for Cognition) & James M. McQueen (Donders Centre for Cognition, MPI for Psycholinguistics)
g.severijnen@donders.ru.nl

One of the challenges in speech perception is that listeners must deal with considerable segmental and suprasegmental variability in the acoustic signal due to differences between talkers. Most previous studies have focused on how listeners deal with *segmental* variability. Listeners do this by adjusting their perception of phonetic categories in a talker-specific fashion[1]. Also, listeners predict upcoming phonological representations based on talker-specific information[2]. *Suprasegmental* variability, for instance in speaking rate, also has large consequences for perception of lexical stress patterns[3] and hence spoken words. In this EEG experiment, we investigated whether listeners track talker-specific usage of suprasegmental cues to lexical stress in order to predict and correctly recognize spoken words.

In a training phase, native Dutch participants (*N*=20) first learned to map novel non-word minimal stress pairs (e.g., *USklot/usKLOT*; capitalization indicates lexical stress) onto different object referents (e.g., the item *USklot* referring to a “lamp”, the item *usKLOT* referring to a “train”). These non-words were produced by two male talkers and, through careful acoustic manipulation, each talker only used one acoustic cue to signal lexical stress (e.g., Talker A only used F0, Talker B only used amplitude; talker-cue mapping counter-balanced across participants). The training phase consisted of a series of two-alternative forced choice (2AFC) and typing tasks. This allowed participants to learn the correct item-to-object mappings as well as, through perceptual learning, which talker used which acoustic cue to signal lexical stress.

At test, participants heard semantically constraining sentences, spoken by either talker, containing the critical non-words in sentence-final position (e.g., *Het woord voor lamp is een [target]* “, The word for lamp is a [target]”). The sentence-final non-word could either be produced using the expected talker-specific cue (e.g., Talker A using F0; *control condition*) or the unexpected cue (e.g., Talker A using amplitude; *cue-switch condition*; see Table 1). If participants learned about the talker-specific cues, they would be able to predict upcoming talker-matching word-forms (e.g., *USklot* produced by Talker A, cued using only F0). We hypothesized that the sentences in the cue-switch condition would lead to longer RTs compared to the control condition. Also, we hypothesized that the cue-switch condition would elicit a relatively larger N200 response, an ERP component that has been found to indicate a phonological mismatch[2,4]. We also included two further conditions. In the first, the sentence-final non-word was a segmentally different non-word, produced using talker-contingent cues (e.g., *BOLdep* produced by Talker A cued using F0; *word-switch condition*). In the last condition, the sentence-final non-word was a suprasegmentally different non-word, produced using talker-contingent cues (e.g., *usKLOT* produced by Talker A, cued using F0; *stress-switch condition*). We hypothesized that these conditions would elicit a relatively larger N400.

For the analyses, incorrect responses were excluded. Results from linear mixed models on log-transformed RTs (see Table 2) showed that the cue-switch stimuli indeed led to longer RTs compared to control: if in training Talker A used F0 to cue lexical stress, participants were slowed down in their responses if, at test, Talker A unexpectedly used amplitude as cue to lexical stress. This suggests that the cue-switch condition created a mismatch between predicted and perceived word-forms based on the learned talker-specific cues. However, in contrast to our predictions, the N200 amplitude was no different between cue-switch and control (see Figure 1). Finally, the N400 was relatively larger in the word-switch and, in a slightly later time-window, in the stress-switch condition. This illustrates that valid ERPs were obtained using the current stimuli and that participants predicted segmental and suprasegmental information in the sentence-final non-words.

Based on the notion that participants predicted suprasegmental information, as illustrated by the larger N400 response in the stress-switch condition, and the behavioral result in the cue-switch condition, we conclude that this illustrates talker-specific prediction of suprasegmental cues, picked up through perceptual learning on previous encounters.

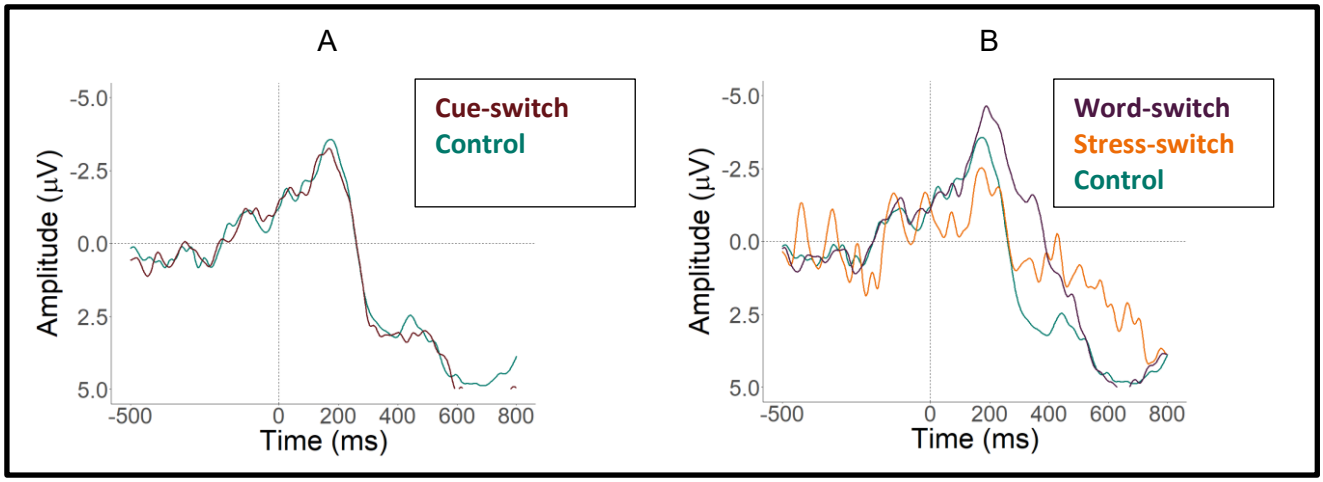
Table 1. Different conditions during the test phase. English translation of the Dutch carrier sentence: “The word for lamp is a [target]” (capitalization in the target word indicates lexical stress). Only Talker A is depicted in Table 1 (and the talker-cue mapping holds for Talker A using F0). Participants heard both talkers in the experiment and talker-cue mappings were counter-balanced across participants.

Condition	Talker	Cue	Cue-switch	Semantic incongruency
Control *Het woord voor lamp is een USklot*”	A	F0	No	No
Cue-switch *Het woord voor lamp is een USklot*	A	Amplitude	Yes	No
Word-switch *Het woord voor lamp is een BOLdep*	A	F0	No	Yes
Stress-switch *Het woord voor lamp is een usKLOT*	A	F0	No	Yes

Table 2. Reaction times and accuracy scores

Condition	Mean RT in ms (SD)	Mean accuracy in % (SD)
Control	1161 (460)	92 (27)
Cue-switch	1221 (496)	90 (31)
Word-switch	894 (349)	99 (11)
Stress-switch	1516 (593)	50 (50)

Figure 1. **ERPs**. The cue-switch condition (dark red in A) and the word-switch and stress-switch condition (purple and orange in B) vs. control (in green)



REFERENCES

1. Eisner, F. & McQueen, J. M. The specificity of perceptual learning in speech processing. *Percept. Psychophys.* **67**, 224–238 (2005).
2. Brunellière, A. & Soto-Faraco, S. The speakers’ accent shapes the listeners’ phonological predictions during speech perception. *Brain Lang.* **125**, 82–93 (2013).
3. Reinisch, E., Jesse, A. & McQueen, J. M. Speaking Rate Affects the Perception of Duration as a Suprasegmental Lexical-stress Cue. *Lang. Speech* **54**, 147–165 (2011).
4. Connolly, J. F. & Phillips, N. A. Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. *J. Cogn. Neurosci.* **6**, 256–266 (1994).