Listeners Track Talker-Specific Prosody to Deal With Talker-Variability

Giulio G.A. Severijnen[a], Hans Rutger Bosker[b], Vitória Piai[a, c], James M. McQueen[a, b]

[a] Donders Centre for Cognition, Radboud University
Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands
[b] Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[c] Donders Centre for Medical Neuroscience, Department of Medical Psychology, Radboudumc
Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands

[Accepted for publication in *Brain Research*, July 27, 2021]

giulio.severijnen@donders.ru.nl
hansrutger.bosker@mpi.nl
vitoria.piai@donders.ru.nl
james.mcqueen@donders.ru.nl

Corresponding author
Giulio G.A. Severijnen
giulio.severijnen@donders.ru.nl
Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands

Declarations of competing interests: none

Abstract

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. In this EEG experiment, we investigated whether listeners track talker-specific usage of *suprasegmental* cues to lexical stress to recognize spoken words correctly. In a three-day training phase, Dutch participants learned to map non-word minimal stress pairs onto different object referents (e.g., *USklot* meant "lamp"; *usKLOT* meant "train"). These non-words were produced by two male talkers. Critically, each talker used only one suprasegmental cue to signal stress (e.g., Talker A used only F0 and Talker B only intensity). We expected participants to learn which talker used which cue to signal stress. In the test phase, participants indicated whether spoken sentences including these non-words were correct ("The word for lamp is…"). We found that participants were slower to indicate that a stimulus was correct if the non-word was produced with the unexpected cue (e.g., Talker A using intensity). That is, if in training Talker A used F0 to signal stress, participants experienced a mismatch between predicted and perceived phonological word-forms if, at test, Talker A unexpectedly used intensity to cue stress. In contrast, the N200 amplitude, an event-related potential related to phonological prediction, was not modulated by the cue mismatch. Theoretical implications of these contrasting results are discussed. The behavioral findings illustrate talker-specific prediction of prosodic cues, picked up through perceptual learning during training.

Keywords: prosody, perceptual learning, lexical stress, phonological prediction, N200

Abbreviations: VOT, voice onset time; F0, fundamental frequency; 2AFC, two-alternative forced choice; SW, strong-weak; WS, weak-strong.

1. Introduction

One of the challenges in speech perception is that listeners must deal with the variability in how different talkers produce speech. That is, even when different talkers produce the exact same sentence, the acoustic realization of this sentence is highly variable between talkers. Still, despite this variability, listeners are able to almost effortlessly recognize utterances spoken by different talkers. In the present study we assess whether and how listeners track talker-specific usage of prosodic cues to lexical stress to facilitate spoken word recognition.

In speech perception, listeners must decode a message by mapping auditory information in the speech signal onto stored knowledge about the sound forms of words in order to recognize each of the words in that message (McQueen, 2005). The acoustic signal consists of both segmental information (such as individual vowels and consonants) and suprasegmental information that signals prosodic structures beyond the segments (such as lexical stress and sentential focus). Hence, speech perception is about combining both sources of information to recognize spoken words (Eisner and McQueen, 2018). For example, consider the phrase "The stranger objects", which depending on the lexical stress on "objects" can be paraphrased as the noun phrase "the more unusual OBjects" (capitalization indicates stress) or the sentence "the newcomer obJECTS". In order to correctly understand this phrase, listeners must use not only information about the vowels and consonants but also suprasegmental information to lexical stress. That is, ignoring either segmental or suprasegmental information would impede correct comprehension of the intended message. We describe two cognitive mechanisms that allow listeners to deal with speech variability: *perceptual learning* and *prediction.*

1.1. Perceptual learning as mechanism to deal with talker variability

Listeners may use talker variability to help them correctly perceive spoken words. That is, while variability in the acoustic signal might lead to difficulties in speech perception, listeners may also exploit talker-specific information to support speech perception. More specifically, identifying a talker's voice, heard on previous encounters, facilitates perceptual processing of phonetic content of a novel utterance from that same talker (Nygaard et al., 1994). Also, listeners can adapt to talker-specific pronunciations of speech sounds by using lexical information (i.e., the word the speech sound appear in) to alter how they map acoustic input to perceptual categories for those talkers (Eisner and McQueen, 2005).

Besides perceptual learning based on lexical information, listeners can also change their reliance on different acoustic cues to perceive speech sounds based on distributional information. Idemaru and Holt (2011) exposed English participants to words starting with plosives in which the canonical relationship between fundamental frequency (F0) and a voiced/voiceless plosive was inversed (a voiceless plosive in English is normally signaled by a high F0 and long voice onset time (VOT)). When this relationship was changed (i.e., a voiceless plosive being cued by a low F0), participants down-weighted their reliance on F0 as a cue and based their responses primarily on VOT. The authors coined this "dimension-based statistical learning". This type of learning has also been found for vowels in English (Liu and Holt, 2015). These findings illustrate that listeners are able to change the weights given to connections between acoustic dimensions and perceptual categories to accommodate to short-term regularities in the acoustic signal.

In a series of experiments, Zhang and Holt (2018) illustrated that these learning effects are indeed adaptations to talkers' speaking styles instead of to the acoustic input in general. They exposed English participants to English minimal pairs (e.g., beer-pier) with ambiguous fundamental frequency (F0) and voice onset time (VOT) values, and measured the proportion of pier-responses (/p/ is normally signaled by a high F0 and a long VOT). While the proportion of pier-responses would have been at chance-level (F0 and VOT were both ambiguous), they found that responses were modulated by the F0 range of accompanying stimuli. More specifically, target words in a low F0 range were perceived as having a higher F0, leading to more /p/-responses and *vice versa* for words in a high F0 range. In two additional experiments following this design, they found that responses were also modulated by talker characteristics (i.e., spoken by either a male or a female talker) or visual presentation of a male or a female talker, even when the F0 values were kept ambiguous. In sum, these findings illustrate that listeners are able to track distinct coevolving regularities (e.g., different talkers with their own speaking style) not only based on acoustic input (as found in the first experiment), but also based on talker characteristics and visual talker identification which allows listeners to rapidly adapt perceptual categories based on talker-specific information.

1.2. Prediction as mechanism to deal with talker variability

The use of prediction in speech perception is not a new proposal. Several studies from the 1970s and 1980s already showed that listeners use prediction in speech perception. For instance, listeners make predictions about upcoming words based on the preceding semantic

and syntactic context (Marslen-Wilson, 1973; Miller et al., 1984) and make predictions about sentence accent based on the intonation of the preceding context (Cutler, 1976).

Listeners also use prediction to deal with talker variability. That is, listeners seem to use talker information that is present in the context to predict upcoming speech that is consistent with that talker. This is done both at the lexical level (Van Berkum et al., 2005) and the prelexical level (Brunellière and Soto-Faraco, 2013). More specifically, Brunellière and Soto-Faraco (2013) found that listeners used information about a talker's regional accent to predict phonological word-forms that are consistent with that talker. In their experiment, Catalan participants listened to semantically constraining sentences spoken in either an Eastern Catalan accent (which applies vowel reduction: [pərmis] for /permis/, "permission") or a Western Catalan accent (no vowel reduction: [permis]). In these sentences, the critical word containing the possible vowel reduction (*permís*) always occurred in sentence-final position, allowing for prediction of the sentence-final word. In some of the sentences, the sentence-final word contained a mismatch between the expected and the actual phonetic realization (i.e., an Eastern Catalan talker producing [permis] without vowel reduction, or *vice versa*). These mismatches elicited a relatively larger N200 response, an event-related potential (ERP) argued to reflect acoustic-phonetic processing in the phonological stage of word processing (Connolly and Phillips, 1994), as compared to sentences in which there was no mismatch. The authors concluded that listeners predicted word-forms based on the regional accent presented in the sentence context.

Taken together, these two mechanisms, perceptual learning and prediction, can help listeners deal with talker variability. First, listeners can adapt their perceptual categories for specific talkers through perceptual learning cued by auditory and visual identification of a talker. Second, based on these altered talker-specific categories, listeners can predict upcoming word-forms that are consistent with that talker, facilitating speech perception on subsequent encounters. However, previous studies have primarily studied these mechanisms in relation to segmental information while suprasegmental variability is also widely present in speech. For example, Clopper and Smiljanic (2011) illustrated that prosodic variation (pause distribution and F0 patterns) in American English was affected by dialect and gender. Similarly, Xie et al. (2021) found individual talker differences in productions of sentence intonation. Furthermore, prosodic variation in Dutch has been found to be affected by dialects (Gussenhoven and Van Der Vliet, 1999) and sex-related differences (Haan and Van Heuven, 1999). It remains unclear however, how listeners deal with variability in suprasegmental information.

1.3. The role of lexical prosody in speech perception

As the earlier "OBject" – "obJECT" example illustrated, suprasegmental information is crucial for speech comprehension and several studies have found that listeners indeed make use of this kind of information in spoken word recognition. For example, in Cutler and Van Donselaar (2001), Dutch participants performed a lexical decision task with minimal stress pairs (*VOORnaam/voorNAAM, "*first name"/"respectable"). Results showed that when participants were previously primed with the exact same word (e.g., *VOORnaam*), RTs were faster when responding to the target (*VOORnaam).* However, this facilitation disappeared when they were primed with the other member of the minimal pair (e.g., *voorNAAM*). The authors concluded that the use of suprasegmental information constrained word activation so that only the correct member of the minimal pair was activated.

Listeners do not use suprasegmental information only when it is strictly required to discriminate between lexical items but also to facilitate perception more generally. Reinisch, Jesse, and McQueen (2010) showed that participants use suprasegmental stress information to recognize spoken words as soon as it becomes available. In an eye-tracking study, they exposed Dutch participants to segmentally overlapping words (*OCtopus*/*okTOber*) and found that when participants were presented with one of these words (e.g., *OCtopus*), they fixated the target word (*OCtopus*) more often as compared to segmentally overlapping competitors (*okTOber*). Critically, they did so before the point of segmental disambiguation. This illustrates that when the words are segmentally identical (until the point of disambiguation), Dutch listeners make use of suprasegmental information to recognize spoken words, even when not strictly necessary. The same effect has also been found in English listeners for primary-stress words (Jesse et al., 2017) and in Italian listeners (Sulpizio and McQueen, 2012).

Just as segmental variability affects word recognition, suprasegmental variability can also have large consequences. For example, perception of lexical tone in Cantonese is influenced by the fundamental frequency (F0) in surrounding (preceding and following) context (Sjerps et al., 2018). Also, the vocal tract size of a given talker, typically differing for male vs. female talkers, can change the perception of vowels (Bosker et al., 2020a). Furthermore, the speaking rate in a preceding context can affect the perception of lexical stress (Reinisch et al., 2011) and can even change whether a given word is heard or not (Bosker et al., 2020b; Dilley and Pitt, 2010). Considering the effect of suprasegmental variability on the perception of lexical stress and word recognition more generally, it is important to find out how listeners deal with this variability. It has previously been found, for instance, that listeners adapt to suprasegmental lexical-stress errors in foreign accented speech (Reinisch and Weber, 2012).

More specifically, when listeners heard words, spoken in foreign accented speech, in which stress patterns were non-canonically produced (i.e., with suprasegmental information signaling the wrong stress pattern), they quickly adapted to those realizations to still correctly perceive the target words. The focus of the present study is to find out whether listeners also adapt to individual talker-specific realizations of lexical stress patterns.

1.4. The present study

The present study was concerned with the following question: Do listeners keep track of how individual talkers cue lexical stress, using this information on subsequent encounters with the same talkers to predict talker-congruent word-forms? We created a set of disyllabic non-word minimal stress pairs (e.g., *USklot* vs. *usKLOT*), produced by two different talkers, and we manipulated which talker used which cue to signal lexical stress (F0 or intensity). That is, in a three-day training program participants were taught novel non-word-to-object mappings (e.g., *USklot* meant "lamp"; *usKLOT* meant "train"), while hearing, for instance, Talker A always use F0 to signal stress, and Talker B use intensity. Participants performed a series of two-alternative forced choice (2AFC) and typing tasks divided over these three days (see Fig. 1). We predicted that they would explicitly learn the meanings of the non-words and implicitly learn which talker used which suprasegmental cue to signal lexical stress. In a final test phase, we recorded participants' reaction times (RTs) and electroencephalogram (EEG) as participants heard semantically constraining sentences, containing the newly learnt non-words, produced by both Talker A and B (e.g., "The word for lamp is *USklot*"). Their task was to indicate whether the spoken stimulus was correct or incorrect by means of a button-press. We predicted that if participants had learned about the talker-specific cues to lexical stress, they would be able to predict talker-congruent word-forms (i.e., *USklot* produced using F0 to cue stress when hearing Talker A, but produced using intensity when hearing Talker B).

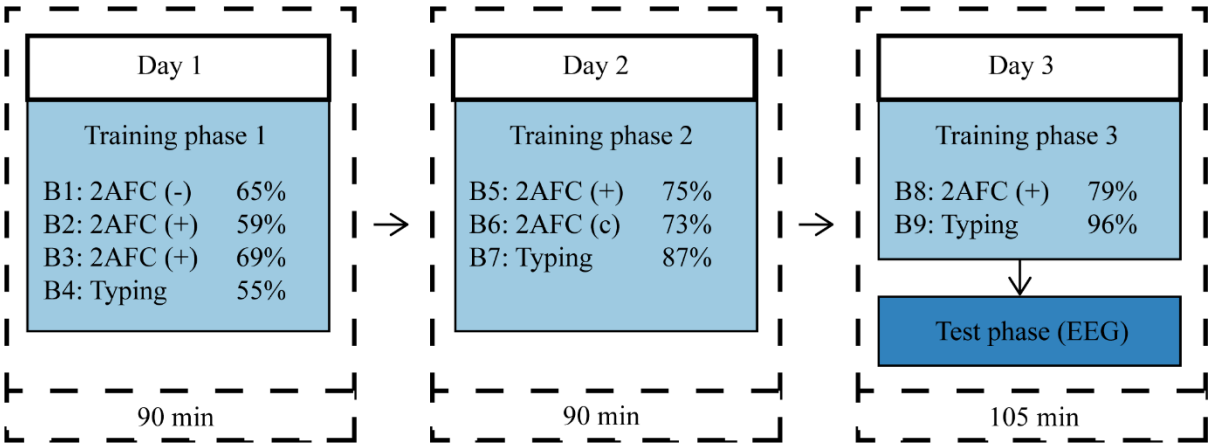

Fig. 1. Schematic overview of the experiment and accuracy scores of the training tasks. The two-alternative forced choice (2AFC) tasks could either contain no minimal pairs within trials (-), only minimal pairs within trials (+) or only contain the items on which an error was made in the previous training block (c).

To test this hypothesis, the test phase consisted of several conditions that differed in the sentence-final target word (see Table 1). First, a control condition contained the correct critical item, produced using the correct cues for a given talker (e.g., *USklot* for "lamp" by Talker A using F0). Second, the cue-switch condition still contained the correct critical item, produced by the same talker, but using the unexpected cues (e.g., *USklot* for "lamp" by Talker A using intensity). Third, the stress-switch condition contained the wrong member of the minimal pair, produced using the talker-congruent stress cue (e.g., *usKLOT* for "lamp" by Talker A using F0). Finally, the word-switch condition contained one of the other previously learned items (e.g., *BOLdep* for "lamp" by Talker A using F0). Importantly, in this fashion, the cue-switch condition did not contain a semantic incongruency (only a cue-incongruency; the sentence-final word in the cue-switch condition only differed from control in the cue that was used to signal lexical stress) while the stress-switch and the word-switch condition did contain semantic incongruencies.

Table 1.

Example test stimuli in the different conditions. Only Talker A is being depicted in Table 1 even though participants heard both talkers (i.e., opposite values hold for Talker B). We counterbalanced which talker used which cue (i.e., talker-cue mappings) across participants. "Yes" and "No" in Cue-switch and Semantic incongruency refer to whether the conditions contain a cue-switch or a semantic incongruency. "Yes" and "No" in Correct response refers to which behavioral response was the correct one.

Condition	Talker	Cue	Cue-switch	Semantic incongruency	Correct response
Control ***Het woord voor lamp is een USklot*** "The word for lamp is a USklot"	A	F0	No	No	Yes
Cue-switch ***Het woord voor lamp is een USklot*** "The word for lamp is a USklot"	A	Intensity	Yes	No	Yes
Stress-switch ***Het woord voor lamp is een usKLOT*** "The word for lamp is a usKLOT"	A	F0	No	Yes	No
Word-switch ***Het woord voor lamp is een BOLdep*** "The word for lamp is a BOLdep"	A	F0	No	Yes	No

Our primary hypothesis was that the sentences in the cue-switch condition would create a mismatch between the predicted word-forms (i.e., the talker-congruent word-forms) and the perceived word-forms. We predicted that this would (1) lead to longer RTs and, (2) as in the study by Brunellière and Soto-Faraco (2013), elicit a relatively larger N200 response in the cue-switch condition as opposed to the control condition. As Connolly and Phillips (1994) point out, the N200 is related to processing at the phonological stage of word processing, which is to be distinguished from the N400 that results from semantic violations. Since the target words in the cue-switch and the control condition are segmentally identical and have the same stress pattern, any difference in processing (either in RTs or ERPs) can be attributed to predicted phonological representations. This would indicate that participants learned about the talker-specific cues to lexical stress in training and used this information in predicting

upcoming speech on subsequent encounters at test. In addition to this learning effect originating from the training phase, the presence of the cue-switch condition can also affect the learned representations throughout the test phase. That is, contrary to the training phase, in which all the items followed the correct cues for either talker, the cue-switch condition (occurring on 25% of the trials) contained incongruent cues. This provided conflicting information, which has previously been found to affect learned representations (Kraljic and Samuel, 2005; Kurumada et al., 2014) and could lead to unlearning in the test phase.

In addition to the cue-switch condition to test the primary hypothesis, we included the stress-switch and word-switch conditions as 'verification conditions' to inform us on the learning behavior of the participants and whether they would predict the sentence-final words in the first place. We hypothesized that since the stress-switch and the word-switch conditions contained a semantic mismatch between the predicted and perceived sentence-final words, these would elicit a relatively larger N400 response as compared to the control condition. The N400 is an ERP reflecting the semantic relationship between a word and the context it appears in (Kutas and Hillyard, 1984). Although the present study does not allow us to distinguish between prediction and integration accounts of the N400 (for discussion, see Mantegna et al., 2019), we interpret it here as reflecting predictive processing.

Concerning RTs, we did not have any specific predictions for these conditions. On the one hand, RTs could increase compared to the control condition because the mismatch between the sentence and the sentence-final word could cause slowing down of the response. On the other hand, RTs could also decrease in the word-switch condition: The decision to reject an incongruent word in the word-switch condition could be faster, since the mismatching segmental information becomes apparent more quickly compared to the control condition. Alternatively, there could also be no difference in RTs: In the stress-switch condition, participants need the same amount of acoustic input as in the control condition to base their decision on. Note that the behavioral task required participants to make a different behavioral response in the word-switch and the stress-switch condition compared to the control condition (see Table 1). More specifically, the behavioral response in the word-switch and the stress-switch condition required a "no"-response (e.g., *BOLdep* is *not* a "lamp") while the control and cue-switch conditions required a "yes"-response (e.g., *USklot* cued by either F0 or intensity is a "lamp"). This needs to be taken into consideration when comparing RTs between the control condition and the word-switch and stress-switch conditions.

2. Results

2.1. Behavioral

The main goal of the behavioral analyses was to find out whether reaction times (RTs) and accuracy scores differed in the experimental conditions compared to the control condition. Mean RTs in ms and accuracy percentages are displayed in Table 2. Mean RTs and RT distributions are also depicted in Fig. 2. The behavioral data (log-transformed RTs; binomial accuracy) were analyzed using linear mixed-effects models (see section 4.4.1 for details). These models tested for effects of Condition, with the control condition mapped onto the intercept. All the following effects were therefore compared to the control condition. We also tested for effects of and interactions with Trial Number to assess changes in Condition effects across the test phase. Finally, we tested for effects of and interactions with Cue to find out whether the prosodic cue used to signal the stress pattern (F0 vs. intensity) affected the results. See Supplementary Table S1 and S2 for the complete results of the models.

We observed significantly longer RTs in the cue-switch condition compared to control (β = 0.04, *SE* = 0.01 t = 3.60, p = .001). This indicated that when participants were presented with the correct word but produced using unexpected prosodic cues for that particular talker, they were slower compared to when the expected cues were used to produce that word. This finding supports our hypothesis that listeners picked up on the talker-specific cues in the exposure phase and used these to predict talker-consistent word-forms at test. When the actual suprasegmental cues to lexical stress then mismatched this prediction, it slowed participants down. Moreover, we observed a main effect of Trial Number (β = -0.04, *SE* = 0.01 t = -3.15, p = .002) and a marginally significant interaction between Trial Number and the cue-switch condition on RTs (β = -0.02, *SE* = 0.01 t = -1.80, p = .07). This indicated that while RTs decreased overall throughout the experiment, there was a tendency for the decrease to be even stronger for the cue-switch condition. Thus, the RTs in the cue-switch and control condition tended to converge. This suggests that participants might also have been 'unlearning' the talker-specific effect during the experiment, presumably as a consequence of hearing the talker-incongruent cue-switch condition at test.

Table 2.

Mean (SD) response times (from correct trials only; in ms) and percentages of correct answers during the test phase.

Condition	RT (ms)	Accuracy (%)
Control	1200 (501)	92 (27)
Cue-switch	1264 (530)	89 (30)
Word-switch	884 (333)	99 (8)
Stress-switch	1525 (610)	57 (50)

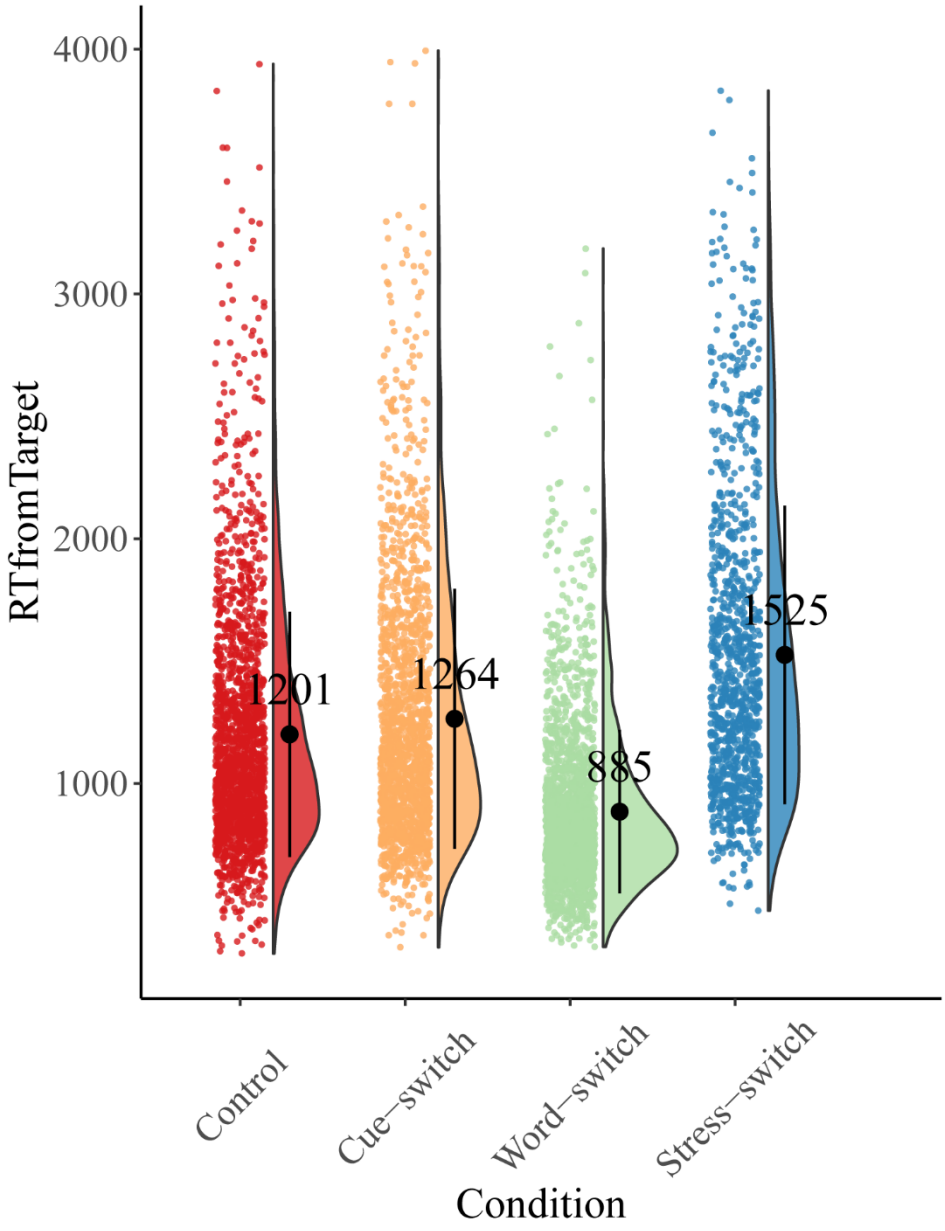

Fig. 2. Violin plots of the reaction times (RTs) in the different experimental conditions during the test phase. In the violin plots, the single dot represents mean RTs in ms, the lines represent

the standard deviation across participants. Individual data points are plotted as raincloud plots for each condition.

We also found that the word-switch condition had overall shorter RTs compared to control (β = -0.29, *SE* = 0.04, *t* = -8.10, $p < .001$), indicating faster responses when participants were presented with an entirely segmentally different word than expected (e.g., *BOLdep*). The stress-switch condition had longer RTs than control (β = 0.27, *SE* = 0.03, *t* = 9.88, $p < .001$), indicating slower responses when participants were presented with the opposite member of a minimal pair. As explained in section 1.4, note that the behavioral task required participants to respond differently to the word-switch and the stress-switch conditions (a "no"–response) compared to the control condition (a "yes"-response) so these RT effects should be interpreted with caution. Moreover, we found an interaction between Trial Number and the word-switch condition on RTs (β = -0.04, *SE* = 0.01 *t* = -3.75, $p < .001$), illustrating that RTs in the word-switch condition became even shorter throughout the experiment.

Finally concerning RTs, we found a main effect of Cue (β = 0.05, *SE* = 0.02 *t* = 3.05, $p = .002$), indicating slower responses for words produced with F0 as cue to stress compared to words produced with intensity. This was the case for all conditions except for the word-switch condition, for which we found a significant interaction (β = -0.07, *SE* = 0.02 *t* = -4.09, $p < .001$). In the word-switch condition, words produced with F0 elicited faster responses compared to words produced with intensity. Considering that these were not effects of main interest for the present study, an interpretation of them is currently lacking.

Next, we analyzed whether the accuracy scores of the categorization responses were different for the four conditions (see Table 2). The model showed that the cue-switch did not differ significantly from the control condition (β = 0.06, *SE* = 0.17 *t* = 0.39, $p = .735$), indicating that participants performed equally well in these conditions. In both these conditions, the target word should elicit the same "yes"–response (i.e., the meaning of the word is correct in both conditions; see section 1.4). This result illustrates that participants were able to correctly perceive the target word, despite the talker-incongruent prosodic cues in the cue-switch condition. Further, the model showed that participants in the stress-switch condition (β = -2.33, *SE* = 0.24 *t* = -9.43, $p < .001$) performed worse compared to the control condition, while participants had higher accuracy scores in the word-switch condition than control (β = 3.04, *SE* = 0.57, *t* = 5.24, $p < .001$). These results suggest that while participants successfully learned the segmental information in the non-words (as indicated by higher accuracy in the word-switch condition), they still struggled with the suprasegmental information, as shown by lower

accuracy in the stress-switch condition. The model revealed no significant effects of (or interactions with) Trial Number, indicating that the accuracy was stable across the experiment.

Finally, the model revealed no significant effect of Cue on accuracy scores (β = -0.26, SE = 0.25 t = -1.05, p = .294) but did reveal a significant interaction with the cue-switch condition (β = -0.64, SE = 0.26 t = -2.48, p < .013) and a marginally significant interaction with the word-switch condition (β = 1.18, SE = 0.68 t = 1.74, p = .082). This suggests that accuracy scores decreased when words were produced with F0 in the cue-switch condition while they had a tendency to increase in the word-switch condition. Although these findings corroborate the higher RTs for F0 in the RT analysis, we currently lack a clear explanation for them.

To control for possible confounds, we ran four additional analyses (see Supplementary Information section 1.1). First, we noticed that performance on Training Block 8 (2AFC) was relatively low compared to Training Block 9 (typing task). We wanted to find out whether this could affect our behavioral results. One notable result emerged from this analysis. This result showed that while there was no main effect of Training Block 8 performance on RTs during the test phase (β = -0.03 , SE = 0.04, t = -0.73, p = .47), a significant interaction with the cue-switch condition was present (β = 0.03, SE = 0.01 t = 2.54, p = .02). This interaction illustrates that as performance on Training Block 8 improved (i.e., on a task requiring acoustic evaluation of the stimuli), the talker-cue mismatch effect increased. Second, to examine the low performance on the stress-switch condition in more detail, we looked into performance of individual participants. Third, it is important to note that our behavioral results might have been affected by the fixed order of presentation of the various conditions (see section 4.4.2). The third additional analysis evaluated this possibility. Fourth, we wanted to ensure that our RT result was not affected by possible outliers in the data. In the final analysis, we thus excluded extreme RT observations and ran the same linear mixed model for RTs. Importantly, the four additional analyses provided some interesting insights but did not alter any of our main conclusions.

2.2. EEG results

The main goal of the EEG analyses was to examine whether the N200 amplitude was modulated by the sentences in the cue-switch condition, relative to control, indicating a perceived phonological mismatch. In addition, as a secondary analysis, we sought to find out whether the N400 amplitude was modulated in the word-switch and stress-switch condition, relative to control, to verify that our manipulations and stimuli could elicit the intended ERPs.

We computed ERPs time-locked to stimulus onset and ran cluster-based permutation analyses (Maris and Oostenveld, 2007) over the entire epoch (i.e., -500 ms to 800 ms relative to stimulus onset; see section 4.4.2). This allowed us to assess significant differences between the conditions that would coincide with both the N400 time-window (between 200 ms and 600 ms poststimulus; Kutas & Federmeier, 2011) and the N200 time-window (between 285 and 335 ms; Brunellière & Soto-Faraco, 2013).

First, the cluster-based permutation analysis revealed no significant difference between the cue-switch condition and the control condition (p = .971). This indicated that both conditions are from the same probability distribution (i.e., the conditions are interchangeable) which implies that – contrary to our expectations – the cue-switch condition did not elicit a relatively larger N200 response compared to the control condition (see Fig. 3A, Fig. 3D).

Second, the cluster-based permutation analysis revealed a significant difference between the word-switch condition and the control condition (p = .002). The difference between the two conditions was most prominent between 174 ms and 458 ms. To illustrate the location and latency of the difference, we plotted various topographical maps of the amplitude difference over time between 174 ms and 458 ms (see Fig. 3E). Fig. 3B shows the ERPs of both conditions from one of the channels in this cluster, where it is evident that the word-switch condition has a larger amplitude than the control condition. The location and latency of the difference between the conditions are consistent with an N400 effect. Given also the experimental manipulation (i.e., a semantic incongruency), we conclude that the difference between the word-switch and control conditions is most likely due to a larger N400 response in the former condition.

Third, the cluster-based permutation analysis revealed a significant difference between the stress-switch condition and the control condition (p = .048). The difference between these two conditions was most prominent between 402 ms and 800 ms. Again, we plotted various topographical maps of the amplitude difference over time between 402 ms and 800 ms (see Fig. 3F) and the ERPs of both conditions from one of the channels in this cluster (see Fig. 3C). It is evident that the stress-switch condition had a larger amplitude than the control condition. Similarly to the word-switch condition, the location and latency (though it occurs in a slightly later time-window) are consistent with an N400-like effect. In addition to the experimental manipulation, this led to the conclusion that the larger amplitude in the stress-switch condition is due to a larger N400-like response compared to the control condition.

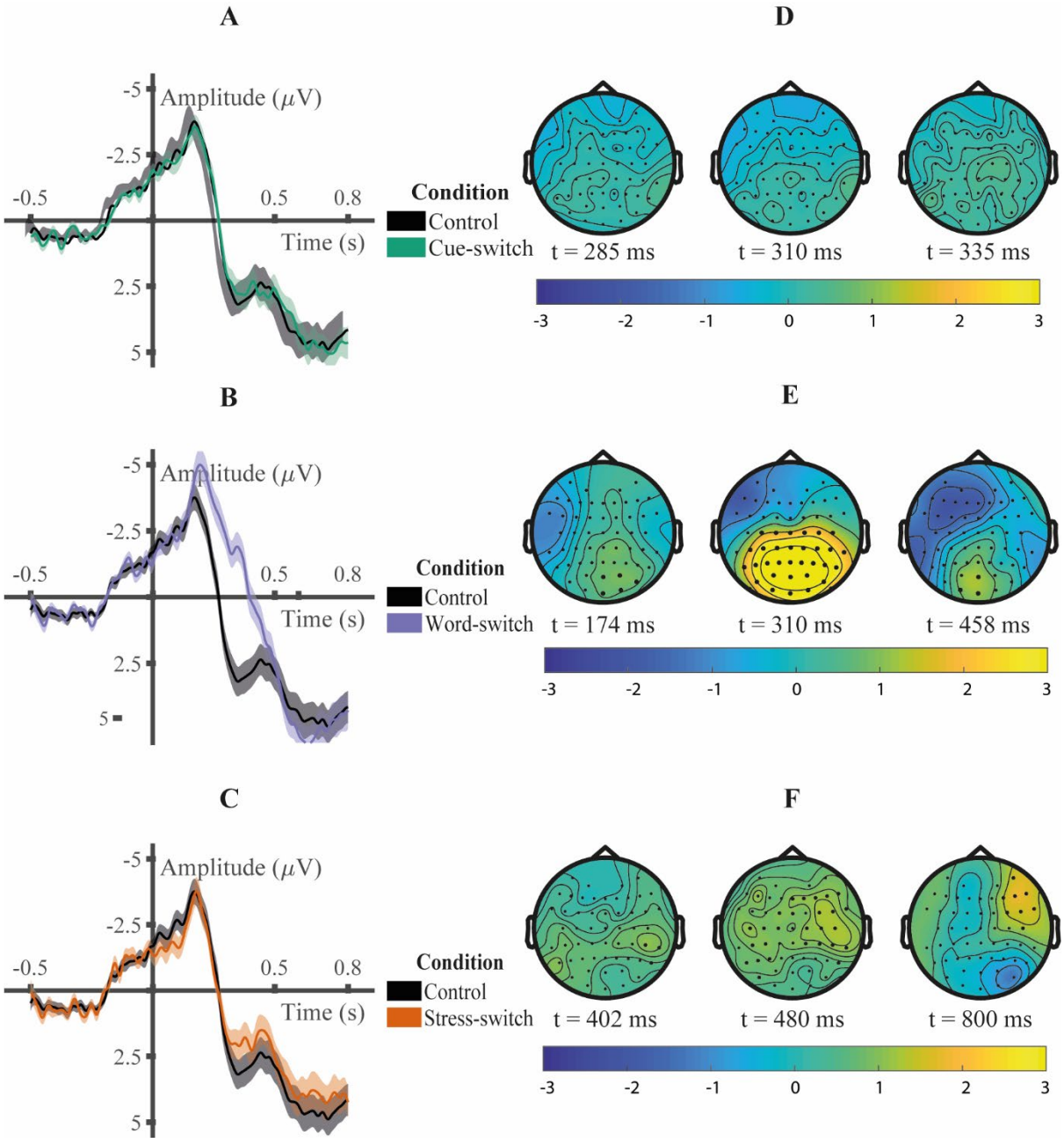

Fig. 3. (Color figure available online) **A-C.** ERPs of channel Pz, comparing the control condition to the cue-switch condition (A), the word-switch condition (B), and the stress-switch condition (C). The shaded areas represent the standard error across participants. **D-F.** Topographical maps of the amplitude difference, comparing the control condition to the cue-switch (D, cf. panel A), the word-switch condition (E, cf. panel B); and to the stress-switch condition (F, cf. panel C). For each contrast, topographical maps for three time-points have been plotted to illustrate the amplitude difference over time.

3. Discussion

The present study tested whether listeners can learn about how two distinct talkers cue lexical stress differently and if they use that talker-specific information to predict talker-congruent word-forms on subsequent encounters. Results demonstrated that when participants were presented with talker-incongruent prosodic cues to lexical stress (i.e., sentences containing unexpected cues for a specific talker), this led to longer RTs compared to the control condition. In contrast, the amplitude of the N200 was not modulated by talker-incongruency in the stress cues. The behavioral findings suggest that participants had learned the talker-specific prosodic cues during training and used that information to predict talker-specific word-forms.

These findings build on those of Eisner and McQueen (2005) and Zhang and Holt (2018), who found evidence for talker-specific perceptual learning of segmental information. We show for the first time that listeners also use talker-specific perceptual learning to deal with talker-variability in *lexical prosody*. In line with the dimension-based learning account in Zhang and Holt (2018), we show that listeners are able to learn which acoustic cues are used by different talkers to signal prosodic structures. We interpret these outcomes to indicate that our listeners adjusted the relative connection weights between the acoustic dimensions (i.e., F0, intensity) and their perceptual categories (i.e., a trochee; Strong-Weak (SW) or a iamb; Weak-Strong (WS)). That is, when listeners learned that one talker only used F0 to signal lexical stress patterns, the weight of that talker-relevant dimension was increased and the weight of the talker-irrelevant dimensions was decreased, which influenced perception on subsequent encounters.

Recall that in the present study, listeners were presented with two different talkers. Listeners were thus not simply required to adapt to one talker but had to track talker-specific usage of prosodic cues of multiple talkers, similar to Zhang and Holt (2018) and Xie et al. (2021) for intonational prosody. Additionally, since the carrier sentences during the test phase did not contain any talker-specific cues to lexical stress (i.e., monosyllables only), listeners were required to re-activate previously formed memories about the speaking styles of both talkers acquired during training. The present study thus illustrates that listeners not only track these regularities while encountering different talkers but also create new memories for these talkers in which talker-information is stored. When they encounter the same talker on a subsequent instance, the formed memory is re-activated and the weights that are given to the acoustic dimensions are adjusted accordingly.

The new memories did not remain stable throughout the experiment, however. As the marginally significant interaction between the cue-switch condition and Trial Number

suggested, there was a tendency for the difference in RTs between the control and cue-switch condition to become smaller over the course of the test phase. This convergence of the RTs in the two conditions shows that the talker-specific effect was gradually reduced over the course of the test phase. A possible explanation for this effect is the exposure to talker-incongruent stimuli during the test phase (i.e., the stimuli in the cue-switch condition). Even though these stimuli were only present in 25% of the trials at test, they may still have caused unlearning. A similar effect of unlearning has been found for segmental perceptual learning (Kraljic and Samuel, 2005) and for prosodic structures (Kurumada et al., 2014). Both studies found that when participants were exposed to conflicting information from the same talker, as opposed to what they had learned before, the learned effect disappeared rather quickly. This illustrates the flexibility of these talker-specific memories, which is confirmed in the present study. That is, even though participants formed memories that were stable enough to initially predict the talker-specific cues, these memories were quickly adapted when conflicting cues were presented at test. An open question for future research concerns how much experience with one specific talker is needed for this memory to become stable over longer periods of time (cf. Eisner & McQueen, 2006).

In addition to perceptual learning, the present study shows that listeners used prediction as a second mechanism to deal with prosodic talker-variability. More specifically, participants predicted which prosodic cues were used to produce the stress patterns in the target words. When the pre-activated word-form did not match the perceived word-forms (i.e., in the cue-switch condition), this led to a processing cost, as shown by the longer RTs compared to the control-condition. In line with Brunellière and Soto-Faraco (2013), this illustrates that listeners make use of information about a talker's speaking style to predict upcoming phonological word-forms that are consistent with that talker.

The present findings are consistent with Bayesian accounts of prediction in speech perception (Norris et al., 2016). Speech perception is Bayesian in the sense that listeners try to construct the best possible model of the world (i.e., upcoming speech). Predictions are thus not fixed but should change when the world changes. Bayes' theorem provides a formal procedure for updating beliefs in the light of new evidence. One of the key factors in Bayesian prediction is the type of feedback that listeners use to change their predictions. That is, Bayesian accounts involve feedback for learning, which should be distinguished from activation feedback. More specifically, while activation feedback flows from higher-level components to lower-level components and is intended to improve perception 'on the spot', feedback for learning improves future perception but does not alter on-line perception (Norris et al., 2003). The

perceptual learning mechanism that listeners use in the present study can be considered to be a form of feedback for learning and thus supports Bayesian accounts of prediction. More specifically, we propose that through exposure to the talker-specific cues, listeners update their prior beliefs about which prosodic cues are used by either talker. In light of this new evidence, listeners change their predictions based on those new beliefs.

In contrast to the behavioral result, there was no evidence for a difference in the N200 response between the cue-switch and the control condition. This was surprising as it conflicts with our behavioral results (longer RTs in the cue-switch condition compared to control) nor with the results in Brunellière and Soto-Faraco (2013), who did find a modulation of the N200 amplitude for segmentally talker-incongruent word-forms. This raises the question whether the present ERP finding is true evidence for the null hypothesis, or whether we were simply unable to modulate ERPs with the present design and stimuli. Note that we had included the word-switch and the stress-switch conditions as verification conditions to inform us about whether participants would successfully learn the non-words in the present study and whether these items would modulate ERPs. In both conditions, we found a relatively larger N400-like response compared to the control condition. These were elicited by mismatching semantic information in the target words. This illustrated that participants learned the segmental information (word-switch condition) and suprasegmental information (stress-switch condition) in the items and that these items modulated ERPs. Hence, the null result for the contrast between the cue-switch condition and control cannot be attributed to a failure to collect adequate EEG data. The issue remains why the N200 amplitude was not modulated as in Brunellière and Soto-Faraco (2013), even though behavioral evidence was found for phonological prediction in the present study. We propose this could be due to (1) the to-be-expected effect size of the hypothesized N200 effect in the present study and (2) the sensitivity of the N200 to phonetic detail.

First, there are several reasons to believe that the effect size of the targeted N200 effect would likely be much smaller in the present study compared to Brunellière and Soto-Faraco (2013). For instance, we used manipulated speech (vs. natural speech in Brunellière and Soto-Faraco, 2013) and participants could only rely on exposure to the talker-specific cue usages within the experiment. That is, in Brunellière and Soto-Faraco (2013), participants were already familiar with the accents prior to the experiment, while in our study participants were exposed to our talkers for the very first time. Still, we believe that the most important difference between the present study and Brunellière and Soto-Faraco (2013) relates to the presence of confirmatory talker-cue usages in the carrier sentences at test. Recall that in the present study,

we avoided talker-specific prosodic cues in the carrier sentences at test. That is, apart from the sentence-final non-word, the sentence stimuli at test contained only monosyllabic words, without any confirmatory prosodic information about how Talker A or B produced lexical stress. However, the carrier sentences used at test in Brunellière and Soto-Faraco (2013) did contain words with vs. without vowel reduction (in line with the talker's regional accent). As such, the mismatch at the target word in Brunellière and Soto-Faraco (i.e., vowel reduction in the carrier sentence vs. no vowel reduction in the sentence-final word) was much more apparent and 'local' compared to the present study, where the mismatch concerned prosodic cues in the target word at test vs. prosodic cues in the training phase. Finally, the acoustic mismatch in Brunellière and Soto-Faraco (2013) resulted in a larger phonological-category mismatch compared to that in the present study. More specifically, the mismatch in prosodic cues in the present study leads to a within-category mismatch (the resulting stimuli still contain the same target lexical stress pattern; it is cued only in a phonetically different way). In contrast, the vowel reduction in Brunellière and Soto-Faraco (2013) leads to a between-category mismatch (the words containing a vowel reduction result in a different vowel compared to the words that do not contain a vowel reduction).

Second, we should consider the possibility that the present study could not have led to a modulation of the N200 at all. Previous studies (Brunellière and Soto-Faraco, 2013; Connolly and Phillips, 1994) relied on a mismatch based on segmental information while the sentences in the present study contain a mismatch concerning prosodic cues only. To our knowledge, an N200 response to mismatching prosodic cues has not been found previously. In addition, previous studies argued that listeners predict the phonological form of an upcoming word and that the N200 is elicited if the predicted phoneme cannot be selected (for review, see Nieuwland, 2019). In the present study, even though there was a mismatch between prosodic cues, the predicted phonemes (and indeed the lexical stress pattern) were still the same and could thus be selected. This could explain the lack of a modulation of the N200 response in the present study. Others have even claimed that the N200 is not sensitive to phonological prediction at all. Namely, even though several studies have reported an N200 response to phonologically mismatching word-forms, others failed to find a modulation of the N200 even for segmentally mismatching word-forms (Diaz and Swaab, 2007). It remains unclear whether the N200 component can be functionally dissociated from the N400 component (Nieuwland, 2019). An alternative account is that the N200 actually reflects an early onset of the N400 instead of a distinct component, as argued based on similar topographies (Poulton and Nieuwland, 2019). In fact, visual inspection of our topographies in the word-switch condition,

which contains both a phonological and semantic mismatch, confirm this view: we observed no difference in scalp topographies in the N200 time-window and the N400 time-window.

It is important to note that some of the behavioral findings in the word-switch and stress-switch conditions were not as expected. That is, while shorter RTs and higher accuracy scores in the word-switch condition compared to control confirmed that our participants correctly distinguished the various items based on segmental information and did so with relative ease, the results in the stress-switch condition are surprising. In the stress-switch condition, RTs were much longer and accuracy scores were considerably lower compared to the control condition. Since the decision in the stress-switch condition was mainly based on the prosodic cues, this could be taken as an indication that the prosodic cues were not learned that well after all. Still, given the solid behavioral performance in the control condition (which required processing of the same information), and the N400-like effects for these two conditions in the EEG data, we conclude that this cannot be the entire explanation.

We offer a potential alternative account for this surprising finding. Namely, performance could have been affected by the presence of the word-switch condition at test. Recall that during the training phase, we always presented two referents of the minimal pairs together in the 2AFC tasks (except for Training Block 1), directing participants' attention to prosodic cues. During the test phase, presentation of segmentally different words (in the word-switch condition) could have led participants to pay less attention to the prosodic cues and base their responses more on segmental information. Similarly, it has previously been found that when segmentally different words are present in the stimulus list, listeners realize that they do not need to focus on suprasegmental information to perceive the words (Sulpizio and McQueen, 2011). If participants indeed followed this strategy, this would have led to more "yes"-responses (which led to incorrect responses in the stress-switch condition but correct responses in the cue-switch and control condition). Note that these findings do not affect our interpretation of the contrast between the cue-switch and control condition, since these two conditions required the same "yes"-response and accuracy scores on these conditions were much higher and more comparable.

Although the present study provides insight into the mechanisms – perceptual learning and prediction – used to deal with prosodic talker-variability, there are some limitations to take into account. First, the behavioral performance on the stress-switch condition at test challenges the idea that participants learned suprasegmental information in the items. On the other hand, this is also countered by the behavioral result in the cue-switch and control condition. In fact, these different findings might be an indication that participants were learning two different

types of information. On the one hand, they were learning the item-to-object mappings, which was disturbed in the stress-switch condition. On the other hand, they were separately picking up on the talker-specific cues, as illustrated by the cue-switch condition. This could suggest that the information about talker-specific prosodic cues is not necessarily attached to only the learned non-words, but is instead represented on a more general, abstract level (Bosker, in press) which may speculatively even be generalized to new words. Future experiments may assess whether the talker-specific learning observed here also generalizes to novel items not encountered in exposure. Second, the present study did not test how these mechanisms act on natural speech. That is, we used stimuli in which only one cue signaled stress patterns while normally, stressed syllables in Dutch are signaled by a combination of higher F0, greater intensity and longer duration (Rietveld and Van Heuven, 2009). The question remains whether the observed effects would apply to the same extent to natural speech. Third, the present study used behavioral and EEG measures to test prediction as a mechanism. While previous studies have also used these measures, more sensitive methods can be used to distinguish prediction from integration accounts. For example, eye-tracking has often been used to measure *anticipatory* eye-movements (Altmann and Kamide, 1999; Kamide et al., 2003), as an index of prediction before perception of the target words.

In conclusion, the present study showed for the first time that listeners can adjust their perceptual categories in a talker-specific manner not only for segmental information, as shown previously (Brunellière and Soto-Faraco, 2013; Eisner and McQueen, 2005; Zhang and Holt, 2018), but also for suprasegmental information. Listeners can predict upcoming word-forms based on those talker-specific categories. Applying this to the aforementioned "The stranger objects" example, the present study illustrates that when listeners encounter two different talkers who produce this phrase, listeners learn about the prosodic cues that each talker uses to signal the lexical stress patterns in the phrase. Then, based on this learned information, listeners can predict how each talker will signal those words on subsequent encounters which helps listeners to correctly perceive the words and the phrase despite the large variability in prosodic cues between talkers.

4. Experimental procedure

4.1. Participants

Twenty-four native speakers of Dutch were recruited from the Max Planck Institute for Psycholinguistics (MPI) participant pool and twelve from the Radboud University participant

pool. Recruitment was divided over two locations due to circumstances related to the COVID-19 pandemic. All participants gave informed consent and were paid for their participation. Five participants were excluded because they made at least 60 (94%) errors on the stress-switch condition. We adopted this criterion because extremely low accuracy scores on this conditions suggested that the task instructions were not understood well. One final participant was excluded because of noisy EEG data. The 30 remaining participants were right-handed and did not have any hearing and/or reading problems (8 male, 22 female, age range: 18-48; M_{age} = *23.8,* SD_{age} = *6.5).*

4.2 Materials

4.2.1. Target words

We created 32 disyllabic minimal non-word pairs (see Supplementary Table S5) that were segmentally identical but differed in whether the first or second syllable was stressed (e.g., *USklot* vs. *usKLOT*). The stimuli were recorded twice in a carrier sentence (e.g., *Het woord voor muis is een...,* "The word for mouse is a...") by two male native speakers of Dutch: once with stress on the first and once with stress on the second syllable.

We used the recordings to measure three prosodic cues (Rietveld & Van Heuven, 2009) using Praat (Boersma and Weenink, 2019). First, we measured mean F0 of the voiced part (including consonants containing voicing) in syllables with suitable F0 settings for male talkers (75 – 250 Hz). Second, we derived mean syllable intensity relative to the auditory threshold ($2 \cdot 10^5$ Pa) using the "Get intensity" function in Praat. The intensity of the recordings was not normalized before measurements so there was some minor natural variability in intensity between recordings (M = 66 dB, SD = 4.48). Third, syllable duration was measured. Note that unlike stress in English, where it is principally cued by vowel quality (Cutler, 1986), stress in Dutch is primarily cued suprasegmentally. We calculated the suprasegmental cues for both syllables, with and without lexical stress, separately for both talkers, and we averaged across stressed and unstressed syllables to derive perceptually ambiguous values for each prosodic cue (Table 3). We applied these ambiguous settings using PSOLA in Praat to recordings from SW members of all pairs from both talkers. These stimuli were subsequently evaluated by the first and second author. Note that an acoustically ambiguous value did not always correspond to a perceptually ambiguous stimulus. In such cases, the ambiguous value of that cue was increased or decreased by the step size corresponding to that cue to obtain a new ambiguous value. For duration, this led to two different ambiguous values (see Table 3). The resulting

sounds were acoustically ambiguous in lexical stress and were taken as midpoint stimuli of all the acoustic lexical stress continua that we created next.

Table 3.

Mean acoustic measures and step sizes (across talkers) of prosodic cues for both syllables. Two duration values are provided as being ambiguous. These correspond to the two different values used for different subsets of non-words (see Supplementary Table S3 for the two subsets). Also, the values of the endpoints (Strong-Weak and Weak-Strong) are the acoustically observed values; the ambiguous values were calculated based on the production data and selected based on evaluation by the first and second author. Step sizes were used to create the 7-step continua.

	Strong-Weak (SW)	**Ambiguous stress**	**Weak-Strong (WS)**	**Step sizes**
First syllable				
Duration (ms)	254	202 288	184	17.5
F0 (Hz)	145.6	134.9	124.1	8.1
Intensity (dB)	70.01	68.1	66.20	2.5
Second syllable				
Duration (ms)	376	395 362	402	6.5
F0 (Hz)	108.3	118.5	128.6	7.6
Intensity (dB)	62.2	64.0	65.9	2.4

We created, for each talker and for each individual non-word pair, two continua from SW to WS by varying either the F0 or intensity. Considering the effects of variation in duration on ERPs, we kept duration at ambiguous values in all the stimuli. Using the information about what values signal clear SW patterns, clear WS patterns and ambiguous patterns as well as plausible step sizes for each cue (see Table 3), we created – for each non-word pair and for each talker – acoustic tokens that cued clear SW and WS patterns using only one cue, while the rest was kept at ambiguous values. For instance, for the non-word pair *USklot* vs. *usKLOT* in the F0 continuum, we manipulated F0 values to signal clear SW and WS patterns while intensity and duration were set to ambiguous values. Similarly, in the intensity continuum, we manipulated intensity while F0 and duration were ambiguous. Note that since only one cue signaled lexical stress, the original step sizes were not large enough to elicit the intended complete switch from SW to WS. Based on auditory evaluations by the first and second author, the step sizes were increased. Manipulations were performed using PSOLA in Praat (Boersma

and Weenink, 2019). Note that intensity and duration were manipulated globally (i.e., a single value for a given syllable). F0 was manipulated by creating F0 contours containing a declination (syllable 1: 23.5 Hz; syllable 2: 38.2 Hz) that varied in mean F0 (i.e., the overall value of F0 rather than the contour). Spectrograms of the most SW-like and the most WS-like stimuli for one non-word are depicted in Fig. 4.

Finally, we ran a pretest on the resulting stimuli in the continua. Recall that during stimulus manipulation, decisions were made based on auditory evaluation of the first and second author. The pretest ensured that those decisions resulted in stimuli that satisfied our aims regarding perception of the stimuli. That is, it verified which steps would signal the clearest SW and WS tokens. Also, it verified that the cues signaled SW and WS tokens to the same extent (i.e., avoiding any dominance of one of the cues) and that the two talkers were perceptually comparable. For acoustic details of the stimuli and the pretest, see Supplementary Information, section 1.2.

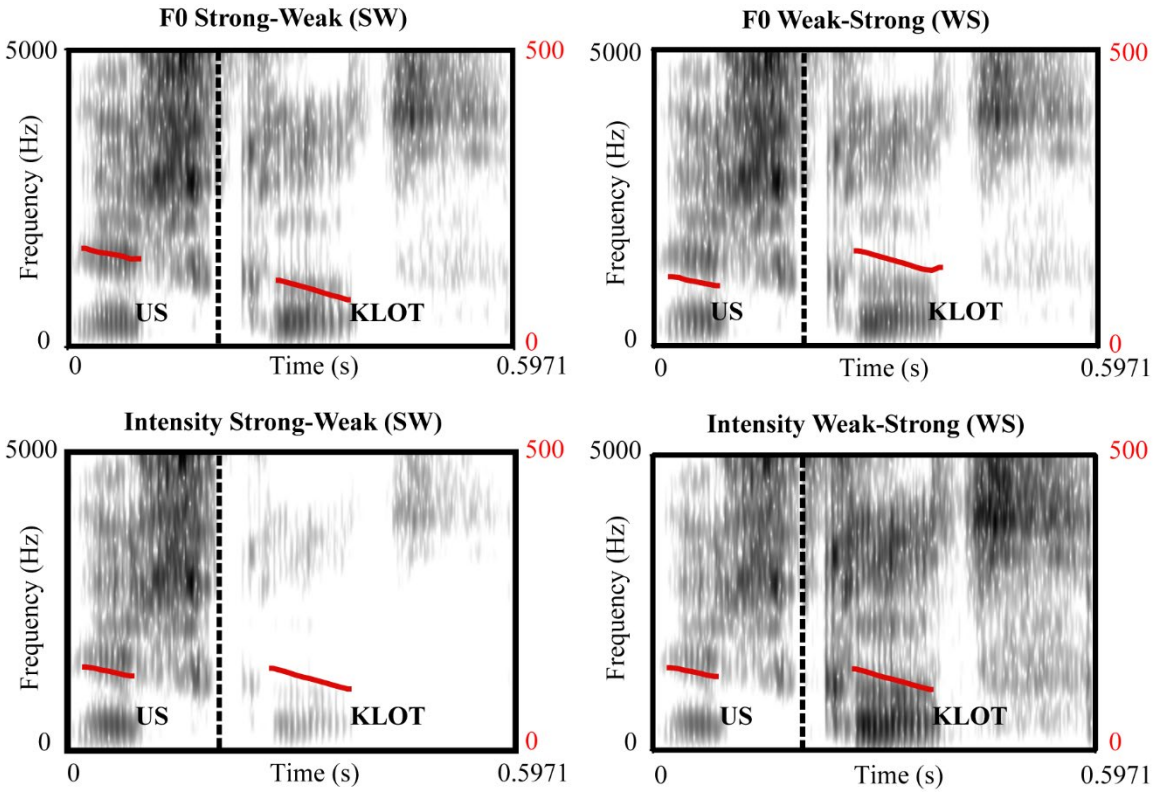

Fig. 4. Spectrograms of the most Strong-Weak (SW)-like and the most Weak-Strong (WS)-like stimuli of one non-word (*usklot*) in the two continua (varying F0, top row; varying intensity, bottom row). The red lines indicate the F0 tracks. The y-axis on the right hand side of each plot, depicted in red, represents the scale for the F0 tracks.

4.2.2. Carrier sentences and visual materials

In addition to the target non-words, we needed various carrier sentences for the training phase and the test phase. In the training phase, we used *Dit is een...,* "This is a..." containing only monosyllabic words, presented on 50% of the trials. In addition, we also presented *Dit object is een voorbeeld van een...,* "This object is an example of a ..." on the rest of the trials, including two disyllabic words, one with initial stress *VOORbeeld* and one with final stress *obJECT*. To increase exposure to talker-specific prosodic cues, we manipulated the suprasegmental cues in these two disyllabic words in line with the particular talker-cue mappings. That is, if Talker A used F0 to cue lexical stress on the target non-words, then Talker A also produced *obJECT* and *VOORbeeld* using F0. F0 and intensity values for *VOORbeeld and obJECT* were derived from the ones used for the non-words (see Supplementary Information for more details). Furthermore, we recorded feedback sentences for the training phase (*Goed, dit is een... / Fout, dit is een...,* "Right, this is a..."/ "Wrong, this is a...").

For the test phase, we needed semantically constraining sentences that allowed for prediction of the sentence-final word (*Het woord voor lamp is een USklot,* "The word for lamp is an USklot"). We thus recorded the carrier sentence ("The word for ... is a ...") and the objects ("lamp") separately, and spliced the objects as well as the sentence-final item into the carrier words. We avoided any lexical stress cues in these sentences (i.e., only monosyllabic words) since we desired participants to predict talker-specific word-forms based on previously learned knowledge acquired in the training phase, instead of based on cues that were present in the test sentence itself. Hence, the words referring to the objects (e.g. *lamp*, "lamp") were all monosyllabic words (see Supplementary Table S6 for complete list).

Lastly, sixty-four colored line drawings of the monosyllabic objects were selected from the Multilingual Picture (MultiPic) databank (Duñabeitia et al., 2018). These pictures were used as visual references for the objects during the training and testing phase. We attempted to minimize phonological as well as semantic overlap between the Dutch labels of the objects. Lastly, we selected colored line drawings of two standing men from the MultiPic databank which would be used to visually cue the two talkers' identities.

4.3. Procedure

The experiment consisted of a training phase (divided over three sessions) and a test phase. To keep the experimental sessions as short as possible, we divided the sessions over three consecutive days (see Fig. 1). The learning process also benefited from this choice since newly learned spoken words are subject to overnight consolidation (Dumay and Gaskell, 2007).

Following the study by Sulpizio and McQueen (2012), participants performed a series of 2AFC tasks and typing tasks during the training phase. These tasks were designed for participants to learn the item-to-object mappings *explicitly* while also learning the talker-specific cues *implicitly* (i.e., without explicit instructions about the cues). After the final training session, participants were immediately tested on the items they had learned during training while we recorded behavioral responses and EEG.

On the first two days, participants were seated in front of a 326 mm × 244 mm sized (Max Planck Institute for Psycholinguistics; MPI) or a 509 mm × 206 mm (Donders Centre for Cognition; DCC) monitor and audio was presented through Sennheiser HD-250 (MPI) or Sony MDR-7506 (DCC) headphones at a fixed comfortable level. On the last day, participants were seated in front of a 337 mm × 270 mm (MPI) or a 531 × 299 (DCC) sized monitor and audio was presented through Canton (MPI) or AudioEngine A2 (DCC) speakers.

4.3.1. Training phase.

Before the start of the experiment, participants were instructed that they would be learning words from an unknown language. Additionally, they were instructed to pay attention to the non-words being minimal pairs of lexical stress (i.e., we stated that just as the Dutch words *CAnon* ("canon") and *kaNON* ("cannon"), the meaning of the members of the pairs depended on which syllable was stressed), as well as to the pronunciation of the two talkers (i.e., that both talkers produced these words in their own way, without explicitly mentioning that this concerned prosodic cues). Before the first block, participants received four practice trials with items that were not included in the experimental list.

Each item was paired with one particular object. To avoid any potential effects of item-specific or cue-specific learning difficulties (e.g., due to some item-to-object mappings or some cue-talker combinations being harder to learn than others), half of the participants were tested on a second experimental stimulus list in which the item-to-object mappings were reversed within each minimal pair (e.g. a second list in which *usKLOT* would refer to "lamp" and *USklot* to "train"*)* and the cue-talker mapping was switched (e.g., Talker A using intensity instead of F0 and *vice versa* for talker B).

All the tasks (except for Training Block 6) consisted of 128 experimental trials and only Training Block 1 was preceded by four practice trials with items that did not appear in the experimental stimulus list. In Training Block 6, the number of trials depended on the number

of errors made in Training Block 5 (see 2AFC tasks). Furthermore, trials were presented in a randomized order.

4.3.1.1. 2AFC tasks

In the 2AFC tasks, participants were auditorily exposed to the items in carrier sentences (e.g., Dutch versions of "This is a…" / "This object is an example of a …"), produced by both talkers, and were visually presented with two colored line drawings after the sentence had finished. They were instructed to choose which of the two line drawings was the correct referent for a particular item (see Fig. 5 for the trial structure). To emphasize which talker produced each sentence, we displayed an image of the talker surrounded by either a blue or a red square (the color was talker-specific) during the carrier sentence. Also, to ensure that participants would learn the correct label for each line drawing and not a synonym or a super- or subordinate word (e.g. "bulb" instead of "lamp"), the correct Dutch labels were presented above the line-drawings. Participants were instructed to respond with button presses (left or right) to indicate which object was the correct referent for the item. If no response was given after 4 s, the trial was recorded as a missing data point. After the response, we presented a feedback sentence (*Goed, dit is een…*, "Correct, this is a…" or *Fout, dit is een…, "*Wrong, this is a…") followed by the correct item. Also, we displayed the correct object together with the correct orthography of the item (e.g. *USklot*) on the screen, with capitals indicating lexical stress. Note that explicit orthographic feedback was given for the correct object while no feedback was given for the talker-specific prosodic cues (participants were supposed to implicitly learn the talker-specific cues). The next trial began 1 s after the feedback sentence of the previous trial.

In Training Block 1, we never presented the colored line drawings of both members of a minimal pair together, allowing participants to familiarize themselves with the segmental information of the non-words. During all the other 2AFC tasks, we did present both members of the minimal pairs together, directing the participants' attention to the suprasegmental information, which has been found to be necessary for participants to be able to learn minimal stress pairs (Sulpizio and McQueen, 2011). In Training Block 6, participants completed a *conditional* 2AFC task in which we presented only the items on which participants made a mistake during Training Block 5. For each item on which participants had made a mistake, they received both versions of the minimal pair (e.g. *USklot and usKLOT*) spoken by both talkers. In all 2AFC tasks, participants would hear each item four times; once in the carrier sentence and once in the feedback sentence, for both talkers.

4.3.1.2. Typing task

In the typing task, participants were presented with a line drawing of one of the objects and were instructed to type out the correct item. The aim of the typing task was for participants to retrieve the item cued by the object (which was also close to the implicit prediction task at test). Since a spoken production task could lead to interference from the prosodic cues in those self-produced spoken productions, we decided to use a typing task. Every trial started with a fixation cross in the middle of the screen. After 500 ms, participants were presented with a line drawing of one of the objects for 2 s. Afterwards, participants were instructed to type out the correct item of that object, critically with the stressed syllable being capitalized. Accuracy was assessed by comparing the response to the case-sensitive correct string for each trial. Additionally, to adjust for small typing errors, the incorrect responses were checked during data preparation and the accuracy scores were adjusted if the intended answer was correct. As in the 2AFC tasks, participants heard feedback sentences together with the correct object and the correct label was displayed after their response. The next trial began 1 s after the feedback sentence of the previous trial.

4.3.2. Test phase

After the final training phase and the electrode preparation for the EEG session, the test phase started in which participants were tested on the items they had learned in the training phase (see Fig. 5 for trial structure). Participants were auditorily presented with semantically constraining sentences containing the target non-words in sentence-final position (e.g. *Het woord voor lamp is een USklot, "*The word for lamp is an USklot"). Before the carrier sentence, we presented a line-drawing of the talker surrounded by a colored square in the middle of the screen for 500 ms. Afterwards, we presented the carrier sentence and also displayed the line-drawing of the object in the sentence (e.g., lamp) in the middle of the screen. At target-word onset, participants were instructed to respond to whether the meaning of the sentence-final word matched the sentence (right button for a correct word, left button for an incorrect word). If no response was given after 4 s, the trial was recorded as a missing data point. After the response (or time out) a blank screen was displayed for 200 ms followed by a 1.5 s window during which participants could blink (cued by four asterisks on the screen). The next trial (starting with the fixation cross) began immediately after this window.

The experimental stimulus list consisted of two test blocks, with 128 trials in each test block (32 trials per condition). In the first test block, we randomly selected half of the non-words for which we presented the SW version of the minimal pair (e.g., *USklot*) and for the

other half of the selection we presented the WS version. In the second test block we presented the other pair such that participants eventually received both pairs of all the non-words.

Recall that we assessed performance on all four conditions (control, cue-switch, word-switch, and stress-switch). To minimize potential effects of unlearning of the talker-specificity of the cues (and affect performance at test), the sentences in the word-switch, stress-switch and the control condition all contained the correct cues for a specific talker. This ensured that the proportion of trials on which participants experienced unexpected talker-incongruent cues to lexical stress amounted to 25% of all the trials.

Furthermore, we wanted to minimize the effects of the different experimental conditions on the participants' representations of the learned items (i.e. only hearing incorrect versions of the item at test, e.g., *BOLdep* for 'lamp', might confuse participants). To achieve this, the trials were presented in a fixed order within items in each test block. In the first block, the order of presentation contained a correct version (i.e., the correct member of the minimal pair and produced using the talker-congruent cues) in between deviant versions of an item (i.e., either the wrong member of the minimal pair talker-congruent cues or the correct member of the minimal pair produced using talker-incongruent cues). That is, for each item participants always first received the cue-switch version (e.g., *USklot* by Talker A using intensity), followed by the control version (i.e. correct version; *USklot* by Talker A using F0) and lastly the stress-switch version (e.g., *usKLOT* using F0). The word-switch version (e.g., *BOLdep* using F0) was not included in this constraint and could thus appear anywhere. Furthermore, since the cue-switch condition was the condition that should elicit the ERP component in which we were most interested, we decided to present those sentences first. This order of presentation restriction was not applied between items (e.g. the talker-incongruent version of *BOLdep* could appear after the talker-congruent version of *USklot*) as long as it did not violate the within-item constraint.

In the second block, we adjusted the fixed order. Considering that the carrier sentence in the cue-switch and the control condition was identical, we wanted to rule out any possible amplitude modulations of the ERPs due to repetition effects (i.e., a smaller amplitude in the control condition caused by repetition of the same carrier sentence and item). For this reason, the order in which the cue-switch and the control trials were presented in the second test block was reversed (i.e., we first presented the control condition, followed by the cue-switch and the stress-switch condition, again within items). The fixed within-item order of presentation thus only applied within blocks.

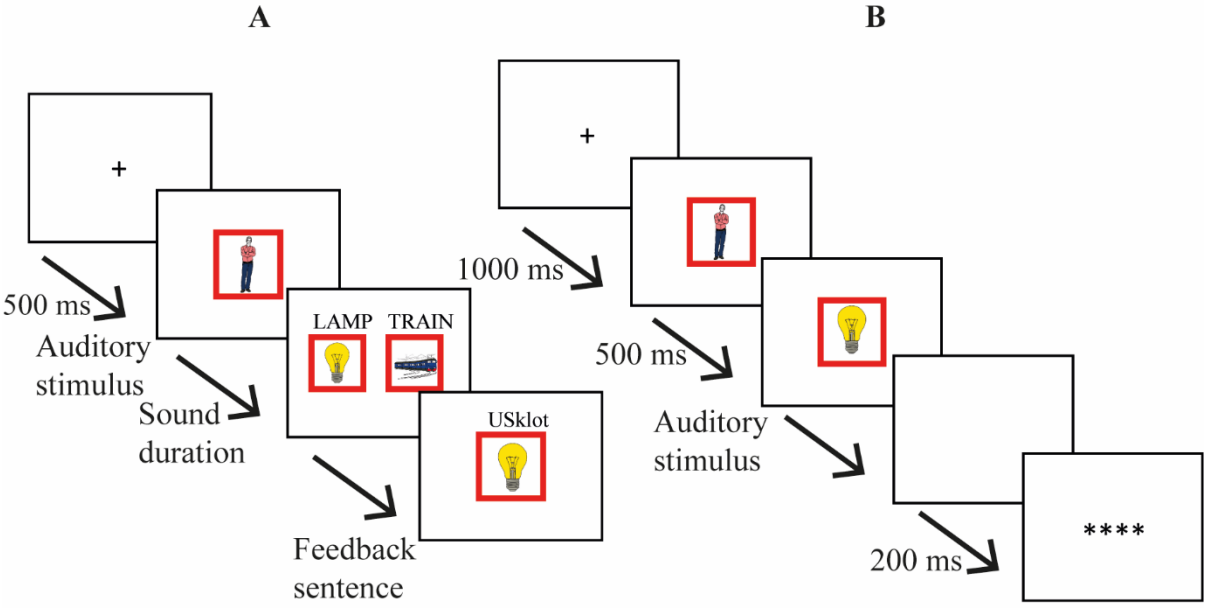

Fig. 5. A. Illustration of a trial in the two-alternative forced choice (2AFC) task in the training phase; **B.** *idem*, but for the test phase.

4.3.3. EEG recording

EEG signals were recorded using 59 electrodes on an Acticap standard 10/20 cap, amplified with a BrainAmps (Brain Products) DC amplifier (500 Hz sampling rate, 0.016-1000 Hz cut-off). We used an on-line reference placed on the left mastoid and electrooculography (EOG) was recorded from two electrodes placed at the temples, one electrode placed below the left eye and the Fp1 electrode. Impedance levels were kept below 25 kΩ.

Preprocessing and analyses were performed using the Fieldtrip toolbox (Oostenveld et al., 2011). The signal was re-referenced offline to the average of the left and the right mastoid and a low-pass filter at 30 Hz was applied. Subsequently, the signal was cut into epochs of 500 ms pre-stimulus and 800 ms post-stimulus (with the onset of the sentence-final target word taken as stimulus onset). Trials with atypical artefacts (i.e., jumps and drifts) and channels that consistently contained these atypical artefacts were rejected prior to independent component analysis (ICA). Eye blinks were removed using ICA (if the number of trials containing eye-blinks exceeded four in at least one condition). Afterwards, the channels and trials that were initially excluded were interpolated based on the weighted average of neighboring channels. Trials still containing eye blinks or noisy channels (that could not be repaired with ICA) were then eventually rejected (2.8% of the total data). Finally, we applied a baseline-correction from 500 to 0 ms before stimulus onset.

4.4 Data analysis

4.4.1. Behavioral analyses

Behavioral analyses were performed on RTs and accuracy scores during the test phase. Since participants were instructed to respond from target word onset onwards, we calculated RTs time-locked to the onset. Also, we log-transformed the RTs (to obtain a more normal distribution in number of observations) and excluded incorrect responses in the RT analysis (15.6%), which resulted in 6467 observations.

The behavioral data were analyzed using a linear mixed-effects model with the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2020). For the RT analysis, the model with the best fit to the data (as tested using log-likelihood model comparisons) contained the following factors: as fixed factors, we included Condition (categorical predictor with four levels, dummy coding with the control condition mapped onto the intercept), Cue (categorical predictor with two levels, deviation coding with intensity coded as -0.5 and F0 coded as 0.5) and Trial Number (continuous predictor that was scaled to z-scores). We finally included their interactions. We also included random intercepts for Participant and Item with by-Participant random slopes for all main effects and by-Item random slopes for Trial Number. The random structures were optimized using Principle Component Analyses (PCA) on the models to obtain the random structure that contained sufficient factors to explain the variance.

Second, we ran a Generalized Linear Mixed Model (GLMM) with a logistic linking function to test whether the accuracy of the categorization responses was different for the four conditions. The binomial dependent variable was the accuracy on the categorization response as to whether the meaning of the sentence-final word was correct given the lead-in sentence (1 for correct answers, 0 for incorrect answers). We included the same predictors as in the linear mixed-effects model we used for the RT data. Also, we added random intercepts for Participant and Item with by-Participant random slopes for Condition and Cue as well as by-Item random slopes for Cue and Trial Number. Similar to the RT models, the random structure was optimized using PCA.

4.4.2. ERP analyses

After baseline correction, we selected the trials on which participants responded correctly and computed average ERPs time-locked to stimulus onset (the sentence-final word) for each subject and condition. To assess the differences between the conditions, we performed cluster-based permutation analyses (Maris and Oostenveld, 2007). This nonparametric method

tests whether two conditions differ significantly from each other by drawing random permutations from the observed data, creating a permutation distribution of a test statistic. We took the sum of *t*-values of the largest cluster as test statistic by performing paired-samples *t*-tests on each data point. Next, we clustered adjacent time-points and electrode sites (thus controlling for multiple comparisons) of data points exceeding a threshold ($\alpha = .05$). The test statistic was then calculated by taking the sum of *t*-values of the largest resulting cluster. All the values of the test statistic that were obtained from 1000 random permutations resulted in the permutation distribution for the test statistic. Next, we calculated the *p*-value under the permutation distribution (using a Monte Carlo estimate) that informed us on the probability (under the null hypothesis that the two conditions are from the same distribution) of observing a cluster-level statistic that is larger than the observed statistic (again, based on a threshold of $\alpha = .05$). In other words, the analysis reveals whether two conditions originate from the same distribution (i.e., are interchangeable) or not while controlling for multiple comparisons.

Acknowledgements

We would like to thank Keanu Kiriwenno, whose voice was recorded for the speech materials used in the study. Also, thanks to Caitlin Decuyper and Abdellah Elouatiq who helped running the EEG session of the experiment. Parts of this work have previously been presented at Rate and Rhythm in Speech Recognition (R3, December 2019, Max Planck Institute for Psycholinguistics, Nijmegen), Architectures and Mechanisms of Language Processing (AMLaP, Potsdam, September 2020; https://amlap2020.github.io/a/41.pdf) and at Middag van de Fonetiek (December 2020; https://www.nvfw.org/content/listeners-learn-and-predict-talker-specific-prosodic-cues-speech-perception).

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The data and the stimuli of this experiment have been made openly available on https://osf.io/8h6xb/?view_only=2a2818bdc29442f38a906e62cebd6dc2.

References

Altmann, G.T.M., Kamide, Y., 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition 73, 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Boersma, P., Weenink, D., 2019. Praat: doing phonetics by computer.

Bosker, H.R., in press. Evidence for selective adaptation and recalibration in the perception of lexical stress. Lang. Speech. https://doi.org/10.1177/00238309211030307

Bosker, H.R., Sjerps, M.J., Reinisch, E., 2020a. Spectral contrast effects are modulated by selective attention in "cocktail party" settings. Atten. Percept. Psychophys. 82, 1318–1332. https://doi.org/10.3758/s13414-019-01824-2

Bosker, H.R., Sjerps, M.J., Reinisch, E., 2020b. Temporal contrast effects in human speech perception are immune to selective attention. Sci. Rep. 10, 5607. https://doi.org/10.1038/s41598-020-62613-8

Brunellière, A., Soto-Faraco, S., 2013. The speakers' accent shapes the listeners' phonological predictions during speech perception. Brain Lang. 125, 82–93. https://doi.org/10.1016/j.bandl.2013.01.007

Clopper, C.G., Smiljanic, R., 2011. Effects of gender and regional dialect on prosodic patterns in American English. J. Phon. 39, 237–245. https://doi.org/10.1016/j.wocn.2011.02.006

Connolly, J.F., Phillips, N.A., 1994. Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. J. Cogn. Neurosci. 6, 256–266. https://doi.org/10.1162/jocn.1994.6.3.256

Cutler, A., 1986. Forbear is a Homophone: Lexical Prosody Does Not Constrain Lexical Access. Lang. Speech 29, 201–220. https://doi.org/10.1177/002383098602900302

Cutler, A., 1976. Phoneme-monitoring reaction time as a function of preceding intonation contour. Percept. Psychophys. 20, 55–60. https://doi.org/10.3758/BF03198706

Cutler, A., Van Donselaar, W., 2001. Voornaam is not (really) a Homophone: Lexical Prosody and Lexical Access in Dutch. Lang. Speech 44, 171–195. https://doi.org/10.1177/00238309010440020301

Diaz, M.T., Swaab, T.Y., 2007. Electrophysiological differentiation of phonological and semantic integration in word and sentence contexts. Brain Res. 1146, 85–100. https://doi.org/10.1016/j.brainres.2006.07.034

Dilley, L.C., Pitt, M.A., 2010. Altering Context Speech Rate Can Cause Words to Appear or Disappear. Psychol. Sci. 21, 1664–1670. https://doi.org/10.1177/0956797610384743

Dumay, N., Gaskell, M.G., 2007. Sleep-Associated Changes in the Mental Representation of Spoken Words. Psychol. Sci. 18, 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

Duñabeitia, J.A., Crepaldi, D., Meyer, A.S., New, B., Pliatsikas, C., Smolka, E., Brysbaert, M., 2018. MultiPic: A standardized set of 750 drawings with norms for six European languages. Q. J. Exp. Psychol. 71, 808–816. https://doi.org/10.1080/17470218.2017.1310261

Eisner, F., McQueen, J.M., 2018. Speech Perception, in: Wixted, J.T. (Ed.), Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 1–46. https://doi.org/10.1002/9781119170174.epcn301

Eisner, F., McQueen, J.M., 2006. Perceptual learning in speech: Stability over time. J. Acoust. Soc. Am. 119, 1950–1953. https://doi.org/10.1121/1.2178721

Eisner, F., McQueen, J.M., 2005. The specificity of perceptual learning in speech processing. Percept. Psychophys. 67, 224–238. https://doi.org/10.3758/BF03206487

Gussenhoven, C., Van Der Vliet, P., 1999. The phonology of tone and intonation in the Dutch dialect of Venlo. J. Linguist. 35, 99–135. https://doi.org/10.1017/S0022226798007324

Haan, J., Van Heuven, V., 1999. Male vs. female pitch range in Dutch questions, in: Proceedings of the 13th International Congress of Phonetic Sciences. San Francisco, pp. 1581–1584.

Idemaru, K., Holt, L.L., 2011. Word recognition reflects dimension-based statistical learning. J. Exp. Psychol. Hum. Percept. Perform. 37, 1939–1956. https://doi.org/10.1037/a0025641

Jesse, A., Poellmann, K., Kong, Y.-Y., 2017. English Listeners Use Suprasegmental Cues to Lexical Stress Early During Spoken-Word Recognition. J. Speech Lang. Hear. Res. 60, 190–198. https://doi.org/10.1044/2016_JSLHR-H-15-0340

Kamide, Y., Altmann, G.T.M., Haywood, S.L., 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. J. Mem. Lang. 49, 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kraljic, T., Samuel, A.G., 2005. Perceptual learning for speech: Is there a return to normal? Cognit. Psychol. 51, 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Kurumada, C., Brown, M., Bibyk, S., Pontillo, F., Tanenhaus, M.K., 2014. Rapid adaptation in online pragmatic interpretation of contrastive prosody. Proc. Annu. Meet. Cogn. Sci. Soc. 36, 791–196.

Kutas, M., Federmeier, K.D., 2011. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). Annu. Rev. Psychol. 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., Hillyard, S., 1984. Brain Potentials During Reading Reflect Word Expectancy and Semantic Association. Nature 307, 161–163. https://doi.org/10.1038/307161a0

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. **lmerTest** Package: Tests in Linear Mixed Effects Models. J. Stat. Softw. 82. https://doi.org/10.18637/jss.v082.i13

Liu, R., Holt, L.L., 2015. Dimension-based statistical learning of vowels. J. Exp. Psychol. Hum. Percept. Perform. 41, 1783–1798. https://doi.org/10.1037/xhp0000092

Mantegna, F., Hintz, F., Ostarek, M., Alday, P.M., Huettig, F., 2019. Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. Neuropsychologia 134, 107199. https://doi.org/10.1016/j.neuropsychologia.2019.107199

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Marslen-Wilson, W., 1973. Linguistic Structure and Speech Shadowing at Very Short Latencies. Nature 244, 522–523.

McQueen, J.M., 2005. Speech Perception, in: Handbook of Cognition. SAGE, London.

Miller, J.L., Green, K., Schermer, T.M., 1984. A distinction between the effects of sentential speaking rate and semantic congruity on word identification. Percept. Psychophys. 36, 329–337. https://doi.org/10.3758/BF03202785

Nieuwland, M.S., 2019. Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. Neurosci. Biobehav. Rev. 96, 367–400. https://doi.org/10.1016/j.neubiorev.2018.11.019

Norris, D., McQueen, J.M., Cutler, A., 2016. Prediction, Bayesian inference and feedback in speech recognition. Lang. Cogn. Neurosci. 31, 4–18. https://doi.org/10.1080/23273798.2015.1081703

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cognit. Psychol. 47, 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech Perception as a Talker-Contingent Process. Psychol. Sci. 5, 42–46.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2011. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput. Intell. Neurosci. 2011, 1:1-1:9. https://doi.org/10.1155/2011/156869

Poulton, V., Nieuwland, M.S., 2019. Can you hear what's coming? An ERP study of phonological prediction.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reinisch, E., Jesse, A., McQueen, J.M., 2011. Speaking Rate Affects the Perception of Duration as a Suprasegmental Lexical-stress Cue. Lang. Speech 54, 147–165. https://doi.org/10.1177/0023830910397489

Reinisch, E., Jesse, A., McQueen, J.M., 2010. Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. Q. J. Exp. Psychol. 63, 772–783. https://doi.org/10.1080/17470210903104412

Reinisch, E., Weber, A., 2012. Adapting to suprasegmental lexical stress errors in foreign-accented speech. J. Acoust. Soc. Am. 132, 1165–1176. https://doi.org/10.1121/1.4730884

Rietveld, A.C.M., Van Heuven, V.J., 2009. Algemene fonetiek, 3rd ed. Coutinho, Bussum.

Sjerps, M.J., Zhang, C., Peng, G., 2018. Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context. J. Exp. Psychol. Hum. Percept. Perform. 44, 914–924. https://doi.org/10.1037/xhp0000504

Sulpizio, S., McQueen, J.M., 2012. Italians use abstract knowledge about lexical stress during spoken-word recognition. J. Mem. Lang. 66, 177–193. https://doi.org/10.1016/j.jml.2011.08.001

Sulpizio, S., McQueen, J.M., 2011. When two newly-acquired words are one: New words differing in stress alone are not automatically represented differently, in: Proceedings of Interspeech 2011. Florence, Italy, pp. 1385–1388.

Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. J. Exp. Psychol. Learn. Mem. Cogn. 31, 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Xie, X., Buxó-Lugo, A., Kurumada, C., 2021. Encoding and decoding of meaning through structured variability in intonational speech prosody. Cognition 211, 104619. https://doi.org/10.1016/j.cognition.2021.104619

Zhang, X., Holt, L.L., 2018. Simultaneous tracking of coevolving distributional regularities in speech. J. Exp. Psychol. Hum. Percept. Perform. 44, 1760–1779. https://doi.org/10.1037/xhp0000569