The Role of Talker-Specific Prosody in Predictive Speech Perception

Giulio G. A. Severijnen[1]
Supervisors: Vitória Piai[1], James M. McQueen[1]

[1]Radboud University Nijmegen, Donders Centre for Cognition, The Netherlands
[2]Max Planck Institute for Psycholinguistics, The Netherlands

One of the challenges in speech perception for listeners is to deal with the huge segmental and suprasegmental variability in the acoustic signal between different talkers. Most studies have focused on how listeners deal with segmental variability. In this electroencephalography (EEG) experiment, we investigated how listeners learn about variability in suprasegmental cues between talkers to recognize spoken words. Participants learned non-word minimal stress pairs (e.g., *USklot*/*usKLOT*), and objects to which the non-words referred (e.g., the item *USklot* referring to a lamp, the item *usKLOT* referring to a train). These non-words were produced by two different talkers and each talker only used one acoustic cue to signal lexical stress patterns (e.g., Talker A only used F0 and Talker B only used amplitude). This allowed participants to learn the correct item-to-object mappings as well as, through perceptual learning, which cues were used by each talker. At test, participants heard semantically constraining sentences, spoken by either talker, containing these non-words in the sentence-final position. The sentence-final word could either be produced using the correct cues (e.g., Talker A using F0; control condition) or the incorrect cues (e.g., Talker A using amplitude; cue-switch condition). If participants learned about the talker-specific cues, they would be able to predict upcoming talker-matching word-forms (e.g., *USklot* cued using only F0). We hypothesized that the sentences in the cue-switch condition would lead to longer RTs and elicit a relatively larger N200 response compared to the control condition. Results showed that the sentences in the cue-switch condition indeed led to longer RTs compared to the control condition. This suggests that these sentences created a mismatch between predicted and perceived word-forms based on the talker-specific cues. In contrast, the N200 amplitude was not modulated by these sentences. We conclude that these results illustrate talker-specific prediction of suprasegmental cues, picked up through perceptual learning on previous encounters.

Keywords: individual differences, prosody, talker-specific learning, prediction, ERP

Corresponding author: Giulio G. A. Severijnen; E-mail: g.severijnen@donders.ru.nl

One of the challenges in speech perception is that listeners must deal with the huge variability in the acoustic signal between different talkers. 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 recognise different utterances. In the current study we assess how listeners deal with this issue. More specifically, we look into whether listeners learn about variability in suprasegmental cues between talkers and use that talker-specific knowledge to recognise spoken words.

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). This acoustic signal consists of both segmental information (such as individual vowels and consonants) and suprasegmental information (such as lexical stress, prosodic focus, etc.) that signal prosodic structures beyond vowels and consonants. As Eisner and McQueen (2018) point out, speech perception is about combining both sources of information to recognise spoken words. To shortly illustrate this, consider the phrase "The stranger objects" and the minimal pair that is present in it: "OBject" and "obJECT" (capitalization indicates lexical stress). Depending on which member of the minimal pair is perceived, the interpretation of the phrase changes (i.e., "The stranger OBjects" with the nominal meaning "the more unusual OBjects" or "The stranger obJECTS" with the verbal meaning "the unknown person objects"). In order to correctly perceive one of these words (and the phrase), listeners must use both information about the vowels and consonants and suprasegmental information to perceive the correct sentence (i.e., ignoring either segmental or suprasegmental information would impede correct recognition).

One of the factors that complicates word recognition is the large variability in the acoustic signal. This variability is caused by several factors, such as coarticulation (i.e., preceding and following sounds influence the realisation of a given phoneme) and the position of words and sounds in the prosodic structure. Furthermore, also the focus of the current study, the speech signal varies due to individual differences between talkers, such as the talker's sex, age and speaking style (McQueen, 2005). This between-talker variation taxes listeners' comprehension efforts: When participants are exposed to multiple talkers in a word identification and naming task, identification and naming latencies for these participants are slower compared to those for participants who only hear one talker (Mullennix et al., 1989).

The challenge for listeners is thus to take talker variabilities into account and use talker-specific information in order to still correctly perceive spoken words. That is, while variability in the acoustic signal might lead to difficulties in speech perception, several studies have found that listeners are also able to exploit talker-specific information that helps them in speech perception. For example, Nygaard, Sommers, and Pisoni (1994) found that word identification in noise was higher when participants heard voices that they had previously been trained on compared to participants who heard a new set of untrained voices. They concluded that listeners who learned to recognise a set of talkers encoded talker-specific information that facilitated subsequent perceptual analysis of words produced by the same talker. But it is not merely the ability to recognise a voice that facilitates subsequent perceptual analysis. As demonstrated by Norris, McQueen, and Cutler (2003), listeners can adapt to talker-specific pronunciations by using lexical information to alter their phonetic categories. Norris et al. (2003) exposed Dutch participants to Dutch words containing an ambiguous word-final fricative between /f/ and /s/ and found that participants adjusted their phonetic categories of /f, s/ according to whether this ambiguous sound appeared in /f/-final words (such as *witlof*, "chicory") or in /s/-final words (such as *naaldbos,* "pine forest"). That is, exposure to the ambiguous sound in /f/-final words biased perception in a subsequent test on an /f-s/ continuum towards /f/, while exposure to the ambiguous sound in /s/-final words biased perception towards /s/. Eisner and McQueen (2005) followed the same design as Norris et al. (2003) and applied this to different talkers. They found that when participants were tested in a categorization task on stimuli in which both the vowels and fricatives originated from a novel talker, the perceptual learning effects concerning ambiguous fricatives did not generalize to that new talker, illustrating talker specificity of perceptual learning.

In contrast to these effects being due to adaptation to the acoustic input, Zhang and Holt (2018) illustrated that these perceptual learning effects are adaptation to talker's speaking styles. In their first experiment, they exposed English participants to English minimal pairs (beer-pier) with manipulated fundamental frequency (F0) and

voice onset time (VOT) values, and measured the proportion of pier-responses (/p/ is normally signalled by a high F0 and a long VOT). Participants were divided into two groups. The first group was exposed to training stimuli in a low F0 range (stimuli containing an F0 of either 130 Hz or 200 Hz) and the second group was exposed to training stimuli in a high F0 range (stimuli containing an F0 of either 200 or 300 Hz). The critical comparison was based on test stimuli that contained an ambiguous VOT value (i.e., participants had to base their responses on F0 values) and an F0 of 200 Hz. Importantly, this F0 was identical in both groups, the only difference was the range in which the stimuli appeared. Results indicated that the proportion of pier-responses was modulated by this frequency range (i.e., a low frequency range elicited more pier-responses since the 200 Hz was perceived as relatively higher). In two additional experiments following this design, they presented all stimuli at 200 Hz but induced the frequency range based on voice quality (i.e., spoken by either a male or a female talker) or visual presentation of a male or a female talker. Results showed that the proportion of pier-responses was again modulated by these manipulations. 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 voice quality and visual talker identification which allows listeners to rapidly adapt phonetic categories based on talker-specific information.

Another mechanism that can help listeners deal with talker variability is prediction. Several studies have already shown that listeners use prediction in speech perception. For example, Marslen-Wilson (1973) found that participants made errors in a sentence shadowing task, that were congruent with the preceding syntactic and semantic context of that sentence. Also, the semantic properties of a precursor sentence can affect the identification of subsequent acoustically ambiguous stimuli (Miller et al., 1984). Finally, Cutler (1976) illustrated in a phoneme monitoring task that response times (RTs) to an initial phoneme of a target word were modulated by whether the target word was predicted to be stressed or unstressed (based on the intonation in the preceding context), illustrating prediction of prosodic structures. Prediction can also help listeners 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. Listeners use prediction both at the lexical level (Van Berkum et al., 2005) and the prelexical level (Brunellière & Soto-Faraco, 2013).

At the prelexical level, Brunellière and Soto-Faraco (2013) found that listeners used information about a talker's 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 ([permis]). In these sentences, the critical word containing the possible vowel reduction (*permís*) always occurred in a sentence-final position (e.g., *És una família molt estricta, abans d'aixecar-te de la taula has de demanar permís*, "It is a very strict family, before getting up from the table, you have to ask permission", allowing for prediction of the sentence-final word. In some of these sentences, however, the sentence-final word contained a mismatch between the expected and the actual phonetic realisation (i.e., an Eastern Catalan talker producing [permis] without vowel reduction, or vice versa). These mismatched sentences elicited a relatively larger N200 response, an event-related potential (ERP) reflecting acoustic-phonetic processing in the phonological stage of word processing (Connolly & 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 context.

Taken together, these two mechanisms, perceptual learning and prediction, can help listeners deal with talker variability. First, listeners can adapt their phonological representations for specific talkers through perceptual learning cued by auditory and visual identification of a talker. Second, based on these altered talker-specific representations, 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, prosodic variation in Dutch has been found to be affected by dialects (Gussenhoven & Van Der Vliet, 1999) and sex-related differences (Haan & Van Heuven, 1999). It remains unclear however, how listeners deal with variability in suprasegmental information.

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 the study by 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*) their response times 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. Furthermore, Reinisch, Jesse, and McQueen (2010) showed that participants immediately use suprasegmental stress information to recognize spoken words as soon as it becomes available. In an eye-tracking study, they exposed Dutch listeners 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 recognise spoken words. The same effect has also been found in English listeners for primary-stress words (Jesse et al., 2017) and in Italian listeners (Sulpizio & McQueen, 2012). In fact, Sulpizio and McQueen (2012) showed that Italian listeners even use abstract knowledge about stress patterns in Italian in newly learned non-words. In an eye-tracking study, participants learned reduced-cue versions (i.e., duration and amplitude were set to ambiguous values) of trisyllabic non-words (*TOlaco/toLAco*). Then, participants were tested on these learned non-words and results showed that participants fixated penultimate-stress target words (which is the default in Italian) more often as compared to antepenultimate-stress target words. Furthermore, performance on the antepenultimate-stress target words (but not for the penultimate-stress target words) was increased when participants were presented with full-cue versions as compared to reduced-cue versions. Since participants were exposed to reduced-cue versions of both stress patterns at training, this result could not be explained through different episodic experiences. Instead, the improved performance on the full-cue versions was due to abstract knowledge about the stress cues that normally signal antepenultimate stress words that participants applied to the newly learned non-words. The authors concluded that Italian listeners have knowledge about stress patterns in Italian (i.e., that the penultimate stress pattern is default) as well as about the acoustic cues that normally signal words without the default stress pattern.

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 speaking rate in a preceding context can affect the perception of lexical stress (Reinisch et al., 2011). Considering the effect of suprasegmental variability on the perception of lexical stress (and consequently word recognition), it is important to find out how listeners deal with this variability. The current study was thus concerned with the following question: How do listeners use information about talker-specific suprasegmental cues that signal lexical stress in predictive speech perception?

To address this question, we investigated whether listeners learned about how different talkers use different acoustic cues to signal lexical stress (prosodic cues) and whether listeners used this information on subsequent encounters with those same talkers to predict talker-matching word-forms. However, we were first required to create and select the stimuli that would be used. In Experiment 1, we thus created a set of disyllabic non-word minimal stress pairs (e.g., *USklot* vs. *usKLOT*), produced by two different talkers, and we manipulated the suprasegmental cues that signal lexical stress in Dutch (F0, amplitude and duration; Rietveld & Van Heuven, 2009) to create cue-specific continua in which only one cue was altered while the others were set to ambiguous values. These stimuli were tested in a categorization experiment.In Experiment 2, participants performed a word-learning experiment in which they were tested on those learned words to find out whether listeners indeed predicted talker-matching word-forms.

Experiment 1

The objective of Experiment 1 was to select the stimuli that would be used for Experiment 2. For this purpose, we created an initial set of 38 minimal pairs of non-words (e.g., *USklot* vs. *usKLOT*). For

Experiment 2, we only used two out of the three cues (i.e., one talker-specific cue for each talker) to avoid a too complicated design. Considering the effects of variation in duration on ERPs in Experiment 2, we decided to drop duration as a talker-specific cue. Instead, we set duration to ambiguous values in all the stimuli. Thus, we created – for both talkers, for both cues (F0 and amplitude) – 7-step continua from a trochaic stress pattern (Strong-Weak; SW) to a iambic stress pattern (Weak-Strong; WS) using only one of the three cues while the other two were set to ambiguous values. Based on the results of a categorisation experiment, we selected the best version along the continua that would serve as a SW item[1] and the best version that would serve as a WS item, for two prosodic cues (i.e., F0 and amplitude, with duration set to ambiguous values), for each talker.

Method

Participants

Thirty native Dutch speakers from the Radboud University in Nijmegen were recruited from the SONA participant pool, aged between 18 and 57 (8 male, 22 female, M_{age} = 24.7, SD_{age} = 8.70) All participated with informed consent and received course credits for their participation. None of the participants reported having any hearing problems.

Materials

The stimuli consisted of 38 minimal disyllabic non-word pairs (see Appendix) 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 Dutch talkers: once with stress on the first and once with stress on the second syllable.

We used the recordings to measure three prosodic cues, that have been shown to be strong perceptual cues to lexical stress in Dutch (F0, amplitude and duration; Rietveld & Van Heuven, 2009). These were calculated for both syllables, with and without lexical stress, separately for both talkers (Table 1), and across both talkers for acoustic manipulation of the stimuli (Table 2). The data in Table 1 provide a small illustration of talker variability. Based on the averaged production data across both talkers, we derived perceptually ambiguous values for each prosodic cue by calculating average values for the first and second syllable separately (Table 2). We applied these ambiguous settings using PSOLA in Praat (Boersma & Weenink, 2019) to recordings from SW members of all pairs from both talkers. The resulting sounds were acoustically ambiguous in lexical stress and were taken as midpoint stimuli of all the acoustic lexical stress continua.

As mentioned earlier, we decided not to include duration as a talker-specific cue in Experiment 2 based on acoustic duration measurements. Hence, a duration-continuum was also not included in Experiment 1 (i.e., the duration issues would persist regardless of perceptual data of Experiment 1). Still, we needed to set duration to an ambiguous value since it could inform listeners about stress patterns. For each non-word separately, we chose the most ambiguous duration value based on evaluation by the first and second author of the current study,

1 Item is used to refer to one of the members of a minimal pair (SW or WS). Non-word is used to refer to (both members of) one minimal pair.

Table 1. Acoustic measures of prosodic cues for two talkers separately.

	Strong-Weak (SW)		**Weak-Strong (WS)**	
First syllable	**Talker 1**	**Talker 2**	**Talker 1**	**Talker 2**
Duration (ms)	233	276	176	193
F0 (Hz)	119	171	116	131
Amplitude (dB)	67	72	65	66
Second syllable				
Duration (ms)	360	392	397	407
F0 (Hz)	94	121	104	152
Amplitude (dB)	58	65	62	68

Table 2. Mean acoustic measures and step sizes (across talkers) of prosodic cues for both syllables.

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

Note. Two duration values are provided as being ambiguous. These correspond to the two different values used for different subsets of non-words. 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.

which resulted in different ambiguous values for two subsets of the stimuli (one set with 202 ms [first syllable] and 395 ms [second syllable], and the second set with 288 ms [first syllable] and 362 ms [second syllable]; see Appendix) and set the duration of all the stimuli to these values.

Starting from these ambiguous stimuli, one 7-step continuum for each remaining prosodic cue (F0, amplitude) was created using Praat ranging from most SW-like (step 1) to most WS-like (step 7), with the ambiguous stimulus in the middle. In each continuum, one prosodic cue was altered while all other cues were kept constant. For each cue, we calculated a step size (separately for the first and second syllable) based on the difference between the average values of stressed vs. ambiguous syllables. The resulting endpoints of the continua were subsequently auditorily evaluated by the first and second author. Manipulations were performed in an inverse manner in the two syllables: To create SW stimuli, we increased the value of the first syllable by its step size and decreased the value of the second syllable by its step size (and did the opposite to create the WS stimuli).

Furthermore, to reduce between-item variability in F0 contours and at the same time preserve F0 declination, we replaced the original F0 contours with linear F0 contours containing an F0 declination. Note that we did not include the F0 declination as a variable but made these contours to make the stimuli sound as natural as possible.

We calculated a plausible magnitude for the F0 declination based on the average maximum and minimum values in each syllable, which yielded a mean F0 declination value for the first and second syllable separately (syllable 1: 23.5 Hz, syllable 2: 38.2 Hz). We then took the mean F0 values in each syllable as midpoint and added half of the mean F0 declination value to obtain the maximum F0 value (set at the first F0 point in the syllable), and vice versa for the minimum F0 value (set at the last F0 point in the syllable). We interpolated the other F0 points, yielding a linear contour between the maximum and minimum value. Spectrograms of the most SW-like and the most WS-like stimuli for one non-word are depicted in Figure 1.

Procedure

Participants were seated in front of a 531 mm 299 mm computer screen. Audio was presented through SONY MDR-7506 headphones at a fixed comfortable level. Every trial started with a fixation cross in the middle of the screen. After 500 ms, a fixed lead-in sentence was presented (*Het woord voor muis is een...,* "The word for mouse is a...") with the target word at the sentence-final position. At sound offset, two response options appeared on the screen (e.g., *USklot vs. usKLOT*; with the stressed syllable capitalized). The response options always appeared in the same place, such that the non-word with word-initial stress always appeared on the left and the

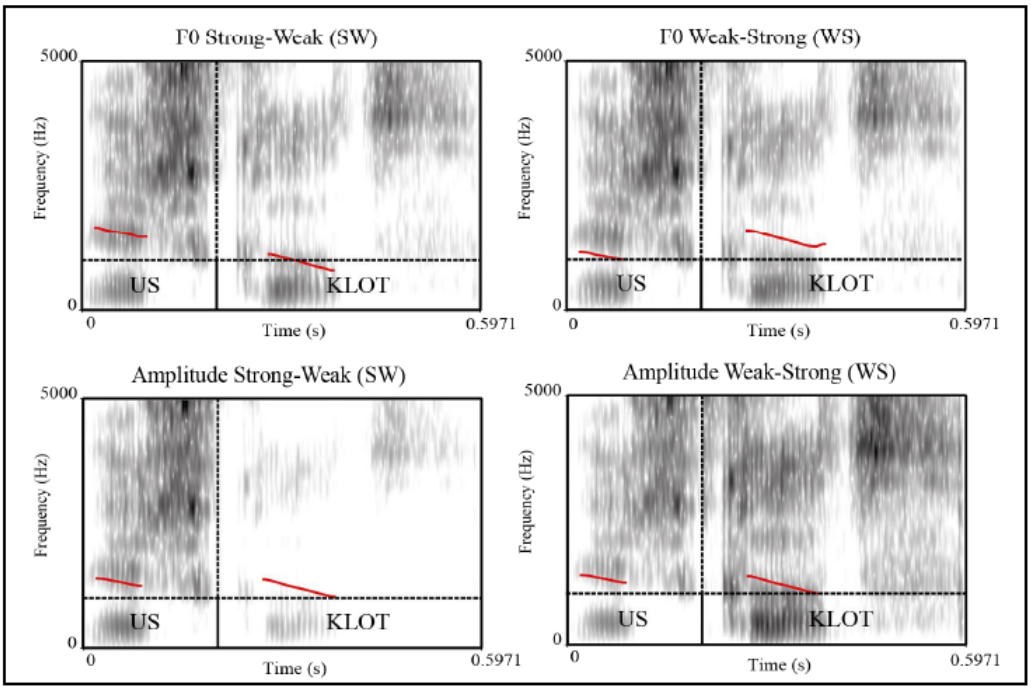


Figure 1. Spectrograms (containing the fundamental frequency (F0) tracks) of the most Strong-Weak (SW)-like and the most Weak-Strong (WS)-like stimuli of one non-word in the two continua.

non-word with word-final stress on the right. If no response was given after 4 s, the trial was recorded as a missing data point. The next trial began 1 s after the response (or time out) of the previous trial. Furthermore, items were presented in a randomised order. To keep the experimental session as short as possible, we divided the non-words into three lists that were rotated across participants. Each participant would thus be presented with every possible version along the continua – for both talkers, for both cues – of a subset of the non-words.

Participants were instructed to listen carefully to the sentences and to respond by pressing a button (left button for the left word, right button for the right word), indicating whether they heard word-initial or word-final stress on the last word of the sentence. Participants received four practice sentences, followed by the experimental trials.

Results

We recorded the classification responses (i.e., SW or WS) on each trial and calculated the mean percentage of SW responses for each step on the continua. Analyses were based on 10640 observations (10 observations for each step on one continuum, separated by talker and cue). Overall percentages (Fig. 2) showed that the number of SW responses decreased (i.e., the number of WS responses increased), with higher steps on the continua (Step 1; SW: M = 79.1%, Step 4; ambiguous: M = 55.1%, Step 7; WS: M = 31.9%). These results indicated that, overall, the acoustic manipulations had the intended effect of creating SW, WS and ambiguous versions of the stimuli.

Next, we calculated the mean percentages of SW responses on each step, for each non-word separately. Results showed that while the mean percentages showed the intended decrease in SW responses, the maximum and minimum values also show variability within the same steps between different non-words (Table 3). Based on the mean percentages for each non-word, six non-words were excluded because they failed to show a perceptual switch across the continua (see Appendix).

Lastly, we required two tokens of each non-word to be used in Experiment 2 as clear SW and WS items for each cue (F0 and amplitude). Moreover, the selected items were required to fulfil two criteria. First, the SW and WS items had to be most distinct from each other for each non-word. Second, the items had to be comparable in perception between both continua (e.g., SW cued by F0 had to be comparable to SW cued by amplitude) and between talkers (e.g., the SW token of Talker A should be comparable to the SW token of Talker B). Since the percentages showed large variability between non-words (Table 3), we made non-word specific selections instead of choosing the same two steps across all non-words. We thus recalculated the previous mean percentages for each non-word, separated by talker and by cue, and for each individual non-word, we manually

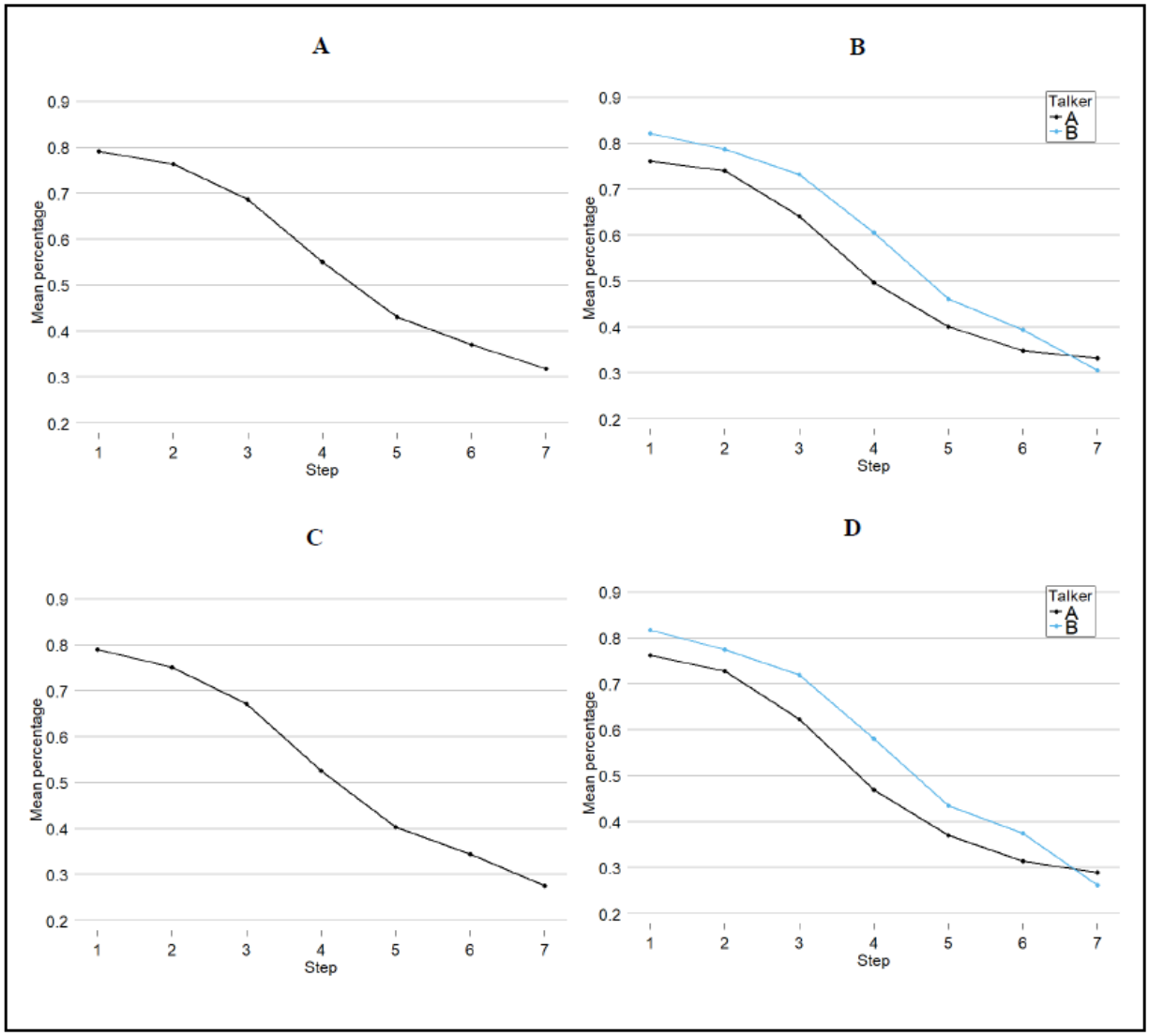


Figure 2. Mean percentages of SW categorization responses on each step across both continua, based on the entire dataset (Figure 2A) and based on the selected non-words (Figure 2B).

inspected the categorization curves and selected the best tokens.

As a final check, we calculated the mean difference of the percentages between the selected SW items and the WS items (M_{dif} = 51.0%, *SD* = 16.8) as well as the mean percentages of SW responses of the selected SW items and the WS items. To check whether this was comparable across talkers and cues, these were calculated separately for each cue, for each talker (see Table 4). The results indicated that overall, there was a clearly perceivable difference between the SW items and the WS items, while the variation between talkers and cues was kept to a minimum.

Discussion

In Experiment 1, we manipulated prosodic cues to create stimuli in which stress patterns were signalled by only one cue (i.e., only using F0 or amplitude while other acoustic cues to stress were set to ambiguous values), and tested these stimuli in

Table 3. Minimum, mean and maximum percentage of Strong-Weak (SW) responses on each step.

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
Max	97.5	92.5	95.0	95.0	82.5	67.5	75.5
Mean	79.1	76.4	68.6	55.1	43.2	37.1	31.9
Min	65.0	45.0	47.5	25.0	20.0	10.0	10.0

Table 4. Mean percentages of Strong-Weak (SW) responses of the selected stimuli.

	SW	WS
Talker 1		
Amplitude	76.9	26.9
F0	78.8	30.0
Talker 2		
Amplitude	81.3	28.4
F0	81.6	28.8

a categorization task. The categorization responses informed us on several issues.

First, the overall percentages showed that the manipulations had the intended effect of creating SW and WS versions by only changing one prosodic cue. This indicated that even when Dutch listeners only had one cue available (since the other two cues were set to ambiguous values), this one cue was enough to pick up on lexical stress patterns, which was crucial for Experiment 2. Second, related to the stimulus selection for Experiment 2, the mean percentages of SW responses on each step showed that there was large variability between non-words. This indicated that even though the manipulations were successful overall, the effects were highly dependent on the segmental information of the non-words. For this reason, we opted for a non-word specific selection of the appropriate steps along the continua, instead of selecting the same steps across all non-words.

Despite this variability, we managed to select 32 minimal pairs that met the two criteria for selection (a large perceivable difference between the SW and WS tokens, which was comparably different across cues and talkers), as illustrated by overall mean percentages of SW responses of the selected stimuli. We used the selected stimuli in Experiment 2 to find out if Dutch listeners can learn about how different talkers use prosodic cues differently and how listeners use this information in predictive speech perception on subsequent encounters with that same talker.

Experiment 2

In Experiment 2, we used the stimuli from Experiment 1 in a word-learning experiment, in which the novel words were signalled using talker-specific cues. This means each talker would signal lexical stress patterns using only one cue (e.g., Talker A only using F0 and Talker B only using amplitude). We then tested whether listeners would learn about the talker-specificity of the suprasegmental cues, and if they would use this newly acquired information in predictive speech processing on subsequent encounters with the talkers producing those words. The experiment consisted of a training phase and a test phase.

In the training phase, participants were exposed to the minimal pairs, produced by both talkers in a series of two-alternative forced choice (2AFC) and typing tasks. The goal of these tasks was for participants to learn two kinds of information. First, the tasks were designed to teach participants all the item-to-object mappings (e.g., that an *USklot* was the word for "lamp" and *usKLOT* was the word for "train"). Second, since participants heard both talkers producing these items, they could learn which prosodic cue was used by either talker to signal lexical stress (i.e., learn that Talker A only used F0 while Talker B only used amplitude).

For the test phase, we followed the design used in the study by Brunellière and Soto-Faraco (2013). Participants were exposed to semantically constraining sentences, allowing for prediction of the sentence-final word (e.g., "The word for lamp is *USklot*") while behavioural responses and their electroecephalography (EEG) data were recorded. We predicted that if participants had learned the correct item-to-object mappings and the talker-specific cues, they would be able to predict talker-matching word-forms (i.e., the correct sentence-final word and which cues would be used by the talker to signal its stress pattern). The experiment contained several conditions that differed in the sentence-final target word (Table 5). 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, there was a cue-switch condition which still contained the correct critical item, produced by the same talker, but using the wrong cues (e.g., *USklot* for "lamp" by Talker A using amplitude). Third, a stress-switch

Table 5. Different conditions in Experiment 2.

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	Amplitude	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

Note. Only Talker A is being depicted in Table 5 even though participants hear both talkers (so the same conditions hold for Talker B). Also, '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, which contained the wrong member of the minimal pair but produced using the correct cues (e.g., *usKLOT* for "lamp" by Talker A using F0). Finally, a word-switch condition containing one of the other learned items (e.g., *BOLdep* for "lamp" by Talker A using F0). Importantly, the cue-switch condition never contained a semantic incongruency (the sentence-final word in the cue-switch condition only differed in which cues were used to signal lexical stress) while the stress-switch and the word-switch condition never contained any cue incongruency (the sentence-final word was always produced using the correct prosodic cues for a given talker).

We had two primary hypotheses. First, we hypothesized that the sentences in the cue-switch condition would create a mismatch between the predicted word-forms (i.e., the talker-matching word-forms) and the perceived word-forms. This would lead to longer RTs and, 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 based on the suprasegmental cues. This would indicate that participants learned about the talker-specific cues and used this information in predicting upcoming speech on subsequent encounters.

Second, we hypothesized that since the stress-switch and the word-switch conditions contain a semantic mismatch between the predicted[2] and perceived sentence-final words, these would elicit a relatively larger N400 response as compared to the control condition. As reported by Kutas and Hillyard (1984), the N400 is an ERP reflecting the semantic relationship between a word and the context it appears in. Concerning RTs, we did not have any specific predictions for these conditions. On the one hand, RTs could increase compared to the control condition since the mismatch between the sentence and the sentence-final word causes slowing down of the response. On the other hand, RTs could also decrease in the word-switch condition: The decision to reject an incongruent

2 Note: The current design does not allow us to distinguish between prediction and integration accounts of the N400. Still, the N400 has previously been found to reflect predictive processing (Mantegna et al., 2019), and we will adopt this view in the current study.

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 behavioural response in the word-switch and the stress-switch condition compared to the control condition (Table 5). More specifically, participants were instructed to respond to whether the meaning of the sentence-final word is correct given the lead-in sentence, which means that the behavioral response in the word-switch and the stress-switch condition required a 'no'- response while the control condition required a 'yes'- response. This needs to be taken into consideration when comparing RTs between the control condition and the word-switch and stress-switch conditions. Together, these results (both RTs and ERPs) comparing the word-switch and the stress-switch conditions to the control condition would serve as a sanity check, and inform us on the learning behaviour of the participants and whether they would predict the sentence-final words in the first place.

Method

Participants

Twenty-three native Dutch participants were recruited from the Max Planck Institute for Psycholinguistics (MPI) participant pool, aged between 18 and 62 (6 male, 17 female, M_{age} = *26.6,* SD_{age} = *10.3*). All participants gave informed consent and were paid for their participation. One participant was excluded because of low accuracy scores on the behavioural task, another because of reported left-handedness and another because of noisy EEG data. The 20 remaining participants were right-handed and did not have any hearing and/or reading problems (5 male, 15 female, age range: 18-48; M_{age} *= 25.5,* SD_{age} *= 7.6).* Note that the intended number of participants was 32. Hence, the results based on the current dataset are preliminary.

Design

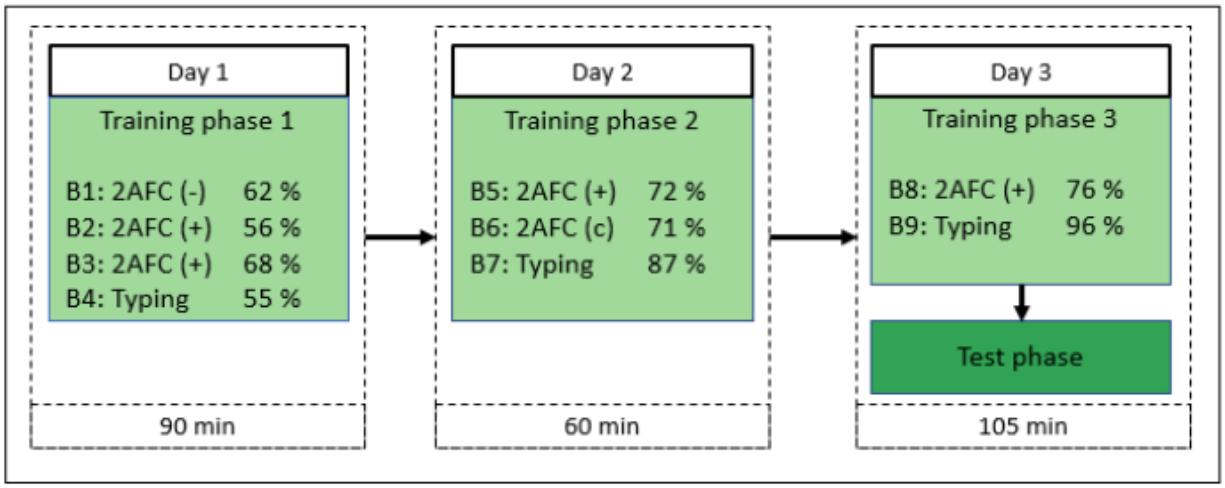
The experiment consisted of a training phase (divided over three sessions) and a test phase. To keep the experimental sessions as short as possible, we divided the sessions over three consecutive days (Fig. 3). The learning process also benefited from this choice since newly learned spoken words are subject to overnight consolidation (Dumay & Gaskell, 2007)its engagement in lexical competition requires an incubation-like period that is crucially associated with sleep. Words learned at 8 p.m. do not induce (inhibitory. Participants performed a series of 2AFC tasks and typing tasks to learn the item-to object mappings and the talker-specific cues. After the third training session, participants were tested on the items they had learned during training while we recorded behavioural responses and EEG.

Materials

The stimuli consisted of the selected 32 minimal pairs of non-words based on the outcomes of Experiment 1, as well as various carrier sentences that were used in the different phases of Experiment 2. All the speech materials were recorded by the same male native Dutch talkers as in Experiment 1 in the same recording session. The carrier sentences for the training phase were recorded separately from the non-words, which allowed for utilisation of one token for all trials, for each talker.

For Experiment 2, we needed different carrier sentences for the training phase and the test phase. First, the training phase required carrier sentences for exposure to the items: *Dit is een...,* "This is a...". The items were placed in the sentence-final position of these sentences. To increase exposure to talker-specific prosodic cues, we also placed the items in an additional carrier sentence containing disyllabic words (*Dit object is een voorbeeld van een...,* "This object is an example of a ...") which allowed for manipulation of the prosodic cues in the two disyllabic words (*object* having word-final stress in Dutch and *voorbeeld* having word initial-stress). The suprasegmental cues in these two words were also manipulated in a talker-specific manner but, in this case, the syllables were set to the corresponding values of syllables in the extreme steps (i.e., step 1 and 7) of the production data obtained in Experiment 1. 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 in the carrier sentence. We avoided any lexical stress



Note. 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).

Figure 3. Schematic overview and accuracy scores of the learning tasks of Experiment 2.

cues in these sentences, since we desired participants to predict talker-specific word-forms based on previously learned knowledge, instead of based on cues that were present in the sentence itself. Hence, the words referring to the objects (e.g. *lamp*, "lamp") were all monosyllabic words.

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

Procedure

As mentioned before, the experiment consisted of a training phase (divided over the first three days) and a test phase that followed the last training session. On the first two days, participants were seated in front of a 326 mm 244 mm sized monitor and audio was presented through Sennheiser HD-250 headphones at a fixed comfortable level. On the last day, participants were seated in front of a 337 mm 270 mm sized monitor and audio was presented through Canton speakers.

Training phase

Following the study by Sulpizio and McQueen (2012), we used 2AFC tasks. Furthermore, we added three typing tasks to the training phase and divided these over the sessions (Fig. 3). Participants did not receive explicit familiarisation onto the items beforehand. 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 (i.e., we stated that just as in "OBject" and "obJECT", 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 being more difficult to learn than others, 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 amplitude instead of F0 and vice versa for talker B).

All the tasks (except for Training Block 6[3]) consisted of 128 experimental trials and only

3 Training Block is used to refer to the different blocks during the training phase, test block is used to refer to different blocks during the test phase

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 randomised order.

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 coloured 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. This allowed participants to gradually learn the correct item-to-object mappings. Every trial started with a fixation cross. After 500 ms, participants heard one of the carrier sentences containing the items (Fig. 4). To emphasize which talker produced each sentence, we displayed an image of the talker surrounded by either a blue or a red square (the colour was talker-specific) during the carrier sentence. At sound offset, two response options (two coloured line drawings) from which participants had to choose appeared on the screen . 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 together with the pictures. 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. Note that explicit 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 coloured line drawings of the minimal pairs together, allowing participants to familiarize themselves with the segmental information of the non-words. During all the other 2AFC tasks, however, we presented 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 & 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. This increased the efficiency of the learning procedure by using a more participant-specific learning task. 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.

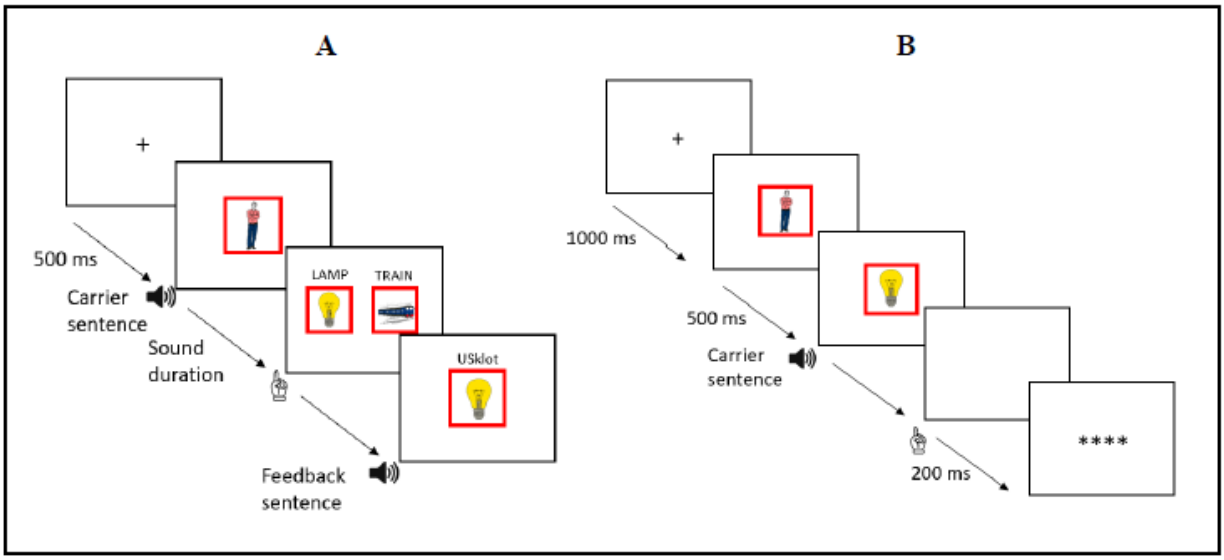


Figure 4. Illustration of a trial in the two-alternative forced choice (2AFC) task (Figure 4A) and in the test phase (Figure 4B).

Typing tasks

In the typing tasks, 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 tasks was for participants to recall the item cued by the object (which was also close to their task at test). Since a spoken production task could lead to interference from the prosodic cues in those spoken productions to the talker-specific prosodic cues in the stimuli, 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 with the stressed syllable being capitalized. Accuracy was assessed by comparing the response to the correct string for each trial. Additionally, to adjust for small typing errors, the incorrect responses were checked afterwards 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.

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. Every trial started with a fixation cross in the middle of the screen (Fig. 4B). After 1000 ms, we displayed the image of the talker producing the sentence during that trial, surrounded by the corresponding coloured square (as in the training phase) for 500 ms. Then, we presented the carrier sentence auditorily (e.g. *Het woord voor lamp is een USklot,* "The word for lamp is an USklot") together with the line drawing of the object in that sentence (e.g., a lamp). Participants were instructed to respond as quickly and accurately as possible with button presses from target word onset onwards (i.e., the sentence-final word; *USklot*). The critical judgement was based on whether the meaning of the sentence-final word matched the sentence (right button for a correct word, left button for an incorrect word). For example, "The word for lamp is an USklot" is correct since *USklot* refers to "lamp". 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.

Recall that we assessed performance on all four conditions (control, cue-switch, word-switch, and stress-switch). However, as opposed to the training phase (in which the prosodic cues were always talker-congruent), participants also received talker-incongruent versions of the items in the test phase. Since these could potentially induce 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 a cue-switch (25% of all the trials) did not exceed the number of trials on which participants did not experience a cue-switch. Furthermore, we wanted to minimise 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 might confuse participants). To achieve this, we made sure that the trials were presented in a fixed order within items. That is, for each item participants always first received the cue-switch version (e.g., *USklot* by Talker A using amplitude), 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. This constraint ensured that participants still heard a correct version of each item in between the conditions in which a deviant version of that item was presented (i.e., the cue-switch and the stress-switch condition). 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.

Lastly, the experimental stimulus list consisted of two test blocks, with 128 trials in each test block (32 trials per condition). All 64 items were divided among these blocks. 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. Furthermore, since 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).

EEG recording

EEG signal was recorded using 59 electrodes on an Acticap standard 10/20 cap, amplified with a BrainAmps (Brain Products) DC amplifier (500 Hz sampling rate, 10-1000 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 the stimulus). Noisy trials and consistently noisy channels 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, noisy channels were interpolated based on the weighted average of neighbouring channels. Trials still containing eye blinks or noisy channels (that could not be fixed) were then eventually rejected (1.5% of the total data). Finally, we applied a baseline-correction from 500 to 0 ms before stimulus onset.

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. To assess the differences between the conditions, we performed cluster-based permutation analyses (Maris & 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.

For all conditions, we performed a cluster-based permutation analysis over the entire epoch (i.e., -500 ms to 800 ms relative to stimulus onset). This allowed us to observe significant differences between the conditions that would coincide with the N400 time-window (between 200 ms and 600 ms poststimulus; Kutas & Federmeier, 2011) or the N200 time-window (between 285 and 335 ms; Brunellière et al., 2013).

Results

Behavioural

To test the behavioural results during the test phase, we analyzed participants' accuracy and reaction times (RTs) of participants' button presses. 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 (17.5%), which resulted in 4216 observations. Mean RTs and accuracy percentages are displayed in Table 6.

The behavioural data were analyzed using a linear mixed-effects model with the lmerTest package (Kuznetsova, Brockhoff & Bojesen Christensen,

Table 6. Mean (SD) response times (from correct trials only) and percentages of correct answers during the test phase.

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

Table 7. Results from the linear-mixed effects model with Log-transformed RTs as dependent variable.

Fixed effect	*β*	*SE*	*t*	*p*
Intercept	6.99	0.04	160.21	p <.001
Word-switch	-0.26	0.01	20.76	p <.001
Stress-switch	0.26	0.02	15.84	p <.001
Cue-switch	0.03	0.01	2.65	p =.008
Trial number	-0.02	0.01	-1.37	p < .171
Word-switch * Trial number	-0.06	0.01	-3.86	p <.001
Cue-switch * Trial number	-0.03	0.01	-2.23	p = .026
Stress-switch * Trial number	-0.03	0.02	-1.69	p = .091

Table 8. Results from the GLMM with the binomial accuracy of the categorization response as dependent variable.

Fixed effect	*β*	*SE*	*T*	*P*
Intercept	2.65	0.21	12.41	p < .001
Word-switch	1.98	0.28	7.19	p < .001
Stress-switch	-2.65	0.13	-20.65	p < .001
Cue-switch	-0.29	0.14	-2.10	p = .037
Trial number	-0.16	0.15	-1.11	p = .269
Word-switch * Trial number	0.29	0.29	1.00	p = .318
Stress-switch * Trial number	0.17	0.16	1.05	p = .292
Cue-switch * Trial number	0.10	0.17	0.66	p = .507

2016) in R (R Core Team, 2014). The model with the best fit to the data (as tested using log-likelihood model comparisons) contained the following factors (see Table 7): as fixed factors, we included Condition (categorical predictor with four levels, with the control condition at the intercept) and Trial number (continuous predictor that has been scaled to z-scores) and their interaction. As random factors, we included Participant and Item.

As mentioned above, the control condition was mapped onto the intercept which means that all the following effects should be compared to the control condition. The model revealed a significant effect for the word-switch condition (β = -0.26, *SE* = 0.01, *t* = -20.76, *p* <.001), indicating faster responses when participants were presented with a segmentally differing word. Also, a significant effect was revealed for the stress-switch condition (β = 0.26, *SE* = 0.02, *t* = -15.84, *p* <.001), indicating slower responses when they were presented with the wrong member of the minimal pair. As noted earlier, the task required participants to respond differently to the word-switch and the stress-switch condition (a "no" – response) compared to the control condition (a "yes"- response) so these RTs should be interpreted with caution. Lastly, and most importantly, the model revealed a significant effect for the cue-switch condition (β = 0.03, *SE* = 0.01 *t* = 2.65, *p* = .008). This indicated that when participants were presented with the correct word but produced using unexpected prosodic cues, they were slower compared to when the expected cues were used to produce that word.

Second, the model did not show evidence for a main effect of Trial number (β = -0.02, *SE* = 0.01, *t* = -1.73, *p* = .171), suggesting that the RTs on the control condition did not change as the experiment progressed. However, we did observe a significant interaction between the word-switch condition and Trial number (β = -0.06, *SE* = 0.01 *t* = -3.86, *p* <.001) and between the cue-switch condition and Trial number (β = -0.03, *SE* = 0.01, *t* = -2.23, *p* = .026). These interactions indicated that the difference in RTs between the word-switch and the control condition grew larger, while the difference between the cue-switch condition and the control condition

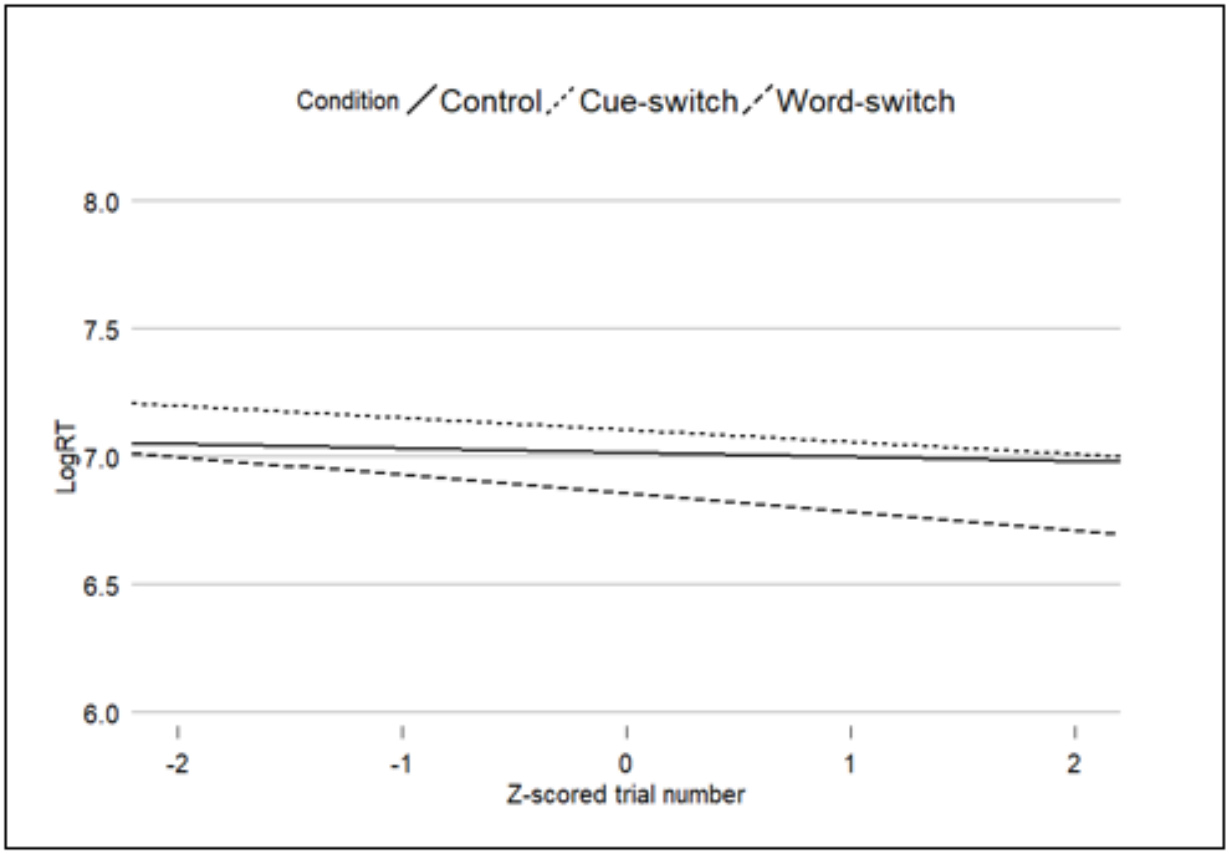


Figure 5. Interaction between Trial number and the cue-switch and the word-switch condition compared to the control condition based on the estimates of the linear mixed effects model.

grew smaller based on Figure 5, we can observe that RTs in the word-switch condition became even lower compared to the control condition, while RTs in the cue-switch condition converged with RTs in the control condition.

Third, we ran a General Linear Mixed Model (GLMM) with a logistic linking function to test whether the accuracy of the categorization responses was different for the four conditions (see Table 8). The binomial dependent variable was the accuracy on the categorization response 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 further included the same predictors as in the linear mixed-effects model we used for the RT data.

The model revealed significant effects for the word-switch condition (β = 1.98, *SE* = 0.28, *t* = 7.19, *p* <.001), the stress-switch condition (β = -2.65, *SE* = 0.13 *t* = -20.65, *p* <.001) and the cue-switch condition (β = -0.29, *SE* = 0.14 *t* = -2.10, *p* <.037). This indicated that the number of correct responses was higher in the word-switch condition, while it was lower in the stress-switch condition and the cue-switch condition compared to the control condition. The model revealed no significant effects of (or interactions with) Trial number, indicating that the accuracy was stable across the experiment.

Fourth, considering the relatively low performance on Training Block 8 compared to Training Block 9, we ran additional analyses to find out whether the test phase results differed depending on performance in Training Block 8. For RTs, we ran the same linear mixed-effects model (see Table 7) but added performance on Training Block 8 (scaled to z-scores) for each participant as a fixed factor to the model (with interactions between Condition and Trial number and between Condition and performance on Training Block 8). The model revealed only a significant interaction between Training Block 8 and the word-switch condition (β = -0.4, *SE* = 0.12 *t* = -3.48, *p* <.001), illustrating that those participants who achieved high accuracy on Training Block 8 also responded faster to the segmentally incongruent word. No such interactions with Training Block 8 were found for the remaining conditions (control, cue-switch and stress-switch), illustrating that the ability to distinguish the minimal pairs acoustically did not affect RTs in these conditions.

Lastly, concerning the accuracy of the categorization responses, we wanted to look into

the unexpectedly low mean correct responses on the stress-switch condition in more detail. To this end, we first correlated performance on Training Block 8 and Training Block 9 during training to the number of errors made in the stress-switch condition. We found that performance in Training Block 8 show a strong negative correlation with the number of errors in the stress-switch condition (r = -.60, p = .004) while the performance in Training Block 9 was only weakly negatively correlated with the number of errors in the stress-switch condition (r = -.16, p = .51). In addition to the correlation analyses, we ran a GLMM with a logistic linking function containing the same dependent variable and predictors as the model for the accuracy on the categorization responses (see Table 8) and we added performance in Training Block 8 (scaled to z-scores) as a fixed factor to the model (again, with interactions between Condition and Trial number and between Condition and performance in Training Block 8). The model revealed only a significant interaction between the accuracy on Training Block 8 and performance on the stress-switch condition (β = 0.42, SE = 0.13, t = 3.05, p =.002). This suggests that with increasing accuracy on Training Block 8, performance in the stress-switch condition at test improved.

EEG results

First, the cluster-based permutation analysis revealed a significant difference between the word-switch condition and the control condition (p = .022). To illustrate the location and latency of the difference, we plotted topographical maps of the statistical analysis which revealed a cluster between 214 ms and 442 ms in which the word-switch and cue-switch condition differed from each other (see Fig. 6A). Based on the ERPs of both conditions from one of the channels in this cluster (see Fig. 6C) we were able to infer the direction of the effect in this time-window. Based on these results, we concluded that the word-switch condition elicited a relatively larger N400 response compared to the control condition.

Second, the cluster-based permutation analysis revealed a significant difference between the stress-switch condition and the control condition (p = .007). Again, we plotted topographical maps of the statistical analysis (see Figure 6B) and ERPs of the channels in the resulting cluster (see Fig. 6D) which illustrate that the stress-switch elicited a relatively larger N400 response compared to the control condition between 300 ms and 700 ms.

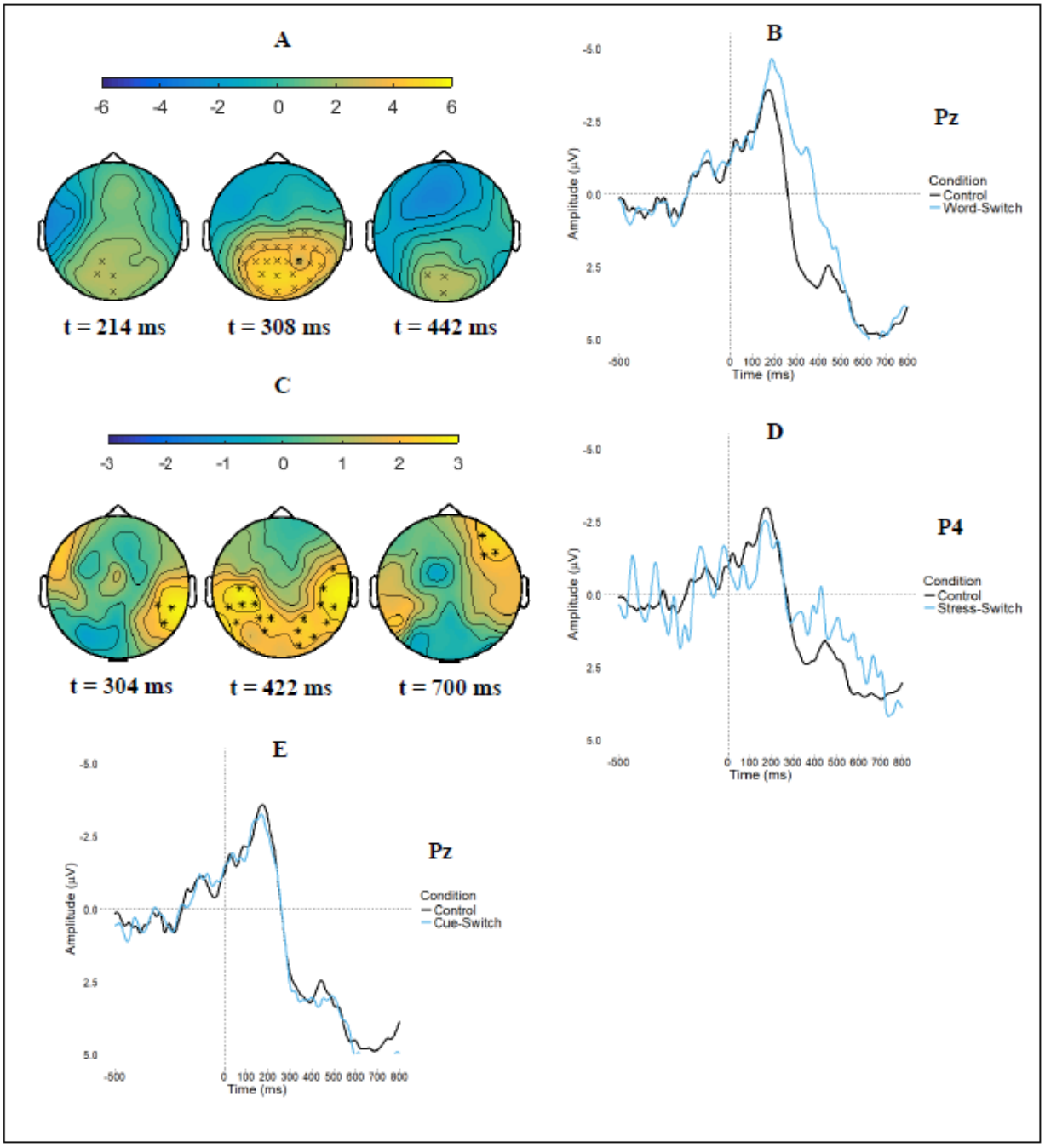
Third, the cluster-based permutation analysis revealed that there was no significant difference between the cue-switch condition and the control condition (p = .713). This indicated that both conditions are from the same probability distribution (i.e., the conditions are interchangeable) which implies that the cue-switch condition did not elicit a relatively larger N200 response compared to the control condition (see Fig. 6E).

Discussion

This study tested whether listeners learned talker-specific prosodic cues that signal lexical stress and if they used that information to predict talker-consistent upcoming word-forms on a subsequent encounter. After a training phase, during which participants learned minimal pairs of non-words and these talker-specific cues, we presented participants with semantically constraining sentences allowing for prediction of talker-specific word-forms (i.e., the correct word given the sentence, and produced using the correct cues given the talker). We hypothesized that when participants were presented with talker-mismatching prosodic cues to lexical stress (i.e., sentences containing the unexpected cues for a specific talker), this would slow down RTs and elicit a relatively larger N200 response compared to control sentences. Results indicated that participants responded more slowly and less accurately to these mismatching sentences. However, the amplitude of the N200 was not modulated by these sentences.

Talker-specific learning of prosodic cues

The central finding of the current study concerns the cue-switch condition. We found that sentences in the cue-switch condition led to longer RTs and more errors compared to the control condition. Since the target words in these two conditions were identical with regard to the stress pattern, the segmental information and which talker produced the words, the only mismatching factor in these sentences were the talker-specific prosodic cues that were used to produce these words. This indicates that participants were slowed down when they experienced a cue-switch. In addition to longer RTs, we found lower accuracy scores in the cue-switch condition compared to the control condition. This suggests that, even though the sentence-final word in the cue-switch condition was correct (given the lead-in sentence), the mismatch in talker-specific prosodic cues led participants to make more errors on those sentences. These behavioural results



Note. No topographical map could be plotted for the cue-switch condition because there was a non-significant result.

Figure 6. Figure 6A and 6B. Topographical maps of the summed t-values. Figure 6C, 6D and 6E. ERP of one of the channels in the resulting cluster.

imply that participants had learned the talker-specific prosodic cues during training and used that information to predict talker-specific word-forms. Furthermore, since there were no talker-specific cues in the carrier sentences at test, this effect did not arise due to a locally induced contrast (between the carrier sentence and the sentence-final target word). Instead, participants activated their talker-specific knowledge based on talker identification and predicted talker-specific word-forms accordingly.

These findings are in line with Eisner and McQueen (2005), who found evidence for talker-specific perceptual learning, and with Zhang and Holt (2018), who illustrated that these talker-specific learning effects are not simply due to adaptation to the acoustic signal. Instead, listeners can track co-evolving regularities (i.e., different talkers with their own speaking style) and adapt perceptual categories based on those talkers. The current study shows for the first time that listeners also

use these talker-specific learning mechanisms with regard to prosodic cues. In other words, we show that listeners are able to learn how different talkers use prosodic cues to a different extent and activate (and use) this information, based on voice quality and visual presentation of (icons of) those talkers, on subsequent encounters.

In addition to these main effects, we found an interaction between the cue-switch condition and Trial number, illustrating that during the course of the experiment, the difference in RTs between the control and cue-switch condition became smaller. This convergence of the RTs in the two conditions shows that the talker-specific effect gradually reduced throughout the experiment, even though we tried to minimize any potential effects of unlearning. A possible source of unlearning is the prosodic cues that were presented in a subset of the trials in the test phase. That is, while participants only heard correct talker-specific prosodic cues during training (i.e., Talker A always using F0 and Talker B always using amplitude), the test phase also contained the opposite cues (i.e., Talker A also using amplitude and Talker B also using F0). These conflicting cues can in turn lead participants to unlearn the talker-specific information they had learned during training. A similar effect of unlearning, but this time for segmental perceptual learning, has also been found by Kraljic & Samuel (2005), who trained participants on perceiving an ambiguous sound on an /s-ʃ/ continuum as either /s/ or /ʃ/. After the training phase, they found that when participants were exposed to conflicting information from the same speaker (hearing the ambiguous sound, learnt as /s/, in an /ʃ/ word), the perceptual learning effect disappeared rather rapidly.

Also, Kurumada, Brown, Bibyk, Pontillo, and Tanenhaus (2014) studied pragmatic interpretation based on prosodic structures presented in sentences. More specifically, they presented participants with two types of sentences in an eye-tracking study. Either sentences containing a high pitch accent on the final noun (e.g., "It looks like a ZEbra) or sentences with a high pitch accent on "looks" (e.g., "It LOOKS like a zebra..." (but it is not)) and found that these different prosodic structures affected the proportion of fixations to target pictures. Crucially, they manipulated how consistently the talker used the pitch accent correctly and found that when the reliability of the use of these cues decreased, the prosodic structures did not affect the proportion of fixations anymore. The authors concluded that the low reliability of the prosodic cues led to down-weighting of these cues. Together, these studies offer an explanation for the interaction between the cue-switch condition and Trial number. That is, since the cue-switch condition contained talker-specific cues that were opposed to what listeners had learned during training, these cues served as 'correcting' cues. Furthermore, since the other conditions (control, word-switch and stress-switch) did contain the correct talker-specific cues, both talkers did not produce either prosodic cue consistently, decreasing the reliability with which both talkers used their talker-specific prosodic cues which could have led participants to down-weight these cues in perception. Importantly, this interaction does not only illustrate that talker-specific perceptual learning effects were unlearned when listeners were presented with correcting cues, but also supports the notion that a learning effect was present in the first place. That is, if participants did not learn the talker-specific cues at training, presenting correcting cues would not have led to unlearning.

Further, we did not find an interaction between performance in Training Block 8 and RTs in the cue-switch condition. At first, this suggests that the talker-specific prosodic cue effect is not affected by performance in the training phase. However, performance in Training Block 8 may not be the ideal predictor for learning of these talker-specific cues. That is, even though there were some differences between participants in performance in Training Block 8, every participant received the same amount of exposure to the talker-specific cues (i.e., next to having to distinguish the minimal pairs acoustically, participants still could pick up on the talker-specific prosodic cues). Also, performance on Training Block 8 can be high even if participants do not learn the talker-specific prosodic cues. Therefore, performance in Training Block 8 is not the best predictor for the RT model. Furthermore, this suggests that participants could be learning two distinct types of information. On the one hand, participants were learning the item-to-object mappings. On the other hand, independently of the correct mappings, participants were supposed to learn the talker-specific prosodic cues.

Concerning EEG results, we did not find a difference in the N200 response between the cue-switch and the control condition. Following the logic in Brunellière et al. (2013) and previous theoretical accounts of the N200 (Connolly & Phillips, 1994), this would suggest that there is no mismatch between the predicted word-forms and the perceived word-forms in the current study. These findings are not in line either with our behavioural results (longer RTs and lower accuracy in the cue-switch condition

compared to the control condition) or with the results in Brunellière et al. (2013), who did find a modulation of the N200 amplitude for segmentally mismatching word-forms. This raises the question whether the current finding is a true null result, or whether we were simply unable to obtain valid ERPs with the learned stimuli in the current study. To be able to look into this latter possibility, we included two conditions as a sanity check that informed us on whether participants would successfully learn the non-words in the current study and whether these items would indeed elicit valid ERPs.

Learning segmental information in the items

The first sanity check would inform us on whether participants learned the segmental information in the items, which was tested in the word-switch condition. Concerning behavioural results, we found that sentences in the word-switch condition led to shorter RTs compared to the control condition. This indicated that participants were faster to reject an incorrect word compared to accepting a correct word based on segmental information. As mentioned in the Introduction, we did not have a specific hypothesis concerning RTs for this condition. Nevertheless, shorter RTs in the word-switch condition could be an illustration that the decision in this condition was easier compared to the control condition. Alternatively, the decision in the word-switch condition could simply be taking place at an earlier time point because of an earlier point of disambiguation. Indeed, 60 out of 64 items in the word-switch condition had a segmental deviation at the first phoneme and the remaining four items deviated at the second phoneme. In contrast, the items in the control condition required participants to process both syllables in order to make a decision on the stress patterns, which resulted in longer RTs. We also found higher accuracy scores in the word-switch condition compared to the control condition. In sum, these results illustrate that participants had correctly learned the segmental information in the non-words.

Next, concerning EEG results, we found a relatively larger N400 response in the word-switch condition compared to the control condition. This indicated that semantic integration of the predicted sentence-final words presented in the word-switch condition was more difficult compared to the control condition. Note that our behavioural result in the word-switch condition might seem in contrast to the ERP result (shorter RTs while we observed a relatively larger N400 response). Indeed previous studies (Brown & Hagoort, 1993; Peeters et al., 2013) found longer RTs for incongruent words (that also elicited a relatively larger N400 response). However, these studies all used a lexical decision task which requires the same behavioural response to congruent and incongruent words (slowing down responses to incongruent words) while in the current study (as mentioned before), the behavioural task was different between these conditions, leading to shorter RTs to incongruent words. Thus, even though RTs in the word-switch were shorter compared to the control condition, they are still in line with the relatively larger N400.

In addition to the behavioural result, this confirms that participants learned the segmental information of the non-words and the correct item-to-object mappings. Also, since the incongruent sentences elicited a relatively larger N400 response, we can conclude that we were able to obtain valid ERPs in our current design and with the current stimuli based on segmental information.

Learning suprasegmental information in items

The second condition that was included as a sanity check concerned the stress-switch condition. This would inform us on whether participants successfully learned the suprasegmental information in the non-words. We found that sentences in the stress-switch condition led to longer RTs compared to the control condition. In contrast to the word-switch condition (that resulted in shorter RTs), this illustrated that participants were slower to reject the wrong member of the minimal pair compared to accepting the correct member in the control condition. The difference between the stress-switch and the word-switch condition can (as with the difference between the word-switch and the control condition) be explained by the type of information the decision was based on. That is, the point of disambiguation of the target words in the word-switch condition (based on segmental information) was much earlier compared to the stress-switch condition (based on suprasegmental information). Still, it is surprising that RTs in the stress-switch condition were much higher than in the control condition (since this condition required the same suprasegmental decision). This increase in RTs in the stress-switch condition compared to the control condition could be due to the order of presentation of the different conditions. More specifically, the within-item controlled order ensured

that participants were first presented with the cue-switch and the control condition of a specific item before receiving the stress-switch condition. Even though this was intended to help participants and slow down unlearning of the talker-specific prosodic cues, this may have caused confusion and slowed down participants in the stress-switch condition. This could also explain why there was no effect of performance in Training Block 8 on RTs in the stress-switch condition. That is, one might expect that with increasing performance in Training Block 8, RTs in the stress-switch condition would decrease since it could be easier to make that decision. Again, participants could have experienced interference from the cue-switch condition leading to longer RTs in the stress-switch condition.

In addition to longer RTs, accuracy scores in the stress-switch condition were also much lower compared to the control condition. This could partly be explained by performance in the training phase. That is, the lower accuracy scores on Training Block 8 (2AFC) compared to Training Block 9 (typing task) suggested that while participants did correctly learn the item-to- object mappings (tested in the typing task of Training Block 9), they still struggled to distinguish the minimal pairs acoustically (tested in the auditory 2AFC task in Training Block 8). In fact, the correlation between the accuracy scores on Training Block 8 and performance on the stress-switch condition, as well as the result of the additional GLMM (containing accuracy scores of Training Block 8) illustrate that participants who performed better in Training Block 8 also performed better in the stress-switch condition. A second possible explanation for the low accuracy scores in the stress-switch condition is the absence of talker-specific prosodic cues in the carrier sentences at test. Recall that during training, the carrier sentences contained disyllabic words that also contained talker-specific prosodic cues. However, at test, the carrier sentences only contained monosyllabic words and thus, the lack of confirmation of talker-specific cues could have impeded perception of stress patterns in the target words which affected the accuracy scores in the stress-switch condition. Third, low accuracy scores on the stress-switch condition could also be explained 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' direction to suprasegmental 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 suprasegmental cues (Sulpizio & McQueen, 2011) and base their responses more on segmental information. If participants indeed followed this strategy, this 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 still, there was a difference between participants who performed better during training and participants who performed worse. Thus, the possible explanations offered above could have had different effects on different levels of proficiency in Training Block 8.

Concerning EEG results, we found a relatively larger N400 response for the stress-switch condition compared to the control condition which illustrates that semantic integration of the predicted sentence-final word was also harder for the wrong member of the minimal pairs. Keep in mind that we only included trials on which participants responded correctly, so the previous conclusion can only be drawn regarding these particular items. Also, the analyses revealed that the time-window of the N400 in the stress-switch condition was slightly later compared to the word-switch condition. As in the behavioural result, this later time-window can be assigned to the point of disambiguation of the target words (i.e., the point of disambiguation being later for the words in the stress-switch condition compared to the words in the word-switch condition). Also, note that the ERPs were much noisier compared to the word-switch condition which is due to the number of trials we were able to include in this condition (being half of the number of trials we included in the word-switch condition because of low accuracy).

In sum, behavioural findings (longer RTs and lower accuracy) illustrate that participants still struggled to distinguish the minimal pairs acoustically. Also, the presence of the word-switch condition and the order of presentation could have confused participants in the test phase and contributed to these results. Still, for the correct trials, we were able to obtain valid ERPs based on suprasegmental information. Together, these sanity checks might suggest that the result in the cue-switch condition (no modulation of the N200 amplitude) was not due to the stimuli that were used in the current study. Instead, other explanations should be considered.

Effect size of the mismatch

An alternative explanation for the lack of the N200 amplitude modulation is that the effect size in the current study is much smaller compared to

the effect size in Brunellière et al. (2013). A possible source for this effect size is that the sample size in the current experiment could have been too small to achieve an appropriate power for our effect size. That is, in this preliminary dataset, we included 20 participants as opposed to the intended 32 participants. Still, Brunellière et al. (2013) included a similar number of participants (21 participants) which did modulate the N200 amplitude. This suggests that in addition to the sample size being too small in the current study, the effect in Brunellière et al. (2013) was also much larger.

This larger effect in Brunellière et al. (2013) can be explained by a more salient mismatch between the carrier sentence and the sentence-final word in Brunellière et al. (2013). There are several differences with the current study that could have contributed to this larger mismatch. First, in the current study, we used acoustically manipulated non-words that contained fewer prosodic cues (since two out of three prosodic cues were set to ambiguous values) as opposed to naturally produced real words in Brunellière et al. (2013). Even though the word-switch and the stress-switch condition illustrated that these stimuli were able to elicit valid ERPs based on semantic information, our acoustically reduced stimuli could still have caused difficulties regarding phonological predictions compared to natural speech (i.e., prior knowledge of stress patterns normally being signaled by multiple cues could interfere with the reduced stimuli in the current study). Second, there is a large difference concerning the amount of experience participants have had with the cues leading to the mismatch. More specifically, in the current experiment, participants went through a three-day learning paradigm while participants in Brunellière et al. (2013), being speakers of Eastern Catalan, already had substantial experience with the accents prior to the experiment. Possibly, the amount of exposure in the current study was too little to modulate the amplitude of the N200 compared to Brunellière et al. (2013). In addition, the mismatch in the current study could be harder for participants to detect (i.e., only one mismatching prosodic cue, related to one specific talker) compared to Brunellière et al. (2013) in which the mismatch is much more general (i.e., a vowel reduction that is general across all talkers of a certain dialect). Third, one important difference relates to the presence of the mismatch in the carrier sentence at test. Recall that in the current study, we avoided talker-specific prosodic cues in the carrier sentence at test because we wanted participants to predict talker-specific word-forms based on previously learned knowledge. However, the carrier sentences in Brunellière et al. (2013) did contain other words with a vowel reduction which means that the mismatch (i.e., vowel reduction in the carrier sentence vs. no vowel reduction in the sentence-final word) is much more apparent compared to the current study. Finally, the mismatch in Brunellière et al. (2013) resulted in a larger mismatch regarding phonological categories to which the sound belongs compared to the mismatch in the current study. More specifically, the mismatch in prosodic cues in the current study leads to a within-category mismatch (the resulting stimuli still contain the same phonemes and lexical stress patterns but are cued differently). In contrast, the vowel reduction in Brunellière et al. (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). This could again have resulted in a larger mismatch in Brunellière et al. (2013) compared to the current study which, in addition to the other differences, could explain the lack of an N200 amplitude modulation in the current study.

Is the N200 sensitive to phonetic detail?

Alternatively, we should consider the possibility that the current study could not modulate the N200 at all. That is, previous studies (Brunellière et al., 2013; Connolly & Phillips, 1994), relied on a mismatch based on segmental information while the sentences in the current study contain a mismatch only concerning prosodic cues (i.e., phonetic detail). To our knowledge, an N200 response to mismatching prosodic cues has not been found in previous studies. Hence, there is no evidence yet indicating that the N200 is sensitive to phonetic detail. In addition, Diaz and Swaab (2007) even failed to modulate the N200 amplitude to segmentally mismatching word-forms in sentence processing (despite there being a phonological mismatch) while there was a modulation of the N200 in word list perception. The authors concluded that phonological processing has a different role in sentence processing due to the presence of a semantic context.

This touches upon an issue that has previously been discussed regarding the N200. Namely, even though several studies have reported an N200 response to phonologically mismatching word-forms, it remains unclear whether the N200 component can be functionally dissociated from the N400 component (for review, see Nieuwland, 2019) various studies claim that word form prediction manifests itself in 'early', pre-N400 brain responses

(e.g., ELAN, M100, P130, N1, P2, N200/PMN, N250. An alternative account is that the N200 actually reflects an early onset of the N400 instead of a distinct component. To shed more light on this possible distinction, Poulton and Nieuwland (2019) examined the scalp distribution of the two components, since a distinct N200 component has a more frontal and widely distributed topography while the N400 component has a more posteriorly distributed topography, and aimed at dissociating both components based on their topography. They measured ERP responses to auditorily presented sentences containing sentence-final words that were either highly predictable, partially overlapping (semantically unpredictable but with phonological overlap) or without overlap (semantically unpredictable and without phonological overlap) which allowed for identification of the different components. The authors then compared the scalp distributions of the early and late time-window and found that both negativities (in the early and in the late time-window) had a more posterior distribution. They concluded that the N200 is not a distinct component that is sensitive to phonological mismatch but is simply an early onset of the N400.

The current study could also contribute to this discussion by examining the scalp distribution of the EEG results in the word-switch condition. More specifically, in addition to the semantic mismatch, the sentence-final words in the word-switch condition also contained a phonological mismatch (the items were segmentally different compared to the control condition). Hence, we can observe whether there is a different scalp distribution (more frontal and widely distributed) for the earlier time-points compared to the later time-points. Keep in mind that the current study was not designed to test the different accounts of the N200. Still, by examining the scalp distributions at different time-points in the word-switch condition, we observed a clear posterior distribution throughout the entire time-window. As in Poulton and Nieuwland (2019), this finding supports the notion that the N200 is an early onset of the N400.

Future directions

Despite these considerations, the behavioural results in the current study do illustrate that listeners experienced a mismatch between predicted and perceived word-forms in the cue-switch condition, demonstrating that listeners are sensitive to these differences between talkers. However, a possible argument against this result concerns performance on the stress-switch condition at test. That is, one might ask whether participants paid attention to the prosodic cues at all considering their low performance on distinguishing the minimal pairs acoustically. Regardless of their performance at test, the accuracy scores on Training Block 8 illustrate that participants did not fully ignore the prosodic cues but used those cues in learning the non-words. Given that participants were required to pay attention to the prosodic cues during the 2AFC blocks in the training phase (i.e., in order to give the correct response, participants had to resort to the prosodic cues), this suggests that participants also picked up on the talker-specific prosodic cues that were used to signal those words. That is, if participants ignored the talker-specific prosodic cues, it would not have been possible to detect lexical stress patterns (since the other cues were set to ambiguous values). Indeed, the difference in RTs and accuracy scores between the control and the cue-switch condition suggest that even though participants struggled to distinguish the minimal pairs acoustically in the stress-switch condition, they still learned how both talkers would signal those non-words. In other words, regardless of whether the non-word is SW or WS, they knew that Talker A would produce either of the non-words with F0 and Talker B with amplitude.

While at first sight the incongruency between these conditions may appear as a limitation to the current study, it could actually provide insights into how talker-specific prosodic information is stored. Namely, it might 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 which may speculatively even be applied to new words. However, the design of the current study was not able to look into these proposed representations since the participants were tested on the same non-words as those learned in the training phase. The question thus remains open whether listeners would also generalize learned information about talker-specific prosodic cues to other instances (i.e., other words). Future experiments could therefore look into whether the obtained result can be replicated in an experiment in which the test phase includes different words than the training phase. This will allow for differentiation between whether listeners had learned about how two talkers use prosodic cues in specific words, or how two talkers use prosodic cues in general, informing us on the level of representation of this information.

Another question that arises, is how the current

findings generalize to natural speech. That is, in natural speech, stress patterns are signaled using multiple cues (the stressed syllable has a higher F0, increased amplitude and longer duration; Rietveld & Van Heuven, 2009). On the other hand, the stress patterns in the current study were signaled using one cue (e.g., F0 or amplitude while the rest was set to ambiguous values). One might wonder how the mechanisms that have been found in the current study, perceptual learning and prediction, are used to deal with variability in multiple prosodic cues. Future experiments could thus extend the current finding by creating stimuli that incorporate these other cues and resemble natural speech to a greater extent.

Other directions for future research could focus on finding converging evidence from other measures (e.g., eye-tracking, EEG) to support the behavioural result in the current study. More specifically, eye-tracking experiments are beneficial since it would allow for use of fewer non-words (since there are fewer issues with repetition of items compared to EEG). In turn, the use of fewer non-words allows for more exposure within a given period of time which optimizes the learning procedure (higher accuracy scores, more exposure to talker-specific cues). In addition, eye-tracking offers a more valid measure of predictive speech perception. That is, eye-tracking paradigms (Altmann & Kamide, 1999; Kamide et al., 2003) have been used to measure *anticipatory* eye-movements which can be taken as a more sensitive measure of prediction in speech perception. Lastly, future experiments could focus on modifications to the experiment that might modulate the N200 amplitude. For example, by using carrier sentences in the test phase in which the talker-specific prosodic cues are present (similar to the sentences in Brunellière et al., 2013), a more salient mismatch is introduced. Hence, this could also shed more light on the sensitivity of the N200 to mismatching prosodic cues. These future experiments could provide more supporting evidence and extend the findings of the current study in discovering how listeners deal with talker variability in prosodic cues.

Conclusion

In sum, the current study investigated how listeners use talker-specific prosodic cues in predictive speech perception. We conclude that the absence of the amplitude modulation of the N200 in the current experiment is due to the effect size being too small. Our behavioural results illustrate that listeners can adjust their perceptual representations in a talker-specific manner not only for segmental information as previously shown (Brunellière & Soto-Faraco, 2013; Eisner & McQueen, 2005; Zhang & Holt, 2018), but also for the first time for suprasegmental information. It appears that listeners can predict upcoming word-forms based on those talker-specific prosodic representations. Applying this to the aforementioned "The stranger objects" example, the current 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 "stranger" and "objects". 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.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer* (Versie 6.065) [Computer software]. www.praat.org

Brown, C., & Hagoort, P. (1993). The Processing Nature of the N400: Evidence fiom Masked Priming. *Journal of Cognitive Neuroscience*, *5*(1), 33–44.

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language*, *125*(1), 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. *Journal of Phonetics*, *39*(2), 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. *Journal of Cognitive Neuroscience*, *6*(3), 256–266. https://doi.org/10.1162/jocn.1994.6.3.256

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, *20*(1), 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. *Language and Speech*, *44*(2), 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 Research*, *1146*, 85–100. https://doi.org/10.1016/j.brainres.2006.07.034

Dumay, N., & Gaskell, M. G. (2007). Sleep-Associated

Changes in the Mental Representation of Spoken Words. *Psychological Science*, *18*(1), 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. (in press). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. https://doi.org/10.3758/BF03206487

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

Gussenhoven, C., & Van Der Vliet, P. (1999). The phonology of tone and intonation in the Dutch dialect of Venlo. *Journal of Linguistics*, *35*(1), 99–135. https://doi.org/10.1017/S0022226798007324

Haan, J., & Van Heuven, V. (1999). Male vs. Female pitch range in Dutch questions. *Proceedings of the 13th International Congress of Phonetic Sciences*, 1581–1584.

Jesse, A., Poellmann, K., & Kong, Y.-Y. (2017). English Listeners Use Suprasegmental Cues to Lexical Stress Early During Spoken-Word Recognition. *Journal of Speech, Language, and Hearing Research*, *60*(1), 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. *Journal of Memory and Language*, *49*(1), 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? *Cognitive Psychology*, *51*(2), 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. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(3), 791–196.

Kutas, M., & Hillyard, S. (1984). Brain Potentials During Reading Reflect Word Expectancy and Semantic Association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Kutas, Marta, & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

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. *Journal of Neuroscience Methods*, *164*(1), 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 K. Lamberts & R. L. Goldstone, *Handbook of Cognition*. SAGE.

Miller, J. L., Green, K., & Schermer, T. M. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics*, *36*(4), 329–337. https://doi.org/10.3758/BF03202785

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. https://doi.org/10.1121/1.397688

Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews*, *96*, 367–400. https://doi.org/10.1016/j.neubiorev.2018.11.019

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 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. *Psychological Science*, *5*(1), 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. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

Peeters, D., Dijkstra, T., & Grainger, J. (2013). The representation and processing of identical cognates by late bilinguals: RT and ERP effects. *Journal of Memory and Language*, *68*(4), 315–332. https://doi.org/10.1016/j.jml.2012.12.003

Poulton, V., & Nieuwland, M. S. (2019, november 19). *Can you hear what's coming? An ERP study of phonological prediction.* [Poster]. Cambridge Language Sciences Symposium Poster Presentations, University of Cambridge.

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal of Experimental Psychology*, *63*(4), 772–783. https://doi.org/10.1080/17470210903104412

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking Rate Affects the Perception of Duration as a Suprasegmental Lexical-stress Cue. *Language and Speech*, *54*(2), 147–165. https://doi.org/10.1177/0023830910397489

Rietveld, A. C. M., & Van Heuven, V. J. (2009). *Algemene fonetiek* (3rd dr.). Coutinho.

Sjerps, M. J., Zhang, C., & Peng, G. (2018). Lexical tone is

perceived relative to locally surrounding context, vowel quality to preceding context. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 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. *Journal of Memory and Language*, *66*(1), 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. *Proceedings of Interspeech 2011*, 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. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(11), 1760–1779. https://doi.org/10.1037/xhp0000569

Appendix

Additional information on the disyllabic non-words used as stimuli for Experiment 1.

Table A1. Non-words (given in Dutch orthography) that were used as stimuli. Capitalization indicates lexical stress.

AMpo [1]	amPO [1]
BLImoek [1]	bliMOEK [1]
ELpat [1]	elPAT [1]
KLAfos [1]	klaFOS [1]
LOsep [1]	loSEP [1]
OELfem [1]	oelFEM [1]
ORlos [1]	orLOS [1]
PLOsim [1]	ploSIM [1]
PRAbop [1]	praBOP [1]
TOEsa [1]	toeSA [1]
USklot [1]	usKLOT [1]
WAsol [1]	waSOL [1]
BOLdep [2]	bolDEP [2]
DEMrof [2]	demROF [2]
DREpos [2]	drePOS [2]
FLARda [2]	flarDA [2]
GAloe [2]	gaLOE [2]
GLErak [2]	gleRAK [2]
NILtaf [2]	nilTAF [2]
NOEfa [2]	noeFA [2]
NORbul [2]	norBUL [2]
KAlom [2]	kaLOM [2]
KESto [2]	kesTO [2]
KLIwo [2]	kliWO [2]
LENdon [2]	lenDON [2]
Odran [2]	oDRAN [2]
PALro [2]	palRO [2]
ROSkil [2]	rosKIL [2]
RUNka [2]	runKA [2]
SKALtra [2]	skalTRA [2]
STOLpaf [2]	stolPAF [2]
STRAdot [2]	straDOT [2]
AALsoi [1, 3]	aalSOI [1, 3]
BANwem [2, 3]	banWEM [2, 3]

MInok [2, 3]	miNOK [2, 3]
POLmo [1, 3]	polMO [1, 3]
SLEvor [2, 3]	sleVOR [2, 3]
TRApal [2, 3]	traPAL [2, 3]

1) Subset of items with 202 ms (first syllable) and 395 ms (second syllable) as ambiguous duration values.
2) Subset of items with 288 ms (first syllable) and 362 ms (second syllable) as ambiguous duration values.
3) Items that were excluded from the Experiment 2.