Accepted for publication in JEP:HPP on Dec 9th, 2022

Tracking Talker-Specific Cues to Lexical Stress:

Evidence from Perceptual Learning

Giulio G.A. Severijnen [a 1], Giuseppe Di Dona [b 1], Hans Rutger Bosker [a,c],

and James M. McQueen [a,c]

[a] *Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands*

[b] *Dipartimento di Psicologia e Scienze Cognitive, Università degli Studi di Trento, Italy*

[c] *Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*

Word Count: 12.022

Author Note

The authors of the present manuscript certify that they have complied with the APA ethical principles regarding research with human participants in the conduct of the research presented in this manuscript and declare no conflict of interests. This study was not preregistered.

Correspondence concerning the article should be addressed to:

Giulio G.A. Severijnen,

giulio.severijnen@donders.ru.nl

Thomas van Aquinostraat 4, 6525 GD Nijmegen, The Netherlands

[1] These two authors contributed equally to this work.

Abstract

When recognizing spoken words, listeners are confronted by variability in the speech signal caused by talker differences. Previous research has focused on *segmental* talker variability; less is known about how *suprasegmental* variability is handled. Here we investigated the use of perceptual learning to deal with between-talker differences in lexical stress. Two groups of participants heard Dutch minimal stress pairs (e.g., *VOORnaam* vs. *voorNAAM*, 'first name' vs. 'respectable') spoken by two male talkers. Group 1 heard Talker 1 use only F0 to signal stress (intensity and duration values were ambiguous), while Talker 2 used only intensity (F0 and duration were ambiguous). Group 2 heard the reverse talker-cue mappings. After training, participants were tested on words from both talkers containing conflicting stress cues ('mixed items'; e.g., one spoken by Talker 1 with F0 signaling initial stress and intensity signaling final stress). We found that listeners used previously learned information about which talker used which cue to interpret the mixed items. For example, the mixed item described above tended to be interpreted as having initial stress by Group 1 but as having final stress by Group 2. This demonstrates that listeners learn how individual talkers signal stress and use that knowledge in spoken-word recognition.

Keywords: lexical stress; perceptual learning; talker variability; suprasegmental cues; cue weighting

Statement of public significance:

- This study demonstrates that listeners can learn how individual talkers produce lexical stress, which helps listeners to deal with those differences during speech recognition.
- When listening to speech originating from different talkers, listeners learn and create memory representations about how those talkers produce lexical stress. These memory representations are then reactivated on future encounters with those talkers, facilitating word recognition.
- These results extend previous theories of word recognition, and highlight the importance of examining how listeners deal with between-talker variability in prosody.

Introduction

Individual differences among talkers lead to highly variable acoustic realizations of speech. For instance, consider the English noun 'IMport' (capitalization indicates lexical stress) being produced by two male talkers. Even though the word itself is identical, individual speaking styles can affect the acoustic realization of that word. Such variability can be found at the segmental level (vowels and consonants) and the suprasegmental level (e.g., intonation, lexical stress), and both types of variability have consequences for correct perception of the intended word. For example, perceiving different suprasegmental information may lead to misinterpreting the word as the verb 'imPORT', impeding successful communication. The present study assessed how listeners can still correctly perceive spoken words despite such talker-driven variability. More specifically, we investigated how the use of a cognitive mechanism, perceptual learning, aids listeners in dealing with talker variability in the production of lexical stress, supporting stable spoken-word recognition.

The presence of acoustic variability in speech has been widely established, focusing primarily on segmental variation driven by talker-specific properties, such as gender, age, and dialect. For instance, vowel formant frequencies vary depending on gender, age and regional dialects (Adank et al., 2004, 2007; Hillenbrand et al., 1995). Also, variability in voice onset time (VOT) of stop consonants has been found between talkers of different age and gender (Allen et al., 2003; Theodore et al., 2009). On top of these differences within specific acoustic cues, talkers also appear to differ in their cue-weighting strategies (for review, see Schertz & Clare, 2020). That is, speech contrasts are often defined by a multidimensional cue space. For example, the /b-p/ contrast in English relies on multiple cues such as VOT, fundamental frequency (F0), and many more (Lisker, 1986). The relative importance of these cues in production differs between talkers depending on their native language (Lisker & Abramson, 1964), dialects (Kang, 2013), and individual speaking styles (Schertz et al., 2015), adding to the acoustic variability in speech.

In addition to these differences in *segmental* structures, talkers also vary in how they produce *suprasegmental* (i.e., prosodic) structures, such as sentence intonation. In Dutch, for example, women produce questions using a wider pitch range compared to men (Haan & Van Heuven, 1999). Moreover,

speech rate in Dutch is affected by regional dialects and gender (Quené, 2008). In American English, Clopper and Smiljanic (2011) found differences in pause distributions and pitch accents between different dialects and genders.

Recently, Xie et al. (2021) found that prosodic variability is not only present between demographic groups (e.g., dialectal or gender groups) but also on an individual talker level. More specifically, they recorded lexically identical declarative statements vs. questions (e.g., 'It's raining.' vs. 'It's raining?'), produced in American English, and measured F0 and duration of the final syllable (i.e., '-ing'). Results indicated that individual talkers differed from each other in the category (i.e., statement vs. question) means and distributions for F0 and duration. In addition, while most talkers used F0 as primary cue, some talkers used both cues in combination with different probabilities, such that the use of F0 and duration could be correlated to different degrees across speakers. In other words, individual talkers seem to produce prosodic information with variability within each cue, but also vary in how the cues are combined to produce the intended structure (i.e., cue weighting in production). In sum, talker variability abounds in speech at both the segmental and suprasegmental level.

The literature on speech perception suggests that listeners are able to exploit this talker-specific cue usage to correctly perceive spoken words. This ability has been attributed to multiple cognitive mechanisms. That is, listeners use normalization to compensate for spectral differences between talkers (Sjerps et al., 2011) and differences in speech rate (Bosker et al., 2020; Reinisch, 2016), scaling the perceptual input to the surrounding acoustic context. Second, speech perception involves phonological abstraction in the lexicon, allowing listeners to map acoustically varying auditory signals to abstract lexical representations (McQueen et al., 2006) which are invariant to talker-related idiosyncrasies. Third, listeners constantly predict upcoming words during speech perception (Van Berkum et al., 2005) as well as how those words will be produced by specific talkers (Brunellière & Soto-Faraco, 2013). Those predicted word-forms will subsequently be easier to process upon perception. Finally, listeners use perceptual learning to change how acoustic input is mapped to prelexical perceptual categories of speech sounds (Eisner & McQueen, 2005) allowing listeners to adapt to varying acoustic input. Previous studies

(Lehet & Holt, 2020; Sjerps & Reinisch, 2015) further illustrated that these mechanisms are applied *in tandem*, showing that normalization and perceptual learning operate at different levels of speech processing. Even though all mechanisms together offer the listener the best solution to the variability problem, the remainder of this study will specifically focus on perceptual learning.

Perceptual learning studies have demonstrated that listeners can change how they map acoustic input to prelexical perceptual categories of speech sounds. More specifically, listeners can use lexical information to change which prelexical category (e.g., an /s/ or an /f/) is activated by the same acoustic token (e.g., an ambiguous fricative [?] in between /s/ and /f/) depending on the word it appears in: hearing [?] in 'platypu[?]' biases perception towards /s/, while 'gira[?]' biases towards /f/ (Norris et al., 2003). Further, listeners can change how they weigh the relative strength of multiple acoustic cues that signal a speech category based on distributional information in the speech input (Idemaru & Holt, 2011, 2014). For instance, Idemaru and Holt (2011) found that listeners can change how much perceptual weight is given to different acoustic cues that signal a speech sound (e.g., /b/ or /p/) based on the distribution of those cues in the speech input. In their experiment, they exposed English participants to words containing voiced/voiceless plosives (e.g., 'beer' vs. 'pier'). They found that when the canonical relation between fundamental frequency (F0) and voice onset time (VOT) was reversed (a voiced plosive is normally signaled with a high initial F0 and long VOT, but voiced plosives were now signaled with a low initial F0), listeners down-weighted their reliance on the unreliable cue (i.e., F0), showing rapid adaptation to short-term deviations in cue distributions. In other words, through perceptual learning listeners changed how each acoustic cue in the auditory signal contributed to perception of a plosive. Importantly, both cases described above involve adapting the strength of the link between the acoustic input and a prelexical category, altering the resulting amount of activation of that category.

In addition to these adaptations to single talkers, listeners can also adapt to speech originating from multiple talkers (Eisner & McQueen, 2005; Kraljic & Samuel, 2007), making it a useful mechanism to recognize words under high-variability contexts. This was also illustrated by Zhang and Holt (2018), who adopted the same paradigm as in Idemaru & Holt (2011, 2014), but crucially included speech originating

from two talkers differing in their F0 range. Results showed that the speech stimuli were perceived relative to the F0 range of each particular talker. More specifically, the same ambiguous F0 value was perceived as being higher in a low F0 range talker, inducing more 'beer'-responses, and *vice versa* for the high F0 range talker. In two subsequent experiments, Zhang & Holt (2018) presented the same stimuli with an ambiguous F0 value, but talker identity was cued by modulating voice characteristics (stimuli spoken by a male or female voice) or by visual presentation of a male or female talker. These experiments similarly resulted in more 'beer'-responses for stimuli spoken by a female talker, or stimuli accompanied by visual presentation of a female talker. In sum, these experiments illustrate simultaneous tracking of speaking styles from multiple talkers. This allows listeners to adjust the links between the acoustic input and perceptual categories for each individual talker, which facilitates perception of those talkers despite the between-talker variability in the signal.

While perceptual learning does indeed appear to be useful for dealing with talker variability, previous experiments have mostly studied it in relation to *segmental* variability. It remains unclear how perceptual learning is applied to *suprasegmental* variability among talkers. One of the few studies looking into this was performed by Xie et al. (2021), who examined the role of perceptual learning in the perception of questions vs. declarative statements. Participants were exposed to segmentally identical phrases (e.g., 'It's cooking {./?}') which, depending on the intonation contour, can either be perceived as a statement or a question. In the training phase, participants heard these phrases with ambiguous intonation contours, midway between a statement and a question, and received feedback on how to interpret them. Crucially, one group learned to perceive these ambiguous stimuli as statements (i.e., statement-biasing) while a second group learned to perceive the same phrases as questions (i.e., question-biasing). In a subsequent test phase, results showed that the statement-biasing group perceived the phrases more as statements while the question-biasing group perceived the phrases more as questions. This confirmed that perceptual learning is used to deal with variability in one type of prosody: sentence intonation. A similar finding for prosodically cued pragmatic structures was demonstrated in Kurumuda et al. (2014).

Prosody can also influence perception at the lexical level, distinguishing different words. For instance, lexical stress in free-stress languages, such as English and Dutch, can distinguish between segmentally identical words with contrastive stress patterns (e.g., 'IMport' vs. 'imPORT'). In Dutch, the target language of the present study, a stressed syllable is usually produced with a higher mean F0, longer duration, and greater intensity (Rietveld & van Heuven, 2009). Moreover, spectral balance (Sluijter & van Heuven, 1996, but see Severijnen et al., 2022) and acoustic vowel reduction (van Bergem, 1993) have also been identified as cues to lexical stress in Dutch. Vowel reduction appears to play a smaller role in Dutch than in English, where vowels in most unstressed syllables are fully reduced to schwa (Cutler, 1986, p. 202; Cutler & Pasveer, 2006). It is important to note that the acoustic cues to lexical stress are not weighted equally in production. In Dutch, for instance, when the word appears in an accented position in the sentence, the strongest cue to lexical stress is F0. When the word does not appear in an accented position, the strongest cue is duration, followed by spectral tilt, overall intensity and spectral expansion (Rietveld & van Heuven, 2009).

Lexical stress information plays an important role in word recognition. First, Cutler and Van Donselaar (2001) showed that, in Dutch, lexical stress is used to constrain lexical activation. They presented Dutch minimal stress pairs (*VOORnaam* vs. *voorNAAM*, 'first name' vs. 'respectable') in a lexical decision task testing repetition priming with stress-matching and stress-mismatching primes (e.g., target: *VOORnaam*; prime: either *VOORnaam* or *voorNAAM*). Results showed that only stress-matching primes facilitated target lexical decision reaction times (RTs). Second, Reinisch et al. (2010) showed that Dutch listeners use lexical stress *immediately* to facilitate word recognition. In an eye-tracking experiment, they exposed listeners to temporarily overlapping word pairs (e.g., *OCtopus* vs. *okTOber*). When participants were presented with one of the word pairs (e.g., *OCtopus*), listeners fixated the target word (*OCtopus*) more often than the competitor (*okTOber*) well before the point of segmental disambiguation (i.e., the onset of the third syllable). This illustrates that even when lexical stress is not strictly necessary to disambiguate different lexical candidates, listeners use it to facilitate perception. Similar effects have been found in English (Cutler, 1986; Jesse et al., 2017) and Italian (Sulpizio & McQueen, 2012).

As with sentence intonation, variability is also present in acoustic realizations of lexical stress. This was illustrated by Eriksson and Heldner (2015), who measured acoustic cues (F0, F0 variation, duration and spectral tilt) to lexical stress in English. They found several differences between talkers. First, the difference in mean F0 between stressed and unstressed syllables was larger for males compared to females. Second, females produced unstressed syllables with greater F0 variation and stressed syllables with longer durations than males. In addition to these gender differences, the speaking context (word lists, phrases, or spontaneous speech) also modulated the abovementioned cues. For example, the effects of stress on mean F0 were smaller in spontaneous speech compared to word lists and phrases. Variability in lexical stress production between genders and between speaking contexts, while in slightly different directions, has also been found in other languages including Italian and Swedish (Eriksson et al., 2013, 2016). Note again that the present study does not examine how listeners deal with variability between genders, but we nevertheless include these gender-related acoustic differences as an illustration of possible sources of between-talker variability.

Talkers also appear to vary on an inter-individual level in how they produce lexical stress. Severijnen et al. (2022) recorded Dutch participants producing segmentally overlapping words but differing in stress pattern (e.g., *VOORnaam*, 'first name' vs. *voorNAAM* ('respectable'). They then measured six acoustic cues that signal lexical stress in Dutch, in stressed and unstressed syllables: mean F0, duration, intensity, spectral tilt, F0 variation, and vowel quality. The analyses involved Linear Discriminant Analyses (LDA), which trained a model for each individual talker to predict whether each observation was a stressed or an unstressed syllable by finding the optimal linear combination of the acoustic cues. This resulted in a set of coefficients, for each talker, indicating how strongly each cue is weighted in production (Schertz et al., 2015). Results from these analyses illustrated that, on top of a general trend to use primarily F0, duration, and intensity, each talker used a unique set of cue-weights to signal lexical stress, illustrating large prosodic variability between individual talkers. Moreover, classes of cue-weighting strategies emerged from the data, differing in which cue was used as the primary cue. For words in an accented position, there was a group of primarily F0-users and a group of intensity-users. For words in an unaccented position, there was a group

of intensity-users and a group of duration-users. These results illustrate the large challenge that listeners are faced with: listeners must be able to perceive the correct stress pattern despite the immense amount of variability that is present between individual talkers.

Nevertheless, an understudied question concerns how listeners deal with talker variability in productions of lexical stress. To our knowledge, only two studies have looked into this. First, Bosker (2021) found evidence for perceptual learning in relation to suprasegmental cues to lexical stress in Dutch. In these experiments, participants heard ambiguous versions of minimal stress pairs (e.g., ambiguous between Dutch *CAnon* 'canon' and *kaNON* 'cannon') in an initial exposure phase. These words were differentially disambiguated for two participant groups by orthographic word forms on the screen. Specifically, presenting the orthographic word form of the Strong-Weak (SW) item (*canon)* induced a SW-bias while presenting the orthographic word form of the Weak-Strong (WS) item (*kanon)* induced a WS-bias. Results from a subsequent test where participants categorized a *CAnon-kaNON* continuum showed that participants in the SW-bias group indeed gave more SW responses while the WS-bias group gave more WS responses. Interestingly, this perceptual recalibration of lexical stress was also found across segmentally differing words. That is, exposure to ambiguous versions of *SERvisch* 'Serbian' vs. *serVIES* 'tableware', that were also disambiguated by orthography, led to similar recalibration effects on *CAnon* vs. *kaNON* test items. In sum, these experiments illustrate that listeners are able to adapt to variability in suprasegmental cues to lexical stress, and these adaptations are not tied to the episodic experiences with those words but seem to generalize across words, and thus imply that spoken word recognition involves abstract prosodic representations.

Second, Severijnen et al. (2021) investigated whether listeners can also adapt to variability in lexical stress *in a talker-specific manner*. In their EEG experiment, consisting of multiple training phases and a final test phase (in which behavioral and EEG data were recorded), native Dutch participants learned to associate non-word minimal stress pairs to object referents (e.g., *USklot* referring to a 'lamp', *usKLOT* referring to a 'train'). The non-words were produced by two male talkers who, importantly, used only one cue to signal lexical stress in the non-words (e.g., Talker 1 used only F0, while Talker 2 used only intensity).

In a subsequent test phase, participants heard semantically constraining carrier sentences (e.g., 'The word for lamp is *USklot'*) containing either talker-congruent versions of the non-words (i.e., produced with the talker-consistent cues; Talker 1 using F0) or talker-incongruent versions, produced with mismatching prosodic cues (e.g., Talker 1 suddenly using intensity). Behavioral results from a yes/no sentence verification task showed that participants were slower to respond to the talker-incongruent versions compared to the talker-congruent versions. The authors concluded that the delayed processing was due to the talker-incongruent prosodic cues, picked up through talker-specific perceptual learning about which talker used which cues to signal lexical stress in the training phase.

Even though Severijnen et al. (2021) provided evidence for talker-specific learning of lexical stress, their results spark several novel questions. First, does talker-specific learning of prosodic cues also have consequences for perception of the intended word, or does it only slow processing down, as observed in Severijnen et al. (2021)? That is, in Severijnen et al. (2021), the target word was identical in both conditions, so the intended word would always be perceived correctly regardless of the talker-cue mismatch. Indeed, the accuracy data in Severijnen et al. (2021) showed no difference between the two conditions. While their critical result (longer response times to the talker-incongruent condition compared to the talker-congruent condition) illustrated that listeners were slowed down in perception when lexical stress was marked with a cue which was not coherent with what was previously learned; it does not inform us on how perceptual learning of talker-specific prosodic cues affects perception of the intended word (i.e., which word is perceived; instead of how it is perceived). While both consequences (slowing down and incorrect perception) are problematic for communication, the latter is more problematic.

Second, can we replicate the behavioral finding in Severijnen et al. (2021), given that there was no modulation of the N200, an ERP related to acoustic-phonetic processing (Connolly & Phillips, 1994)? Specifically, a modulation of the N200-response would have provided electrophysiological evidence for a mismatch between the predicted prosodic cue and the perceived prosodic cue. The lack of an N200 modulation calls into question the replicability of the obtained behavioral results.

Third, do listeners apply the same learning mechanisms when acoustically richer test stimuli are used (i.e., stimuli involving multiple cues to lexical stress)? Specifically, the test stimuli in Severijnen et al. (2021) always contained only one cue to lexical stress. While this provided experimental control, it leaves open the possibility that listeners could employ the talker-specific learning mechanisms only because of the relative simplicity of the stimuli. Examining whether acoustically more complex stimuli elicit similar effects is crucial in examining speech perception closer to real-life situations.

Fourth, and following the previous argumentation, are the same learning mechanisms at work with existing words, compared to the non-words in Severijnen et al. (2021)? While using non-words as stimuli had the benefit of removing any episodic experiences with the words prior to the experiment, this leads to the possibility that since there was no previous experience with the non-words, the talker-specific effects could be easier to pick up on. In contrast, previous experience with how existing words are normally produced could interfere with storage of how newly encountered talkers produce those words. Using existing words as test stimuli could thus shed light on whether listeners are still able to pick up on the talker-specific cues despite previous potentially interfering episodic experiences.

The present study tried to answer these questions, thus aiming to provide further evidence for talker-specific perceptual learning of lexical stress in Dutch. Here we first outline how each question will be addressed. First, we measured categorization responses instead of RTs, which is a more direct measure to assess how talker-specific perceptual learning of lexical stress can affect which word is perceived. Second, we aimed at providing converging evidence, using a different measure, for the behavioral result in Severijnen et al. (2021). Third, the test stimuli in the present study contained multiple cues to lexical stress instead of only one cue in Severijnen et al. (2021). This allowed us to examine whether similar results as in Severijnen et al. (2021) could be observed with multidimensional stimuli that more closely resemble real-life speech. Fourth, the present study used existing words instead of non-words. This allowed us to examine these talker-specific learning effects in stimuli for which listeners already have pre-existing knowledge about and experience with how those words are normally produced.

We ran an online experiment consisting of a training phase and a test phase. In the training phase, participants heard Dutch minimal stress pairs (e.g., *VOORnaam* vs. *voorNAAM*, 'first name' vs. 'respectable'; SW and WS, respectively), produced by two male talkers. Similar to Severijnen et al. (2021), the stimuli were acoustically manipulated such that each talker cued lexical stress using only one acoustic cue. For instance, Talker 1 used only F0 (with intensity and duration set to ambiguous values) while Talker 2 used only intensity (talker-cue mappings were counterbalanced across participants). In a two-alternative forced choice (2AFC) task, participants were instructed to identify the correct member of the minimal pair, after which they received feedback on their responses. Based on the feedback, we expected participants to learn which cue was used by each talker (note that no explicit feedback was given regarding the cues; we expected participants to learn them implicitly). After the training phase, participants were tested on the same word pairs in another 2AFC task. This test included, next to perceptually 'clear' (i.e., unambiguous) control items with the talker-matching cue, also 'mixed items' in a different condition. These mixed items contained two conflicting cues to lexical stress, with F0 signaling one stress pattern, while intensity cued another. The crucial comparison then was how the perception of these mixed items was influenced by the talker-cue mappings learned in the training phase.

We predicted that participants would interpret the conflicting stress cues in the mixed items at test based on the learned information about which cue each talker tended to use. For example, if participants had learned that Talker 1 used F0 in training and then heard a mixed item produced by Talker 1 at test (e.g., F0 signaling SW, intensity signaling WS), they should prioritize in perception the stress pattern being signaled by F0 (e.g., SW). In contrast, if participants learned that Talker 1 used intensity, they should – when presented with the exact same test word – prioritize the stress pattern signaled by intensity (e.g., WS).

Method

Participants

We recruited 85 native speakers of Dutch from the Radboud University participant pool. All participants gave informed consent and were paid or received course credits for their participation. Five participants were excluded because they responded before target word onset on 75% of the trials. We excluded these participants because responses before target word onset could not reflect any perceptual processes related to the targets. The remaining 80 participants did not report having any hearing and/or reading problems (71 female, 9 male, age range: M_{age} = 21.81, SD_{age} = 3.76). We estimated the sample size through a power analysis by which we estimated a power of .858 (95% CI [.836 .879]) with 80 participants (see section *Power analysis* in Supplementary Information). The study was approved by the Ethical Committee of the Faculty of Social Sciences of Radboud University Nijmegen (Project Code ECSW2016-1403-391).

Stimuli

The stimulus set consisted of Dutch minimal stress pairs that were segmentally identical but differed in stress pattern. The set consisted of four disyllabic (e.g., *VOORnaam* vs. *voorNAAM*, 'first name' vs. 'respectable'; capitalization indicates lexical stress) and four trisyllabic word pairs (e.g., *VOORkomen* vs. *voorKOmen, '*to appear' vs. 'to prevent'). In all eight pairs, lexical stress lay on either the first syllable (i.e., Strong-Weak; SW words) or the second syllable (i.e., Weak-Strong; WS words). The words were identified through the CELEX database (Baayen et al., 1996) with matched word frequency between SW and WS words ($t(14)$ = - 0.69, p = .49). See Supplementary Table S2 for the complete stimulus set.

Recordings

The stimuli were recorded by two male native talkers of Dutch, naïve about the experiment's purpose. By selecting two same-gender talkers, we reduced acoustic variability in the stimuli. Moreover,

this promoted the formation of talker-specific instead of gender-specific representations. The talkers were instructed to produce each member of each minimal pair twice, once with stress on the first syllable, once with stress on the second syllable. Considering that the words would be presented in short carrier sentences in the experiment, the talkers were further instructed to produce each word as if it occurred at the end of a sentence (i.e., covertly producing the carrier sentence *Het woord is...* 'The word is...' in one's mind followed by overt production of the target word). This was meant to induce sentence-final prosodic properties in the recordings, such as F0-declination, intensity drop, and sentence-final lengthening. The talkers were allowed to practice carrying out this instruction and were successful in so doing after a few attempts. By recording the words separately from the carrier sentences, we avoided coarticulation with material in the carriers, which facilitated speech editing.

In addition to, and separately from, these words, we recorded three carrier sentences from both speakers. More specifically, we recorded one semantically neutral sentence (*Het woord is…*, 'The word is…') and two feedback sentences (*Goed, het woord is…,* 'Correct, the word is…'; *Fout, het woord is…*, 'Wrong, the word is…').

Stimulus manipulations

We required two types of stimuli. First, we needed clearly-stressed "control items" in which only one cue signaled lexical stress (e.g., only F0 or intensity while the other cues were set to ambiguous values). These items were used in the training phase to allow participants to learn the talker-cue mapping (i.e., which talker used which cue to signal lexical stress). Second, we needed ambiguous "mixed items" that contained two conflicting cues to lexical stress. In these stimuli, the cues appeared in opposing directions such that one cue (e.g., F0) signaled a SW pattern while the second cue (e.g., intensity) signaled a WS pattern. These items were used in the test phase to test whether participants learned the talker-cue mapping during the previous training phase, and used it to categorize these ambiguously-stressed words. In other words, if listeners learned that for instance Talker 1 always used F0 in the training phase and heard a word uttered

from Talker 1 with F0 signaling an SW pattern and Intensity signaling a WS pattern in the test phase, they should be more likely to categorize such words as having an SW pattern. Note that, in the test phase, the control items were also presented together with mixed items to reinforce the previously learned talker-cue mapping and to test that that mapping was still in operation. Duration was always kept at an ambiguous value in all stimuli.

Control items

In control items, only one cue was set to its optimal value to mark lexical stress (e.g., F0) while the remaining cues (e.g., intensity and duration) were set to ambiguous values. For the manipulation of control items, we followed the procedure in Severijnen et al. (2021) while also performing an extensive piloting stage (see Supplementary Information for all the details). The results of the pilot studies informed us about which acoustic values were required for control items to be identified as clear SW or WS words. After a careful selection, the SW tokens were correctly identified as SW with a mean proportion of SW responses of 0.83 (SD = 0.10) and the WS tokens were correctly identified as WS with a mean proportion of SW responses of 0.25 (SD = 0.14) in the pilot studies. From each word pair (N = 8), eight control items were generated, one for each combination of talkers (1, 2), cues (intensity, F0) and Pattern (SW, WS). The final number of control items was sixty-four. The acoustic properties of the control items are summarized in Table 1. See Figure 1 for the spectrograms of the control items for one of the words in the stimulus set and Figure 2a for a schematic representation of the control stimuli.

Table 1.

Mean acoustic measures (SD) of the prosodic cues in all syllables of control items. Two values are provided for duration in each syllable. These correspond to the duration values in disyllabic and trisyllabic words.

Pattern		**SW**									**WS**								
Syllable		**1st Syllable**			**2nd Syllable**			**3rd Syllable**			**1st Syllable**			**2nd Syllable**			**3rd Syllable**		
Cue		**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**
Talker 1 : F0 Talker 2: Intensity	**Talker 1**	154.29 (5.28)	68.66 (1.37)	[1]262.25 (27.75) [2]228.75 (1.25)	101.00 (1.42)	61.91 (1.32)	[1]361.50 (28.72) [2]180.00 (1.11)	105.96 (1.61)	44.93 (0.05)	248.00 (0.00)	106.35 (5.51)	68.66 (1.37)	[1]262.25 (27.75) [2]228.75 (1.25)	127.95 (6.42)	61.94 (1.33)	[1]361.50 (28.72) [2]180.00 (1.11)	103.26 (1.93)	44.91 (0.06)	248.00 (0.00)
	Talker 2	128.05 (5.98)	69.19 (9.68)	[1]288.80 (64.05) [2]232.00 (1.30)	121.58 (2.28)	50.63 (4.31)	[1]338.80 (56.52) [2]177.34 (1.19)	143.41 (5.29)	43.80 (1.97)	248.00 (0.00)	126.97 (7.07)	60.99 (5.18)	[1]288.80 (64.05) [2]232.00 (1.30)	122.08 (2.28)	65.36 (8.50)	[1]338.80 (56.52) [2]177.34 (1.19)	127.73 (2.36)	46.21 (5.48)	248.00 (0.00)
Talker 1: Intensity Talker 2: F0	**Talker 1**	127.10 (2.20)	69.55 (1.02)	[1]262.25 (27.75) [2]228.75 (1.25)	113.36 (3.80)	47.36 (7.72)	[1]361.50 (28.72) [2]180.00 (1.11)	124.41 (3.41)	44.91 (0.06)	248.00 (0.00)	128.53 (2.31)	55.84 (8.74)	[1]262.25 (27.75) [2]228.75 (1.25)	113.14 (5.55)	68.77 (2.50)	[1]361.50 (28.72) [2]180.00 (1.11)	124.41 (3.41)	44.91 (0.06)	248.00 (0.00)
	Talker 2	148.68 (1.27)	68.57 (1.08)	[1]288.80 (64.05) [2]232.00 (1.30)	112.35 (2.48)	59.16 (5.78)	[1]338.80 (56.52) [2]177.34 (1.19)	124.74 (1.72)	43.94 (1.99)	248.00 (0.00)	106.05 (5.86)	68.56 (1.06)	[1]288.80 (64.05) [2]232.00 (1.30)	129.24 (1.87)	59.23 (5.82)	[1]338.80 (56.52) [2]177.34 (1.19)	124.74 (1.72)	43.94 (1.99)	248.00 (0.00)

Note. [1]Duration values in disyllabic words, [2]Duration values in trisyllabic words

Mixed items

In mixed items, Intensity and F0 marked two conflicting stress patterns while duration was put to ambiguous values. For instance, the word *voornaam* could have F0 pointing towards a SW word (i.e., *VOORnaam*) while intensity pointed towards a WS word (i.e., *voorNAAM*). In this example the pattern is defined as 'F0-Intensity' because F0 is the strong cue on the first syllable while Intensity is the strong one on the second. For the 'Intensity-F0' pattern the reverse applies. All the manipulation procedures are described in Supplementary Information. Final stimuli were selected after extensive pilot testing to verify that these mixed items were ambiguous with respect to stress. However, pilot results demonstrated that the manipulations did not result in perfectly ambiguous words (i.e., falling precisely around a mean proportion of 0.5 SW responses). Instead, we found that F0 was weighed slightly more heavily than intensity, with 'F0-Intensity' patterns receiving an average proportion of 0.66 SW responses, and 'Intensity-F0' patterns receiving on average 0.48 SW responses (for details on the acoustic manipulations and the pilot studies, see Supplementary Information section 1.2). Nonetheless, the selected mixed items were perceptually considerably more ambiguous compared to the control items (i.e., falling roughly in between the clear SW and WS control patterns), indicating that the two cues to lexical stress were indeed conflicting with each other. From each word pair (N = 8), two mixed items were generated, one for each combination of Talker (1, 2) and Mixed Pattern (F0-Intensity: in which F0 signaled a SW pattern and intensity a WS pattern; Intensity-F0: vice versa) for a total of thirty-two mixed items. The acoustic properties of the mixed items are summarized in Table 2. See Figure 1 for the spectrograms of the mixed items for one of the words in the stimulus set and Figure 2b for a schematic representation of the mixed stimuli.

Table 2.

Mean acoustic measures (SD) of the prosodic cues in all syllables of mixed items. Two values are provided for duration in each syllable. These correspond to the duration values in disyllabic and trisyllabic words.

Pattern	**F0-Intensity**									**Intensity-F0**								
Syllable	**1st Syllable**			**2nd Syllable**			**3rd Syllable**			**1st Syllable**			**2nd Syllable**			**3rd Syllable**		
Cue	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**	**F0 (Hz)**	**Int (dB)**	**Dur (ms)**
Talker 1	153.51 (7.44)	63.16 (4.12)	[1]262.25 (27.75) [2]228.75 (1.25)	97.49 (6.57)	67.94 (2.20)	[1]361.50 (28.72) [2]180.00 (1.11)	102.11 (1.69)	44.91 (0.08)	248.00 (0.00)	106.81 (8.32)	69.50 (1.02)	[1]262.25 (27.75) [2]228.75 (1.25)	127.11 (5.03)	52.15 (3.04)	[1]361.50 (28.72) [2]180.00 (1.11)	100.75 (1.33)	44.92 (6.52)	248.00 (0.00)
Talker 2	148.66 (12.36)	65.15 (2.71)	[1]288.80 (64.05) [2]232.00 (1.30)	103.81 (8.46)	67.17 (1.67)	[1]338.80 (56.52) [2]177.34 (1.19)	132.00 (2.53)	44.96 (0.03)	248.00 (0.00)	107.69 (9.63)	69.54 (0.99)	[1]288.80 (64.05) [2]232.00 (1.30)	126.83 (13.53)	48.94 (5.24)	[1]338.80 (56.52) [2]177.34 (1.19)	122.74 (2.22)	44.94 (0.06)	248.00 (0.00)

Note. [1]Duration values in disyllabic words, [2]Duration values in trisyllabic words

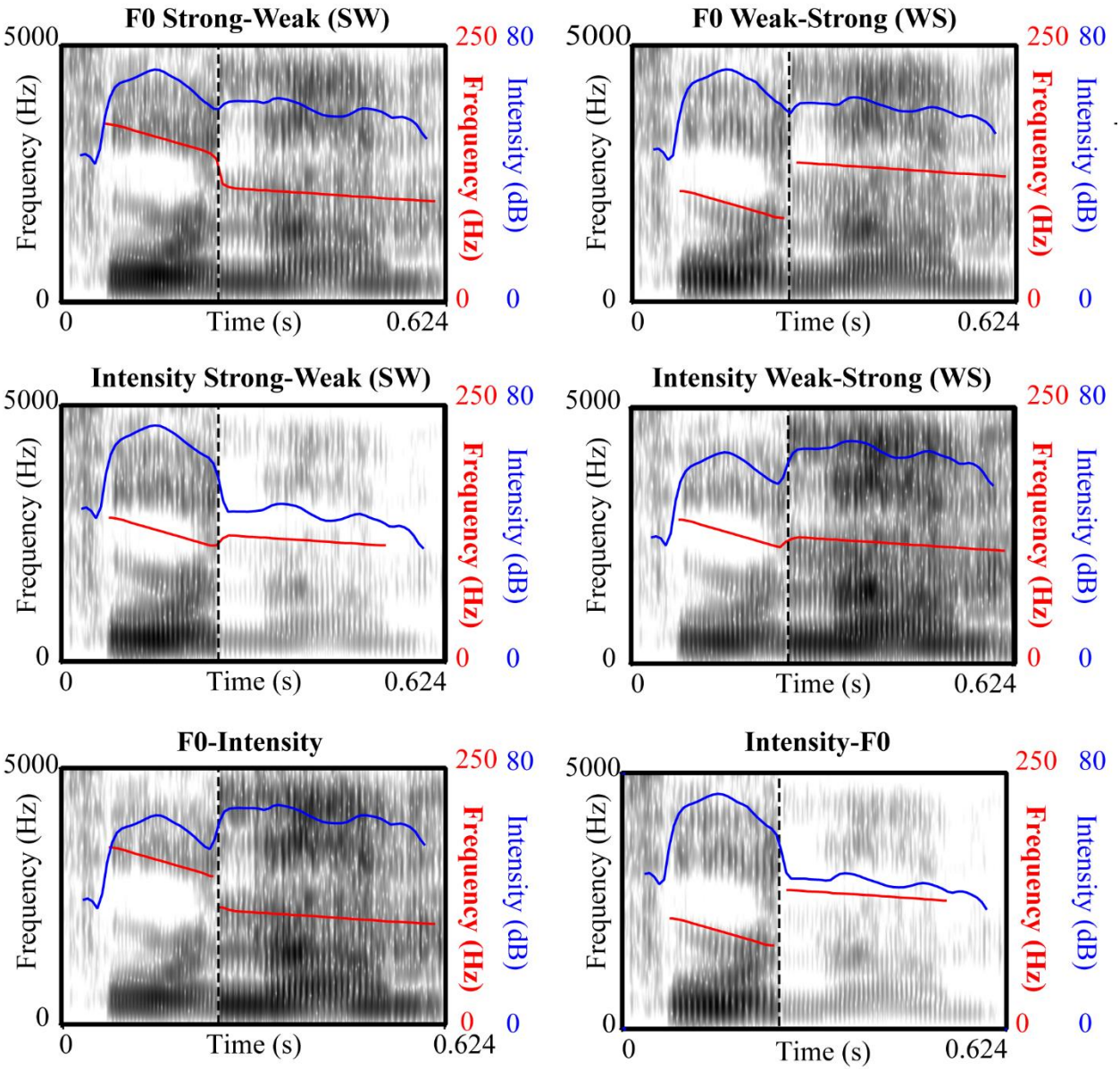


Figure 1. Spectrograms with F0 (red line) and Intensity contours (blue line) of one exemplary stimulus (the item *voornaam*) in the control (1st and 2nd row) and mixed versions (3rd row) produced by Talker 1. Vertical black dashed line indicates syllable boundary. The 1st row shows a control item in which F0 clearly marks lexical stress in a SW word (left column) or a WS word (right column) while intensity is put to fixed ambiguous values. The 2nd row shows the opposite: here it is Intensity that clearly cues lexical stress while F0 is set to fixed ambiguous values. The 3rd row shows the spectrograms of one exemplary mixed item in

which F0 and Intensity each mark two conflicting stress patterns at the same time: the left pane shows the F0-Intensity mixed pattern in which F0 signals an SW word while Intensity signals a WS word, while the right pane shows the opposite mixed pattern Intensity-F0.

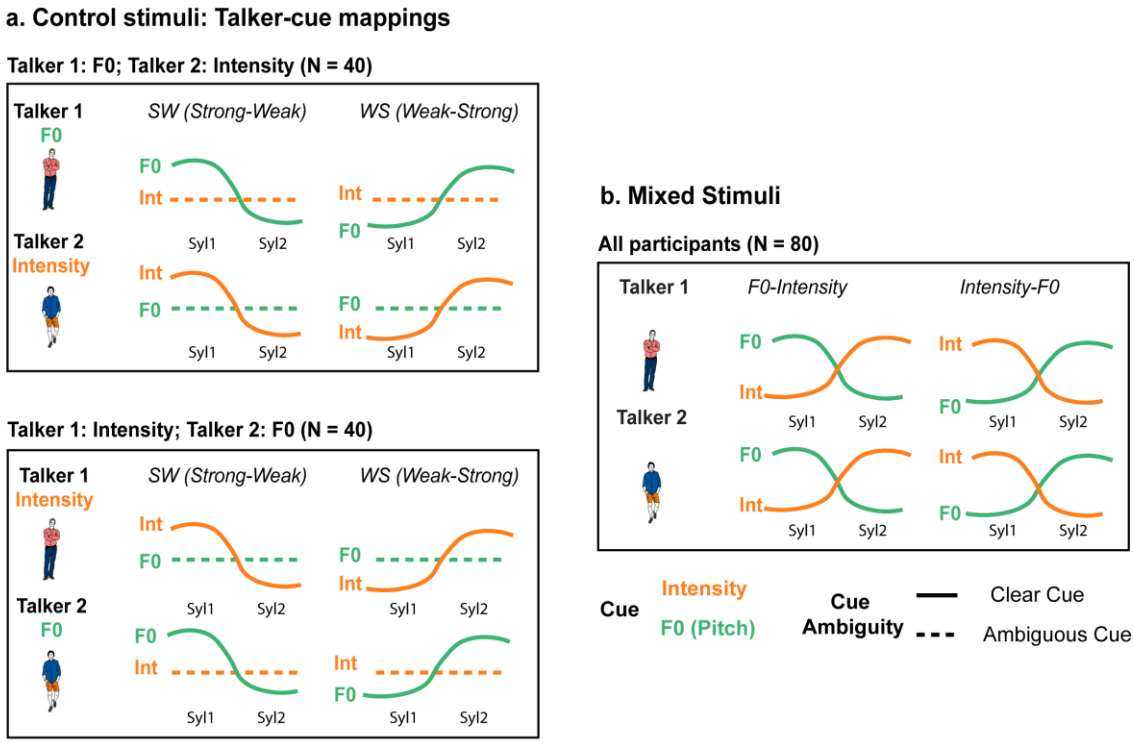


Figure 2. a. Schematic representation of the manipulation of Control stimuli divided by Mapping, Talker and Pattern. Continuous lines indicate a clear cue to lexical stress while dashed lines indicate an ambiguous cue to lexical stress. Green lines indicate the F0 contour while orange lines indicate the Intensity contour across the 1st and the 2nd syllable. The upper panel shows the talker-cue mapping in which Talker 1 always used F0 to signal lexical stress (green solid line) while Intensity was put to average values (orange dashed line) and Talker 2 always used Intensity to signal lexical stress (orange solid line) while F0 was put to average values (green dashed lines). The lower panel shows the reverse talker-cue mapping. **b.** Schematic representation of the manipulation of Mixed stimuli divided by Mixed Pattern and Talker. Continuous lines indicate a clear cue to lexical stress. Green lines indicate the F0 contour while orange lines indicate the Intensity contour. In mixed stimuli, both talkers used both cues within each word at the same time. The two

cues always indicated conflicting lexical stress patterns (e.g., while F0 could signal a Strong-Weak pattern, Intensity signaled a Weak-Strong pattern).

Procedure

The experiment was built and hosted on the Gorilla Experiment Builder (www.gorilla.sc). Participants first performed a headphone screening (Woods et al., 2017), in which participants heard three dichotically presented pure tones, with one of the dichotic tones presented 180° out of phase across the two stereo channels, and were instructed to identify the quietest one. This task is intended to be easy over headphones but difficult over loudspeakers due to phase-cancellation. In previous work, this task achieved 80% accuracy in detecting headphone users vs. speaker users (Milne et al., 2021). Thus while this task is not perfect at detecting headphone users, it still ensured that the majority of participants were likely to have been wearing. These participants could continue with the experiment proper, which consisted of a familiarization, training and a test phase. The brief familiarization phase ensured that participants were familiar with the pronunciations of the words and their meaning. The aim of the training phase was for participants to learn, implicitly, which talker used which cue to signal lexical stress in the control items. After the training phase, participants were tested on the mixed items in a final test phase, which would allow us to observe how perception of the mixed items was affected by the training phase. Participants were randomly assigned to one of the talker-cue mappings (e.g., half of the participants heard Talker 1 use F0 and Talker 2 use intensity; and *vice versa* for the other half) and response position (e.g., SW response items always appearing on the left side, WS response items on the right side; and *vice versa*), with all possible combinations counterbalanced across participants.

Familiarization phase

In the familiarization phase, participants were visually presented with orthographic word forms representing each member of a minimal pair, their definitions, and example sentences with the words, and auditory presentations of the control items of the corresponding words, spoken by both talkers. The auditory

presentations followed the talker-cue mappings of the training and test phases. For example, if a participant heard Talker 1 using F0 and Talker 2 using intensity in the training and test phases, this was also the case during familiarization. This ensured that participants were familiar with the stimuli before the training phase started. The trial structure was as follows. First, we visually presented the SW member of a minimal pair (e.g., *VOORnaam*) on the top left corner of the screen and auditorily presented its corresponding control stimulus (see Figure 3a for the trial structure). After 1500 ms, we presented the WS member on the top right corner with its corresponding control stimulus. Afterwards, we visually presented their definitions below the orthographic depictions of the words, and below that two example sentences. Participants indicated using button presses whether they did or did not know either of the two words.

Training phase

In the training phase, two groups of participants were exposed to the control items embedded in carrier sentences (e.g., *Het woord is VOORnaam*, 'The word is first name'), produced by both talkers (see Figure 3b for the trial structure). Furthermore, a cartoon image (see Figure 3) of the respective talker producing that sentence was visually presented, appearing 700 ms before sentence onset. This was done to strengthen the acquisition of talker-specific cue usage while also reducing the risk of potentially considering that the two talkers were the same person. Two response options (i.e., the two members of the minimal pair; *VOORnaam* and *voorNAAM*) were orthographically presented on the lower left and right corners of the screen 200 ms before sentence onset. Talker images and response options remained on the screen until a response was given. Participants were instructed to respond with button presses ([Z] or [M] responding to the left or right response options, respectively) after target word onset. If no response was given after 5 s from target word onset, the trial was recorded as a missing data point. After the response, we presented a feedback sentence (e.g., *Goed, het woord is VOORnaam*, 'Correct, the word is first name', or *Fout, het woord is VOORnaam*, 'Wrong, the word is first name'). Participants were then visually instructed to press the correct button again based on the feedback. After their second response, they heard the target word one final time in isolation. Participants thus heard the same target word three times in each trial. The next trial

started 1 s after the final auditory presentation of the target word in the previous trial. Each group of participants listened to different control items on the basis of the assigned talker-cue mapping. One group heard Talker 1 using only F0 to signal stress (intensity and duration were ambiguous), while Talker 2 used only intensity (F0 and duration were ambiguous). The other group heard the reverse talker-cue mapping. In this way we counterbalanced the assignment of cues to each talker to prevent possible confounds related to cues or talkers alone. Each group listened to 32 control items: 8 target words (e.g., *voornaam*), with 2 stress patterns (SW or WS) uttered by 2 talkers (Talker 1 and Talker 2). Each item was repeated 6 times for a total of 192 experimental trials. The experimental trials were preceded by 8 practice trials in which the same stimuli from the experimental trials were presented to familiarize participants with the task before starting the actual experiment. Practice trials were then excluded for statistical analyses. The trials were presented in pseudo-randomized order in 4 different counterbalanced lists and no word pairs were ever repeated in two consecutive trials.

Test phase

The test phase was similar to the training phase but differed in two aspects (see Figure 3c for the trial structure). First, participants did not receive feedback on their responses. Instead, the next trial began 1 s after participants gave their response. Second, next to control items (played on 50% of the test trials), the target words in the test phase also included mixed items (on the other 50% of the test trials). Note that we still included control items in this test phase to provide solid anchors of unambiguous items with congruent talker-cue mappings to participants. That is, each group of participants heard the same control items they heard in the training phase, following the same talker cue mapping. Again, these were 32 control items which were repeated 3 times each for a total of 96 trials. In addition, *both* groups, regardless of the talker-cue mapping, listened to the same 32 mixed items: 8 target words (e.g., *voornaam*) with 2 stress patterns (F0-Intensity, Intensity-F0), uttered by 2 talkers (Talker 1, Talker 2) which were repeated 3 times each for a total of 96 trials. The test phase thus consisted of 192 experimental trials in total and was not

preceded by practice trials. The trials were presented in pseudo-randomized order, without repeating word pairs in consecutive trials.

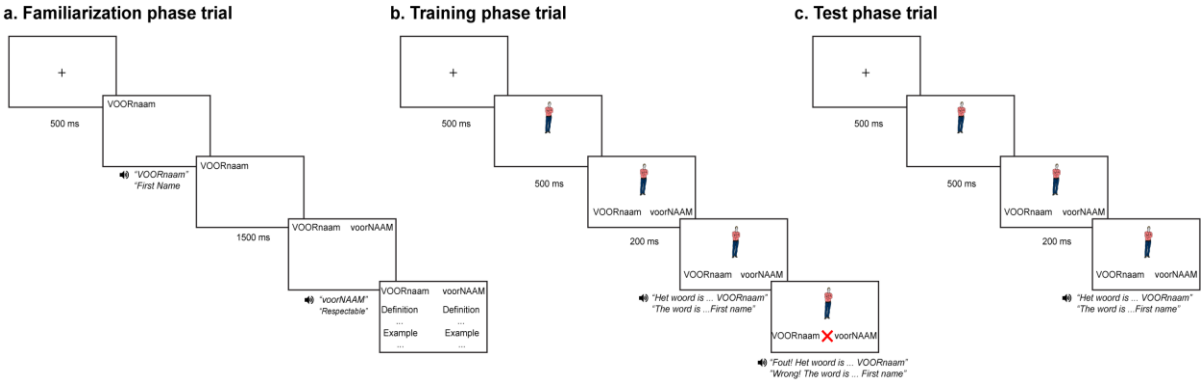


Figure 3. a. Trial structure of the familiarization phase. **b.** Trial structure of the training phase. **c.** Trial structure of the Test phase.

Data analysis

Prior to data analysis, we calculated the percentage of timed-out trials (0.7%) and we further excluded any trials with RTs below 100 ms relative to target word onset (0.5% of the trials), reaching 1.2% of removed trials overall. The latter was done for two reasons. First, due to an error in the way the experiment was programmed, it was possible for participants to respond before target word onset (i.e., before they heard the word). Since such responses do not represent any perceptual processes related to the target words, we decided to exclude them. Second, RTs below 100 ms also include trials on which the majority of the first syllable had not yet been heard in its entirety (shortest first syllable duration was 171 ms). As a result of these exclusions, the final number of analyzed trials was 30,329 (22,775 trials for control items and 7,554 trials for mixed items). Furthermore, we analyzed the familiarization data in order to check whether participants knew all the word stimuli they were presented during the experiment. This analysis showed that 93% of participants knew at least 14 of the 16 words (41.3% knew all words, 33.8% knew 15/16 words and 18.8% knew 14/16 words) while only 7% of participants knew fewer than 13/16 words. Given these results, we ran the analyses of the behavioral data on the complete dataset and a dataset in

which the unknown words were excluded (see Supplementary Tables S9-S12). Since the results from the two analyses were comparable, we base our conclusions on the complete dataset.

We ran two separate models, one for the mixed items and one for the control items. The model for the mixed items tested our primary research question, namely whether the talker-cue mapping (e.g., Talker 1 using F0, Talker 2 using intensity) affected responses depending on the Mixed Item Pattern (e.g., F0-Intensity, Intensity-F0) and who produced the mixed items (e.g., Talker 1 or Talker 2). The model for the control items verified whether the intended stress pattern (SW or WS) was correctly perceived across both the training and test phase. In both models, we analyzed the binomial categorization responses (SW coded as 1; WS as 0) using a Generalized Linear Mixed model (GLMM) with a logistic linking function with the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2020). When needed, post-hoc tests were performed via the emmeans R package (Lenth et al., 2018).

For the mixed items, we obtained an initial model through forward modeling (tested using likelihood-ratio tests, factors were entered in the order in which they are described below) containing the following fixed factors: Mapping (categorical predictor with two levels, deviance coded with Talker 1 using F0 and Talker 2 using intensity coded as -0.5; Talker 1 using intensity and Talker 2 using F0 coded as 0.5), Mixed Pattern (categorical predictor with two levels, deviance coded with F0-Intensity coded as -0.5 and Intensity-F0 coded as 0.5), and Talker (categorical predictor with two levels, deviance coded with Talker 1 coded as -0.5 and Talker 2 coded as 0.5). The model also contained random intercepts for Participant and Item. This model showed a significant three-way interaction between Mapping, Pattern, and Talker (β = -1.74, SE = 0.20, z = -8.85, $p < .001$). While this very broadly illustrates that the categorization responses are indeed dependent on the exact combination of the Mapping, Mixed Pattern and Talker, the exact interpretation of this interaction is not as straightforward. Moreover, testing for further interactions with other factors would require four-way interactions, which are even more difficult to interpret. Therefore, to simplify the analyses, we created a new categorical variable with two levels (Predicted Response: Predicted SW or Predicted WS) that coded for this three-way interaction. Specifically, Predicted Response coded for what we expected the predicted response to be (Predicted SW or Predicted WS), depending on the three

factors Mapping, Mixed Pattern, and Talker. For example, for the Mixed Pattern F0-Intensity produced by Talker 1, our hypothesis was that participants who had learned that Talker 1 used F0 during the training phase should perceive this item as SW. In contrast, participants who had learned that Talker 1 used intensity should perceive the exact same item as WS. We continued model selection through forward modeling using this new categorical variable.

The final model with the best fit to the data included the following factors: Predicted Response (categorical predictor with two levels, deviance coded with Predicted SW coded as -0.5 and Predicted WS coded as 0.5), Talker (categorical predictor with two levels, deviance coded with Talker 1 coded as -0.5 and Talker 2 coded as 0.5), Mixed Pattern (categorical predictor with two levels, deviance coded with F0-Intensity coded as -0.5 and Intensity-F0 coded as 0.5), and Trial Number (continuous predictor). This last predictor was obtained by normalizing the original trial number for mixed items ranging from 1 to 96, obtaining a measure of overall proportion of trials ranging from 0 to 1 within each individual participant. Furthermore, we included interactions between Predicted Response and Talker and Predicted Response and Trial Number. We also included random intercepts for Participant and Item, by-Participant random slopes for all the fixed factors and by-Item random slopes for Predicted Response and Talker. Following the procedure in Bates et al. (2015), we optimized the random structure using Principal Component Analysis (PCA) on the models to obtain the structure that contained the minimally required factors to explain the largest variance. This avoided overfitting problems. The full model syntax was Response ~ Predicted Response * Talker + Mixed Pattern + Predicted Response * Trial Number + (1 + Predicted Response + Talker | Participant) + (1 + Predicted Response + Talker | Item).

For the control items, the model with the best fit to the data (obtained through forward modeling, tested using likelihood-ratio tests, factors were entered in the order in which they are described below) included the following fixed factors: Pattern (categorical predictor with two levels, deviance coded with SW coded as -0.5 and WS coded as 0.5), Phase (categorical predictor with two levels, deviance coded with training phase coded as -0.5 and test phase coded as 0.5), Talker (categorical predictor with two levels, deviance coded with Talker 1 coded as -0.5 and Talker 2 coded as 0.5) and Trial Number (continuous

predictor). Trial Number was normalized with the same method applied to mixed items but separately within each phase (Training, Test) containing 192 trials and 96 trials respectively. Furthermore, we included interactions between Pattern and Phase, Pattern and Trial Number, Phase and Trial Number, Pattern and Talker, and a three-way interaction between Pattern, Phase, and Trial Number. We also included random intercepts for Participant and Item with by-Participant random slopes for the factors Pattern, Talker and Phase and by-Item random slopes for Pattern and Talker. The random structure was optimized using the same approach as used for the mixed items. The full model syntax was Response ~ Pattern * Phase + Talker + Pattern * Phase * Trial Number + (1 | Participant) + (1 + Pattern + Talker + Phase| Item).

Transparency and Openness

We report how we determined our sample size (*Power Analysis section in Supplementary Information*), all data exclusions (see section *Participants* and *Data Analysis*), all manipulations (see sections *Stimulus Manipulations* and Procedure), and all measures in the study (see section *Data Analysis)* and we followed JARS (Kazak, 2018). Across the whole Methods section, all the employed software and packages are reported. All data, analysis code, and research materials are available at https://osf.io/dczx9/?view_only=44f227db3c134685ad1db9cf46e317f7. This study's design and its analysis were not pre-registered.

Results

Mixed items

The analysis of mixed items crucially tested whether participants applied their learning about how the two talkers signaled lexical stress to perceive spoken words with conflicting stress cues. Results for Mixed items are summarized in Figure 4 (e, f, g). Qualitative plots showing the results for Mixed items divided by Talker, Pattern/Predicted Response and Mapping are depicted in Figure 4b. The complete model output is given in Supplementary Table S12. The main effect of Predicted Response (β = -0.74, SE = 0.14,

$z = -5.20$, $p < .001$) revealed a significant difference between Predicted SW and Predicted WS trials. As depicted in Figure 4f, participants showed a higher proportion of SW responses (light red bar) for the Predicted SW trials (Mean prop. of SW resp. = .59; SE = .01) and a lower proportion of SW responses (light blue bar) for the Predicted WS trials (Mean prop. of SW resp. =.49; SE = .01). This result illustrates that perception of identical Mixed items was affected by the learned information about which talker used which cue to signal lexical stress.

Further, a significant main effect of Pattern was found ($\beta = 1.42$, SE = 0.27, $z = 5.34$, $p < .001$), showing that the F0-Intensity pattern (left bar in Figure 4e) was perceived as being more SW-biased compared to the Intensity-F0 pattern (right bar in Figure 4e). This demonstrates that participants weighed F0 as a cue to lexical stress more heavily than intensity, corroborating outcomes from pilot study 3 (see Supplementary Table S8). Importantly, our main effect of interest (i.e., the effect of Predicted Response) was still present regardless of the effect of Pattern. The model also revealed a main effect of Talker ($\beta = 0.44$, SE = 0.18, $z = 2.47$, $p = .014$), showing that Talker 2 was perceived as being more SW-biased than Talker 1.

Lastly, a marginally significant interaction effect between Predicted Response and Trial Number ($\beta = 0.36$, SE = 0.19, $z = 1.92$, $p = .055$) was found. As shown in Figure 4e, while the Predicted SW responses were characterized by a negative tendency towards fewer SW-biased responses in later trials (Mean Slope = -0.14, SE = 0.14; descending light red line in Figure 4e), the Predicted WS responses showed the opposite trend (Mean Slope = 0.22, SE = 0.14; ascending light blue line in Figure 4e). This indicates that the effect of the talker-cue mappings learned during the training phase was shrinking as the test phase went on, as the Predicted SW and the Predicted WS responses became less SW- and WS-biased, respectively.

No further effect reached significance. The results of this analysis were replicated also when the trials with words that participants reported not to know prior to the experiment were excluded (see Table S12 in Supplementary Information for complete model outputs).

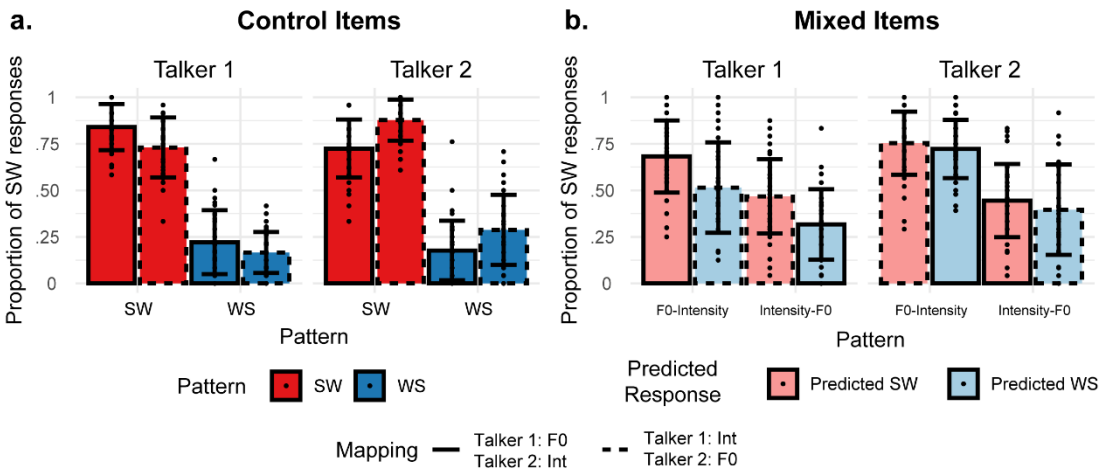
Control items

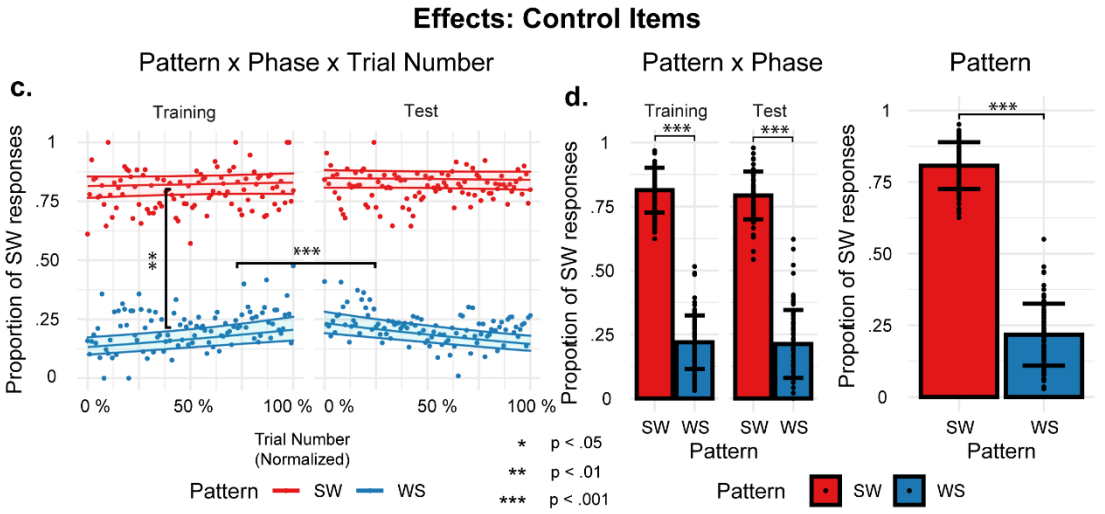
The analysis of control items tested whether participants could categorize words with one clear cue to stress with acceptable accuracy in both the training and test phase. Results for control items are summarized in Figure 4 (a, c, d). Qualitative plots showing the results for Control items divided by Talker, Pattern and Mapping are depicted in Figure 4a. Complete model output and results from post-hoc tests are given in Supplementary Tables S9, S10, and S11. The model revealed a significant effect of Pattern (β = -3.14, SE = 0.21, z = -15.00, $p < .001$), showing that participants could correctly perceive the stress cues for SW (Mean prop. of SW resp. = .80, SE = .009; red bar in the right plot of Figure 4d) and WS (Mean prop. of SW resp. = .22, SE = .01; blue bar in the right plot of Figure 4d) patterns.

A small interaction effect between Pattern and Phase (β = -0.44, SE = 0.15, z = -2.95, p = .003) was also found, suggesting a slightly reduced Pattern effect in the test phase compared to the training phase (see left plot in Figure 4d). Post-hoc tests confirmed the presence of a strong difference between the SW and the WS pattern both in the training (β = 3.18, SE = 0.20, z = 15.91, $p < .001$) and in the test phase (β = 3.15, SE = 0.20, z = 15.37, $p < .001$). Moreover, in the training phase, participants gave slightly more SW-biased responses for both the SW pattern (β = -0.17, SE = 0.06, z = 2.64, p = .011) and the WS pattern (β = -0.13, SE = 0.06, z = 2.11, p = .035). Another interaction effect between Phase and Trial Number (β = -0.64, SE = 0.13, z = -4.95, $p < .001$) was found as well as a main effect of Phase (β = 0.47, SE = 0.08, z = 5.70, $p < .001$).

The model of control items further showed a significant interaction effect between Pattern, Phase and Trial Number (β = -0.95, SE = 0.26, z = 3.68, $p < .001$) represented in Figure 4c. Post-hoc comparisons were performed on the slope of Trial Number comparing the levels of Pattern (SW, WS) at each level of Phase (Training, Test) and comparing the levels of Phase within each level of Pattern. These tests revealed a significant difference in the slope of Trial Number between the SW and WS patterns in the Training (β = 0.52, SE = 0.15, z = 3.52, $p < .001$) but the same test was only marginally significant during the Test phase (β = -0.38, SE = 0.21, z = -2.00, p = .06). During Training, while the SW pattern was stable throughout the phase (Mean slope = -0.06, SE = 0.10; red line in left plot of Figure 4c), the WS pattern showed a negative

slope (Mean slope = -0.58, SE = 0.10; descending blue line in left plot of Figure 4c). This revealed that as the training phase went on, participants gave more WS responses to the items with a WS pattern. Conversely, in the test phase, the opposite tendency was found: while SW pattern was still stable (Mean slope = 0.10, SE = 0.15; red line in right plot of Figure 4c), showing no differences between training and test ($\beta = 0.16$, SE = 0.18, $z = -0.90$, $p = .368$), the WS pattern showed a positive slope (Mean = 0.53, SE = 0.15; ascending blue line in right plot of Figure 4c) revealing that participants gave fewer WS responses as the test phase went on, differently from the training phase ($\beta = -1.11$, SE = 0.18, $z = -6.07$, $p < .001$). No further effect reached significance. The results of this analysis were replicated when the trials including words that participants reported not to know prior to the experiment were excluded (see Tables S9, S10, S11 in Supplementary Information for complete model outputs and post-hoc tests).


a.
Control Items
Talker 1
Talker 2
Proportion of SW responses
1
.75
.50
.25
0
SW
WS
Pattern
Pattern
SW
WS
b.
Mixed Items
F0-Intensity
Intensity-F0
Predicted Response
Predicted SW
Predicted WS
Mapping
Talker 1: F0
Talker 2: Int
Talker 1: Int
Talker 2: F0


Effects: Control Items
c.
Pattern x Phase x Trial Number
Training
Test
**
***
0 %
50 %
100 %
Trial Number
(Normalized)
* p < .05
** p < .01
*** p < .001
d.
Pattern x Phase
Pattern

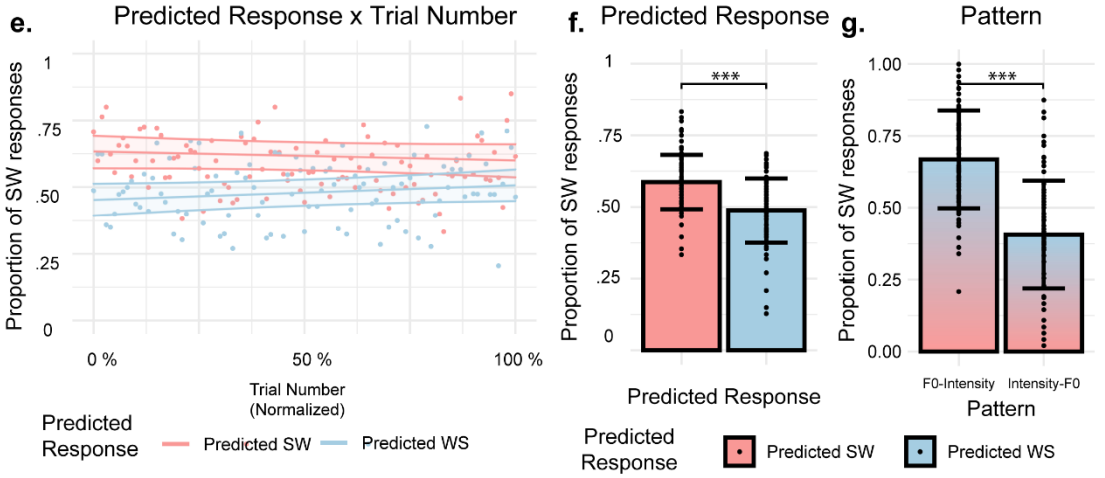

Effects: Mixed Items
e.
Predicted Response x Trial Number
f.
Predicted Response
g.
Pattern

Figure 4. Qualitative plots (1st row) and plots of the significant effects (2nd row) of the proportion of SW responses (y-axis) split by different factors (x-axis). **a.** Qualitative plots for Control items averaged across participants, words and phases, separately for each Pattern (SW in red, WS in blue), Talker (1, 2) and Mapping (solid line indicates Talker 1: F0, Talker 2: Intensity; dashed line indicates Talker 1: Intensity, Talker 2: F0). Points indicate individual participants and error bars represent the Standard Error. **b.** Qualitative plots for Mixed items averaged across participants and words divided by Pattern (F0-Intensity, Intensity-F0), Talker (1, 2), Mapping (solid line indicates Talker 1: F0, Talker 2: Intensity; dashed line indicates Talker 1: Intensity, Talker 2: F0) and Predicted Response (Predicted SW in light red, Predicted WS in light blue). Points represent individual participants and error bars represent the Standard Error. **c.** Interaction effect between Pattern, Task and Trial Number for Control items. Proportion of SW responses split by Task (Training, Test) and Pattern (SW in red, WS in blue). Individual points represent proportions of SW responses averaged across trials separately within each word and each participant. Superimposed lines represent the slope predicted by the model with hued 95% Confidence Intervals. **d.** Interaction effect between Pattern and Task (left plot) showing the proportion of SW responses split by Phase (Training, Test) and Pattern (SW in red and WS in blue). Main effect of Pattern (right plot) showing proportion of SW responses averaged across phases and split by Pattern. **e.** Interaction effect between Predicted Response, and Trial Number for Mixed items. Proportion of SW responses split by Predicted Response (Predicted SW in light red, Predicted WS in light blue Individual points represent proportions of SW responses averaged across trials separately within each word and each participant. Superimposed lines represent the slope predicted by the model with hued 95% Confidence Intervals. **f.** Main effect for Predicted Response. Proportion of SW responses divided by Predicted Response (Predicted SW in light red, Predicted WS in light blue) **g.** Main effect of Pattern. Proportion of SW responses divided by Predicted Response (F0-Intensity, Intensity-F0).

Discussion

We investigated whether listeners could adapt to between-talker variability in lexical stress by learning to associate specific stress cues to specific talkers. Our study showed that this was the case: through perceptual learning, participants mapped different cues to lexical stress to two specific talkers and used this information to differentially categorize words with conflicting stress cues (i.e., mixed items) depending on

this talker-cue mapping. This was shown in our statistical model by a main effect of Predicted Response, indicating that participants gave responses biased towards the stress pattern category consistent with the talker-contingent cue.

Our findings are in line with previous studies showing talker-specific perceptual learning of segmental (Eisner & McQueen, 2005; Theodore & Miller, 2010; Zhang & Holt, 2018) and suprasegmental information (Severijnen et al., 2021; Xie et al., 2021). The use of only one clear cue to lexical stress by different talkers in the training phase, which is a pattern that differs from the usual tendency in Dutch, where F0 and intensity co-occur as stress cues (Rietveld & van Heuven, 2009), pushed participants to recalibrate the perceptual weights of suprasegmental cues in a talker-contingent way. It is worth mentioning that these speech production patterns, in which a given talker prioritizes one main cue (either F0, intensity, or duration) to produce lexical stress have been found in an experiment examining individual differences in the acoustic correlates of lexical stress in Dutch (Severijnen et al., 2022). That is, in Severijnen et al., (2022) it was shown that some talkers indeed have a strong preference for primarily using one cue among others (e.g., F0), while other talkers may prefer using another cue (e.g., intensity). The present study showed that listeners are able to exploit these talker-specific patterns. More specifically, after participants learned that Talker 1 used only intensity as cue to stress, they increased the weight of this cue in subsequent perception, while down-weighting F0. This interpretation is consistent with the dimension-based statistical learning account (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015; Zhang & Holt, 2018) which states that listeners exploit short-term acoustic regularities to adjust the efficiency of specific physical dimensions in signaling speech categories. Importantly, this interpretation extends the domain of the account to suprasegmental cues. As seen in Zhang & Holt (2018) and Xie et al. (2021), despite acoustic cues (i.e., intensity and F0) being equally distributed at the global level of the experiment (i.e., the number of trials in which intensity or F0 was the main cue to stress was identical), participants managed to track the regularities of both cues at the same time in a talker-contingent way, separating them into distinct distributions. In our experiment, this talker-contingent cue tracking may have been supported by other

acoustic talker differences (e.g., segmental pronunciation idiosyncrasies) in the carrier sentences and/or target words themselves, the visual talker cues (different cartoon images), or both (cf. Zhang & Holt, 2018).

In line with the dimension-based learning approach, it appears that listeners built talker-contingent weight sets based on the cue distributions picked up in the training phase, in which unambiguous words were presented. However, as illustrated by the results in the test phase, these cue weights did not remain fixed after the training phase, but were re-adjusted during the test phase. That is, even though the interaction between Predicted Response and Trial was only marginally significant, the difference between Predicted SW and Predicted WS responses seemed to be gradually attenuated, at least numerically, as the test phase went on (see Figure 4e). It is possible that the presence of mixed items weakly altered the talker-specific cue-distributions as they provided two conflicting cues to stress. Previous studies have shown significant and more robust "unlearning" effects for talker-specific segmental information (Kraljic & Samuel, 2005), prosodic information (Kurumada et al., 2014), and most importantly for lexical stress (Severijnen et al., 2021). All these studies showed that providing new talker-specific information at test, which may have been fully or partially incompatible with that presented during the training phase, encouraged further updating of previously acquired talker-specific perceptual weights.

It is important to mention that, despite extensive piloting, the mixed stimuli were still characterized by a perceptual imbalance. Specifically, the effect of Pattern in the mixed stimuli analysis showed that when F0 and Intensity appear as conflicting cues to lexical stress in mixed items, F0 seems to have more weight in general than Intensity in driving categorization responses. This result is also in line with one of the pilot studies. On the one hand, this effect is consistent with the cue hierarchy in Dutch (Rietveld & van Heuven, 2009) for which, in words in an accented position, F0 is thought to be a stronger cue to lexical stress than intensity. On the other hand, the presence of this effect highlights the difficulty to obtain completely balanced multidimensional stimuli. Either way, the presence of this imbalance does not invalidate the effect of Predicted Response. That is, during the test phase, participants in both groups responded to acoustically identical mixed items, so any imbalance in the stimuli was present for both groups. The only difference between the two groups was the talker-cue mapping that they learned during training.

Another important consideration concerns the variability between items in the stimulus pilots. That is, while we managed to select an SW-token and a WS-token for all items, there was considerable variability in the quality of the continua (i.e., some items showed a clear perceptual switch while others less so). Moreover, the pilots were conducted on a different participant sample than the main experiment, which might raise the question of whether a new participant sample would perceive the items in a similar manner. However, responses on the control items in training and test confirmed that the new participant sample correctly perceived the intended stress pattern in those items. Nevertheless, the fact that we observed the talker-specific learning effect regardless of the between-item variability (which we took into account with by-Item random intercepts and slopes in our models), speaks to the robustness of the learning effect. More specifically, it illustrates that it is not just a learning mechanism that listeners can use in an experimental setting with perfectly balanced stimuli, but that listeners can abstract away from the variability and extract the information that is important in dealing with the between-talker variability in the experiment.

The results of the present study are well explained by speech perception models that include a belief-updating mechanism (Kleinschmidt & Jaeger, 2015; Norris et al., 2016; Norris & McQueen, 2008) that allows listeners to recalibrate perception in a talker-specific way. These kinds of models address the variability problem by describing speech perception as a probabilistic process. In these models, listeners behave not only as *optimal recognizers* (Norris & McQueen, 2008) that use all their prior and present knowledge to understand speech, but also as *ideal adapters* (Kleinschmidt & Jaeger, 2015), able to recalibrate their prior knowledge to optimize recognition in future situations. Considering these two notions, listeners appear to have prior beliefs about the statistical distributions of phonetic cues in speech built through a lifetime's experience. In our specific case we can think about prior experience as pertaining to the canonical distribution of stress cues in Dutch. Listeners can then learn the talker-specific cue distributions of novel talkers they have not encountered before and update their prior beliefs about the general distribution of cues by exploiting the structured variability (e.g., the consistent use of one or more cues to stress) in the utterances these novel talkers produce. The belief-updating feature of these frameworks relies on the need of listeners to update their prior knowledge. In other words, if the encountered lexical

stress pattern differs from listeners' prior beliefs (e.g., containing non-canonical cue distributions), they should be pushed to change their knowledge about stress cues by recalibrating perceptual weights towards more optimal word recognition. Interestingly, belief-updating models were developed to explain results from studies exploring adaptation to segmental information. Our results indicate that these models are also appropriate to explain how listeners adapt to suprasegmental information and hence deal with suprasegmental variability in a talker-specific way.

Note that the talker-specific perceptual learning in the present study occurred in the absence of any particular perceptual need imposed by the experimental design of the training phase. That is, the present study did not employ particularly ambiguous items in training that, unlike in the classical perceptual learning paradigm (Norris et al. 2003), guide recalibration. This might suggests that it is not necessary to have ambiguous items in the training phase for listeners to adapt. Instead, the mismatch between the cue distributions of the talkers in the experiment and participants' expectations about those distributions (picked up through previous experience with Dutch male talkers) is enough to guide recalibration. This is in line with the Ideal Adapter Framework (Kleinschmidt & Jaeger, 2015), which predicts that listeners need to adapt when new situations deviate from previous experience.

The present study extends the findings of a similar study (Severijnen et al., 2021), and answered four open questions. First, does talker-specific learning of prosodic cues also have consequences for perception of the intended word, or does it only slow down perception, as observed in Severijnen et al. (2021)? The present study illustrated that talker-specific learning of prosodic cues indeed affects which word is perceived. This has important implications for speech perception, as it shows that perceptual learning can be used to reduce the risk of miscommunication by helping listeners to navigate their way through different prosodic realizations across talkers. Second, can the behavioral finding in Severijnen et al. (2021) be replicated, given that even though the behavioral results illustrated perceptual learning of prosodic cues, there was no modulation of the N200, an ERP related to acoustic-phonetic processing (Connolly & Phillips, 1994)? The present study provided converging behavioral evidence, using a different behavioral measure, for talker-specific perceptual learning of lexical stress, strengthening the robustness of

the effect found in Severijnen et al. (2021). Third, do listeners apply the same learning mechanisms when acoustically richer test stimuli are used? Results illustrated that talker-specific learning was not impeded by the use of stimuli with two acoustic cues in opposing directions, suggesting that listeners can apply talker-specific learning to richer, more complex stimuli. Fourth, are the same learning mechanisms at work with existing words, compared to non-words in Severijnen et al. (2021)? The present study showed that talker-specific learning can also be applied to existing words, illustrating that talker-specific learning is a mechanism that can exploit short-term regularities and supersede long-term information about previously known words.

It would be interesting to know whether talker-specific perceptual learning of suprasegmental cues generalizes to previously unheard words (i.e., test words which are not included in the training stimuli) as seen in previous work on segmental information (McQueen et al., 2006). This kind of generalization is considered as an index of a pre-lexical abstraction process through which adjustments of perceptual weights based on training to ambiguous words can then be used to recognize other words (McQueen et al., 2006). It is incompatible with strictly episodic accounts of word recognition which postulate that listeners store only detailed acoustic instances of heard words (Goldinger, 1998). Sulpizio & McQueen (2012) showed that listeners form abstract representations of lexical stress and recently Bosker (2021) provided evidence for generalization of perceptual learning of lexical stress cues to new words. Our design, however, did not test for generalization because the same lexical items were used in the training and test phases. Nevertheless, our results do provide some indications of generalization of the learning process across word episodes, as mixed stimuli were not encountered during training. In this regard, it is important to recall that there were physical differences between the control and mixed items. Mixed items were not synthesized by directly splicing syllables of control items together (e.g., one intensity-driven strong syllable and one pitch-driven weak syllable). In fact, physical levels of intensity and pitch in control items were drawn from different steps of the pilot-tested continua with respect to the mixed items. This was done to raise the level of ambiguity of mixed items for which less-extreme steps (i.e., less SW or WS) were used relative to those used to make the control items. Second, while control items had one clear cue to stress (e.g., intensity or

F0) and two other cues set to ambiguous levels (e.g., F0 and duration or intensity and duration), mixed items had two conflicting cues to stress and only one ambiguous cue (i.e., duration). Thus, at test, participants were presented with words that were physically different from the ones they heard in training in which additional conflicting cues were present. If participants were to learn episodic instances of stressed words in the training phase without extracting talker-specific cue weights, they would not have shown differences between the Predicted SW and the Predicted WS patterns. Further research is required to explicitly test for generalization across words, but if the hints of its occurrence in the present study were to hold up, it would suggest that the observed learning effects reflect the uptake of abstract information about how talkers speak.

In sum, we showed that listeners can learn how two specific talkers signal lexical stress and apply that learning in recognizing subsequent tokens spoken by the same talkers. These results fit well with Bayesian models that predict that listeners can adjust their prior beliefs about phonetic cues on the basis of short term regularities. Importantly, while such models have been developed to explain how listeners deal with segmental variability, the present study suggests that they can also account for the way in which listeners deal with suprasegmental variability.

Data Availability Statement

The data and the stimuli of this experiment have been made openly available at https://osf.io/dczx9/?view_only=44f227db3c134685ad1db9cf46e317f7 under a CC-BY 4.0 International license.

Acknowledgements

We would like to thank Keanu Kiriwenno and Jelle Gerritsma whose voices were recorded for the speech materials used in this study and the Summer Meeting Series on Talker Variability Across Levels of Speech Categories organized by the Kinder Lab at the University of Rochester for the helpful discussion of results.

Funding

This research was jointly funded by a PhD internal round grant of the Donders Centre for Cognition at Radboud University [J.M., H.R.B.], the Doctoral School of the Department of Cognitive Science of the University of Trento [G.D.D.], and an ERC Starting Grant (HearingHands; #101040276) from the European Research Council of the European Union [H.R.B.].

References

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America*, *116*(3), 1729–1738. https://doi.org/10.1121/1.1779271

Adank, P., van Hout, R., & Velde, H. van de. (2007). An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties. *The Journal of the Acoustical Society of America*, *121*(2), 1130–1141. https://doi.org/10.1121/1.2409492

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552. https://doi.org/10.1121/1.1528172

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The CELEX lexical database (cd-rom)*.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. http://arxiv.org/abs/1506.04967

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

Bosker, H. R. (2022). Evidence for selective adaptation and recalibration in the perception of lexical stress. *Language and speech*, *65*(2), 472-490.

Bosker, H. R., Sjerps, M. J., & Reinisch, E. (2020). Temporal contrast effects in human speech perception are immune to selective attention. *Scientific Reports*, 10(1), 5607. https://doi.org/10.1038/s41598-020-62613-8

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain 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. (1986). Forbear is a Homophone: Lexical Prosody Does Not Constrain Lexical Access. *Language and Speech*, *29*(3), 201–220. https://doi.org/10.1177/002383098602900302

Cutler, A., & Pasveer, D. (2006). Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. *3rd International Conference on Speech Prosody*.

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

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

Eriksson, A., Barbosa, P., & Åkesson, J. (2013). The acoustics of word stress in Swedish: A function of stress level, speaking style and word accent. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 778–782.

Eriksson, A., Bertinetto, P. M., Heldner, M., Nodari, R., & Lenoci, G. (2016). *The Acoustics of Lexical Stress in Italian as a Function of Stress Level and Speaking Style*. 1059–1063. https://doi.org/10.21437/Interspeech.2016-348

Eriksson, A., & Heldner, M. (2015). The Acoustics of Word Stress in English as a Function of Stress Level and Speaking Style. *Proceedings of Interspeech*, 41–45.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251. https://doi.org/10.1037/0033-295X.105.2.251

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.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939.

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009.

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

Kang, K.-H. (2013). F0 Perturbation as a Perceptual Cue to Stop Distinction in Busan and Seoul Dialects of Korean. *Phonetics and Speech Sciences*, *5*(4), 137–143. https://doi.org/10.13064/KSSS.2013.5.4.137

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. https://doi.org/10.1037/a0038695

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01546-0

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive Science*, *41*, 885–912.

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, 104328. https://doi.org/10.1016/j.cognition.2020.104328

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R Package Version*, *1*(1), 3.

Lisker, L. (1986). "Voicing" in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees. *Language and Speech*, *29*(1), 3–11. https://doi.org/10.1177/002383098602900102

Lisker, L., & Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. WORD, *20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126. https://doi.org/10.1207/s15516709cog0000_79

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

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

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18. https://doi.org/10.1080/23273798.2015.1081703

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, *123*(2), 1104–1113. https://doi.org/10.1121/1.2821762

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reinisch, E. (2016). Natural fast speech is perceived as faster than linearly time-compressed speech. Attention, Perception, & Psychophysics, 78(4), 1203–1217. https://doi.org/10.3758/s13414-016-1067-x

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

Rietveld, T., & van Heuven, V. J. (2009). *Algemene Fonetiek (3e geheel herziene druk)*. Bussum: Coutinho.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *WIREs Cognitive Science*, *11*(2). https://doi.org/10.1002/wcs.1521

Severijnen, G., Bosker, H. R., & McQueen, J. M. (2022). How do “VOORnaam” and “voorNAAM” differ between talkers? A corpus analysis of individual talker differences in lexical stress in Dutch. In the 18th Conference on Laboratory Phonology (LabPhon 18).Severijnen, G. G. A., Bosker, H. R., & McQueen, J. M. (2022). Acoustic correlates of Dutch lexical stress re-examined:

Spectral tilt is not always more reliable than intensity. *Proceedings of the 11th International Conference on Speech Prosody*. Speech Prosody, Lisbon, Portugal.

Severijnen, G. G. A., Bosker, H. R., Piai, V., & McQueen, J. M. (2021). Listeners track talker-specific prosody to deal with talker-variability. *Brain Research, 1769*. https://doi.org/10.1016/j.brainres.2021.147605.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia, 49*(14), 3831–3846. https://doi.org/10.1016/j.neuropsychologia.2011.09.044

Sjerps, M. J., & Reinisch, E. (2015). Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 41*(3), 710. https://doi.org/10.1037/a0039028

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America, 100*(4), 2471–2485. https://doi.org/10.1121/1.417955

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.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America, 128*(4), 2090–2099. https://doi.org/10.1121/1.3467771

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America, 125*(6), 3974–3982. https://doi.org/10.1121/1.3106131

van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication, 12*(1), 1–23. https://doi.org/10.1016/0167-6393(93)90015-D

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

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

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

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