UNIVERSITÀ
DI TRENTO

Department of Psychology and Cognitive Science

Doctoral School in Cognitive Science

XXXIV Cycle

Abstraction, retrieval, and perceptual learning in the integrated processing of linguistic and talker-related information

Advisors

dr. Simone Sulpizio

dr. Michele Scaltritti

PhD Candidate

Giuseppe Di Dona

Academic Year 2020-2021

Contents

Overview

During speech perception listeners receive both linguistic information about the speech content as well as information regarding the identity of the talker. While these two aspects have been traditionally studied in isolation, with a dominant interest for linguistic information over talker identity, it is now a widely accepted notion that these two kinds of information are processed in an integrated way. The inclusion of talker-related information in the domain of speech perception highlighted both benefits and challengers for listeners. On the one hand, linguistic and talker-identity information appear to be mutually beneficial for the extraction of both kinds of information from the speech signal. On the other hand, listeners must take care of the great acoustic variability that characterizes the physical dimensions linked to the two kinds of information. The aim of the present dissertation is to study three specific cognitive mechanisms that listeners can use to access the benefits of the integrated processing of linguistic and talker-related information as well as to deal with their intrinsic variability. Three empirical studies employing both behavioural and neurophysiological techniques highlight peculiar aspects of *abstraction*, *memory retrieval* and *perceptual learning* mechanisms in relation to the consequences of including the talker in the study of speech perception.

Chapter 1: General Introduction

Speech perception refers to a wide set of cognitive processes by which human listeners can understand what is being said. All these processes occur in a plethora of social contexts and in all of them we inevitably find at least one talker. This consideration may seem trivial except for the fact that cognitive and neurophysiological models of speech perception leave the talker in second place. For these accounts, talker-related information is considered either as an information to remove from the speech signal or is not taken into consideration at all (Gaskell & Marslen-Wilson, 1998; Halle, 2013; Hickok & Poeppel, 2007; Jacobsen, Schröger, & Alter, 2004; Strange, 1989). Although, while understanding *what* is being said is often presented as the one and only goal of speech perception, it should be remembered that identifying *who* is speaking is fundamental for communicating effectively in the complex social world (Kuhl, 2011). From this apparently strict separation between the *what* and the *who*, the study of speech perception developed across separate paths.

Starting from the acoustic-phonetic level, linguistic and talker-identity information are thought to be indexed by different physical features: while phonemes (i.e., vowels) are often ascribed to their formant frequencies (Hewlett & Beck, 2013), talker-identity (i.e., voice identity) is often reduced to the mean fundamental frequency (Baumann & Belin, 2010). Consequently, also the study of the cognitive processes underlying the extraction of linguistic and talker-identity information has inherited this functional and neurobiological segregation (Belin & Zatorre, 2003; DeWitt & Rauschecker, 2012; Zäske et al., 2017).

Despite this apparent division, the extraction of linguistic and talker-identity information can often be co-dependent, given the fact that the physical dimensions in which they can be encoded are intrinsically overlapping (Hirahara & Kato, 1992; Ladefoged & Broadbent, 1957). From such considerations, several studies started to outline the possibility that linguistic and talker-identity information could be processed in an integrated way. Different behavioural and neurophysiological studies provided converging evidence about phoneme identification being impaired when the talker keeps changing and likewise, talker-identification being hampered by sudden variations in phonemes (Kaganovich et al., 2006; Mullennix & Pisoni, 1990; Zhang et al., 2016), suggesting that the extraction of linguistic and talker-identity information are computationally interdependent. The integration of linguistic and talker-identity information not

only emerges at an acoustic-phonetic level of processing but also during higher and more complex cognitive operations. Van Berkum et al. (2008) showed that listeners infer social characteristics based on the talkers' voices and use this information to re-evaluate the meaning of utterances in real time, suggesting that understanding speech engages a constant integration of semantic information emerging from the linguistic content and pragmatic information about the communicative context (i.e., the social information inferred from the talker's voice). These two examples may suggest that the integration of linguistic and talker-identity information is intrinsic to speech perception at multiple levels of computational complexity and abstraction.

The inclusion of talker-related information in the study of speech perception, and its strong binding with linguistic information, stimulated several studies from which two main points can be drawn. On the one hand, several studies have shown that talker-identity information as well as linguistic information are mutually beneficial for the extraction of one another. Prior exposure with a talker's voice was shown to facilitate phoneme perception (Eisner & Mcqueen, 2005; K. Johnson, 1990; Kraljic & Samuel, 2006, 2007; Norris et al., 2003), word recognition in ideal (Creel et al., 2008; Nygaard et al., 1994; Nygaard & Pisoni, 1998) and noisy situations (Johnsrude et al., 2013; Newman & Evers, 2007; Souza et al., 2013). Likewise, identifying the talker is easier when this occurs in the listeners native language (Bregman & Creel, 2014; Perrachione & Wong, 2007; Zarate et al., 2015).

On the other hand, both talker-identity and linguistic information are subject to high degree of acoustic variability. Multiple theoretical accounts agree about some key computational steps that listeners encounter during speech perception (Eisner & McQueen, 2018; Kemmerer, 2015) in order to take care of such variability. Considering the extraction of linguistic information, listeners essentially map a continuous signal to discrete linguistic units (i.e., phonemes) that – when combined – form meaningful words and sentences. Nevertheless, talkers produce phonemes in phonetically different ways (Allen et al., 2003; Peterson & Barney, 1952) thus listeners never hear the same acoustic realization of a phoneme twice. What is even more outstanding, is that one phoneme produced by one talker could be very acoustically similar to another phoneme produced by a different talker (Repp & Liberman, 1987). Known as the *lack of invariance problem* (Liberman et al., 1967), this posits an arduous computational challenge for listeners. All these problems may equally concern talker-identification, as listeners can recognize talkers by hearing different words which entail variation induced by linguistic information.

Chapter 1

The remaining sections of this introductory chapter will describe three different cognitive mechanisms that listeners may use to benefit from the integration of linguistic and talker-identity information and/or to deal with their congenital variability. First, *abstraction* will be introduced as one of the fundamental mechanisms by which listeners can solve the lack of invariance problem while dealing with talker-variability through the formation of talker-invariant phoneme representations. Importantly, the possibility of extending the use of this mechanism to deal with linguistic variability in the talker-identity domain will be addressed. Second, *memory retrieval* will be described as a mechanism that allows listeners to take advantage from familiar linguistic or talker-identity information during speech perception. Further, the issue about shared vs. segregated retrieval mechanisms across linguistic and talker-identity domains will be addressed. Third, *perceptual learning* will be framed as the potential mechanism that allows listeners to map the idiosyncratic ways by which speakers produce speech to abstract representations of segments and suprasegmental structures. More specifically the focus will be on lexical stress.

1.1 Abstract voice representations

Abstractionist models of speech perception entail abstract phonological or prelexical representation (Gaskell & Marslen-Wilson, 1997; McClelland & Elman, 1986; Norris, 1994) that listeners form by abstracting away from the acoustic-phonetic variability of the context. Regarding the issue of talker variability, abstract phonological representations might be crucial in supporting the storage and retrieval of information about how talkers produce speech. For instance, through the *normalization* process, listeners can compute an acoustic-phonetic model of the talker's voice that can be used online (i.e., within the same instance) to facilitate the mapping between talker-specific phonetic realization of phonemes to abstract phonological representation (Sjerps et al., 2011a, 2011b). Additionally, through *perceptual learning*, listeners can directly extract talker-specific details about the phonetic realization of phonemes while storing them within abstract representations of phonemes and generalize it to future encounters with the same phoneme, also when it is embedded in previously unheard words (Eisner & Mcqueen, 2005; McQueen et al., 2006).

While the works presented in this dissertation were developed following the abstractionist accounts of speech perception that critically depend on the existence of abstract representations of speech units, it is important to consider that other theoretical accounts focus on different

perspectives. For instance, episodic models of speech perception hypothesise that listeners store independent episodic memories for every word they encounter. These episodic traces would encode both linguistic and talker-related acoustic information. When a new word is heard, its memory trace is compared to all the other stored traces and the outcome of this process results in a weighted activation pattern of all the traces in memory, acting as a kind of distributed representation (Goldinger, 1998). While abstractionist accounts frame talker-related variability as *a problem to solve* by removing talker-related information through the normalization and abstraction processes, episodic accounts frame it as a valuable resource to listeners which can favour the word recognition process.

One important aspect of such models that partly characterize the studies of this dissertation is the value of talker-related information for speech perception. In fact, the abstraction process has been mainly studied for its role in dealing with talker-variability allowing listeners to grasp the linguistic content while discarding talker-related information. Nevertheless, listeners could face a similar problem when they must recognize the talker from different speech tokens: mapping different voice tokens to an abstract talker's voice representation (Belin et al., 2004; Latinus & Belin, 2011) might be a similar issue as reconducting different speech tokens to abstract phonological representations. Hence, the abstraction process might also favour talker identification when listeners are faced with linguistically-driven phonetic variability.

Numerous studies employing neurophysiological techniques showed that the cognitive system can spontaneously form abstract phoneme representations very early in time (Eulitz & Lahiri, 2004; Jacobsen, Schröger, & Alter, 2004; Shestakova et al., 2002), while no study showed that this might also be the case for voices. Yet, different studies showed that listeners should be able to form abstract representations of both phonemes and talker-identities, respectively invariant to the talker's voice or to the speech content, but only as a consequence of explicit task demands (Bonte et al., 2009, 2014; Zäske et al., 2014, 2017).

It is thus still unclear whether the abstraction process that leads to the formation of phoneme-invariant talker voice/identity representation is spontaneously activated or follows environmental demands and/or the listener's intentions. Investigating these issues may clarify whether the cognitive system is preferentially tuned towards linguistic or talker-identity information during speech perception. In Chapter 2, this and other related issues are addressed, as well as the possibility that listeners may form abstract representations via a more general-domain

abstraction mechanism. Understanding whether this aspect is specific for speech or broadly pertains to the auditory domain might provide important insights about its scopes and constraints.

1.2 Memory Retrieval of familiar information

Perceiving speech is easier in a familiar context, being the context a familiar talker or a familiar/native language. The *familiar talker advantage*, that is the facilitation listeners get during speech recognition while hearing a familiar talker, and the *language familiarity effect*, that refers to the enhanced talker-recognition performance in native linguistic contexts (Perrachione, 2017), clearly highlight the connection of linguistic and talker-identity information and their mutual influence during speech perception. Talker familiarity might be an important aid when immersed in a multi-talker situation, where listeners are requested to track speech from specific talkers while ignoring others to understand what is being said. Johnsrude et al. (2013) showed that this is the case, and additionally provided evidence that talker familiarity is beneficial for listeners also when familiar talkers must be ignored. In their experiment, listeners heard two simultaneously presented sentences. In one condition, one sentence was produced by a familiar voice (i.e., participants' spouses) while the other by an unfamiliar one. In another condition, both sentences were produced by unfamiliar voices. Participants were asked to recognize specific words either from the familiar or the unfamiliar voices on the basis of cue-words contained in the same sentences, which informed participants about what voice would have given the target information. Participants not only showed higher accuracy when target words were produced by the familiar voices but also when the familiar voice was to be ignored and the unfamiliar one to be attended with respect to when both voices were unfamiliar.

Language familiarity, instead, might be a natural consequence of listeners familiarizing with voices in their native language. Zarate et al. (2015) familiarized native English participants with the voice of 5 different multilingual talkers through a voice recognition task and then tested their recognition capacity. Both the familiarization and the test phases were repeated in English, German, Mandarin, Pseudo-English (i.e., English pseudowords) and non-speech (e.g., laughs, cry). At test, participants showed gradually descending voice recognition accuracy when shifting from English to pseudo-English, German, Mandarin and non-speech, indicating that one possible source of the language familiarity effect might be the familiarity with phonetic and phonological information.

These two effects described above might reflect the ability of the listener to jointly use linguistic and talker-related information both when recognizing words or identifying talkers that may essentially rest upon shared cognitive mechanisms. In fact, different neurophysiological studies showed that representations of native phonemes or familiar talkers are automatically retrieved by the cognitive system very early in time (Beauchemin et al., 2006; Dehaene-Lambertz, 1997; Näätänen et al., 1997) and possibly by similar memory retrieval mechanisms. Within this perspective, the similarities between the memory retrieval processes for known linguistic and talker-identity information might suggest that listeners exploit the information they know to better parse the information they do not know. This may lead to an efficient management of cognitive resources as listeners may focus their attention onto the unknown information once they have retrieved the familiar one. As mentioned above, one caveat is that listeners possibly familiarize with voices in a context where linguistic information is understood, hence familiar. Similarly, the acquisition of phonemes in children initially develops in contexts where the voices they hear are indeed familiar. Therefore, in Chapter 3 the memory retrieval processes for familiar voices and phonemes which have been previously learned in isolation are investigated and compared.

1.3 Talker-specific perceptual learning

As previously mentioned, different talkers may produce the same phonemes in phonetically different ways, as well as different phonemes in phonetically similar ways (Adank, Smits, et al., 2004; Peterson & Barney, 1952). Nonetheless, when listeners familiarize with the talker's voice via a training procedure, word recognition is facilitated (Nygaard et al., 1994), suggesting that the exposure to a talker's voice might help listeners in dealing with lack of invariance and talker-variability. Further, the talker-familiarity effect generalizes to previously unheard words (Yonan & Sommers, 2000), possibly indicating that the advantage that listeners get by being exposed to the talker's voice acts at a pre-lexical level. This particular feature inspired researchers to study whether such an effect could stem from listeners' ability to learn how specific talkers produced specific linguistic units. To this regard, different studies showed that listeners can learn how different talkers produce phonemes by readjusting the perceptual weights of physical cues by which such phonemes can be identified (Eisner & Mcqueen, 2005; Kraljic & Samuel, 2007; Norris et al., 2003; Samuel & Kraljic, 2009). In Eisner & Mcqueen (2005), Dutch participants heard a talker producing an ambiguous fricative sound between [f] and [s] that was inserted at the end of

different words during a lexical decision task. One group of participants heard the ambiguous sound in place of an [f] sound (as in *witlof*, "chicory"), while the other group of listeners heard it in place of an [s] sound (as in *radijs*, "radish"). Subsequently, participants were tested on an ambiguous continuum of [εf]–[εs] syllables. The [f]-biased group gave more [εf] responses while the [s]-biased group gave more [εs] responses. The same effect was recorded also when the test fricatives were spliced after vowels of novel talkers but not when novel talkers produced the whole [εf]–[εs] continuum unless the novel talker's fricative sounds were put into the words of the exposure talker, suggesting that listeners learned in a talker-specific way. Importantly in a subsequent similar study, McQueen et al. (2006) showed that this kind of talker-specific learning generalized across words. This means that participants could grasp how the talkers produced specific phonemes by hearing one set of word during a learning phase but then applied such information also when new words were presented at test. This is considered a crucial feature of this particular learning process called *perceptual learning* that is thought to entail abstract phoneme representations: by hearing real words in which an ambiguous phoneme has been inserted, listeners can activate the corresponding phoneme representation and act at a pre-lexical level to readjust the perceptual weights associated to its primary cues.

On the other hand, a series of studies showed that when listeners encounter talkers with non-native accents they *relax* their acceptance criteria of what would be a correct (i.e., native) linguistic production along different levels (Reinisch & Weber, 2012; Witteman et al., 2013; Zheng & Samuel, 2020). *Relaxation of criteria* could be an alternative mechanism to deal with talker-related variability, compared to perceptual learning. Nonetheless, this adaptation mechanism seems to be restricted to non-native accents, and it is not yet clear if the same mechanism could be applied to individual native talkers and the related variability across different linguistic levels.

While the remainder of this section focuses on perceptual learning, it is worth mentioning that the studies on non-native accents highlight the value of suprasegmental information for speech perception: when hearing non-native speakers, their prosody is possibly the first thing a native speaker notices. Suprasegmental information though, is also crucial to perceive the message content. In several languages (e.g., Italian, Dutch, German, Spanish) suprasegmental information, and more specifically *lexical stress*, encodes precious linguistic information, allowing the distinction of minimal pairs as in the English words FOREarm (i.e., the part of the arm extending

from the elbow to the wrist) – foreARM (i.e., to prepare for trouble). While the number of minimal *stress pairs* is quite limited in the mentioned languages, it was shown that knowing the position of stress can drastically reduce the number of activated lexical competitors, providing an information as useful as the segmental one (Cutler & Pasveer, 2006). Moreover, suprasegmental information can also be affected by between-talker variability (Eriksson & Heldner, 2015; Xie et al., 2021), thus listeners might be prone to learn how different talkers produce it, in order to correctly perceive speech. Studying how listeners learn the idiosyncratic ways in which talkers produce speech, and in this specific case, how they produce lexical stress is particularly relevant for a deeper understanding of the beneficial relationship between linguistic and talker-related information. In Chapter 4, perceptual learning is presented as an effective mechanism by which listeners learn to adapt to talker-related variability by updating the connection between phonetic cues and linguistic categories on the basis of short-term exposure. Moreover, in the General Discussion section, the possibility that talker-specific perceptual learning hinges upon abstraction and memory retrieval of both linguistic and talker-identity information is addressed.

Chapter 2: Formant-invariant voice and pitch representations are pre-attentively formed from constantly varying speech and non-speech stimuli.[1]

The present study aimed at determining if listeners can form abstract voice representations while ignoring constantly changing phonological information and if they can use the resulting information to facilitate voice change detection. The study also aimed at understanding whether the use of such abstraction mechanism is restricted to the speech domain, or whether it can be deployed also in non-speech contexts. We ran an EEG experiment including a passive and an active oddball task, each featuring a speech and a rotated-speech condition. In the speech condition, participants heard constantly changing vowels uttered by a male speaker as standard stimuli which were infrequently replaced by vowels uttered by a female speaker with higher pitch. In the rotated-speech condition, participants heard rotated vowels, in which the natural formant structure of speech was disrupted. In the passive task, the Mismatch Negativity was elicited after the presentation of the deviant voice in both conditions, indicating that listeners could successfully group together different stimuli into a formant-invariant voice representation. In the active task, responses were faster and more accurate and elicited an enhanced P3b responses compared to the rotated-speech condition. Results showed that whereas at a pre-attentive level the cognitive system can track pitch regularities while abstracting away from constantly changing formant information both in speech and in non-speech, at a volitional level the use of such information is facilitated for speech. This facilitation was also testified by a stronger synchronization in the theta band (4-7 Hz), potentially pointing towards differences in encoding/retrieval processes.

[1] This chapter was submitted for publication and is currently under review. The study has been conducted in collaboration with dr. Simone Sulpizio (Department of Psychology, University of Milano Bicocca) and dr. Michele Scaltritti (Department of Psychology and Cognitive Science, University of Trento).

Chapter 2

2.1 Introduction

The speech signal encodes both linguistic and vocal information. These two types of information can be selectively extracted and used for different communicative and social goals. In fact, listeners can understand the message content irrespectively of who is speaking and can identify the talker's voice regardless of what is being said. However, these operations are not undemanding as they may seem and, in order to perform them, speakers need to orient their attention accordingly.

In an ERP study, Kaganovich et al. (2006) asked participants to listen to different vowels uttered by different talkers. In one task, participants were asked to identify the talker notwithstanding changes in the unattended vowel dimension, whereas in another task they had to identify vowels while ignoring changes in the unattended talker dimension. The Garner paradigm (Garner, 2014) employed by the authors predicts that if two dimensions are processed together, sudden changes in the unattended dimension would hamper the processing of the attended one. Consistently, when compared with a baseline task (i.e., a task where no changes in the unattended dimension occurred), both tasks were characterized by a sustained negativity surfacing in the N100 time-window and spreading until the P3 time window. These findings were interpreted as evidence of the involvement of two attention-based processes allowing for the dissociation of phonological vs. vocal information. Specifically, a low-level filtering process, occurring in the N100 time window, would isolate the physical dimension of interest, whereas a second higher-level one, occurring in the P3 time-window, would be responsible for matching the output of the filtering process to the correct response representation in working memory.

This result suggests that when listeners are asked to extract information from a complex signal by orienting their attention toward a target information, they need to take care of physical variability both in the attended and in the unattended dimensions. Speech tokens embedding phonological and vocal information are produced in different ways by different talkers. Thus, regardless of the specific type of information to select or ignore, listeners need to use their cognitive resources to model and summarize variability within a stable percept.

One way by which listeners can facilitate the extraction of relevant information from speech and deal with physical variability is by forming abstract representations which are selectively invariant to changes along specific dimensions of the speech signal (Belin et al., 2004; Norris & McQueen, 2008). Concerning this issue, Bonte et al. (2009) ran an EEG experiment in which participants listened to different vowels uttered by different talkers which were randomly

presented across different blocks. In separate blocks, they were asked to detect consecutive repetitions of either the same vowel or the same talker. In each task (i.e, detect vowel repetitions or talker repetitions), the alpha phase realignment surfacing ~250 ms after stimulus presentation was stronger for the task-dependent (phonemic or vocal) dimension. According to the authors' interpretation, alpha phase alignment is induced by selective attention guiding the temporal binding of information contained in abstract representations previously formed in auditory cortices. The interpretation of this result provides a characterization of the neural implementation of the task-induced attentional processes reported in Kaganovich et al. (2006), which require abstract representations to work correctly. Still, it is not clear how or when such abstract representations can inform and orient the attentional processes, nor if their formation occurs pre-attentively or needs the involvement of attention-based processes.

There is evidence that abstract (i.e., talker-invariant) representations of phonemes are automatically formed by the cognitive system. For example, Jacobsen, Schröger, and Alter (2004) ran an EEG experiment with a passive oddball paradigm, in which participants heard one vowel as standard stimulus with fixed first (F1) and second formant (F2) values – which are cues for vowel identification –, but with continuous variation in F0, which is a cue for voice identification. The presentation of a deviant vowel featuring different F1/F2 values yielded an MMN, notwithstanding the constant variation along non-linguistic information (i.e., F0 and intensity). The finding suggests that listeners automatically abstract away from non-linguistic cues (i.e., F0) while focusing on phonological information (i.e., F1 and F2). The results were replicated using speech-like stimuli (i.e., complex tones synthesized with the same F0, F1 and F2), but not with non-speech stimuli (i.e., simple tones lacking of formant structure, Jacobsen, Schröger, & Sussman, 2004). This suggests that abstraction mechanisms are speech-specific and get activated only in presence of a formant structure.

Crucially, no evidence about the potential involvement of these abstraction mechanisms in the formation of phoneme-invariant voice representations has been shown yet. However, such mechanisms can be reasonably hypothesized, as i) talker-related information is highly relevant during communication (Van Berkum et al., 2008), and ii) the cognitive system shows a domain-general ability to detect the violation of abstract regularities occurring within various physical features in acoustic stimuli. Consistently, many EEG studies used the "abstract-feature" oddball paradigm (e.g., Saarinen et al., 1992), in which standard stimuli differ along several physical

dimensions while being similar with respect to another one. These experiments demonstrated a reliable elicitation of the MMN, indexing the ability to automatically group together different sounds on the basis of the similarity with respect to one physical dimension, regardless of other constantly changing ones (for a review, see Paavilainen, 2013). These kind of abstract regularities in sound streams seem to be captured already by the brain of new-borns (Carral et al., 2005). These results may thus indicate that the cognitive system is able to extract invariant sound features in constantly varying acoustic contexts via a general-purpose auditory abstraction process, which can be subsequently used to process different kinds of regularities in several domains such as speech (Eulitz & Lahiri, 2004) and music (Virtala et al., 2011).

Nonetheless, although listeners may be able to track different acoustic regularities in sounds and store them within abstract representations via general-purpose mechanisms, they might be influenced by their prolonged experience with speech and voices. Consistently, the identification of the linguistic (i.e., words) or vocal component (i.e., talker identity) of speech is facilitated when one of the two information is familiar to the listener (Johnsrude et al., 2013; Nygaard et al., 1994; Zarate et al., 2015), suggesting that even if listeners are focusing on one specific dimension of the speech signal, being familiar with the ignored dimension(s) is still beneficial. The influence of linguistic and voice-related experience surfaces early in time, as the MMN shows larger amplitude when native phonemes (Dehaene-Lambertz, 1997; Näätänen et al., 1997) and words (Pulvermüller et al., 2001, 2004) or familiar voices (Beauchemin et al., 2006), are presented as deviant stimuli. This effect has commonly been considered as an index of a memory trace retrieval process (Näätänen et al., 2007), which occurs in a time window compatible with the one where the cognitive system forms abstract regularities representations. Thus, listeners may be facilitated in detecting regularities when they hear speech by retrieving representations of known linguistic/vocal information in which both the attended and the unattended information can be encoded.

This study has two main aims. The first aim is to establish whether the abstraction mechanism is information-specific within the speech domain, that is whether listeners can spontaneously form abstract representations of the talker's voice irrespectively of phonological information, exactly as they do with phonemes irrespectively of physical variations in the talker's voice (Jacobsen, Schröger, & Alter, 2004; Jacobsen, Schröger, & Sussman, 2004; Shestakova et al., 2002). To achieve this goal, the "abstract-feature" oddball paradigm was used. In a first

condition, different vowels uttered by a male voice were presented as standard stimuli. While F1/F2 values were constantly changed, the F0 value was kept fixed. Standard stimuli were infrequently replaced by deviant stimuli, that were produced by a female voice, characterized by a higher F0. If listeners can automatically form an abstract representation of the talker's voice irrespectively of the constant variation in phonological information (i.e., F1/F2 values of different phonemes), a MMN is expected. This result would indicate that listeners can form phoneme-invariant representations of the talker's voice similarly as they build talker-invariant phoneme representations. The absence of any MMN, instead, would suggest that the cognitive system is preferentially tuned to detecting variation along the phonological dimension, leaving the vocal one in second place. If this is the case, the abstraction mechanism under investigation could then be considered as information-specific, at least within the speech domain (as suggested by Jacobsen et al., (2004) results).

Since the possible presence of the MMN could also be due to an acoustic-based abstraction mechanism, as suggested by the studies reviewed by Paavilainen (2013), the second aim of the present study was to understand whether the abstraction mechanism is speech-specific or whether it represents a general-purpose mechanism which is then employed across different domains, including speech perception. To do this, in a second condition, another "abstract-feature" oddball block was implemented, but this time the stimuli corresponded to the spectrally rotated version of the speech stimuli presented in the first condition. Spectral rotation consists in manipulating the spectrum of a specific sound by selecting a mirroring frequency (e.g., 2000 Hz) and exchanging the power values of the high frequencies with those of the low frequencies and *vice versa* (Blesser, 1972). This procedure results in auditory stimuli with implausible formant values, disrupting any possible recognition of phonological information, while keeping both the spectral complexity and the pitch contour intact (Marklund et al., 2018; Sjerps et al., 2011a). If a MMN is successfully elicited in this condition, this would indicate that listeners can form abstract representations also from non-speech sounds, suggesting that the abstraction mechanism under investigation is not speech specific. Additionally, in case the MMN is elicited in both conditions, phonological information might still be pre-attentively extracted to facilitate the detection of vocal changes. In this case, the MMN should be stronger for the speech condition, indexing the automatic retrieval of native phoneme representations. (Dehaene-Lambertz, 1997; Näätänen et al., 1997).

Additionally, an active version of the oddball task was conducted, in order to understand whether the output of the abstraction mechanisms facilitates the detection of changes within specific stimulus features (i.e., pitch) while other constantly varying dimensions (i.e., F1 and F2) are disregarded. If this is the case, for the conditions in which a MMN is elicited in the passive oddball task, a P3b is expected following the correct detection of deviant stimuli in the active oddball task. Moreover, since the amplitude of P3b is sensitive to the amount of cognitive and attentional resources deployed to stimulus processing independently of its physical features (Duncan et al., 2009), it represents a good index to assess differences in the task demands of volitional target detection across speech and non-speech contexts. Therefore, in the possibility that a MMN is elicited in both conditions, we would reasonably expect the P3b having a smaller amplitude for the condition with higher task demands.

Finally, we also explored the oscillatory activity in the theta (4-7 Hz), alpha (8-12 Hz) and beta (13-30 Hz) frequency bands. Power modulations in the theta band are often found in correspondence to the presentation of deviant events in both passive (Jin et al., 2014; Ko et al., 2012; Koerner et al., 2016) and active oddball tasks with speech and non-speech stimuli (Citherlet et al., 2020; Kolev et al., 1997; Spencer & Polich, 1999). These modulations appear to be sensitive to pitch variations (Hsu et al., 2015; Li & Chen, 2018) and have been associated to processes of encoding (Klimesch, 1999), retrieval (Bastiaansen et al., 2005; Klimesch et al., 2001) and working memory load (Fuentemilla et al., 2008; Jensen & Tesche, 2002; Kolev et al., 1997). Power modulation in the alpha and in the beta bands are also commonly found in passive and active oddball tasks (Hsu et al., 2015; Mazaheri & Picton, 2005; Öniz & Başar, 2009) but while alpha activity is associated to attentional control (Wöstmann et al., 2017) and informational gating (Strauß et al., 2014), beta modulations are informative about the temporal dynamics of maintenance and disruption of perceptual and cognitive sets (Engel & Fries, 2010), which in our experiment are induced by the presentation of deviant events. Therefore, the study of oscillatory activity within the theta, alpha and beta bands may extend the functional characterization of non-phase-locked activity underlying fundamental cognitive processes that subserve the extraction of regularities in the auditory and in the speech domain, while possibly providing complementary evidence with respect to the underlying mechanisms.

Chapter 2

2.2 Method

2.2.1 Participants

Seventeen healthy Italian native speakers were recruited. Two participants were excluded from the final sample because of excessive noise in the EEG data. The final sample included 11 female and 4 male participants (M_{age} = 22.60, SD_{age} = 2.74), all right-handed (Edinburgh Handedness Inventory: M = .78, SD = .13). The sample size was decided on the basis of previous studies that used the abstract oddball paradigm and reliably recorded both the MMN and/or the P3b responses (Bendixen & Schröger, 2008; Escera et al., 2014; Escera & Malmierca, 2014). Participants reported to be neurologically healthy and to have normal hearing[2]. Participation was compensated either with course credit or with 10€ per hour. The study was approved by the Ethical Committee of The University of Trento. Participants signed an informed consent document prior to the experiment.

2.2.2 Stimuli

One female and one male Italian native speaker respectively aged 38 and 36 were recruited to record the experimental stimuli. They were asked to read aloud 5 isolated Italian vowels (/a/, /e/, /ɛ/, /i/, /ɔ/) three times each. Their voice was recorded at 44100 Hz with a professional recorder in a silent room. The best tokens were selected based on quantitative and qualitative evaluation. Noisy tokens and tokens with abnormal pitch contours (e.g., list-reading intonation) were discarded. After this, the tokens with the smallest difference of F1 and F2 between the two talkers were selected in order to minimize any possible attentional shift caused by large F1-F2 differences between the talkers. The central 100 ms part of each vowel was extracted. The pitch contour in each token was adjusted to a flat line to prevent participants from confounding idiosyncratic pitch shifts as talker-identity changes. The pitch was set to an average value that was calculated as the mean across all tokens within each speaker. Stimuli were low-pass filtered at the cut-off frequency of 4000 Hz in order to match the spectral dimensions of the rotated speech stimuli, which require to be low-pass filtered before applying spectral rotation (Blesser, 1972). Intensity was put to an

[2] Participants' musical experience was assessed with the Ollen Musical Sophistication Index (Ollen, 2006) in order to avoid confounds in the interpretation of possible amplitude modulation of the MMN component as pitch changes were shown to elicit stronger MMNs in musically trained listeners (Chandrasekaran et al., 2009). None of the participants was musically trained.

average value of 70 dB with linear slopes of 10 ms at the onset and the offset in each token to avoid any harsh transition between silence and sound in the EEG experiment.

Rotated speech stimuli were created by rotating the spectrum of speech stimuli using a spectral rotation function in MATLAB with a cut-off frequency of 4000 Hz (available at https://www.phon.ucl.ac.uk/downloads/matlab/Blesser.zip); the same function and other similar implementations of the spectral rotation algorithm were used in several studies to produce non-speech control stimuli in attempt to contrast acoustic and speech-specific perceptual processes (Azadpour & Balaban, 2008; Marklund, Gustavsson, Kallioinen, & Schwarz, 2020; Scott, 2000; Steinmetzger & Rosen, 2017). The result of this procedure is a sound with a mirrored spectrogram along a mirroring frequency (i.e., 2000 Hz corresponding to half of the cut-off frequency) with respect to the input sound. This means that the point-by-point power of lower frequencies (e.g., 0 Hz, 500 Hz, 1000 Hz) is transferred to higher frequencies (4000 Hz, 3500 Hz, 3000 Hz) and *vice versa*. The physical characteristics of the experimental stimuli are summarized in Table 1.

Chapter 2

Table 1. Pitch (F0), First and Second Formant (F1, F2) values of the experimental stimuli for each talker and each condition.

		Condition					
		Speech			Rotated Speech		
Talker's Sex	Vowel	F0	F1	F2	F0	F1	F2
Male	a	121 Hz	816 Hz	1252 Hz	121 Hz	768 Hz	1623 Hz
	e	121 Hz	384 Hz	2141 Hz	121 Hz	653 Hz	1360 Hz
	i	121 Hz	360 Hz	2039 Hz	121 Hz	795 Hz	1402 Hz
	ɔ	121 Hz	561 Hz	862 Hz	121 Hz	772 Hz	1007 Hz
	ɛ	121 Hz	571 Hz	1782 Hz	121 Hz	1049 Hz	1717 Hz
Female	a	184 Hz	981 Hz	1469 Hz	184 Hz	1269 Hz	2081 Hz
	e	184 Hz	368 Hz	1698 Hz	184 Hz	803 Hz	1332 Hz
	i	184 Hz	329 Hz	1209 Hz	184 Hz	780 Hz	1113 Hz
	ɔ	184 Hz	733 Hz	1169 Hz	184 Hz	964 Hz	1976 Hz
	ɛ	184 Hz	695 Hz	1599 Hz	184 Hz	934 Hz	1675 Hz

2.2.3 Procedure

First, participants were asked to fill questionnaires collecting demographic information, handedness, and musical expertise. Then, they were prepared for the EEG recording in a dimly lit room. The experiment consisted of a passive and an active version of the oddball task. During the passive oddball task, participants were asked to watch a silent video depicting drone footage of different landscapes while auditory stimuli were delivered via Etymotic ER-1 headphones at fixed volume (70 dB) using E-prime 2.0 Software (Schneider & Zuccoloto, 2007). Speech and rotated-speech stimuli were presented across two different blocks (speech vs rotated-speech conditions). Each block included 680 standard events and 120 deviant events. At the end of each block, the 120 deviant stimuli were presented in random order to serve as control events. These latter stimuli were included in the experiment to control for the effects induced by the physical properties of the stimuli. Normally, the MMN component is calculated by subtracting standard ERPs from deviant ERPs (Näätänen et al., 2007), but the result of this computation is also influenced by physical differences between standard and deviant events. By using control events, which are physically identical to deviant events but are presented with the standard events' distribution, the MMN calculated by subtracting control from deviant events is uncontaminated by differences in terms of physical features and thus better highlights the cognitive processes of interest (Tuninetti et al., 2017). Between the two blocks, each of which lasted approximately 11 minutes, participants could take a small break.

In the speech condition, all the vowels produced by the male speaker were equiprobably presented in random order as standard stimuli with a fixed Interstimulus Interval (ISI) of 700 ms. All the vowels produced by the female talker were equiprobably presented as deviant stimuli (probability of occurrence = .15) with the constraint that a minimum of two standard events occurred before the presentation of a deviant event. The same vowel was never repeated in two consecutive trials, irrespectively of its standard/deviant status. In the rotated speech condition, the same presentation paradigm was applied. The rotated speech condition was always presented first as presenting the speech condition first could have made participants aware of the stimulation paradigm structure, possibly leading to unwanted attentional modulations in the subsequent block.

After the passive oddball task, the active oddball task took place. This task was identical to the previous one, with the only exception that participants were asked to press a button on a joypad as fast as possible when they heard a deviant event. Before doing so, participants were

debriefed on what they heard in the passive task to ensure that they understood which stimuli were the deviant ones. They were told that the speech stimuli were produced by human voices while rotated speech stimuli were produced by guessing how aliens' voices could have sounded like. Participants were to press a button every time they heard the human-female or the alien-female voice. Before each experimental block, a practice block was presented. For the first 10 practice trials, participants were helped in performing the task by a graphical representation of the stimulus list presented on the screen where the information about the standard/deviant status of each upcoming stimulus was specified. For the subsequent 20 practice trials participants performed the task as in the experimental part, that is with no graphical help and by watching the silent video that was presented in the passive task. At the end of the practice block, they received feedback on their performance. After this, the experimental blocks started and lasted approximately the same amount of time as in the passive task. The whole experiment lasted approximately 1.30 h.

2.2.4 EEG recording and preprocessing

The EEG was recorded with an eego sports system (ANT Neuro) at a sampling frequency of 1000 Hz (filters: DC to 130 Hz, third- order sinc filter), from 64 Ag/AgCl shielded electrodes referenced to CPz and placed in the standard 10-10 locations on an elastic cap. Electro-oculograms were acquired with an additional electrode placed under the left eye. Impedance was kept < 20 kΩ. Data pre-processing was performed with the MATLAB toolboxes EEGLAB (Delorme & Makeig, 2004), ERPLAB (Lopez-Calderon & Luck, 2014) and FieldTrip (Oostenveld et al., 2011). The signal was re-referenced offline to the average reference. Data were high-pass filtered at 0.1 Hz using a 2nd order Butterworth filter (12 dB/oct Roll-off). A Notch filter at 50 Hz was then applied to attenuate line noise. Independent Component Analysis was run on the continuous signal using the Infomax algorithm (Bell & Sejnowski, 1995). Eye-blink and eye-movement components were identified with ICLabel algorithm (Pion-Tonachini et al., 2019) and removed. Excessively noisy channels were interpolated via spherical interpolation. Mastoid and Electro-Oculogram channels were excluded from the analyses.

2.2.4.1 ERP data pre-processing

Data were low-pass filtered at 30 Hz using a 2nd order Butterworth filter (12 dB/oct Roll-off). Epochs were extracted from -200 ms before stimulus onset until 800 ms after stimulus onset and a baseline correction was applied by subtracting the mean voltage of the -200 - 0 pre-stimulus

period from the entire epoch. Epochs containing signal with an amplitude exceeding ± 100 μV in any of the 62 EEG channels were rejected. An average of 1.03% ± 0.81% of the total number of epochs per participant were rejected and the percentage of rejected epochs was similar across conditions for the passive oddball task (Control Speech 0.66% ± 1.48%, Deviant Speech 0.83% ± 0.94%, Control Rotated 2.28% ± 3.72%, Deviant Rotated 0.94% ± 1.21%) and the active oddball task (Standard Speech 0.83% ± 1.52%, Deviant Speech 1.00% ± 1.84%, Standard Rotated 0.34% ± 0.53%, Deviant Rotated 0.38% ± 0.76%).

For the passive oddball task, separate ERPs were computed by averaging epochs within each participant and within all the combinations of the factors condition (speech, rotated speech) and stimulus type (control, deviant). The differential waveforms of the MMN were calculated within each participant and within each condition, by subtracting the control ERP from the deviant ERP. The same averaging method was used for the active oddball task with the factors condition (speech, rotated speech) and stimulus type (standard, deviant), but this time only the events with a correct response were considered. All the epochs corresponding to standard events coming immediately after deviant events were removed from the analysis, to avoid any contamination from late potentials triggered by deviant events.

2.2.4.2 Time-Frequency data pre-processing

Data were low-pass filtered at 80 Hz using a 2nd order Butterworth filter (12 dB/oct Roll-off). Epochs were extracted from -800 ms before stimulus onset until 1200 ms after stimulus onset. Epochs containing signal with an amplitude exceeding ± 100 μV in any of the 62 EEG channels were rejected. An average of 5.3% ± 3.35 % of the total number of epochs per participant were rejected and the percentage of rejected epochs was similar across conditions for the passive task (Control Speech 4.39% ± 8.91%, Deviant Speech 4.33% ± 5.30%, Control Rotated 5.83% ± 7.42%, Deviant Rotated 4.06% ± 4.56%) and the active task (Standard Speech 6.36% ± 7.01%, Deviant Speech 7.78% ± 8.77%, Standard Rotated 4.26% ± 4.63%, Deviant Rotated 4.11% ± 4.77%). The time-frequency representation was computed via Morlet wavelets with 10 ms steps from -300 to 800 ms with respect to stimulus onset in each epoch for the 4-30 Hz frequencies (1 Hz step) with a linearly increasing number of cycles (range 2-10) in order to balance spectral and temporal precision (Cohen, 2014). The Event-Related Spectral Perturbations (ERSPs) for both active and passive oddball tasks were computed in the whole spectrum by averaging epochs within

each participant and within all the combinations of the factors condition (speech, rotated speech) and stimulus type (standard, deviant). All the epochs corresponding to standard events coming immediately after deviant events were removed from the analysis, to avoid any contamination from later potentials triggered by deviant events. For the active oddball task, only the events with a correct response were considered.

2.2.5 Statistical Analyses

Behavioural Data

Accuracy and RTs were both analyzed using the "lme4" package (Bates et al., 2015) in R Software (R Core Team, 2013). Participants' accuracy in the active task was analyzed by means of a Generalized Linear Mixed Model (GLMM) with a logit link-function. The best model was selected by adding each predictor one by one. For each added predictor, the model was tested via Chi-Square test against the model without the predictor. The predictor was kept in the model only when the Chi-Square test showed significant differences with respect to the model without it. The final model included the fixed factors Condition (speech, rotated speech) and Stimulus type (standard, deviant) as well as by-participants and by-items random intercepts. Reaction times (RTs) of correct deviant events were analyzed by means of a Linear Mixed Model (LMM). Model selection was performed with the same method used for the model of accuracy data. The final model included Condition (speech, rotated speech) as fixed factor as well as by-participants and by-items random intercepts All factors in all models were deviance coded with the numerical values 0.5 and -0.5 following the factors' levels order presented in this section. With this coding, the model's coefficients represent the main effect, coded as the difference between the levels of each factor. Post-hoc comparisons were implemented via "emmeans" R package.

EEG Data

Nonparametric cluster-based permutation tests were used for both ERPs and time-frequency analyses. In this approach, conditions are compared via multiple paired t-tests performed at each time point within each channel. T-values with a p-value $< .05$ are selected and clustered on the basis of temporal and spatial adjacency. All the t-values within each cluster are then summed and compared with the distribution of the t-values under the null hypothesis which is obtained by

calculating the test statistic several times (N = 2,500) on the data points shuffled across conditions. The proportion of random permutations where the observed cluster's t-value is larger than the t-value drawn from the actual data represents the cluster p-value. When analyzing ERP components for which the literature provides robust temporal coordinates (e.g., MMN) and specific directions (i.e., positive or negative), one-tailed tests were restricted to an a-priori defined time-window (see below). For every statistical test, 95 % Confidence Intervals of the p-value are reported. Cohen's d is also reported and was calculated by dividing the mean of the differences between conditions by the standard deviation of the differences between the conditions at test and obtained from the individual values of the dependent variable (i.e., voltage or power). Individual values were computed separately for each condition by averaging the dependent variable across channels and time samples of significant clusters within every individual participant following the indication of FieldTrip's authors (see https://www.fieldtriptoolbox.org/example/effectsize/ for additional information).

In the passive oddball task, the presence of the MMN component within each condition was assessed by comparing deviant and control events via a one-tailed test in the 110-225 ms time window as suggested in Kappenman et al. (2021)[3]. Visual inspection of the ERPs also showed the presence of a sustained negative component emerging starting ~350 ms after stimulus onset and lasting until the end of the epoch, mostly distributed along Frontal and Fronto-Central electrode sites (see Figure S1, Appendix A for the ERP waveforms on a large set of channels). This component was tentatively identified as the Late Discriminative Negativity (LDN), which was also reported in another study encompassing the abstract-feature paradigm as "Late Mismatch Negativity" (Zachau et al., 2005). Previous studies that used the canonical oddball paradigm reported the presence of this component in different time windows scattered across the 350-600 ms interval (Choudhury et al., 2015; David et al., 2020; Honbolygó et al., 2020). Given the absence of a-priori hypotheses on its presence and/or modulation, the analysis of this component must be considered explorative. For this reason, and in order not to select an ad-hoc time window based on visual inspection, we performed a one-tailed test in a wider 350-800 m time-window, which safely started after the offset of the MMN and lasted throughout the whole epoch. Finally, to assess the presence of a P3b component in the active oddball task, a broad time-window was considered, by

[3] In the cited study, the measurement windows for the MMN and P300 were identified by cross-validating the time windows generally reported in the literature with the results of a cluster-based permutation analysis.

comparing deviant and standard events via a one-tailed test between 300 and 600 ms after stimulus onset. The time window was selected following the same logic used for the MMN (Kappenman et al., 2021). The difference between conditions (speech, rotated speech) was then tested by comparing the two differential waveforms calculated by subtracting the control ERP from the deviant ERP for the MMN and the LDN, and the standard ERP from the deviant ERP for the P3b.

Statistical analyses on time-frequency data were conducted on theta (4-7 Hz), alpha (8-12 Hz) and beta (13-30 Hz) frequency bands by averaging power values within each band within the same combination of factors by which the ERP analyses were computed. The whole 0-800 ms epoch was used in the analyses as we had no specific hypotheses about the temporal unfolding of possible power modulation following non-phase locked activity. The two differential ERSPs calculated within each condition (i.e., speech vs rotated-speech) by subtracting power of control/standard events from the one of deviant events were directly compared. The test ascertains the difference between speech and rotated speech conditions and is equivalent to testing for an interaction effect between probability (i.e., standard/control and deviant) and condition (speech, rotated speech). Post-hoc tests were then performed by directly comparing the ERSPs of standard/control events with the ERSPs of deviant events within every condition.

2.3 Results

2.3.1 Behavioural Results

The mean proportion of accurate responses in the speech condition was .99 (SD = .002) for standard and .98 (SD = .01) for deviant events, whereas in the rotated speech condition it was .97 (SD = .06) for standard and .83 (SD = 0.16) for deviant events. The model of accuracy data revealed the main effect of Condition ($\beta = 3.24$, SE =0.18, z =-17.67, $p < .001$) showing a higher accuracy in the speech condition (M = .99, SD = .004) with respect to the rotated speech condition (M = .95, SD = .06). A main effect of Stimulus Type ($\beta = 2.49$, SE = 0.10, z = 24.89, p = < .001) showed higher accuracy for standard events (M = .98, SD = .03) with respect to deviant events (M = .90, SD = .09). The mean reaction times for correctly identified deviant events was 414 ms (SD = 86) in the speech condition and 457 ms (SD = 110) in the rotated speech condition. The model of reaction times data revealed only the main effect of condition ($\beta = -45.70$, SE =3.08, z =-14.82, $p < .001$), showing that participants responded significantly faster in the speech than in the rotated speech condition. Behavioural results are summarized in Figure 1.

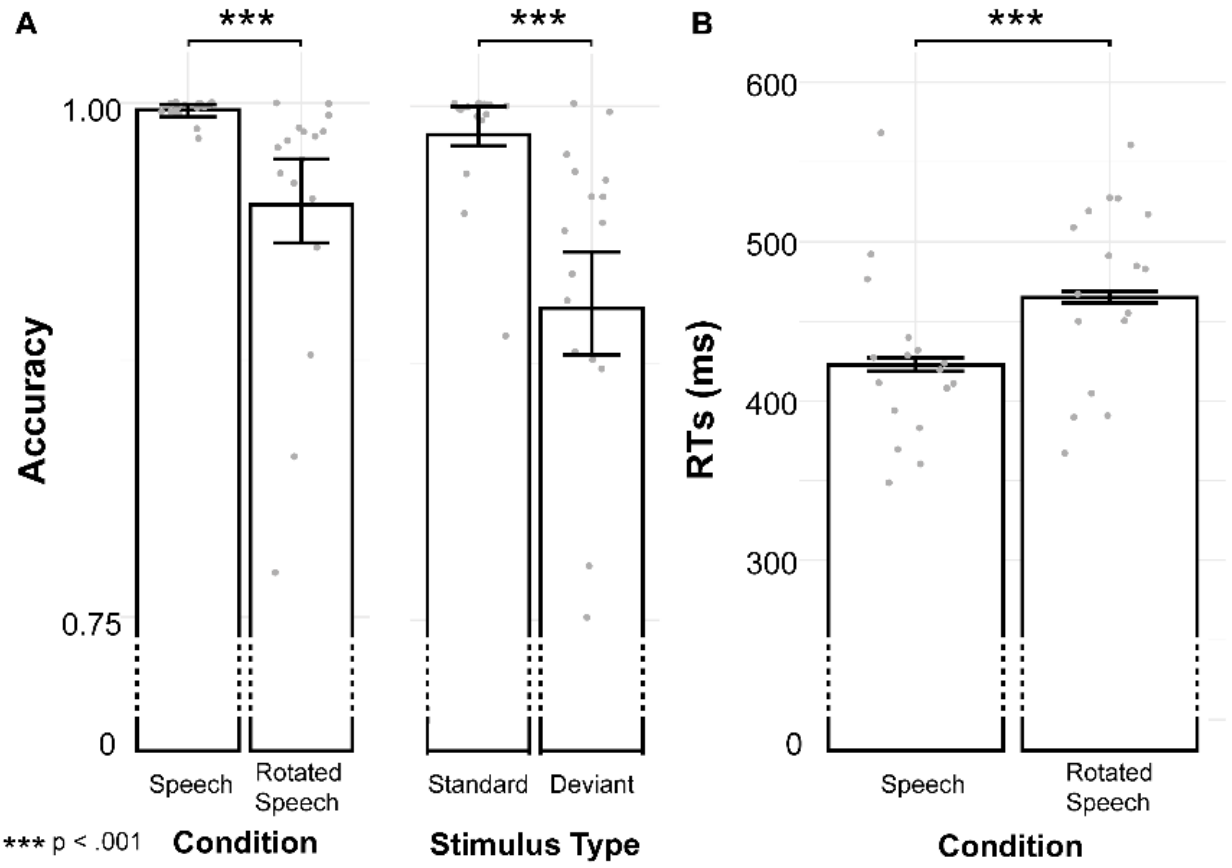


Figure 1. Behavioural results of the active oddball task. (A) Proportion of correct responses broken down by Condition (1st column) and by Stimulus type (2nd column). (B) Reaction times of correct responses to deviant events only. Error bars represent the SE and grey points represent individual observations. For illustrative purposes, only the relevant portion of the y axis is shown in both plots (dashed lines indicate the discontinuity of the axis).

2.3.2 ERP Results

In the passive oddball task, the presence of the Mismatch Negativity in the 110-225 ms time window was revealed by a significant difference between control and deviant ERPs for both the speech (one negative cluster encompassing the whole window duration, $p < .001$, 95% CI [.000 .001], $d = 1.646$), and the rotated speech condition (one negative cluster surfacing between 138-225 ms, $p < .001$, 95% [.000 .001], $d = 1.741$). Both clusters showed a topographical distribution coherent with that of the MMN, being mostly pronounced over Frontal, Fronto-Central and Central channels. The test of the interaction did not reveal any difference between conditions in the 110-225 ms time window.

The significant difference in the 350-800 ms between control and deviant ERPs confirmed the presence of a LDN component, which was represented by a stronger negativity in the deviant than in the control ERPs for both the speech ($p < .001$, 95% [.000 .002], $d = 1.371$) and the rotated speech condition ($p < .001$, 95% CI [.000 .001], $d = 1.701$), respectively captured by negative clusters emerging in the 350-800 ms and in the 460-800 ms time window. The test of the interaction showed that, in the 350-800 ms time window a stronger LDN response surfaced for the speech condition compared to the rotated speech condition, mostly distributed over right frontal electrodes as highlighted by the presence of a negative cluster in the 631-733 ms time window ($p = .021$, 95% CI [.014 .027], $d = 1.710$). ERP results for the passive oddball task are summarized in Figure 2 (see Figure S1 in Appendix A for additional descriptive plots).

In the active oddball task, a significant positive difference surfaced between standard and deviant ERPs in the P3b time window for the speech ($p < .001$, 95% CI [.000 .001], $d = 2.070$) and rotated speech condition ($p < .001$, 95% CI [.000 .001], $d = 1.7891$), captured by two positive clusters emerging in the 300-600 time window, spatially distributed over Central, Centro-parietal, Parietal and Parieto-occipital channels. The test on the differential ERPs, calculated by subtracting standard ERPs from deviant ERPs, revealed a stronger P3b effect in the speech condition with respect to the rotated speech condition ($p = .001$, 95% [.000 .002], $d = 1.490$), highlighted by a positive cluster mostly distributed over Central and Centro-Parietal channels in the 300-565 ms time window. ERP results for the active oddball tasks are summarized in Figure 2 (see Figure S2 in Appendix A for additional descriptive plots).

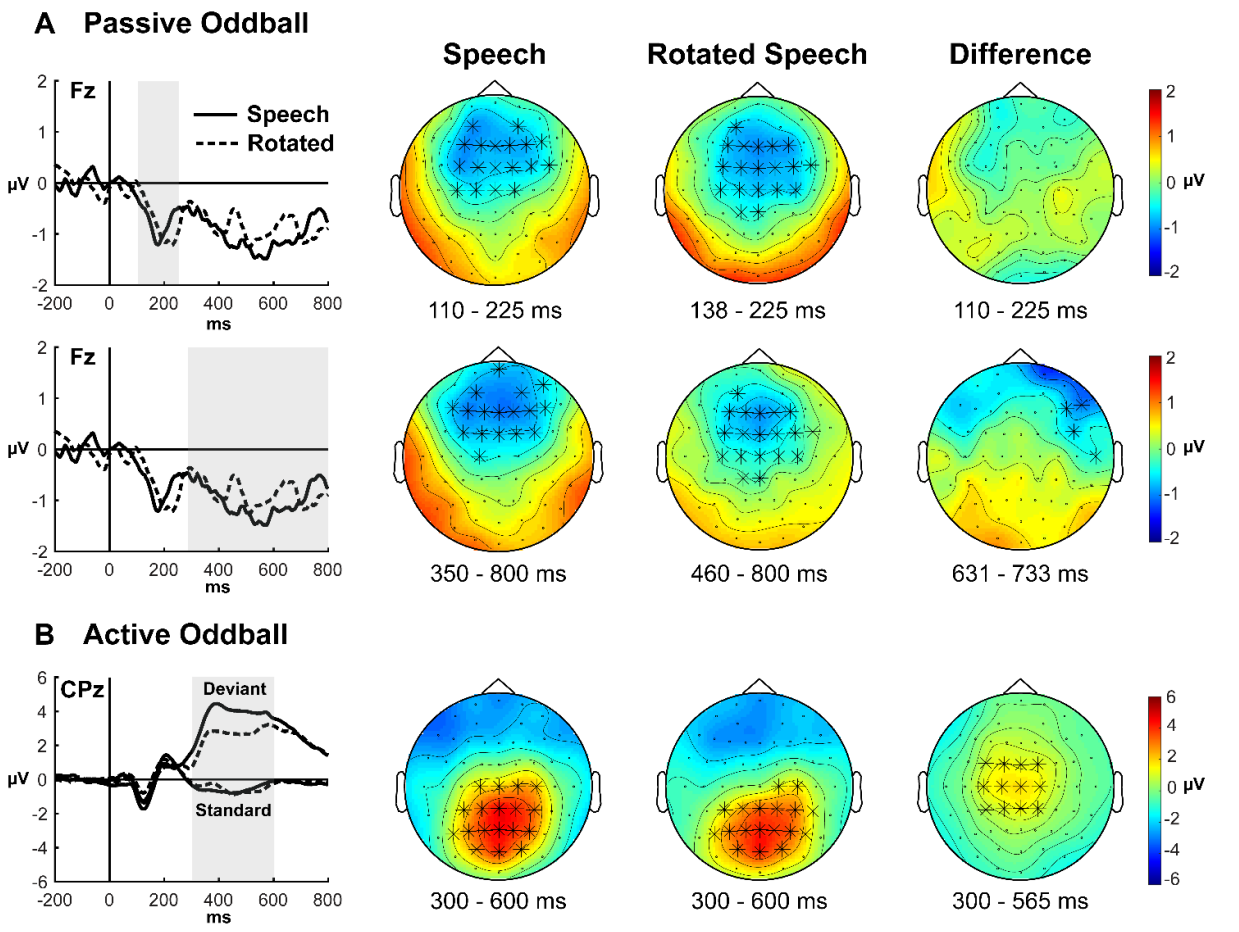


Figure 2. ERP results. (A) Passive oddball task. The first column displays the differential waveforms at a representative channel (Fz) for the speech (continuous line) and the rotated speech (dashed line) condition. The grey rectangles indicate the time-window used in the analyses (MMN, first row; LDN, second row). In the subsequent columns, topographies show the spatial distribution of the MMN (first row) and LDN (second row) in the time windows where significant differences emerged. The last column represents the voltage difference between conditions, calculated by subtracting the differential waveforms in the rotated speech condition from the ones calculated in the speech condition. Electrodes that were included in the clusters for more than 50% of the samples within the cluster time windows (reported below the topographies) are represented by black asterisk marks superimposed to the maps. (B) Active oddball task. The first column represents the ERPs of standard and deviant events at a representative channel (CPz) for the speech (continuous line) and the rotated speech (dashed line). In the subsequent columns, topographies show the spatial distribution of the differential P300 waveforms, calculated by subtracting the standard ERP from the deviant ERP in the time windows where significant differences emerged for each condition. The last column represents the voltage difference between conditions, calculated by subtracting the differential waveforms in the rotated speech condition from the ones calculated in the speech condition. Electrodes are marked as in A.

2.3.3 Time-Frequency Results

In the passive oddball task, the test on differential ERSPs across the speech and rotated-speech conditions within the beta band showed the presence of a negative cluster distributed on Central, Centro-Parietal and Parietal electrode sites between 310 and 540 ms ($p = .022$, 95 % CI [.015 .028], $d = 1.748$). The source of this effect was attributed to a significant difference between deviant and control events surfaced in the rotated speech condition, as revealed by two spatiotemporally distinguishable clusters (see Figure S3, Appendix A). One positive cluster unfolded over left Fronto-central and Central channels ($p = .009$, 95 % CI [.005 .012], $d = 1.559$), ranging between 140 and 540 ms, apparently indexing both an early-emerging desynchronization in control events and a later occurring synchronization in deviant events (see Figure 3). A second positive cluster was detected ($p = .017$, 95 % CI [.012 .022], $d = 1.399$) between 630 and 800 ms signaling another ERS in deviant events distributed over right Parieto-occipital and Occipital channels. Although, the test for the interaction between the speech and the rotated speech condition did not provide a statistically reliable result (the upper limit of the p-value 95% C.I. surpassed the critical alpha level of 0.025). Given the explorative nature of the time-frequency analysis and the absence of similar patterns in previous studies employing the same paradigm, this last result was not further interpreted. No significant differences between controls and deviants were found for the speech condition in the beta band. The tests on differential ERSPs did not show significant differences between conditions either in the theta or in the alpha frequency bands.

In the active oddball task, the test on differential ERSPs across speech and rotated-speech conditions within the theta band revealed the presence of a positive cluster ($p = .013$, 95 % CI [.009 .018], $d = 1.160$) surfacing between 320 and 800 ms on right Central, Centro-Parietal and Parietal electrodes. Post-hoc tests operated between standard and deviant events within each condition, showed that deviant events yielded a stronger synchronization in the theta band compared to control events, as highlighted by the presence of a positive cluster both in the speech ($p < .001$, 95 % CI[.000 .001], $d = 1.274$) and the rotated speech ($p < .001$, 95 % CI [.000 .001], $d = 1.2679$) conditions, widely distributed from Pre-Frontal to Parietal electrode sites, in the 130-800 ms and in the 150-660 ms time windows, respectively (see Figure S4, Appendix A). Therefore, the effect found for differential ERSPs substantially reflected a stronger theta synchronization occurring in deviant evets for the speech condition. In the beta band, the test on differential ERSPs, operated between conditions revealed the presence of a positive cluster ($p =$

.015, 95 % CI [.010 .019], d = 1.247), emerging between 590 and 800 ms across Central, Centro-Parietal and Parietal electrode sites. Post-hoc tests operated between standard and deviant events within each condition showed a desynchronization in deviant with respect to standard events both in the speech (p = .010, 95 % CI [.006 .014], d = 1.360) and the rotated speech condition (p = .004, 95 % CI [.002 .007], d = 1.242), captured by negative clusters unfolding over Central and Centro-Parietal channels, in the 250 -590 ms and in the 250 -710 ms time windows, respectively. The speech condition was also characterized by a stronger beta synchronization for deviant events with respect to standard ones, surfacing right after the earlier-occurring desynchronization and widely distributed on the scalp between 570 and 800 ms (p = .010, 95 % CI [.006 .014], d = 1.154), which presumably induced the effect highlighted by the test on differential ERSPs (see Figure S5, Appendix A). Results are summarized in Figure 3.

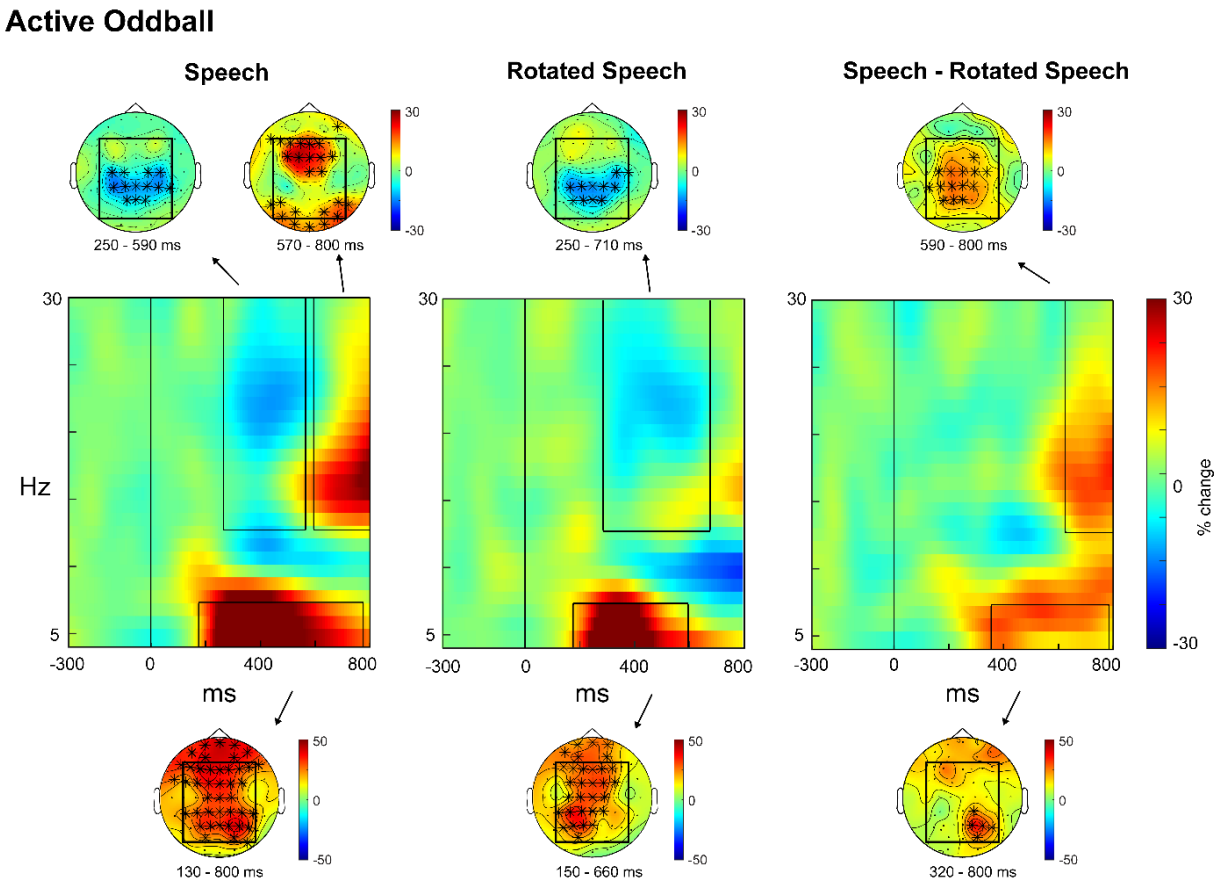


Figure 3. Time-Frequency results for the active oddball tasks. The time-frequency power spectra show the power modulations (% change) characterizing the differential ERSPs for each condition (1st and 2nd columns) as well as the difference between them, corresponding to the interaction effect (3rd column). Spectra were obtained by averaging activity for the electrodes F5, F3, F1, Fz, F2, F4, F6, FC5, FC3, FC1, FCz, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6, P5, P3, P1, Pz, P2, P4, P6, PO5, PO3, PO1, POz, PO2, PO4, PO6. Black squares represent the temporal distribution of the significant clusters within theta (4-7 Hz) and beta (13-30 Hz) bands. The mean number of channels included in each cluster represented in the power spectra was calculated across all time-samples and only the time-bins including at least half of the mean number of channels are enclosed in black squares. Topographies in the lower and higher row show the spatial distribution of theta and beta ERDs/ERSs characterizing the differential ERSPs for each condition (1st and 2nd columns) as well as the difference between them, corresponding to the interaction effect (3rd column). Electrodes that were included in the clusters for more than 50% of the samples within the cluster time windows (reported below each topography) are represented by black asterisk marks superimposed to the maps. Black squares on topographies represent the channels that were included in the averaged spectral plots.

2.4 Discussion

The aim of this EEG study was to understand whether listeners can pre-attentively form phoneme-invariant voice representations from constantly changing vowel stimuli. The same test was further performed when using rotated speech stimuli, in order to clarify whether the phenomenon is restricted only to the speech domain. Secondly, the volitional usage of the abstract information was examined through an active version of the task. On the basis of our results, we argue that while listeners can represent abstract regularities in sounds via a presumably domain-general mechanism, the extensive experience they have with speech and voices can facilitate the volitional use of the represented information. Particularly, the latter hinges upon pre-existing voice representations in which information can be encoded more efficiently and finally matched with response categories.

2.4.1 Passive Oddball Task

The ERP data showed that the MMN was clearly elicited with both speech and rotated speech stimuli, with no sizeable differences between these two conditions. Note that the experiment was designed so that the MMN could be triggered by the presentation of a deviant stimulus only if the preceding standard stimuli were grouped into an abstract regularity representation of the invariant F0 despite the constant variations along F1 and F2. In line with previous studies showing that listeners can track different regularities in many stimulus features at the same time (Huotilainen et al., 1993; Pakarinen et al., 2010), the elicitation of the MMN across both the speech and the rotated speech condition indicates that the cognitive system is able to represent abstract regularities via a domain-general mechanism. By using this mechanism, the cognitive system can equally form talker-invariant phoneme representations as shown by previous studies (Eulitz & Lahiri, 2004; Jacobsen, Schröger, & Sussman, 2004, 2004; A. Shestakova et al., 2002), and phoneme-invariant voice representations as suggested by our results.

It is reasonable to think that, in the present experiment, the regularities-extraction mechanism did not retain phonological information. In fact, the influence of phonological information should have yielded a stronger MMN for the speech condition which instead was undistinguishable from the MMN generated in the rotated-speech condition. However, the amplitude of MMN can reflect both acoustic and linguistic differences (Näätänen et al., 2007) between standard and deviant stimuli. To isolate the contribution of these two sources, previous

studies (Christmann et al., 2014; Marklund et al., 2018) contrasted the MMNs generated by vowel contrasts in speech and rotated speech using the classic oddball paradigm. These studies showed a stronger MMN for speech than for rotated speech stimuli. and suggested that such difference reflect the specific contribution of the presence of phonological information to the final amplitude. The absence of any difference in the MMN elicited in the speech and the rotated speech condition we reported might speculatively suggest that phonological information was stripped away in order to form voice representations at a pre-attentive level.

Interestingly, the phonological/formant information presumably ignored by these early-occurring mechanisms may have been considered during later processes. In fact, within the passive oddball task, a sustained negativity surfaced right after the offset of the MMN, in a 350-800 ms time-window and featuring a fronto-central spatial distribution. We identified this sustained negativity as an instantiation of the LDN, an automatic response with an unsettled functional significance, which occasionally occurs after the MMM (Datta et al., 2010). The LDN has been consistently recorded in children (Cheour et al., 2001; Ervast et al., 2015; Shestakova et al., 2003) and less often in adults (Bishop et al., 2011; Mueller et al., 2008). Zachau et al. (2005) reported the presence of the LDN in adults following abstract-rules violation with simple tones stimuli and suggested that the LDN is an index of a transferring mechanism, allowing the formation of representations of sound regularities in memory. The authors suggest that this mechanism could provide the computational basis for the segmentation of speech signals, further clarifying the reasons for which the LDN is consistently found in children (Bishop et al., 2011), who are still developing linguistic abilities. This notion was further strengthened by similar results obtained by Liu et al. (2014) with consonant and lexical tone contrasts in pre-school and school-aged Mandarin speaking children. David et al. (2020) also reported a larger LDN in children with respect to adults, elicited by phonologically complex rather than simple multisyllabic non-words. Although this regularities-transferring mechanism could be of relevant use for language learning, our findings together with previous studies (Zachau et al., 2005) suggest that it is not necessarily language-specific.

Even though the activation of the regularities-transferring mechanism may not be restricted to the speech domain, it could still be modulated by the presence of meaningful phonological information. In fact, we found a stronger LDN for the speech condition, and the difference was mainly distributed over right frontal electrode sites. This difference does not stem from differences

in terms of spectral complexity – speech and rotated speech are thought to be equally complex (Maier et al., 2011) –, nor in terms of physical properties of speech and rotated speech stimuli, as the differential waveforms were calculated by subtracting the averaged ERPs of deviant events from the ERP of physically identical control events. Therefore, this effect seems to be related to the presence of high-level information encoded in speech. If this effect is an actual index of an information-transferring mechanism subserving learning processes, we could speculate that, when hearing natural sounding voices from speech (i.e., containing meaningful phonological information), listeners can store the regularity information into a pre-existing voice representation. In fact our cognitive system is thought to prototypically represent male and female voices, and update those voice models throughout lifetime (Latinus et al., 2013; Petkov & Vuong, 2013; Yovel & Belin, 2013). However, even if previous studies might provide sufficient information to interpret this result, considering the a-posteriori nature of the analysis and the instability of the test operated between speech and rotated speech conditions (upper limit of the p-value 95% C.I. surpassed the critical alpha level of .025), the interpretation provided here only represents a tentative proposal.

2.4.2 Active Oddball Task

While at a pre-attentive level, abstract pitch/voice regularities seem to be easily extracted from sounds irrespectively of the presence of phonological information, at an attentive level it appears that the information about previously formed regularities can be transferred to working memory and matched to response categories more efficiently when phonological information is present. In the active oddball task, participants performed better in the speech than in the rotated speech condition. Further, the EEG data showed that the correct detection of the deviant stimuli was associated with a clear P3b response, with a stronger amplitude for the speech condition. The P3b component is commonly thought to reflect a range of cognitive processes subserving the revision of a mental representation induced by incoming stimuli (Donchin, 1981): When new or target stimuli are detected, attentional processes are thought to update the stimulus representation held in working memory (Polich, 2007). Additionally, previous studies showed that the amplitude of the P3b component is also modulated by task difficulty, being lower in the context of higher task demands. The latter would directly determine the amount of cognitive and/or attentional resources required to revise mental representations (Kok, 2001; Polich, 1987, 2007). Additionally, as shown in previous P300 studies (Başar-Eroglu et al., 1992; Demiralp et al., 2001; Yordanova et al., 2000), the correct detection of deviant events was also associated with an increased theta synchronization,

which was present both in the speech and in the rotated-speech conditions, albeit enhanced in the former compared to the latter. Oscillatory activity within the theta band has a primary role in neurophysiological models of memory (Backus et al., 2016; Lisman & Buzsaki, 2008). Consequently, synchronization within the theta band is commonly associated with working memory (WM) capacity/load (Dong et al., 2015; Moran et al., 2010; Scharinger et al., 2017) and more specifically with the encoding (Klimesch, 1999) and retrieval processes (Bastiaansen et al., 2005; Klimesch et al., 2001). Thus, looking at behavioural and electrophysiological data together, it seems that volitionally detecting an interruption of the pitch/voice regularity required less cognitive resources when hearing speech.

One possibility is that listeners needed more cognitive resources for the acoustic analysis of the pitch dimension, given the smaller number of available cues to pitch changes in the rotated speech condition. In fact, despite spectral rotation preserves the pitch contour, it disrupts the relationship occurring between formant frequencies and pitch in natural speech (Assmann & Nearey, 2007). To this regard, enhanced theta ERS over Frontal electrode sites has also been linked to higher spectral quality, indicating that the quantity of available spectral information directly promotes speech intelligibility (Obleser & Weisz, 2012). Yet, the differences in theta ERS start to emerge at ~300 ms across parietal and parieto-occipital electrode sites, suggesting that the source of the effects could lie in differences pertaining to higher and later-occurring cognitive levels of processing.

In the context of the neurocognitive model of regularities extraction portrayed by Paavilainen (2013), while at a pre-attentive level the auditory cortex automatically represents regularities about different acoustic features, at an attentive level explicit awareness about the rules governing deviant stimuli presentation is necessary to reach high-levels of accuracy in deviant detection. In our study, we made sure participants had explicit knowledge about the task structure and the stimuli by directly describing the stimulation paradigm of the active oddball task and providing extensive practice. However, despite this, participants had life-long experience with speech produced by male and female voices, but certainly not with rotated speech produced by "alien male and female voices". Relatedly, sound regularities appear to be extracted without particular attentional focus (Batterink & Paller, 2019; D. Duncan & Theeuwes, 2020), but extensive experience with particular auditory material may facilitate top-down processing of the extracted regularities, especially with speech stimuli (Monte-Ordoño & Toro, 2017; Sun et al.,

2015). The specific functional role of experience in facilitating the volitional processing of abstract regularities are not yet fully understood and have been linked with enhanced statistical learning abilities (Pesnot Lerousseau & Schön, 2021) or with the development of more efficient strategies in information encoding (Monte-Ordoño & Toro, 2017). In our experiment, the enhanced theta ERS for the speech condition suggests that the presence of intelligible speech and/or human-like voices may have promoted a more efficient encoding strategy of the extracted regularities. Subsequently, regularities could be stored within pre-existing voice representation, thus requesting smaller amounts of cognitive resources, as signalled by the larger P3b amplitude.

The consequences of this facilitation effect may also be tracked in the pattern of beta modulations found for the active task. Oscillatory activity in the beta band is thought to be tied to the status of a cognitive and/or perceptual set Engel & Fries, (2010): While synchronization signals the maintenance of a set dictated by endogenous top-down processes, desynchronization indexes the disruption of cognitive/perceptual sets following exogenous bottom-up sensory components. In line with this interpretation, the ERD associated with the presentation of deviant stimuli may index a disruption of the previous stable cognitive set in which several different instances of speech or non-speech stimuli were being accumulated into one voice/pitch representation. While in the rotated speech condition beta ERD appeared to be longer lasting, in the speech condition it was readily followed by a synchronization. Qualitatively, a beta synchronization with similar spatial distribution seemed to emerge also for the rotated speech condition, but later in time with respect to the speech condition (see Figure S5 in Appendix A). This temporal dynamic might further suggest that the efficient encoding of regularities in speech also allowed for a faster reestablishment of the cognitive set that characterized listeners' activity prior to the presentation of deviant events.

2.4.3 Final Remarks

In conclusion, we show that listeners pre-attentively track pitch regularities in a context of constantly changing formant information irrespectively of the presence of phonological information by possibly using a domain-general mechanism which encodes abstract representations. Regularities representations are then transferred to long-term memory, while encoding additional vocal information in the case of human-like speech. At an attentive level, the presence of phonological information facilitates the volitional use of the previously abstracted

information, suggesting that the output of pre-attentive abstraction mechanisms is not transferred to working memory without effort. ERP and the time-frequency results offer converging evidence that the source of the facilitation driven by the presence of phonological information may be provided by the extensive experience listeners have with speech and voices. This could substantially provide listeners with more efficient encoding strategies which need smaller amounts of cognitive resources to encode information into pre-existing voice representations, ultimately promoting faster and more accurate behavioural response

Chapter 3: Early differentiation of memory retrieval processes for newly learned voices and phonemes as indexed by the MMN[4]

Linguistic and vocal information are thought to be differentially processed since the early stages of speech perception, but it remains unclear if this differentiation also concerns automatic processes of memory retrieval. The aim of this ERP study was to compare the automatic retrieval processes for newly learned voices vs phonemes. In a longitudinal experiment, two groups of participants were trained in learning either a new phoneme or a new voice. The MMN elicited by the presentation of the two was measured before and after the training. An enhanced MMN was elicited by the presentation of the learned phoneme, reflecting the activation of an automatic memory retrieval process. Instead, a reduced MMN was elicited by the learned voice, indicating that the voice was perceived as a typical member of the learned voice identity. This suggest that the automatic processes that retrieve linguistic and vocal information are differently affected by experience.

[4] This chapter has been published in *Brain & Language*, and should be cited as *Di Dona, G., Scaltritti, M., Sulpizio, S. (2021) Early differentiation of memory retrieval processes for newly learned voices and phonemes as indexed by the MMN. Brain & Language. 10.1016/j.bandl.2021.104981*. The study has been conducted in collaboration with dr. Simone Sulpizio (Department of Psychology, University of Milano Bicocca) and dr. Michele Scaltritti (Department of Psychology and Cognitive Science, University of Trento).

Chapter 3

3.1 Introduction

Albeit linguistic and vocal information are naturally intertwined in the speech signal, these two types of information can be selectively extracted to achieve different goals. Indeed, we can understand what is said irrespectively of who is saying it, but we can also identify who is speaking regardless of what she/he is saying. This selectivity becomes possible due to the way in which the cognitive system stores, retrieves and combines different kinds of information that are indexed by different physical features of the signal. Psycholinguistic (Norris & McQueen, 2008) and psychoacoustic models (Belin et al., 2004) consider phonemes and voices as the fundamental information units for speech perception and talker identification, respectively (Formisano et al., 2008). Phonemes can be described on the basis of their first and second formant frequencies (F1 and F2) (Obleser et al., 2003) whereas voices are usually reduced to their fundamental frequency (F0) (Latinus & Belin, 2011). As their identification relies on different acoustic indexes and is performed for different purposes, phonemes and voices are considered to be independently and asymmetrically processed by different brain networks. While phoneme identification predominantly relies on the left superior temporal gyrus (DeWitt & Rauschecker, 2012) voice identification predominantly relies on its right homologous site (Belin & Zatorre, 2003; Zäske et al., 2017).

Despite the aforementioned functional and neurobiological segregation, some evidence suggests that linguistic and vocal information are dynamically integrated at different levels. Behaviourally, neither linguistic nor vocal information can be purposefully ignored without active effort during identification tasks (Mullennix & Pisoni, 1990). Kaganovich et al. (2006) showed that the attentional effort required to filter out either one information or the other is also indexed by the modulation of electrophysiological activity at the level of the Event-Related Potentials (ERPs), across the N1, N2 and P3 components. Authors suggested that the early onset of this effect in the N1 time window indicates that the effort originates during low-level filtering processes. Instead, the modulation of the N2 and P3 components was interpreted as being due to a reduced amount of attentional resources available to support the activation and selection of high-level representations in working memory.

Further, the integration between vocal and linguistic information also characterizes the retrieval processes from long-term memory. When one of the two types of information is retrieved, the identification of the other seems facilitated. Word identification is in fact easier when listeners

hear familiar voices (Nygaard et al., 1994). Similarly, talker identification is easier when they hear native speech (Perrachione & Wong, 2007). These two effects indicate that past experiences with either the linguistic (Zarate et al., 2015) or the vocal component (Johnsrude et al., 2013) of the speech signal aid the identification of the other type of information. In this perspective, the parallel between these two phenomena suggests that they may originate from shared processes that automatically retrieve linguistic and vocal information from memory that is then used to orient attentional resources to the content of interest (Lakatos et al., 2013). Whereas abstract representations of phonemes and voices can be spontaneously formed in a similar way during passive listening (see Chapter 2 and Eulitz & Lahiri, 2004; Jacobsen, Schröger, & Alter, 2004; Shestakova et al., 2003), it is still unknown whether these two kinds of information are also similarly retrieved from long term memory. Addressing this issue will contribute to shed light on how top-down processes funnel former linguistic or vocal knowledge into the processing stream of the upcoming auditory signal. Here, we used ERPs and focused on the Mismatch Negativity (MMN) to investigate how learned voices and learned phonemes are retrieved from long-term memory.

MMN is a highly informative electrophysiological response that can signal not only physical changes in the auditory environment, irrespectively of the listener's attention (Näätänen & Michie, 1979), but also the automatic activation of high-level representations such as memory traces (Näätänen et al., 2007; Pulvermüller & Shtyrov, 2006). In the passive-oddball paradigm, a sound is repeatedly presented (standard stimulus) and is infrequently replaced by a different sound (deviant stimulus). The EEG signal related to deviant events shows a negative displacement from the one related to standard events in the N2 time window, usually around 150-250 ms from the onset of the deviant sound (Näätänen, 1995). This effect is due to a violation of the representation of the standard sound in short term memory (Näätänen et al., 2005). Interestingly, the MMN response is sensitive to linguistic experience, being larger when the deviant stimulus is a known phoneme (or word) compared to when it is an unknown one (Dehaene-Lambertz, 1997; Pulvermüller et al., 2001; Shtyrov & Pulvermüller, 2002). This *enhancement effect* has been interpreted as indexing the retrieval process of native speech material from long-term memory (Näätänen et al., 2005). The same pattern has been reported for familiar voices: Beauchemin et al. (2006) found that the MMN was larger when the deviant phoneme was produced by a familiar talker (i.e., a relative or a friend of the participant), than by an unknown one. The authors suggested

that the enhanced MMN reflects the presence of a memory trace retrieval process for familiar voices. Interestingly, voice familiarity also affected the P3a, a positivity peaking around 300 ms after the onset of the deviant stimulus and usually associated to the automatic reorientation of attention (Comerchero & Polich, 1999). With regard to the P3a, Beauchemin et al. (2006) suggested that, once retrieved, familiar voices appear as more salient to the listener with respect to unknown voices, thus triggering an automatic re-orientation of attention.

Although scanty evidence mentioned above seems to suggest that memory traces for familiar voices and native phonemes are automatically retrieved by means of shared retrieval processes as indicated by the presence of an enhanced MMN, there are at least two crucial aspects that need to be considered. First, apart from individual acoustic features, the representation of a familiar voice could also conceal linguistic information, as such representation would result from several meaningful linguistic interactions with a specific talker. In fact, listeners are able to learn how specific talkers produce phonemes (Eisner & McQueen, 2005) or whole words (Perrachione et al., 2015) by establishing talker-specific phonetic and linguistic representations. A representation of a voice could then entail information about how such voice produces specific speech sounds (Perrachione, 2017; Perrachione & Choi, 2016). Therefore, to study the similarities between the retrieval processes for known phonemes and familiar voices one should isolate the two types of information by investigating memory traces selectively built for either linguistic or vocal information.

A second critical aspect is related to the use of electrophysiological measures to study high-level cognitive processes and the need to account for the dramatic impact that physical properties of experimental stimuli may have on the EEG signal. Amplitude and peak latency of MMN are extremely sensitive to such changes (Näätänen et al., 2007), hence comparisons between MMNs originated by physically different stimuli must be interpreted with caution.

In the present longitudinal study, we overcame the two above crucial issues and trained two groups of Italian native-speakers in learning a new phoneme and becoming familiar with a new voice, and measured their MMN response in both a pre-training and a post-training EEG session. In the pre-training session, participants were exposed to two conditions, both featuring the same standard stimulus – i.e., the syllable /pi:/ produced by an unfamiliar German native speaker. The deviant stimulus varied as a function of the condition. In the phoneme-change condition, it was the syllable /py:/ produced by the same unfamiliar talker that produced the deviant stimulus.

In the voice-change condition, the deviant stimulus was the same syllable /piː/ of the standard stimulus but produced by a different unfamiliar German native speaker. After this first EEG session, participants were divided in two groups and were randomly assigned either to a syllable-identification training or to a talker-identification training. The former group learnt the German phoneme /yː/ presented in the phoneme-change condition, whereas the second one familiarized with the unfamiliar German voice from the voice-change condition. After the training, participants underwent the second EEG session, that was identical to the first one. The use of differentiated training procedures allowed for the isolation of different encoding strategies: the focus of attention during speech encoding – being directed towards linguistic or talker-related information – increases the salience of specific features of the speech signal's representation. Depending on which kind of information is encoded, the application of such strategies results in enhanced behavioural performances in tasks where the encoded information is needed (McAuliffe & Babel, 2016; McGuire & Babel, 2020; Theodore et al., 2015).

Additionally, by learning foreign speaking voices, participants cannot retain any linguistic information, and similarly participants learning a new phoneme from an unfamiliar voice cannot form a voice identity representation of the talker. Testing participants on identical stimuli in both sessions allowed us to control for the influence of physical features and to isolate the high-level processes of interest, i.e., the presence of the enhancement effect as a marker of long-term memory trace retrieval.

On the basis of the previous literature, we sketched two clear-cut predictions. First, we expected that, in both sessions, an MMN is elicited by all the conditions, as the acoustic changes between the standard and deviant stimuli should be clearly detectable. Second, and most importantly, we tested whether memory traces for newly learned voices and newly learned phonemes are retrieved by means of shared retrieval processes and thus would show similar electrophysiological responses. If this is the case, the two different training procedures are expected to trigger the same enhancement effect on MMNs: At the post-training EEG session, the group involved in the talker-identification training should show enhanced MMN when the learned voice is presented as the deviant stimulus whereas the group enrolled in the syllable-identification training should show enhanced MMN when the deviant stimulus is the learned phoneme. An exploratory analysis of P3a was also carried out as it seems to be differently modulated by the

presentation of familiar voices (Beauchemin et al., 2006) or more generally by passive exposure to speech sounds (Kurkela et al., 2019).

3.2 Method

3.2.1 Participants

Thirty-two healthy Italian native speakers were recruited. Two participants were excluded from the analyses as their performance in the talker-identification training did not reach the requested threshold. The final sample included thirty participants (26 females and 4 males, M_{age} = 21.53, SD_{age} = 2.69), all right-handed (as established by the Edinburgh Handedness Inventory, Oldfield, 1971; M = .70, SD = .12). Participants reported to be neurologically healthy and to have normal hearing. Participants' foreign language knowledge and use was assessed with a questionnaire (Sulpizio et al., 2019), in which participants were asked to: a) state which languages they knew, b) estimate the average amount of hours they spent using those languages in a day, c) evaluate their written and oral proficiency on a scale from 1 (*really low*) to 10 (*really high*) and d) indicate whether they had any language certificate. Twenty-nine participants reported English as L2, 1 participant reported English as L3 and French as L2. With respect to L3 and L4, 15 participants reported French, 9 Spanish, 1 Japanese, 1 Chinese and 1 Russian (for further details, see Table S1, Appendix B). Importantly, all participants reported no prior knowledge of German, nor any attendance to lectures/courses of German throughout their lifetime. Participants' education (in years) was also collected (M = 15.66, SD = 2.20).

Participation was compensated either with course credit or with 10€ per hour. The study was approved by the Ethical Committee of The University of Trento. Participants signed an informed consent document prior to the experiment.

3.2.2 Stimuli

Six male native speakers of German (M_{age} = 24, SD = 7*)* were recruited to record the experimental stimuli. They were asked to read aloud two brief texts and several isolated words (n = 23) and syllables (n = 8) in German. Their voice was recorded at 48000 Hz with a professional recorder in a silent room. The texts were two descriptions of two German cities: Hamburg (“Hamburg,” 2019) and Saarbrücken (“Saarbrücken,” 2019). Word stimuli were selected among German minimal

pairs. This was done to force participants to focus on vowels to retain pitch information during the talker-identification training rather than attending to possible idiosyncratic productions of consonants. Syllable stimuli were composed of the phoneme /p/ + a German vowel. Specifically, the syllables were: /pyː/, /pʊ/, /piː/, /pʏ/, /pɐ/, /pə/, /pøː/, /pœ/. To elicit the correct sound without the use of phonemic transcription, talkers were asked to read a priming word containing the desired syllable before reading the actual isolated syllable. Texts, words, and syllables were presented in a random order, and recorded three times each. The best tokens – i.e., those showing, in a qualitative assessment, the lowest of noise and the least number of prosodic irregularities – were selected.

One talker was excluded from the subsequent analyses because of a high level of external noise in the recording. Following Baumann & Belin, (2010) and using Praat software (Paul Boersma & David Weenink, 2018), a voice analysis was performed on the vowels of every syllable token in order to understand which physical characteristics differentiated the speakers' voices. For each talker, mean pitch (F0) and mean F4/F5 formant dispersion in all syllables were calculated. Mean values and standard deviations are reported in Table 2. As only four talkers were needed for the experiment, talker 5 was excluded as his mean F4/F5 dispersion value (718 Hz) was the most distant from the mean F4/F5 dispersion value calculated across all talkers (M = 969 Hz, SD = 147). This was done to reduce the number of physical features by which talkers may be identified. Texts, words, and syllables produced by Talker 1, Talker 2, Talker 3 were selected as stimuli for the talker-identification training. Instead, syllables /piː/ and /pyː/ produced by Talker 4 were used for the syllable-identification training: /iː/ and /yː/ are phonologically contrasting in German. By means of the syllable-identification training Italian participants were supposed to learn the phoneme /yː/, which is not present in the Italian phonological repertoire.

A continuum between /piː/ and /pyː/ was created to test categorical perception of /iː/ and /yː/. The two syllables were morphed with each other using the TANDEM-STRAIGHT MATLAB toolbox (Kawahara et al., 2008). TANDEM-STRAIGHT decomposes speech into fundamental frequency, formant frequencies, aperiodicity, spectro-temporal density, and time. Anchor points across time on the spectrogram were selected to mark onset, midpoint and offset of segments. For every anchor point in time, frequency anchors were set on the first and the second formant frequencies to obtain smoothly morphed stimuli. The morphing continuum was synthesized through linear interpolation of time and aperiodicity parameters and through logarithmic

interpolation of pitch (F0), formant frequencies and spectro-temporal density across time-frequency anchors. A 29-step continuum was generated, producing weighted morphed syllables going from 0% /pi:/ and 100% /py:/ to 100% /pi:/ and 0% /py:/.

For the EEG experiment, the syllable /pi:/ produced by Talker 4 was used as standard stimulus. To create the phonological contrast, the syllable /py:/ produced by Talker 4 was selected as deviant stimulus. Instead, to create a voice contrast, the syllable /pi:/ produced by Talker 1 was selected as deviant stimulus. These critical tokens were selected on the basis of duration similarity (the exact values are reported in Table S2, Appendix B). The duration of the syllable was set at 250 ms for all the tokens by cutting the last offset part of the stimuli and inserting a 50 ms fade-out in amplitude. The physical characteristics of the stimuli used in the EEG experiment are summarized in Table 3. Finally, all the syllable tokens were resynthesized using TANDEM-STRAIGHT to ensure that the stimuli had the same quality overall the whole experiment. The intensity of all the tokens was finally set to 60 dB.

Table 2. Mean values and Standard Deviations (SD) of fundamental frequency (F0) and dispersion across the fourth and the fifth formant (F4/F5) for every talker[5]

Talker	Mean F0 (SD)	Mean F4/F5 Dispersion (SD)
1	100.78 Hz (8.59)	1060.08 Hz (161.07)
2	126.27 Hz (6.46)	962.75 Hz (401.37)
3	111.88 Hz (23.64)	1017.16 Hz (428.98)
4	112.88 Hz (17.66)	1087.01 Hz (256.90)
5	118.48 Hz (12.30)	718.03 Hz (305.98)

[5] The data of the talker that was excluded for the high level of external noise is not reported in Table 2.

Table 3. Physical characteristics of Standard and Deviant stimuli used in the EEG experiment.[6]

	Standard stimulus	Deviant Stimulus	
		phoneme-change	voice-change
Syllable	/pi:/	/py:/	/pi:/
Talker	4	4	1
F0	120 Hz	118 Hz	103 Hz
F1	345 Hz	433 Hz	276 Hz
F2	2292 Hz	1591 Hz	2377 Hz
Duration	250 ms	250 ms	250 ms
Intensity	60 dB	60 dB	60 dB

3.2.3 Procedure

The experiment lasted several days and included two EEG recordings that took place before and after a behavioural training, which differed among the experimental groups.

Pre-training EEG session

During the first day, participants were asked to fill in the questionnaire collecting demographic information, handedness, and language background. Then, participants were prepared for the EEG recording in a dimly lit room and took part in the pre-training session of the EEG experiment. During the experiment, participants were asked to watch a silent video documentary about deep sea creatures while auditory stimuli were delivered via Etymotic ER-1 headphones at fixed volume (60 dB) using E-prime 2.0 Software (Schneider & Zuccoloto, 2007).

Stimuli were presented using the passive oddball paradigm. The syllable /pi:/ produced by Talker 4 was repeatedly presented as standard stimulus with a fixed Interstimulus Interval (ISI) of 550 ms. The standard stimulus was infrequently replaced by the deviant stimulus with a probability

[6] F0, F1 and F2 were measured on the voiced part of the final vowels of the syllables.

of occurrence of .15. The order of presentation of standard and deviant events was randomized, but a minimum of two standard events occurred before the presentation of a deviant event. In the voice-change condition the syllable /piː/ produced by Talker 1 was used as deviant, while in the phoneme-change condition the syllable /pyː/ produced by Talker 4 was used. The conditions were separately presented, one per block and the order of presentation was counterbalanced across participants. The two blocks included 850 standard events and 150 deviant events that were randomly presented to each participant. Participants took a small break between the two blocks.

At the end of the EEG experiment, participants were randomly assigned either to the talker-identification training or to the syllable-identification training, forming two groups of 15 participants each. The two groups were matched for age, sex and years of education.

Talker-Identification Training

Participants assigned to this group took the online version of the Glasgow Voice Memory Test (Aglieri et al. 2017; available at https://experiments.psy.gla.ac.uk/index.php) to assess the individual ability to memorize and recognize unfamiliar voices. This test was administered in order to identify potential phonoagnosic participants in the sample, as indicated by a performance scoring below 2 SD from the group-mean (Roswandowitz et al., 2014). No participant showed a performance below the selected threshold.

Then, the talker-identification training started. The training procedure was modelled on former studies in literature that were successful in establishing representations of voice identity for the trained voices (Fontaine et al., 2017; Latinus et al., 2011). In this kind of trainings, the use of multiple talkers can provide an acoustic space in which voices can be physically represented (Andics et al., 2013). This helps listeners to grasp the physical features by which voices can be discriminated from each other in the first place. Once a physical substrate is provided, listeners are facilitated in pinning idiosyncratic vocal features to identity labels (i.e., personal names) and limit the perceptual space around them, solidifying voice representations. This is not the case with phonemes, which are already contrastively represented on a common acoustic and perceptual space with familiar physical dimensions.

In the first training block, participants familiarized with the 3 voices (Talker 1, Talker 2, Talker 3) by listening to two brief recorded texts for each talker. A fake name and a number (1, 2 or 3) for each talker was presented at the centre of the screen while the recorded texts were played

via headphones at fixed volume (60 dB). To form the stimulus-response mapping, at the end of every recording, participants were asked to press the indicated keyboard button corresponding to one of the three talkers, following written instructions on the screen. All the recorded texts were presented once in a fixed order.

In the second training block, participants performed a talker identification 3-AFC task: Twenty-three words (see Table S3, Appendix B) were then auditorily presented in a random order via headphones and participants were asked to identify the talker by pressing button 1, 2 or 3 on the keyboard. The names of the talkers and the associated buttons were always visible on the screen while the task was performed. After any incorrect answer, the stimulus was presented again, and the correct answer was given on the screen. In the third training block participants performed the 3-AFC task on isolated syllables. All the recorded syllables (/pyː/, /pʊ/, /piː/, /pʏ/, /pɐ/, /pə/, /pøː/, /pœ/) were presented 5 times for each talker (n = 3) in a random order, for a total of 120 trials. Participants received feedback on their performance as they did in the previous block. Successively, the test block was presented: This block was identical to the third training block, but no feedback was given. After the test block, participants went home and came back in the following days to repeat the training, once a day, until their performance at test exceeded the discrimination threshold level of 66% in a 3-AFC (Prins, 2016). Two participants that still showed a performance below the threshold at the fifth day of training were not invited to take part to the second EEG session and were thus excluded from the sample. The day after the criterion was met, participants came to the lab for the post-training EEG session. Before the post-training EEG session, they repeated the training and the test phase once more to ensure that the identification was consolidated (i.e., the discrimination threshold was again above 66%). The training lasted on average 3.33 days (*SD* = 0.72, range 3 - 5).

Syllable-identification Training

Participants took part in a Syllable Identification and Goodness Rating task, and a Listen-and-Repeat task. The procedure was the same used by Tamminen et al. (2015) who ran an MMN study in which they trained Finnish participants to learn a phoneme not present in their phonological repertoire. Here, this procedure was used to teach Italian participants the German phoneme /yː/. The training started with a familiarization phase, during which, via headphones, participants could listen to the /piː/ and /pyː/ syllables recorded from Talker 4 as many times as they wanted by

pressing buttons 1 and 2 on the keyboard. The two stimuli corresponded to the endpoints of the 29-step continuum. Then, the Syllable Identification and Goodness Rating task started. To be sure that participants understood the task, they were presented with a practice block in which all the 29 variants of the syllables from the continuum were presented once. For every stimulus presentation, participants were asked to state which syllable they heard by pressing button 1 or 2 on the numerical keyboard. Afterwards, they were asked to express a goodness rating of the stimulus on the basis of how much it was representative of the selected syllable category (/piː/ or /pyː/) by pressing a button from 1 (*bad representative of the category*) to 7 (*good representative of the category*) on the keyboard. After the practice, the test blocks were presented. In the test blocks participants performed again the Syllable Identification and Goodness Rating task for each of the 29 variants of the syllables. Each variant was presented 10 times for a total of 290 trials divided into 2 blocks, with a small break between them. Afterwards, participants started the Listen-and-Repeat task. During this task, the stimuli at the two endpoints of the continuum (i.e., /piː/ and /pyː/ syllables) were presented via headphones 30 times each and participants were asked to repeat aloud each sound as precisely as possible. In the subsequent day, the Listen and Repeat task was repeated twice, interleaved by the Syllable Identification and Goodness Rating task. On the third day, the Syllable Identification and Goodness rating task was repeated, followed by one last session of the Listen-and-Repeat task. Afterwards, the post-training EEG session took place.

Post-training EEG session

This recording session was identical to the first one, with the exception that no questionnaire was administered to the participants.

3.2.4 EEG recording and processing

The EEG was recorded with an eego sports system (ANT Neuro) at a sampling rate of 500 Hz (filters: DC to 130 Hz, third- order sinc filter), from 64 Ag/AgCl shielded electrodes referenced to CPz and placed in the standard 10-10 locations on an elastic cap. Electro-oculograms were acquired with an additional electrode placed under the left eye. Impedance was kept < 20 kΩ. The signal was re-referenced offline to the average reference. Data was filtered between 0.01 and 30 Hz using a 4^{th} order Butterworth passband filter (24 dB/oct Roll-off) and resampled to 250 Hz. A Notch filter at 50 Hz was applied to attenuate line noise. Independent Component Analysis was

run on the continuous signal using the Infomax algorithm (Bell & Sejnowski, 1995), and eye blink components were identified and removed. Epochs were extracted from 100 ms before stimulus onset until 500 ms after stimulus onset and a baseline correction was applied. The baseline was corrected by subtracting the mean voltage of the pre-stimulus period (-100 to 0 ms) from the waveform of the entire epoch. Epochs containing signal with an amplitude exceeding 100 μV in any of the 64 channels were rejected. An average of 2.16 epochs (SD = 4.82) epochs per participant were rejected. All the epochs corresponding to standard events coming immediately after deviant trials were removed from the analysis, to avoid any contamination from later potentials triggered by deviant events.

3.2.5 Statistical Analyses

Behavioural Data - Talker-identification training

The accuracy data was analysed by means of a Generalized Linear Mixed Model (GLMM) with a logit link-function using the 'lme4'package (Bates et al., 2015) in R Software (R Core Team, 2013). Data was fitted to the full model with fixed factors of session (pre-training, post-training), talker (Talker 1, Talker 2, Talker 3) and their interaction, and by-participants and by-item random intercepts. The best model was selected by implementing backward elimination on the full model via likelihood-ratio Chi-squared tests implemented with the drop1 R function.

Behavioural Data - Syllable-identification training

For each participant, the proportion of /py:/-answers was fitted to a logistic psychometric function with the R package 'quickpsy' (Linares & López i Moliner, 2016) which estimates the Point of Subjective Equality (PSE) and the slope of the identification response. The PSE is the predicted level of morphing where the proportion of answers is at chance level (.5 for 2-AFC tasks). The slope value refers to the steepness of the response curve and represents the subjective degree of certainty: The steeper the slope the more defined are the two categories. Individual PSE were then analysed across sessions to evaluate the effect of training by means of paired t-tests. As slope values violated the normality assumption (tested via Shapiro-Wilk test, $W = 0.79$, $p < .001$), they were analysed via Wilcoxon Signed rank test. Mean goodness ratings associated to the stimulus at the PSE were calculated within every participant and within every session. Paired Wilcoxon signed rank tests were performed to confront mean goodness ratings at PSE with those at the endpoints

of the continuum and z-values were reported. The same statistical test was then used to evaluate the possible changes in mean goodness ratings at PSE between the pre-training, mid-training, and post-training sessions. All the t-tests and Wilcoxon signed rank tests were then corrected with False Discovery Rate (FDR) adjustment.

EEG data

Separate ERPs were computed by averaging epochs within each participant and within all the combinations of the factors condition (phoneme-change, voice-change), probability of occurrence (standard, deviant) and session (pre-training, post-training). The MMN was calculated within each participant and within each of the combinations of factors condition and session, by subtracting the standard ERP from the deviant ERP. Fz, FCz and Cz channels were selected for statistical analyses as indicated by previous works on the enhancement effect (Beauchemin et al., 2006; Shtyrov et al., 2010; Tamminen et al., 2015). The mean peak latency of MMN was separately measured for the phoneme-change and the voice change conditions (Gu et al., 2013) to prevent possible influences of overlapping components (i.e., P3a) that could impact the precision of measurement of the enhancement effect of the MMN. This last methodological aspect is critical in our experiment as latency differences are likely to occur between two separate MMN components that are generated by changes in different physical dimensions (Näätänen et al., 2007).This was done by averaging the latency values of the most negative peak between 150 and 350 ms of each participant across all sessions and channels. The mean amplitude of the MMN was measured on a 40 ms time window that was centred on the mean peak latency (Steinberg et al., 2011).

Paired t-tests were run to compare the mean amplitude of standard and deviant events to check that MMN was correctly elicited in the selected time window. Then, a four-way mixed ANOVA was performed on the amplitude of MMN with group (talker-identification training, syllable-identification training) as between-participants factor and condition (voice-change, phoneme-change), session (pre-training, post-training), and channel (Fz, FCz, Cz) as within-participants factors.

To verify the presence of the enhancement effect, paired t-tests were performed on the mean amplitude of MMN, comparing the pre-training with the post-training session, within every group and condition. The amplitude of the enhancement was then calculated by subtracting the

mean amplitude of MMN of the pre-training session from the one measured at the post-training session. A three-way mixed ANOVA was performed on the amplitude of the enhancement effect with the group (talker-identification training, syllable-identification training) as between-participants factor, and condition (voice-change, phoneme-change) and channel (Fz, FCz and Cz) as within-participants factors.

The qualitative inspection of differential waveforms clearly indicated the presence of a P3a component in a scalp area extending from fronto-central to centro-parietal electrode sites. The mean amplitude of P3a was calculated on FCz, Cz and CPz on an 80 ms time window (Beauchemin et al., 2006) that was centred on the mean peak latency of the most positive peak in the 250-500 time window. The mean peak latency was calculated using the same method that was used for MMN but this time irrespectively of the condition as the use of a relatively large time window reduces the influence of other contiguous components (i.e., MMN).

A five-way mixed ANOVA was run with group (talker-identification training, syllable-identification training) as a between-participants factor, and condition (voice-change, phoneme-change), session (pre-training, post-training), probability of occurrence (standard, deviant), and channel (FCz, Cz, CPz) as within-participants factors. Greenhouse-Geisser correction was applied to degrees of freedom when sphericity assumptions were violated. P-values of post-hoc t-tests were corrected applying the FDR correction.

3.3 Results

3.3.1 Behavioural data

Talker-Identification Training

The pre-training and post-training accuracy scores are represented in Figure 4A, B. At the end of the post-training session, the mean accuracy in the 3-AFC identification task was 85% (SD = 0.10) across all talkers. The mean accuracy for Talker 1, 2 and 3 were 86% (SD = 0.08), 90% (SD = 0.09), and 79% (SD = 0.09), respectively. The final GLMM included session and talker as fixed factors and participants and item as random factors. The model showed a significant effect of session, revealing a higher identification accuracy in the last than in the first session (β = 1.33, SE = 0.08, z = 15.18, p < .001). The effect of Talker was also significant, with Talker 3 being recognized less accurately than Talker 1 (β = -0.63, SE = 0.09, z = -6.36, p < .001) and Talker 2

being recognized more accurately than Talker 1 ($\beta = 0.27$, $SE = 0.10$, $z = 2.51$, $p = .011$) and Talker 3 (. $\beta = 0.90$, SE = 0.10, z = 8.72, p < .001).

Syllable-Identification Training

The pre-training and post-training identification responses and goodness ratings are represented in Figure 4C, D. PSE values shifted from a location that was approximately at the physical centre of the continuum in the pre-training session ($M_{PSE} = .53$), towards a morphing level nearer to the syllable /py:/ in the post-training session ($M_{PSE} = .61$), t (14) = 4.02, $p = .003$. PSE values also shifted between pre-training and mid-training session t (14) = 3.28, $p = .005$ and between mid-training and post-training session t (14) = 3.46, $p = .004$, showing a constant increase. Slope values showed a significant increase in steepness only from mid-training session to post-training session $z = 2.78$, $p = .01$. The mean goodness rating values associated to the endpoints of the continuum calculated across sessions were higher with respect to the ones at PSE both at the 0% /py:/ end $z = 5.80$, $p < .001$ and at the 100% /py:/ end $z = 5.77$, $p < .001$, indicating that participants judged the endpoints of the continuum as better representatives of the respective syllable categories. The mean goodness rating values calculated at PSE did not differ across sessions (all $p > .2$) meaning that the overall perceived quality of the stimuli at PSE did not change after training.

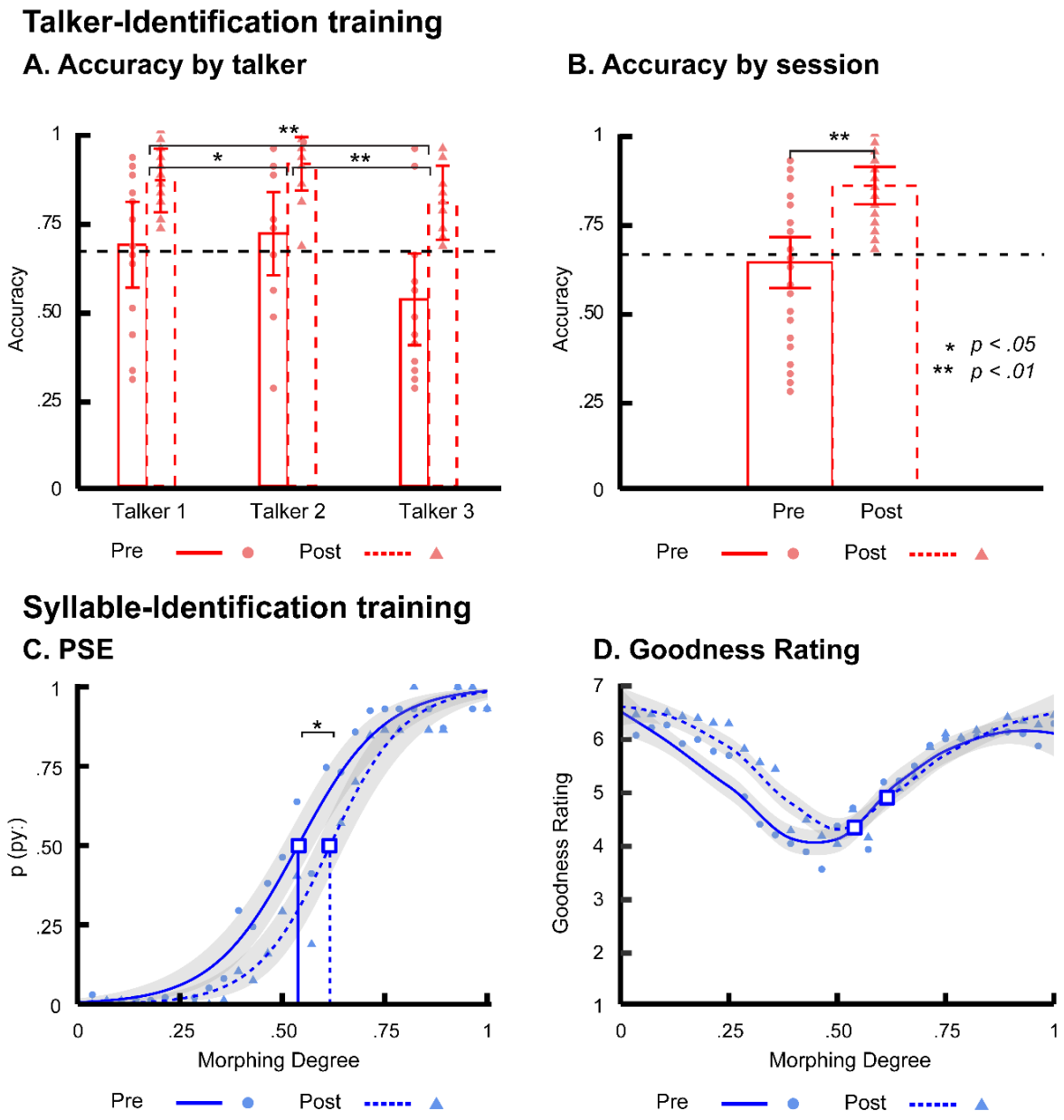


Figure 4. Behavioural results of the talker-identification (red) and the syllable-identification training (blue) in the pre- (continuous line) and in the post-training (dashed line) sessions. (A) and (B) show the proportion of accurate responses for the 3-AFC task of the talker-identification training broken down by talker and by session. Error bars represent the standard error. The dashed horizontal line represents the behavioural discrimination threshold of .66. Small circles and triangles indicate respectively individual scores in the pre- and the post-training sessions. (C) Probability of answering /py:/ as a function of the morphing degree across the pre- and post-training session. The small squares represent the PSE. (D) Goodness ratings as a function of the morphing degree across the pre- and post-training sessions. Small squares represent the goodness rating at PSE. Shaded grey areas represent the standard error. Small circles and triangles indicate respectively individual scores in the pre- and the post-training sessions.

3.3.2 EEG

Mismatch Negativity component

Following the peak detection algorithm described above, the mean amplitudes of MMN was measured in the 215-255 ms time window for the voice-change condition and in the 199-239 ms time window for the phoneme-change condition. The difference between standard and deviant events was significant at every channel (all *ps* < .01) within all the combinations of group, condition, and session factors (see Table S4, Appendix B) confirming that MMN was successfully elicited. MMN waveforms are displayed in Figure 5A, B.

The ANOVA on the mean values of MMN showed a three-way interaction between group, condition, and session F (1, 28) = 5.37, p = .028, η_p^2 = .161. Follow-up 2-way ANOVAs conducted separately within each group indicated that participants enrolled in the syllable-identification training only showed a main effect of session F (1,14) = 11.78, p = 0.004, η_p^2 = .457, with larger MMN for the post-training than the pre-training session. Differently, the group enrolled in the talker-identification training showed a two-way interaction between condition and session F (1,14) = 9.92, p = .007, η_p^2 = .415. Although post-hoc comparisons for the talker-identification training failed to show any significant difference (all *ps* > .1) between the sessions, the inspection of the means suggested that while the amplitude of MMN decreased (i.e. became less negative) in the voice-change condition *(*M_{pre} = -1.363, SD_{pre} = 1.116; M_{post} = 1.057, SD_{post} = 0.755) it increased (i.e. became more negative) in the phoneme-change condition (M_{pre} = -0.813, SD_{pre} = 1.051; M_{post} = 1.065, SD_{post} = 0.975) after the training. Finally, the main effect of channel was also significant, F (1.26, 35.28) = 24.00, p < .001. No further effect reached significance (all *Fs* < 3.727, *ps* > .063).

The enhancement effect

The analyses on the enhancement effect (Figure 5C) showed an interaction between group and condition F (1,28) = 5.36, p = .028, η_p^2 = 0.161, with the two groups showing two patterns going in opposite directions for the conditions that were targeted by the respective training procedures. While the amplitude of MMN in the voice-change condition unexpectedly decreased for the group enrolled in the talker-identification training, it increased in the phoneme-change condition for the group enrolled in the syllable-identification training, yielding a significant difference between the two t (28) = 3.03, p = .014. The two groups also differed in the voice-change condition, as in the group enrolled in the syllable-identification training this condition yielded an increase in the

amplitude of MMN with respect to the decrease recorded in the other group t (28) = 3.09, p = .014. Additionally, the group enrolled in the talker-identification training showed a significant difference between conditions t (14) = 3.149, p = .014, with the MMN amplitude decreasing in the voice-change condition, but increasing in the phoneme-change condition after training (all other |t|s < 0.06, *ps* > .973). No further effect reached significance (all *Fs* < 3.728, *ps* > .063).

P3a

The mean amplitude of P3a was measured in the 282-362 ms time window. The inspection of the grand-averaged ERPs suggested that the amplitude recorded for deviant events increased between the pre- and post-training session across both groups, and both conditions, but apparently more in the group enrolled in the syllable-identification training (Figure 6). The ANOVA showed a significant interaction between group and session, F (1,28) = 7.77, p = .009, η_p^2 = .217. Post-hoc comparisons revealed that only the group enrolled in the syllable-identification training showed a larger P3a in the post-training than in the pre-training session, t (14) = 3.43, p = .016 (all other |t|s < 1.98, *ps* > .113). Additionally, the three-way interaction between condition, channel and probability of occurrence was significant, F (1.514, 42.392) = 3.76, p = .042, η_p^2 = .119. The analysis of the voice-change condition showed an interaction between probability of occurrence and channel, F (1.364, 39.556) = 7.83, p < .001, η_p^2 = .212. The same interaction also emerged in the analysis of the phoneme-change condition F (1.268, 36.772) = 6.47, p = .002, η_p^2 = .182. No further effect reached significance (all *Fs* < 3.84, *ps* > .059).

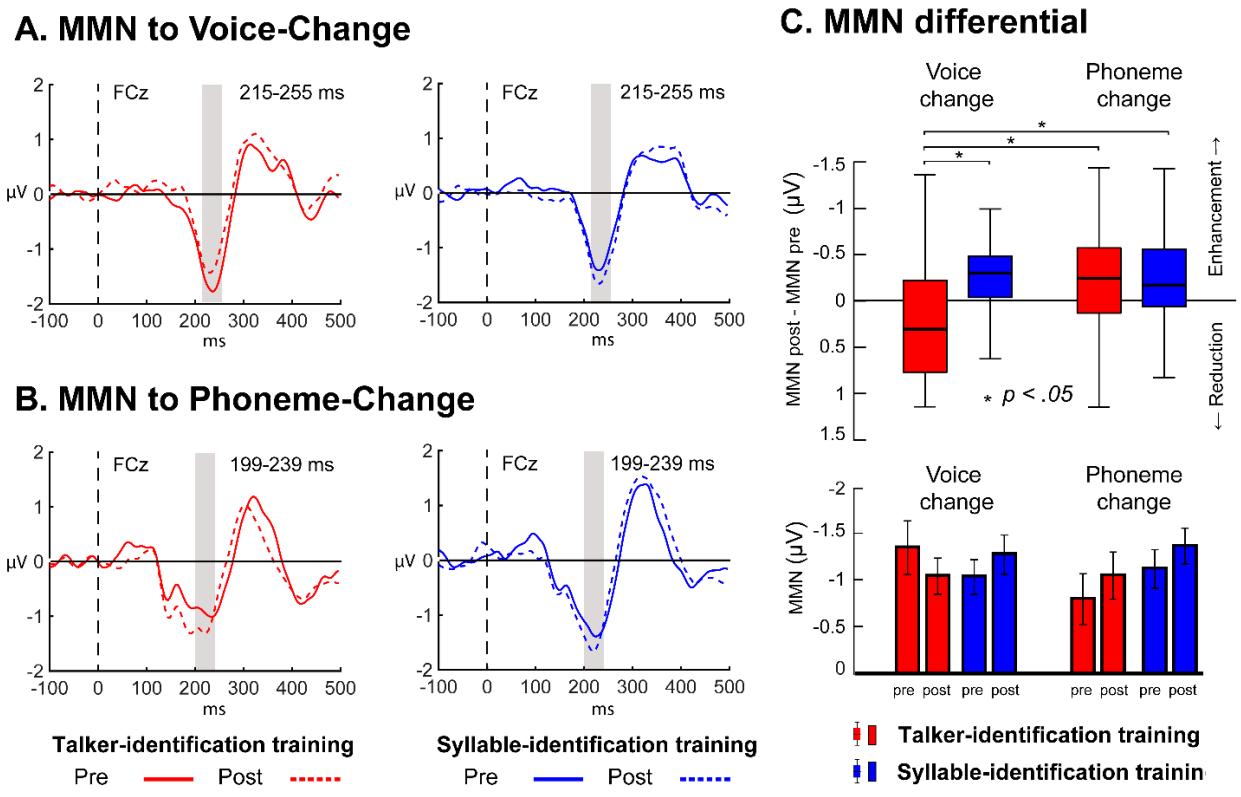


Figure 5. MMN for the different conditions in the group enrolled in the talker-identification (red) and the group enrolled in the syllable-identification training (blue). MMN was calculated in the pre- (continuous line) and in the post-training (dashed line) sessions at a representative channel (FCz) for the voice-change condition (A) and the phoneme-change condition (B). The grey rectangle indicates the time-window used in the analysis. (C) Boxplots (upper part) represent the differential amplitude calculated by subtracting the MMN measured at the post- from the one measured at the pre-training session in both conditions. Barplots (lower part) represent the mean amplitude value of MMN (± SE) divided by session (x axis). Boxplots and barplots represent signal amplitude averaged across Fz, FCz and Cz channels for the Voice-change condition (left) and the Phoneme-change condition (right) in the group enrolled in the talker-identification (red) and the syllable-identification training (blue).

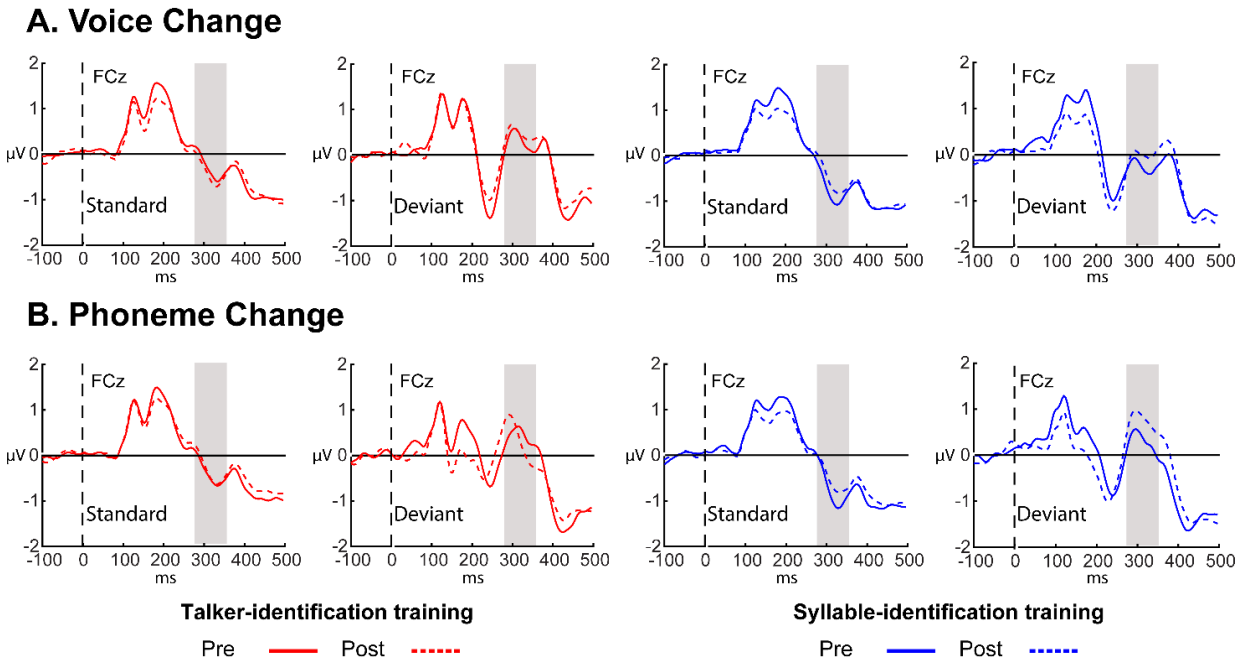


Figure 6: P3a for the different conditions in the group enrolled in the talker-identification (red) and the group enrolled in the syllable-identification training (blue). ERPs for standard and deviant events calculated in the pre- (continuous line) and the post training (dashed line) sessions at a representative channel (FCz) for the voice-change condition (A) and for the phoneme-change condition (B). The grey rectangle indicates the 282-362 ms time-window used in the analysis.

3.4 Discussion

This longitudinal study investigated how listeners automatically retrieve familiar voices and phonemes from memory. We trained one group of participants to identify a foreign-speaking voice, and the other one to identify and produce a new phoneme without providing any talker related information nor different speech samples from which to retain additional voice-specific acoustic features. In this way we controlled the influence of linguistic and vocal information during the formation of the memory traces for a voice and a phoneme, respectively.

Behavioural data confirmed that participants learned the trained materials (i.e., voice or phoneme). For the talker-identification training, the accuracy improved across days, indicating that participants formed a voice representation in memory that helped them to identify the talker independently of linguistic information. Similarly, for the syllable-identification training, the shift of the PSE (i.e., the category boundary) and the increase in the steepness of the slope indicated that the formation of a phonemic representation in memory reshaped the perceptual boundaries between the known and the newly-learned phoneme independently of talker's voice identity.

The ERP data showed that both voice and phoneme changes successfully elicited an MMN, indicating that listeners were able to preattentively detect the acoustic differences that characterized the two conditions (Tuninetti et al., 2017). However, with respect to the training-induced changes in the amplitude of MMN, the enhancement effect was visible for the learned phoneme, but not for the learned voice, suggesting that voices and phonemes are retrieved from memory via different mechanisms. Below we argue that the automatic retrieval processes elicited by the presentation of learned phonemes and voices are differently modulated by experience, suggesting that the processing stream of linguistic and vocal information are at least in part functionally dissociated since the early stages of speech perception.

3.4.1 Learning and retrieving a new phoneme

In the behavioural task, the PSE at baseline (i.e., pre-training session) was approximately located at the physical centre of the continuum. This suggests that participants initially relied on the acoustic features to identify the syllables, but then recalibrated the identification response on the basis of what they learned. Within these circumstances, the shift in the PSE towards the /py:/ category possibly reflects the surfacing of a top-down categorization driving the processing of acoustic information (Dehaene-Lambertz et al., 2005). Moreover, the increase in the steepness of

the slope indicates that the categorization criterion became sharper over time. Goodness ratings were not influenced by the training and this suggests that qualitative evaluation processes of newly learned phonemes may rely on mechanisms that take more time to develop (Tamminen et al., 2015) with respect to the ones responsible for identification and memory retrieval. Nonetheless, the learning of a phonological category is also testified by the electrophysiological results: In line with previous findings, the group enrolled in the syllable-identification training showed an enhancement effect for the learned phoneme which is thought to represent an automatic memory retrieval process (Dehaene-Lambertz, 1997; Näätänen et al., 1997). We can exclude that this effect may have been determined by an accidental familiarization with the voice of Talker 4 which was constantly presented during the training, as – in sharp contrast with the results for the syllable-identification training - the group enrolled in the talker-identification training showed a reduction of MMN as a result of the familiarization with the voice of Talker 1.[7]

Taken together, our results are in line with previous works that used listen-and-repeat tasks to teach participants foreign vowels and consonants. In these studies, new phonemes are learned by exploiting their contrastive nature with native phonemes for different physical features (e.g., duration, voice onset time, formant frequencies; (Saloranta et al., 2020;Tamminen et al., 2015; Ylinen et al., 2010). Considering the replication of these findings, new phonemes appear to be learned even if they are phonetically defined by different physical features and this is a convincing clue that points towards the formation of abstract phonemic representations (Shestakova et al., 2002).

As an additional finding we reported that, independently of the condition, the group enrolled in the syllable-identification training showed a larger P3a in the post-training than in the pre-training session. P3a is thought to index an early reallocation of attention that follows the detection of change in auditory stimulation and its amplitude increases as a function of both the physical differences between the standard and deviant stimuli (Wronka et al., 2012), and the target status – i.e., the P3a is larger for target than non-target stimuli (Comerchero & Polich, 1999).

[7] An additional analysis was performed to further ascertain the absence of any talker familiarity effect due to the exposure to the voice of Talker 4 during the syllable-identification training. Paired t-tests (FDR corrected) on the amplitude of standard ERPs averaged across Fz, FCZ and Cz channels were performed in the time window used for the analysis of the MMN between the pre and the post-training session. No significant difference was found (all $ps > .2$).

The listen-and-repeat procedure required participants to attend to the presented stimuli before repeating them out aloud. As a result of the attentional request of this procedure, the presentation of the /piː/ and /pyː/ syllables may have induced a target-like response to previously non-target sounds also during the EEG experiment, resulting in an enhanced P3a component irrespectively of the talker's voice or the probability of presentation. In fact, during speech production multiple stages – as, e.g., self-monitoring (Levelt et al., 1999), phonetic encoding and articulation (Jongman et al., 2015, 2020) – require the allocation of sustained attention. Also, speech production may enrich the auditory representation with articulatory and motor features (Grabski & Sato, 2020; Scott & Perrachione, 2019). For these reasons, it is likely that the specific attentional demand enhanced the attentional engagement elicited /piː/ and /pyː/ syllables in the EEG recording. This resulted in a stronger P3a, which was generalized to all the instances of /piː/ and /pyː/ (i.e., standard and deviant syllable /piː/ across talkers and deviant syllable /pyː/).

3.4.2 Learning and retrieving a new voice

When comparing the post-training with the pre-training session, for the voice-change condition, the amplitude of MMN increased when untrained (i.e., in the group enrolled in the syllable-identification training), but unexpectedly decreased when it was trained (i.e., in the group enrolled in the talker-identification training). Therefore, learning a new voice induced an apparent reduction – instead of an enhancement – of the MMN.

Within the neural voice space, voices are thought to be represented as a function of the acoustic distance from a prototypical voice model, which is built and updated throughout the life-course of individuals (Latinus et al., 2013). While the voice space is fundamental for the comparison between different voices, the training-based acquisition of familiarity with a voice results in the formation of a within-voice space in which the intra-talker variability is represented in relation to a mean voice identity representation (Lavan et al., 2019). Two fMRI studies showed that after voice identification training, right inferior frontal cortex and left superior temporal sulcus respond more weakly to identity-typical voices vs identity-atypical voices, indicating that the more a voice stimulus is near to the hypothetical value of the learned mean voice identity, the less these areas will be activated, independently of the position of the voices in the acoustic voice space (Andics et al., 2010, 2013). It is possible that in the context of our study, the presentation of the learned voice in the post-training EEG session may have triggered the activation of an acquired

mean voice representation to which the presented auditory instance was perceived as more identity-typical than it was at the pre-training session, thus determining a reduction of amplitude of the MMN. Yet, given the differences between indirect and direct measures of neurophysiological activity this hypothesis only represents an educated proposal that needs further testing.

The absence of the enhancement effect is in contrast with one particular study that investigated automatic memory retrieval processes for familiar voices, in which Beauchemin et al., (2006) showed larger MMN responses for the French vowel /a/ pronounced by familiar than unknown voices. This inconsistency could be attributable to the different nature of the voice representations investigated in the two studies. While Beauchemin et al., (2006) used voices of family members or friends of the participants, in the present study, participants were familiarized with previously unknown voices through training. Recently familiarized voices acquired through training protocols are not fully akin to ecologically acquired voice identities (Maguinness et al., 2018; Zäske et al., 2017) and appear to be dependent on separate neural networks (Birkett et al., 2007; Zäske et al., 2017). Another crucial difference between the two studies is the linguistic context in which learning occurred: while it was native Beauchemin et al., (2006), in the present study a non-native linguistic environment prevented the influence of known linguistic information during voice learning. Different studies report enhanced MMN contingent to the presentation of native phonemes or words (Dehaene-Lambertz, 1997; Näätänen et al., 1997; Pulvermüller et al., 2001; Shtyrov & Pulvermüller, 2002) and it was shown that listeners are able to learn how specific talkers produce phonemes (Eisner & McQueen, 2005) or whole words (McLaughlin et al., 2015) by establishing talker-specific phonetic and linguistic representations. Thus, it is also possible that the finding reported in Beauchemin et al., (2006) rather reflects the activation of a talker-specific phonetic memory trace for the deviant native phoneme.

Two other similar studies showed no differences between MMN to familiar vs unfamiliar voices (Gustavsson et al., 2013; Plante-Hébert et al., 2017). In these two studies, authors used multiple different utterances as experimental stimuli and this methodological aspect may suggest that the enhancement effect indeed depends on the presence of specific linguistic information. This explanation would also be in line with the unexpected enhancement effects reported for the untrained stimuli of our experiment (i.e., the voice-change condition for the group enrolled in the syllable-identification training and the phoneme-change condition for the group enrolled in the

talker-identification training), which may reflect the retrieval process of a talker-specific phonetic memory trace for the tested phonemes.

Considering the discrepancies between the results of the present study and the ones of Beauchemin et al., (2006) further research seems needed in order to better characterize the nature of recently familiarized and familiar voice representations as well as the impact of linguistic information on voice learning.

3.4.3 Limitations

The unexpected enhanced MMN for the untrained stimuli in both groups were possibly induced by passive exposure to the stimuli in the EEG recording. Different studies attempted to capture the effect of passive exposure on auditory change detection mechanisms. Studies with word stimuli showed that passive exposure may lead to enhanced MMN for novel tonal contrasts or tonal word-forms within a single experimental session (L. Liu et al., 2018; Yue et al., 2014). Contrastively, other studies showed that while different training tasks can modulate the amplitude of MMN (Kraus et al., 1995; Tremblay et al., 1997), passive exposure alone is not sufficient to do so (Elmer et al., 2017; Sheehan et al., 2005). As described in Kurkela et al., (2019), the role of passive exposure in the modulation of the electrophysiological activity related to auditory change detection is still unclear. Therefore, the interpretation offered here only represents a speculative proposal that needs to be adequately addressed with further empirical inquiries.

3.4.4 Final remarks and conclusion

The different modulation in the amplitude of MMN responses for trained voices and phonemes challenge the idea that phonemes and voices are retrieved from memory via shared retrieval processes. Interestingly, Schall et al. (2015) showed that, electromagnetic responses during active recognition of native speech and familiar voices start to diverge as early as 200 ms after speech presentation, irrespectively of the physical properties of the stimuli. Our data show that this functional dissociation may characterize also automatic memory retrieval processes as they occur in a compatible time window (i.e., ~200-250 ms). Moreover, these processes are possibly influenced by the way linguistic and vocal information are represented in the brain.

In conclusion, our results clearly show that the brain handles newly learned voices and phonemes differently. The automatic processes that retrieve vocal or linguistic information from

memory appear to be affected by experience in a different way, suggesting the presence of a functional dissociation since the early stages of speech perception.

Chapter 4: Listeners Deal with Between-Talker Variability by Learning Talker-Specific Cues to Lexical Stress[8]

A key challenge for listeners during speech perception is to deal with variability in the acoustic signal caused by differences between talkers. While previous research has mainly focused on segmental talker variability, less is known about how suprasegmental variability is handled. In the current experiment, we assessed/probed the use of a specific cognitive mechanism, perceptual learning, to deal with between-talker differences in lexical stress. In a learning experiment, participants heard Dutch minimal stress pairs (e.g., VOORnaam vs. voorNAAM, 'first name' vs. 'respectable') spoken by two talkers. Talker 1 used only F0 to signal stress (with intensity and duration set to ambiguous), the second talker used only intensity. Crucially, a second group learned the reverse talker-cue mapping. Participants were then tested on words containing two conflicting cues to stress (mixed items). For example, F0 signaled initial stress (e.g., VOORnaam) while intensity signaled final stress (e.g., voorNAAM) in the same word. We found that, despite these conflicting cues, listeners used previously learned information about which talker used which cue, to correctly recognize the spoken words. That is, when listeners learned that Talker 1 used F0, the mixed item described above was perceived as bearing initial stress when produced by Talker 1. If in contrast, listeners learned that Talker 1 used intensity, the word was perceived as bearing final stress (and similarly for words produced by Talker 2). This confirmed the use of perceptual learning as a mechanism to deal with suprasegmental variability between talkers

[8] This chapter will be soon submitted to be evaluated for publication. The study has been conducted in collaboration with Giulio Severijnen (PhD student at Donders Centre for Cognition) dr. Hans Rutger Bosker (Max Planck Institute for Psycholinguistics) and Prof. James McQueen (Donders Centre for Cognition).

Chapter 4

4.1 Introduction

Individual differences between talkers lead to highly variable acoustic realizations of speech. For instance, consider the English noun 'IMport' (capitalization indicates lexical stress) being produced by a male and a female talker. Even though the word itself is identical, biological differences (e.g., size of the vocal tract), but also 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 (e.g., perceiving the other member of the minimal stress pair: the verb 'imPORT') may impede successful communication. The present study assessed how listeners deal with such variability. More specifically, we investigated the use of a cognitive mechanism, perceptual learning, to deal with between-talker variability in lexical stress.

The presence of acoustic variability in speech has been widely established, and can be found at the level of individual segments (i.e., consonants and vowels). For instance, productions of vowels contain different formant frequencies depending on gender, age and regional dialects (Adank et al., 2007; Adank, van Hout, et al., 2004; Hillenbrand et al., 1995). Also, variability in voice onset time (VOT) of stop consonants has been found between talkers (Allen et al., 2003; Theodore et al., 2009). On top of these differences within acoustic cues, talkers even appear to differ in their cue-weighting strategies (for review, see Schertz & Clare, 2020). That is, speech sounds can be 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 is affected by regional dialects in Dutch and gender (Quené, 2008). In American English, Clopper & Smiljanic (2011) found differences in pause distributions and pitch accents between different dialects and genders. Finally, Xie et al. (2021) found that prosodic variability is not only present between large scale groups (e.g., dialects or gender) 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 (statement vs. question) means and distributions for F0 and duration. In addition, talkers also differed in how correlated the cues were. In other words, individual talkers seem to produce prosodic information with variability within each cue, but also in how the cues are combined to produce the intended structure (i.e., cue-weighting). In sum, talker variability is thus highly present 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 use to correctly perceive spoken words, which has been attributed to multiple cognitive mechanisms. That is, listeners use normalization (Sjerps et al., 2011b), abstraction (McQueen et al., 2006) prediction (Brunellière & Soto-Faraco, 2013; Van Berkum et al., 2005), and perceptual learning (Eisner & McQueen, 2005) to deal with variability in the acoustic signal. Lehet & Holt, (2020) 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 offer interesting solutions to the variability problem, the remainder of this study will focus on perceptual learning.

Perceptual learning studies have demonstrated that listeners can change how they map acoustic input to perceptual categories of speech sounds. One of the many factors that can drive such perceptual learning is distributional information of acoustic cues (Idemaru & Holt, 2011, 2014). For instance, Idemaru & Holt, (2011) found that listeners can change how much perceptual weight is given to different acoustic cues 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. peer). They found that when the canonical relation between fundamental frequency (F0) and voice onset time (VOT) was reversed (a voiced plosive is normally signalled with a high F0 and long VOT, but voiced plosives were now signalled with a low F0), listeners down-weighted their reliance on the unreliable cue (i.e., F0), showing rapid adaptation to short-term deviations in cue distributions. 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 deal with between-talker variability. This was

also illustrated by (X. Zhang & Holt, 2018), who adopted the same paradigm as 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. Similar results were induced by modulating talker characteristics (i.e., stimuli spoken by a male or female voice) or by visual presentation of a male or female talker. In sum, these experiments illustrate simultaneous tracking of speaking styles from multiple talkers.

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 largely unclear how perceptual learning is applied to suprasegmental variability between 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. In their experiment, 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 perceptual learning to deal with variability in one type of prosody: sentence intonation.

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, a stressed syllable is usually produced with a higher mean F0, longer duration, and greater intensity (Rietveld & Heuven, 2009). Moreover, spectral balance (Sluijter & van Heuven, 1996) and acoustic vowel reduction (van Bergem, 1993) have also been identified as cues to lexical stress in Dutch, but less so compared to English, where vowels in most unstressed syllables are fully reduced to schwa (Cutler, 1986). It is important to note that the acoustic cues to

lexical stress are not weighted equally in perception. Rather, 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 & Heuven, 2009).

The importance of lexical stress in word recognition has been emphasized by a number of studies. First, Cutler & Van Donselaar, (2001) showed that, in Dutch, lexical stress is used to constrain lexical activation. They presented Dutch minimal stress pairs (e.g., 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 as depicted by shorter 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 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 & 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 style (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. Variation between gender and speaking styles has also been found in other languages such as Italian and Swedish (Eriksson et al., 2013, 2016).

Nevertheless, an understudied question concerns how listeners deal with between-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 his 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 (e.g., canon induces a Strong-Weak (SW) bias while kanon induces a Weak-Strong (WS) bias). Results showed that participants in the SW-bias group indeed gave more SW responses while the WS-bias group gave more WS responses on a subsequent categorization test. Interestingly, perceptual recalibration was also found across segmentally differing words (e.g., using ambiguous versions of SERvisch "Serbian" vs. serVIES "tableware" in exposure). 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.

Second, Severijnen et al. (2021) investigated whether listeners can also adapt to variability in lexical stress in a talker-specific manner. In their experiment, consisting of multiple training phases and a final test phase, native Dutch participants learned to associate non-word minimal stress pairs to object referents (e.g., USklot referring to a lamp, usKLOT referring to 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 using intensity). Results from a yes/no sentence verification task showed that participants were slower to recognize 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 uses 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 leave some open questions. First, while the critical result (longer response times to the talker-incongruent condition compared to the talker-congruent condition) illustrated that listeners were slowed down in perception, it does not necessarily inform us on how perceptual learning of prosodic cues actually influences which word is recognized. That is, the

target word in both conditions was identical, so the intended word would always be perceived regardless of the talker-cue mismatch. In fact, accuracy data in Severijnen et al. (2021) showed no difference between both conditions. An open question concerns how perceptual learning of prosodic cues can have consequences for word recognition (i.e., which word is perceived instead of how it is perceived). Second, even though the behavioral results illustrated perceptual learning of prosodic cues, there was no modulation of the N200, and ERP related to acoustic-phonetic processing (Connolly & Phillips, 1994), which might question the reliability of the obtained results. Third, the test stimuli in Severijnen et al. (2021) always contained only one cue to lexical stress. While this provided large experimental control, it does not address whether same learning mechanisms are at work when richer stimuli are used (i.e., stimuli involving multiple cues to lexical stress).

The present study tried to fill these gaps, and aimed at providing further evidence for perceptual learning of lexical stress in Dutch. 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'), produced by two male talkers. Similar to Severijnen et al. (2021), each talker used only one acoustic cue to lexical stress. 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 task (2AFC), 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 either talker. Note that no explicit feedback was given on the cues, we expected participants to learn them implicitly. After the training phase, participants were tested on the same word pairs in a 2AFC task. This test included, next to perceptually 'clear' (i.e., unambiguous) control items, also "mixed items". These mixed items contained two conflicting cues to lexical stress, with F0 signalling 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 had two hypotheses for the present study. First, 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 in test (e.g., F0 signalling

SW, intensity signalling WS), they should prioritize the stress pattern being signalled 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 signalled by intensity (e.g., WS).

Three important differences should be noted between the present study and Severijnen et al. (2021). First, Severijnen et al. (2021) used non-words as stimuli, which has the benefit of removing any episodic experiences with the words prior to the experiment. While this adds experimental control, it leaves an open question of whether the same mechanism is at work with existing words. Therefore, the present study used existing Dutch words as stimuli. Second, the test items (i.e., mixed items) in the present study were acoustically identical, as opposed to the test items in Severijnen et al. (2021). This in turn, adds considerable experimental control to the present study by taking away any acoustic influences from the stimuli. Third, the mixed items in the present study contain multiple cues to lexical stress compared to the test items in Severijnen et al. (2021), which allowed us to observe whether the same learning mechanisms are at play in multidimensional stimuli. Fourth, the present study uses a different measure compared to Severijnen et al. (2021). That is, the present study examines the categorization responses instead of RTs, which serves as a more direct indication of word recognition. In sum, the present study tries to provide converging evidence using more realistic and controlled stimuli, with a more direct index of word recognition compared to Severijnen et al. (2021).

4.2 Method

4.2.1 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 would not reflect any perceptual processes related to the task. The remaining 80 participants did not report to have any hearing and/or reading problems (71 male, 9 female, age range: M_{age} = 21.81, SD_{age} = 3.76).

4.2.2 Stimuli

The stimulus list consisted of Dutch minimal stress pairs that were segmentally identical but differing in stress pattern. The list consisted of four disyllabic (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 which lexical stress lays on the first syllable (i.e., Strong-Weak; SW words) or on 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 Table S5 in Appendix C for the complete stimulus list.

4.2.3 Recordings

The stimuli were recorded by two male native talkers of Dutch, naïve about the experiment purpose. The talkers were instructed to produce each member of the minimal pairs 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. This was meant to induce sentence-final prosodic properties in the recordings, such as F0-declination, intensity drop, and sentence-final lengthening. In addition to, and separately from these words, we recorded several carrier sentences. 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…").

4.2.4 Stimulus manipulations

We required two types of stimuli in the experiment. First, for the training phase, we needed stimuli in which only one cue signaled lexical stress (e.g., only F0 or intensity while the other cues were set to ambiguous values). These would be used as "control items". Second, for the test phase, we needed stimuli that contained two conflicting cues to lexical stress, henceforth: "mixed items". 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. Duration was always kept at an ambiguous value in all stimuli.

Control items

The first step was to use the recordings to create the control items. The recordings were annotated using the automatic WEBMaus Basic Tool (Kisler et al., 2017) and were manually checked. For the manipulations of the control items, we followed the procedure in Severijnen et al. (2020). We used the recordings to measure three prosodic cues that signal lexical stress in Dutch: F0, intensity and duration (Rietveld & Heuven, 2009) for each syllable separately using Praat (Boersma & Weenink, 2019). First, we measured the mean F0 in the voiced part of the syllable (pitch settings: 75 – 250 Hz). Second, we measured the mean intensity over the entire syllable using the 'Get intensity' function in Praat. Third, we calculated the syllable duration. Each cue was measured in stressed and unstressed syllables, for all items and both talkers. Next, we averaged across talkers and items to obtain mean values for clear SW and WS patterns. Additionally, we averaged across stressed and unstressed syllables to obtain a value of an ambiguous syllable for each cue. Lastly, we calculated step sizes for each cue by subtracting a clear stressed (SW or WS, depending on the first or second syllable) value from an ambiguous value (see Table 4). Given the large differences in syllable duration between disyllabic and trisyllabic words, we calculated the acoustic measures separately for those words. We then took the stressed syllable from each member of the minimal pairs (e.g., *VOOR* in *VOORnaam, NAAM* in *voorNAAM, VOOR* in *VOORkomen, KO* in *voorKOmen*) and applied the syllable-specific ambiguous settings using PSOLA in Praat (Boersma & Weenink, 2019). The syllables were then concatenated to create ambiguous versions of the words. In the trisyllabic words, the third syllable was also set to ambiguous duration and intensity values while the original F0 contour was kept. No further manipulations were performed on the third syllable, and it was then concatenated with the first two syllables. The resulting ambiguous stimuli were then taken as midpoint of the lexical stress continua that were created next.

We created two 7-step continua for each minimal pair and for each talker by altering one prosodic cue (F0 or intensity) and keeping the rest ambiguous. The continua ranged from step 1 (SW) to step 7 (WS) with the ambiguous stimulus in the middle (step 4). Note that we did not alter duration as a cue to stress but kept it at an ambiguous value in all the stimuli. We only needed two cues as variables, so we decided to drop duration to minimize durational variability in the stimuli. To create the F0 and intensity continua, we took the ambiguous stimulus as midpoint and created SW stimuli by increasing one cue in the first syllable by its step size and decreasing that cue in the second syllable (and *vice versa* for WS stimuli). For example, for the word pair *VOORnaam* vs.

voorNAAM, we created an F0 continuum in which F0 signaled clear SW and WS patterns with intensity and duration at ambiguous values. Similarly, we created an intensity continuum in which intensity signaled the stress patterns with ambiguous F0 and duration values.

Based on auditory evaluation of the resulting continua by the first authors, we noticed that the acoustic manipulations did not always have the intended perceptual effect on all items. First, we noticed that an acoustically ambiguous value did not always correspond to a perceptually ambiguous stimulus. Second, we noticed that the used step size did not always result in clear stressed syllables. Note that, since only one cue signaled lexical stress, more extreme step sizes might be needed. Given these between-item differences, we opted for an item-specific manipulation. For each item, we increased/decreased the ambiguous value by the step size of that corresponding cue to obtain a new ambiguous stimulus. Also, the step size was increased if necessary (see Supplementary Table S6 and S7 in Appendix C for item-specific values).

To make the stimuli sound as natural as possible, we performed two further manipulations. First, we replaced the original F0 contours with contours containing a linear F0-declination that was fixed across items. To find a plausible magnitude for the declination, we calculated the difference between the maximum and minimum F0 value in each syllable. We then averaged across words and talkers to obtain a mean value for each syllable separately. This yielded a plausible declination for each syllable (syllable 1: 28 Hz; syllable 2: 14 Hz). To create the contours, we calculated a new starting and end value for the pitch points. We took the mean F0 value in each syllable and added half of the declination magnitude to obtain the starting value (i.e., the new highest F0 value in the syllable). We subtracted half of the magnitude from the mean to obtain the end value (i.e., the new lowest F0 value). We then set the first pitch point in the syllable to the starting value, the last pitch point to the end value and interpolated the values between those. This resulted in linear contours with the magnitude of the declinations and with the mean F0 of that syllable in the middle. Second, the manipulations often led to jumps in F0 between syllables which resulted in artificially sounding speech. To reduce this, we smoothed the transitions between the syllables by inserting a 5 ms fade out at the end of a syllable and a fade in at the beginning of the following syllable.

Finally, we ran two pretests on the stimuli to select the best steps along the continua following two criteria. First, the selected steps should signal clear SW and WS tokens of the words. Second, the selected steps should be comparable across talkers and cues. For example, the SW

token for Talker 1 using F0 should be comparable to the SW tokens for Talker 2 using intensity (and similarly for the rest of the talker-cue combinations). Given all the item-specific manipulations that were performed, the pretests thus verified whether the performed manipulations allowed us to select items following these criteria. For details on the pretests, see the first section of Appendix C.

Table 4. Mean acoustic measures and step sizes of the prosodic cues in all syllables. No step sizes are given for the third syllable since no manipulations were performed on that syllable that required a step size. Also, two values are provided for duration in each syllable. These correspond to the duration values in disyllabic and trisyllabic words.

Syllable			Strong-Weak (SW)	Ambiguous	Weak-Strong (WS)	Step sizes
1st	Duration(ms)	*Disyllabic*	255	225	195	30
		Trisyllabic	212	199	186	13
	F0 (Hz)		130	122	113	8
	Intensity(dB)		71	69	66	3
2nd	Duration (ms)	*Disyllabic*	369	399	429	30
		Trisyllabic	171	202	233	31
	F0 (Hz)		116	118	122	4
	Intensity (dB)		64	65	67	2
3rd	Duration (ms)		324	233	265	NA
	F0 (Hz)		108	112	114	NA
	Intensity (dB)		58	59	61	NA

Mixed items

Next, we used the data on the control items to create the mixed items. For each word and talker, we created items in which F0 signaled a SW pattern and intensity a WS pattern (F0-Intensity), and *vice versa* (Intensity-F0). This was done by combining two cue steps along the lexical stress continua of the control items. For example, for the mixed item *voornaam*, we took step 1 on the intensity continuum (syllable 1: 69 dB, syllable 2: 46 dB) and step 9 on the F0 continuum (syllable 1: 94 Hz, syllable 2: 130 Hz). In the mixed item, these values were thus combined (syllable 1: 69 dB, 94 Hz; syllable 2: 46 dB, 130 Hz). See Figure 7 for the spectrograms of the control and mixed items for one of the words in the stimulus set.

Similar to the control items, we also ran a pretest on the mixed items to select the optimal combination of cue steps based on two criteria. First, when combining both cues (e.g., F0 signaling SW, intensity signaling WS) into one item, the overall stress pattern should be ambiguous. Second, when focusing on only one of the cues, a clear stress patterns should still be perceived (e.g., focusing on F0 should result in perception of SW). This pretest verified whether the items indeed met the two criteria and were thus suitable for the experiment. Results showed that the manipulations did not result in completely ambiguous words. Instead, we found that F0 was the more dominant cue in perception (for details on the acoustic manipulations and the pretest, see first section in Appendix C). Nonetheless, the mixed items resulted in more ambiguous responses compared to the control items, indicating that the two cues to lexical stress were interfering with each other. Therefore, we judged the items to be suitable for the main experiment.

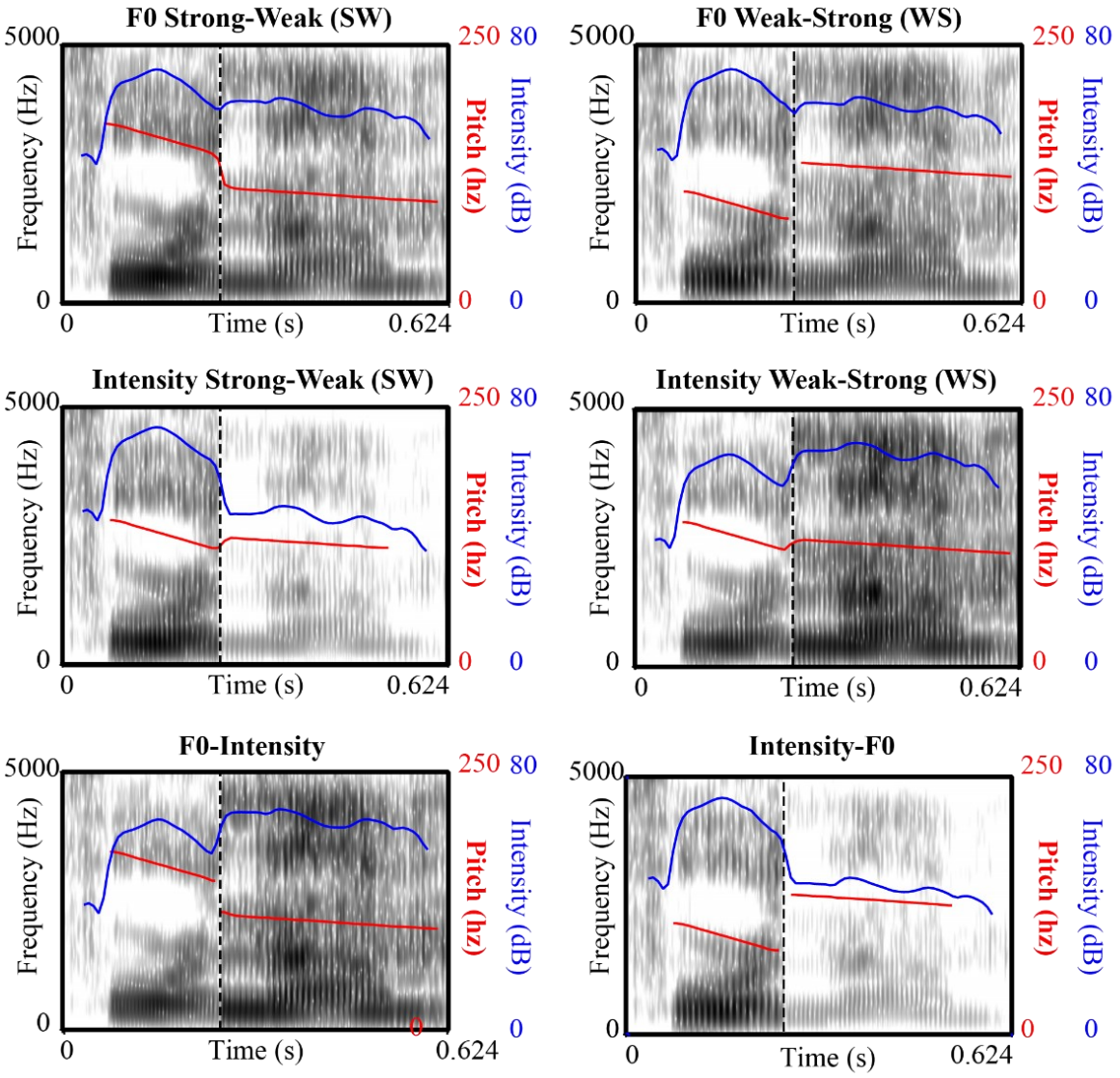


Figure 7. Spectrograms of control items (F0-continuum, top row; intensity-continuum; middle row) and mixed items (bottom row) of one of the target words (*voornaam*), produced by Talker 1. The red lines indicate F0 tracks and the y-axis on the right-hand side, depicted in red, represents the scale for the F0 tracks. The blue lines indicate intensity contours and the y-axis on the right-hand side, depicted in blue, represents the scale for intensity.

4.2.5 Procedure

The experiment was built and hosted online using the Gorilla Experiment Builder (www.gorilla.sc). Participants first performed a headphone screening, allowing only participants who wore headphones to continue with the experiment, which consisted of a familiarization, training and a test phase. The familiarization phase ensured that participants were familiar with the pronunciations of the words and their definitions. The aim of the training phase was for participants to learn which talker used which cue to signal lexical stress in the control items. After the training phase, participants were tested on the mixed items, which would allow us to observe how perception of the mixed items was affected by the training phase. Participants assigned to one of the talker-cue mappings (e.g., Talker 1 used F0, Talker 2 used intensity and *vice versa*) and response position (e.g., SW item appearing on the left side, WS item on the right side and *vice versa*), all possible combinations were counterbalanced across participants.

Familiarization phase

In the familiarization phase, participants were visually presented with orthographic word forms, definitions of each member of the minimal pair, example sentences with the words, and auditorily presented with the control items of the corresponding words. 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 the minimal pair (e.g., *VOORnaam*) on the top left corner of the screen and auditorily presented its corresponding control stimulus. After 1500 ms, we presented the WS member on the top right corner with its corresponding control stimulus. Below the orthographic depictions of the words, we visually presented their definitions and below that an example sentence. After presentation of the orthographic word forms and control items, the task was self-paced and all the visual stimuli remained on the screen for the entire trial. Before continuing to the next trial, participants had to indicate using button presses whether they did not know one (or both) of the words. That information could be taken into account in the data-analysis of the results.

Training phase

In the training phase, 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

8a for the trial structure). Furthermore, we visually presented an image of the talker producing that sentence 700 ms before sentence onset and visually presented two response options (i.e., the two members of the minimal pair; *VOORnaam* and *voorNAAM*) 200 ms before sentence onset. Both 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) from target word onset to which member of the minimal pair they had heard. If no response was given after 5s, the trial was recorded as a missing data point. After the response, we presented a feedback sentence (*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. Participants were presented with three repetitions of the target word in each trial. The next trial started after 1 s after the final auditory presentation of the target word. The task consisted of 192 experimental trials and was preceded by eight practice trials. The trials were presented in randomized order in 4 different counterbalanced lists and no word pairs were ever repeated in two consecutive trials.

Test phase

The test phase was identical to the training phase but differed on two aspects (see Figure 8b for the trial structure). First, the target words in the test phase were the mixed items instead of the control items. Note that we still included the control items in this phase on 50% of the trials to provide solid anchors of unambiguous items to participants. Second, participants did not receive feedback on their responses. That is, the next trial began 1 s after participants gave their response. The test phase consisted of 192 experimental trials and was not preceded by practice trials. The trials were presented in a pseudo-randomized order. That is, we ensured that a control and mixed items were not preceded by each other for the same word.

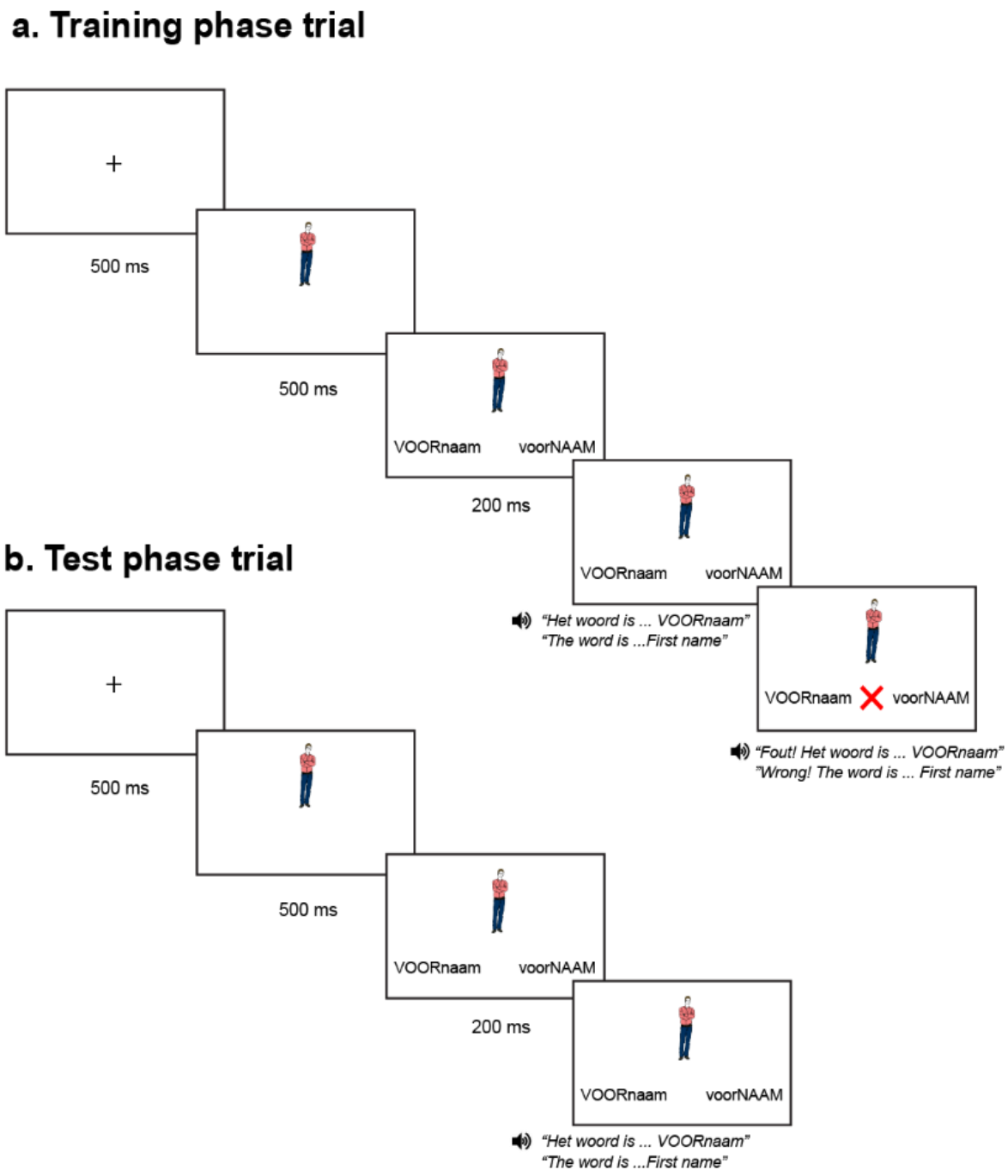


Figure 8. a. Trial structure of one trial in the training phase. b. idem, but for the test phase.

4.2.6 Data analysis

Prior to data analysis, we calculated the percentage of timed-out trials (0.7%) and we excluded any trials with RTs below 100 ms relative to target word onset (0.5% of the trials). The latter was done for two reasons. First, due to an error in the experiment, it was possible for participants to respond before target word onset (i.e., before they heard the word). Since the response then does not represent any perceptual processes related to the target words, we decided to exclude these items. Second, RTs below 100 ms also capture trials on which the majority of the first syllable had not yet been heard in its entirety (shortest first syllable duration was 171 ms). Secondly, we analyzed familiarization data in order to check whether participants knew all the word stimuli they were presented during the experiment. This analysis showed that the 93% of participants knew at least 14 of 16 words (41.3 % knew all words, 33.8 % knew 15/16 words and 18.8% knew 14/16 words) while only the 7% of participants knew less than 13/16 words.

The analyses of categorization responses were twofold. First, we analyzed the mixed items in the test phase as a measure of how perception was affected by the training 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, 2013). Second, we analyzed the control items from both the training and test phase to confirm that our participants correctly perceived the intended stress patterns in these items across both phases. Given the varying results on the familiarization task, we ran these analyses on the complete dataset and a dataset in which the unknown items were excluded.

For the mixed items, we wanted to test whether the talker-cue mapping (e.g., Talker 1 using F0, Talker 2 using intensity) affected responses on mixed items 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). This would entail testing a three-way interaction in the model between Mapping, Pattern, and Talker. However, 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, 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 uses F0 should perceive this item as SW. In

contrast, participants who had learned that Talker 1 uses intensity should perceive the exact same item as WS.

The 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 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 analyses were based on 7,554 observations.

For the control items, the model with the best fit to the data (tested using log-likelihood model comparisons) 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 for the mixed items. The analyses were based on 22,775 observations in total.

4.2.7 Power analysis

A power analysis was computed to estimate the minimal sample size required to reach a power of .80 by means of data simulation (Kumle et al., 2021). The final dataset of the experiment was simulated 1000 times each for different sample sizes (N=20,40,60,80) using the Mean and the Standard Deviation values for control and mixed items drawn from our pre-tests. The effect of Predicted Response was implemented as a small .05 difference in proportion of SW responses between the "Predicted SW" and the "Predicted WS" factor levels. For each simulation, the model including only "Predicted Response" as fixed effect and a by-participant random intercept was contrasted with a null model without the predictor including only a by-participant random intercept via a Chi-Square test and the p-value was retained. The proportion of significant tests represents the estimated power which corresponded to .858 (95% CI [.836 .858]) for the simulation with N=80.

4.3 Results

4.3.1 Mixed items

The analysis of mixed items crucially tested whether participants applied their learning about how the two talkers signalled lexical stress to perceive spoken words with conflicting stress cues. Results for Mixed items are summarized in Figure 9 (e, f, g). Qualitative plots showing the results for Mixed items divided by Talker, Pattern/Predicted Response and Mapping are depicted in Figure 9b. 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 9f, 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 (β = 1.42, SE = 0.27, z = 5.34, $p < .001$), showing that the F0-Intensity pattern (left bar in Figure 9e) was perceived as being more SW-biased compared to the Intensity-F0 pattern (right bar in Figure 9e). This demonstrates that participants weighed F0 as a cue to lexical stress more heavily than intensity, corroborating outcomes from pre-test 3 (see Supplementary Table S9). 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 with respect to 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 9e, while the predicted response Predicted SW was characterized by a negative tendency towards less SW-biased response (Mean Slope = -0.14, SE = 0.14; descending light red line in Figure 9e), the predicted response Predicted WS showed the opposite trend (Mean Slope = 0.22, SE = 0.14; ascending light blue line in Figure 9e).

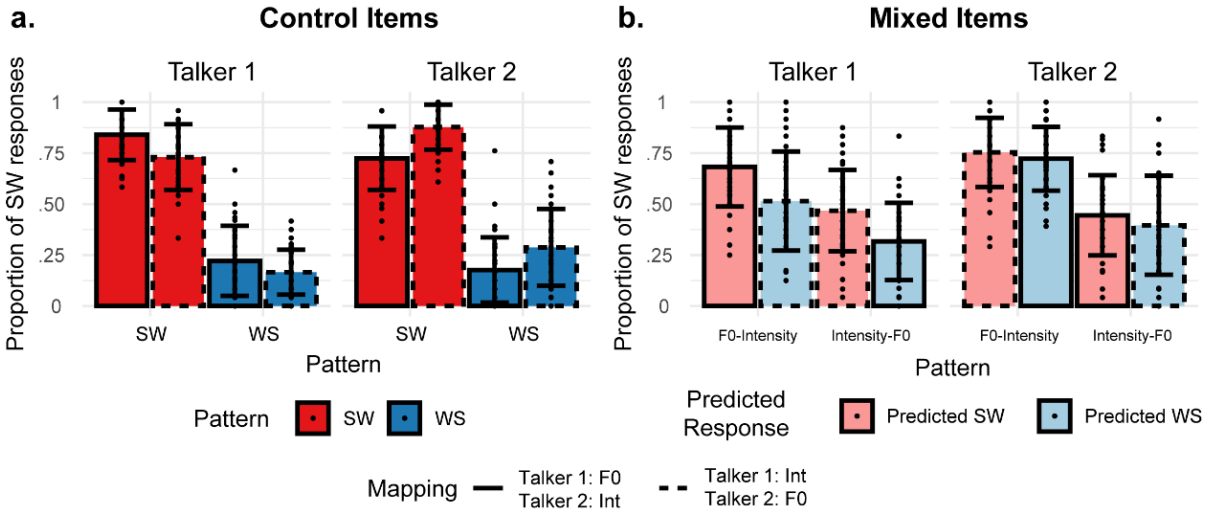
No further effect reached significance. The results of this analysis were replicated in total also when the trials including words that participants reported not to know prior to the experiment were excluded (see Table S15 in Appendix C for complete model outputs).

4.3.2 Control items

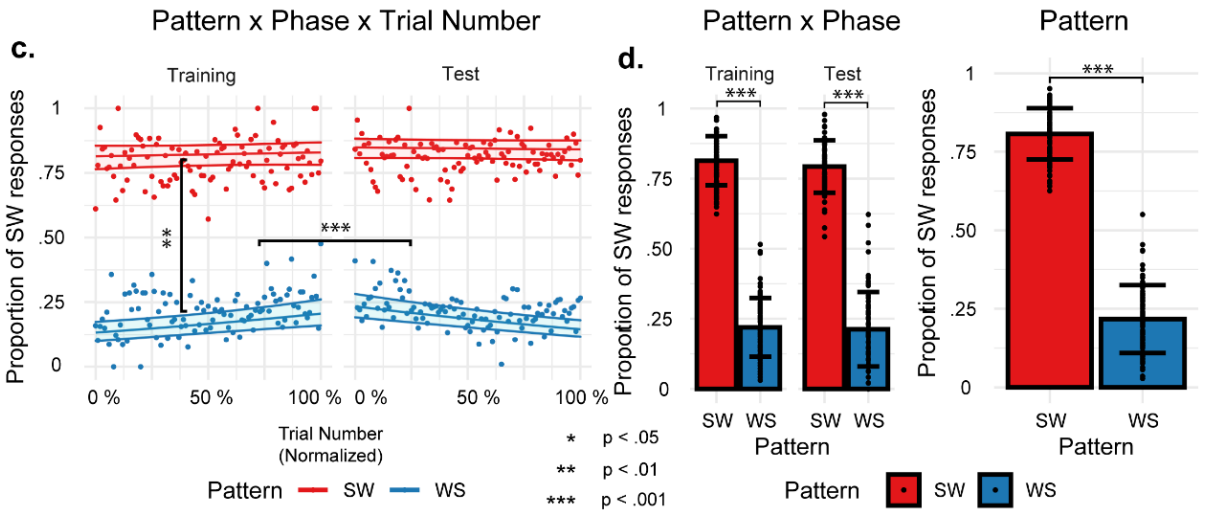
The analysis of control items tested whether participants could categorize words with one clear cue to stress with acceptable accuracy in both the training and test phase. Results for control items are summarized in Figure 9 (c, d). Qualitative plots showing the results for Control items divided by Talker, Pattern and Mapping are depicted in Figure 9a. The model revealed a significant effect of Pattern ($\beta = -3.14$, 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 9d) and WS (Mean prop. of SW resp. = .22, SE = .01; blue bar in the right plot of Figure 9d) patterns.

A small interaction effect between Pattern and Phase ($\beta = -0.44$, 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 9d). Post-hoc tests confirmed the presence of a strong difference between the SW and the WS pattern both in the training ($\beta = 3.18$, SE = 0.20, $z = 15.91$, $p < .001$) and in the test phase ($\beta = 3.15$, 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 ($\beta = -0.17$, SE = 0.06, $z = 2.64$, $p = .011$) and the WS pattern ($\beta = -0.13$, SE = 0.06, $z = 2.11$, $p = .035$). Another interaction effect between Phase and Trial Number ($\beta = -0.64$, SE = 0.13, $z = -4.95$, $p < .001$) was found as well as a main effect of Phase ($\beta = 0.47$, 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 ($\beta = -0.95$, SE = 0.26, $z = 3.68$, $p < .001$) represented in Figure 9c. 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 ($\beta = 0.52$, SE = 0.15, $z = 3.52$, $p < .001$) but the same test was only marginally significant during the Test phase ($\beta = -0.38$, 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 9c), the WS pattern showed a negative slope (Mean slope = -0.58, SE = 0.10; descending blue line in left plot of Figure 9c). This revealed that as the Trial Number 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 9c), 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 9c) 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 in total also when the trials including words that participants reported not to know prior to the experiment were excluded (see Tables S12, S13, S14 in Appendix C for complete model outputs and post-hoc tests).



Effects: Control Items



Effects: Mixed Items

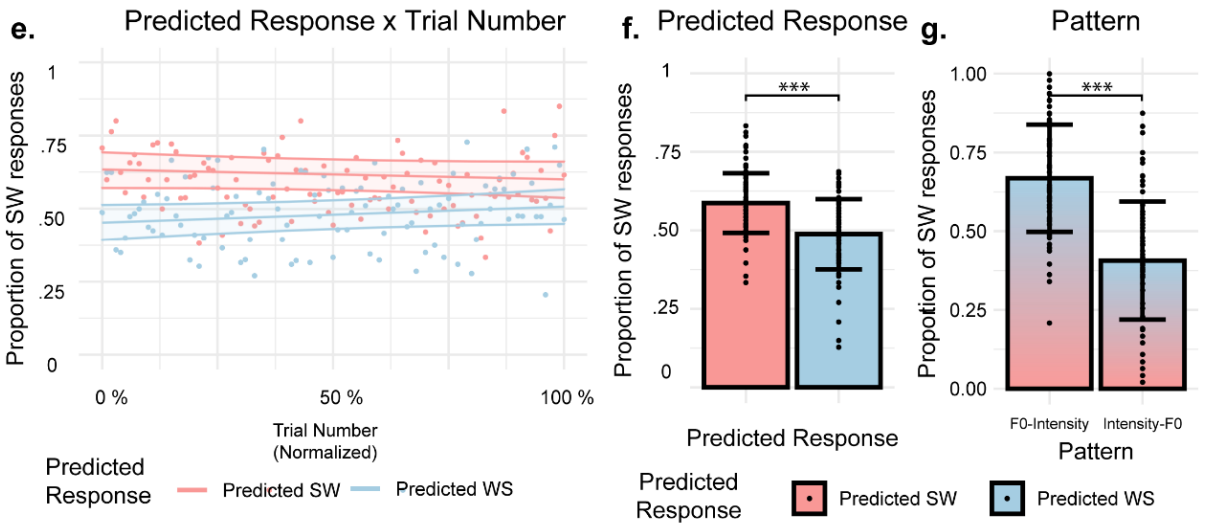


Figure 9. Qualitative plots (1st row) and plots of the effects (2nd row) of the proportion of SW responses (y-axis) split by different factors (x-axis). a. Qualitative plots for Control items averaged across participants, words and phases, separately for each Pattern (SW in red, WS in blue), Talker (1, 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).

4.4 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 specific talkers and used this information to differentially categorize words (i.e., mixed items) with conflicting stress cues depending on this talker-cue mapping. This was evidenced in our statistical model by a main effect of Predicted Response, showing that participants gave responses biased towards the stress pattern category following 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; X. 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 canonical stress patterns in Dutch where F0 and intensity co-occur as stress cues in (Rietveld & Heuven, 2009), pushed participants to recalibrate the perceptual weights of suprasegmental cues in a talker-contingent way. Specifically, as they 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 when categorizing minimal stress pairs. This interpretation is coherent with the dimension-based statistical learning account (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; R. Liu & Holt, 2015; X. Zhang & Holt, 2018) which states that listeners exploit short-term acoustic regularities to adjust the efficiency of specific physical dimensions in signaling speech categories and 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 particular experiment, this talker-contingent cue tracking may have been driven by acoustic talker differences (e.g., pronunciation idiosyncrasies) in the carrier sentences and/or target words themselves, the visual talker cues (different cartoon images), or both (see Zhang & Holt, 2018).

Following the dimension-based learning approach, listeners built talker-contingent weight sets based on the cue distributions picked up in the training phase, in which unambiguous words were presented. However, the learning process possibly continued throughout the whole

experiment, as illustrated by the results in the test phase. Even though the interaction between Predicted Response and Trial was only marginally significant, the difference between Predicted SW and the Predicted WS responses seemed to be gradually attenuated as the test phase went on (see Figure 9e). It is possible that the presence of mixed items weakly altered the talker-specific cue-distribution as they provided two conflicting cues to stress. Previous studies showed 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 of these studies showed that providing new talker-specific information at test, which may have been fully or partially incompatible with the one presented during the training phase, reactivated the learning process itself, inducing listeners to update perceptual weights.

It is worth pointing out that minimal stress pairs are rare in Dutch, and sentential contexts in which both members of a pair are equally semantically and syntactically acceptable are even rarer. This contrasts with languages such as Spanish, where there are many minimal stress pairs with shared semantics and syntax (e.g., *CANto* 'I sing' vs. *canTÓ* 'she sang'). Therefore, it could be considered surprising that Dutch listeners track talker-specific cues to lexical stress at all, as it may not provide a major advantage in Dutch spoken word recognition. However, the fact that we do find evidence for talker-specific tracking of suprasegmental cues to lexical stress in Dutch suggests that it is a relatively robust perceptual mechanism, which can likely also be found in other languages with free-stress such German and Spanish, presumably independent of the number of minimal stress pairs in a given language. This emphasizes the central role of lexical stress in spoken word recognition as well of perceptual learning in the perception of lexical stress.

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 either 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 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 an optimal word recognition level.

Note that the talker-specific perceptual learning mechanism in the present study might have occurred regardless of a need imposed by the experimental task. That is, the present study did not employ particularly ambiguous items in training that, unlike in classical perceptual learning paradigm, guide recalibration. To this regard, the Ideal Adapter Framework (Kleinschmidt & Jaeger, 2015) also predicts that listeners have beliefs about the amount of variability across different situations, and consequently between talkers. This prediction about a talker-specific uncertainty in cue distributions might suffice to listeners as an "internalized" need to adapt. Crucially, these class of models that postulate a belief-updating feature were developed to explain results from studies capitalizing on segmental information, and no model was yet developed to explain how listeners might deal with variability at the suprasegmental level. Our results might indicate that these models are appropriate to also explain how listeners perceive suprasegmental information and deal with variability in a talker-specific way.

The present study goes beyond the findings of another similar study with non-word stimuli (Severijnen et al., 2021), suggesting that talker-specific learning of cues to lexical stress also applies to existing words. This is an important aspect to highlight as the results of the present study show that talker-specific learning is a mechanism that can exploit short-term regularities and supersede long-term information about previously known words. In addition, while in Severijnen et al (2021) word recognition was indirectly assessed with RTs, that when delayed may suggest a processing difficulty induced by the presentation of an unexpected acoustic word form, in the present study spoken word recognition was directly assessed by studying word categorization. Moreover, in the present study we employed richer multidimensional test stimuli with two different cues with respect to Severijnen et al (2021).

It would be interesting to understand whether talker-specific perceptual learning of suprasegmental cues generalized to previously unheard words (i.e., test words which are not included in the training stimuli) as seen in previous work on segmental information (Eisner & McQueen, 2005). This kind of generalization is considered as an index of a pre-lexical abstraction process by which the perceptual weights can be adjusted based on exposure to ambiguous words and then used to recognize new words (Cutler et al., 2010) and is incompatible with the episodic accounts of word recognition which postulate that listeners store 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 did not test for generalization to new words in detail as the same lexical items were used for training and test phases but testing this possibility might be of interest for future studies. Nevertheless, our results do provide some indications of generalization of the learning process across word episodes, as mixed items were not encountered during training. To this regard, it is important to recall the physical differences between 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 with respect to control items. Second, while control items had one clear cue to stress (e.g., Intensity or Pitch) and two other cues put to ambiguous levels (e.g., Pitch 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 and possibly only the Pattern effect would have emerged, as it can be strongly linked to the physical characteristics of mixed items as shown in our pre-test (see Table S11 and Figure S11 in Supplementary Information).

In sum, we showed that listeners can learn how two specific talkers signal lexical stress and apply that learning in recognizing subsequent tokens from the same talkers. 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 shows that they might also account for the way by which listeners deal with suprasegmental variability.

Chapter 5: General Discussion

The aim of the present dissertation was to study three different cognitive mechanisms that are fundamental for listeners to benefit from the integration of linguistic and talker-identity information or to deal with their intrinsic variability. The investigation of how such mechanisms work could be crucial to understand how listeners can correctly perceive speech while retaining both linguistic and talker-identity information as well as to clarify how these kinds of information interact.

Previous studies on the abstraction mechanism in the domain of speech perception provided evidence that the listeners need active attentional effort in order to extract either linguistic or talker-identity information from the speech signal (Bonte et al., 2009, 2014; Kaganovich et al., 2006; Mullennix & Pisoni, 1990; Zhang et al., 2016). Instead, other studies focused on the ability of the cognitive system to automatically extract linguistic information, while discharging listeners from dealing with (i.e., talker-related) non-linguistic information (Eulitz & Lahiri, 2004; Jacobsen, Schröger, & Alter, 2004; Jacobsen, Schröger, & Sussman, 2004; Shestakova et al., 2002). Following this last series of studies, the EEG and behavioural experiment reported in Chapter 2 investigated whether listeners could form representations of the talker's voice which remain invariant with respect to phonological information.

First, EEG results of the passive task showed that listeners can pre-attentively form abstract representations of the talker's voice regardless of constantly changing phonemes, as suggested by the elicitation of the MMN. Extending the results of previous studies on abstract (i.e., talker-invariant) phoneme representations (Jacobsen, Schröger, & Alter, 2004; Jacobsen, Schröger, & Sussman, 2004), this result shows that the abstraction mechanism can also be used to abstract from phonological information while retaining the talker's voice.

Second, this mechanism appeared to operate even when the formant structure encoding phonological information was disrupted, as suggested by the elicitation of the MMN with rotated speech. This suggests that the presence of meaningful phonological information is not required for the abstraction mechanism to work, thus indicating that this mechanism is not speech-specific, as also suggested by similar results obtained with complex tones (Huotilainen et al., 1993; Pakarinen et al., 2010).

Third, as indicated by enhanced behavioural performance and larger P3b in the speech condition of the active task, while at a pre-attentive level listeners appear to form abstract representations of voice irrespectively of the presence of meaningful phonological information, at an attentive level the presence of phonological information facilitated the detection of changes occurring in pitch, the primary dimension indexing talker-identity (Baumann & Belin, 2010). As indicated by enhanced power in the theta band for the speech condition, phonological information might have facilitated this task by providing access to pre-existing prototypical voice representations, which allowed listeners to encode talker-related information more efficiently, reducing the amount of cognitive resources needed.

Similarly to abstraction, memory retrieval seems to be automatically deployed during speech perception. Indeed, the two mechanisms seem to unfold within a similar time-window. However, in contrast to abstraction which is portrayed as a general-domain ability (in Chapter 2), memory retrieval shows a certain degree of functional segregation depending on what kind of information is retrieved. Previous studies on the automatic retrieval of memory traces for phonemes and voices showed that familiar voices and native phonemes could be retrieved by means of shared processes when phonemes and voices were naturally acquired in a native linguistic environment (Beauchemin et al., 2006; Dehaene-Lambertz, 1997). The longitudinal EEG and behavioural study in Chapter 3 investigated whether newly learned phonemes and voices acquired in a non-native linguistic environment are retrieved by shared or segregated processes.

First, as suggested by the different amplitude modulation of the MMN, the automatic retrieval process appears to be functionally dissociated for newly learned phonemes vs. voices, similarly as it also occurs when these two kinds of information are actively extracted from the speech signal with the support of selective attention (Schall et al., 2015). Second, the study successfully replicated EEG and behavioural findings of previous experiments that employed the listen-and-repeat paradigm to teach non-native phonological contrasts (Saloranta et al., 2020; Tamminen et al., 2015; Ylinen et al., 2010), corroborating its effectiveness.

In Chapter 4, we studied perceptual learning from the behavioural point of view. This mechanism is rather complex and may depend on the synergic activation of more simple processes such as the ones investigated in previous Chapters 2 and 3. Previous studies on perceptual learning showed that listeners can learn how different talkers produce speech sounds. The majority of these studies focused on segmental information (Eisner & McQueen, 2005; Idemaru & Holt, 2011, 2014;

Kraljic & Samuel, 2007), but few studies also investigated perceptual learning of suprasegmental structures (Bosker, 2021; Xie et al., 2021). The behavioural study in Chapter 4 investigated whether listeners could learn how different talkers mark words lexical stress.

First, results showed that listeners can indeed learn how different talkers mark lexical stress also when hearing existing words. As suggested by the Ideal Adapter Framework (Kleinschmidt & Jaeger, 2015), by exploiting the short-term distributions of talker-specific cues to stress, listeners could adjust the perceptual weights associated to those cues and use the updated weights for word categorization.

Second, the learning process started in the training phase, when participants were exposed to words with clear cues to stress, but continued through the test phase, which included words with ambiguous cues. This highlights the flexibility of the cognitive system in re-adapting perceptual weights as soon as the distributions of cues are altered with respect to the a-priori distributions built during lifelong exposure to language.

5.1 Temporal and computational features of the three mechanisms

The results of the three empirical studies described in Chapters 2, 3 and 4 provide important insights about the temporal and computational features of abstraction, memory retrieval and talker-specific perceptual learning, which may be informative for models of speech perception in which talker-related information is relevant. While we studied the three different mechanisms in isolation, some considerations can be made about their interplay, as all of them subserve similar purposes but with different functional specificities and domains.

Abstraction is specifically useful for listeners when dealing with physical variability: Its deployment results in forming a representation based on one main physical dimension that remains invariant with respect to changes occurring in other non-relevant dimensions. The study presented in Chapter 2 extends the domain of this mechanism: Listeners can automatically form phoneme-invariant voice representations as they can form talker-invariant phoneme representations (Eulitz & Lahiri, 2004; Jacobsen, Schröger, & Alter, 2004; Jacobsen, Schröger, & Sussman, 2004; Shestakova et al., 2002). Hence, this result shows that listeners are equipped to face the lack of invariance problem (Liberman et al., 1967) also when they must identify the talker while hearing variable linguistic structures.

In line with our results, the automaticity of such processes was highlighted in a relevant fMRI study (Formisano et al., 2008). In this study, different vowels uttered by different speakers were passively presented to listeners while neurophysiological activity (i.e., BOLD signal) was recorded. By employing machine learning algorithms, authors managed to isolate the neural regions that contributed the most to a phoneme-wise classification of neural activity, independently of the talkers' voices, but also to a talker-wise classification, independently of phonemes. Results showed that primary auditory cortices, which are associated with the early stages of speech perception (Kemmerer, 2015), were involved in the vowel-wise (i.e., phoneme-invariant) or in the talker-wise (i.e., talker-invariant) classification of stimuli. This study shows that clear patterns in neural signals indicating the formation of abstract phoneme and voice representations can be recorded also in absence of task demands and can be attributed to specific brain regions which should be involved in the early stages of sound perception. In Chapter 2, thanks to the high temporal sensitivity of EEG, we gathered direct evidence about the early occurrence of such processes also for the talker dimension.

Moreover, the study in Chapter 2 also indicated that the formation of abstract voice representations could occur in absence of meaningful phonological information (i.e., with rotated speech) suggesting that this mechanism works in a similar way across different auditory domains. In line with our results, Huotilainen et al. (1993) and Pakarinen et al. (2010) showed that listeners can automatically track sudden changes related to different and co-occurring physical dimensions at the same time, even in the case of complex tones (i.e., non-speech sounds). Specifically, Pakarinen et al., (2010) reliably recorded the MMN by implementing a *no-standard oddball paradigm*. In this oddball version, only deviant events are presented, as every stimulus is characterized by a variation in one specific physical dimension (e.g., intensity, duration, noise level). The elicitation of the MMN to each of the deviant types indicates that the stimuli were constantly grouped and re-grouped together based on one specific dimension (e.g., intensity) despite continuous changes occurring in other dimensions (e.g., duration). Additionally, this occurred for different dimensions within the same experiment, suggesting that listeners can form abstract representations of multiple sound features at the same time.

This interpretation might lead the discussion towards two odd conclusions. If listeners can automatically take care of physical variability by forming abstract representations of different sound features very early in time, variability should not impact their behavioural performance

during phoneme categorization or talker identification tasks. Further, if listeners can isolate and track different sound features at the same time via bottom-up processes that depend on physical features of auditory stimulation, there could be no need for prior knowledge to provide categories in which information can be encoded to support such processes.

Clearly, the first conclusion is in sharp contrast with an extensive literature showing how the extraction of linguistic and talker-related information interact at a behavioural level with positive (Bregman & Creel, 2014; Perrachione & Wong, 2007; Zarate et al., 2015) or negative (Kaganovich et al., 2006; Mullennix & Pisoni, 1990; C. Zhang et al., 2016) outcomes in terms of performance. Relatedly, in Chapter 2 we also show that when listeners are actively monitoring sound streams looking for changes in the talker's voice, their behavioural performance is enhanced in the presence of meaningful phonological information with respect to when such information is disrupted in rotated-speech stimuli. This suggests that while abstraction may automatically occur early in time providing useful information to listeners, the active use of such information still depends on later controlled processes in which linguistic and talker-related information appear to interact.

With respect to the second conclusion, in the discussion section of Chapter 2, we suggest that abstraction per se might relate to a primitive ability that represents the statistical distribution of different sound features within auditory streams (Batterink & Paller, 2019; Carral et al., 2005; Paavilainen, 2013; Saarinen et al., 1992). While prior experience may have limited influence on this low-level and general-domain process, it may show its influence when the extracted information must be linked to response categories as for phonemes and voices. To this regard, when listeners were actively asked to detect voice changes while ignoring phonological information (Chapter 2), we showed enhanced behavioural performances and P3b amplitude in the speech condition with respect to when phonological information was disrupted (i.e., rotated speech).

With respect to the timeline of the integration of linguistic and talker-identity information, in the context of our study, their interaction starts to improve performance from ~300 ms from target onset onwards. The analysis of oscillatory activity suggested that this improvement might be due to better encoding strategies for speech, promoted by pre-existing voice and/or phoneme representations. Going back to the second conclusion, the influence of top-down processes driven

by prior knowledge is not absent but simply surfaces later in time and, more importantly, under the control of the listener.

In Chapter 3, the influence of prior knowledge (instantiated via controlled training procedures) in supporting the extraction of phonological and talker-related information is observed earlier with respect to the results highlighted in Chapter 2, in a time window contingent to the one in which abstraction takes place. Such inconsistency might relate both to the complexity and the range of operations that the cognitive system was implementing in the two tasks, or to temporal differences in the activation of automatic vs volitional memory retrieval processes. In fact, in Chapter 2, to possibly access pre-existing voice representations in the active oddball task, listeners had to detect abstract regularities in the audio stream first, an operation that takes a certain amount of time. It is reasonable to assume that the cognitive system has computational limitations on the number of automatic processes operating on audio streams that can occur in parallel before a certain point in time without consequences on performance (Muller et al., 2005; Nager et al., 2003). These temporal limitations of automatic processes may consequently call for the intervention of selective attention mechanisms, and related cognitive resources, to orient listeners towards task-relevant aspects of speech (Snyder & Alain, 2007; Sussman et al., 2014). The attentional process would thus take over the automatic mechanisms that are not completed before its activation. Conversely, when the cognitive system is simply presented with a sound as an instance of either a familiar talker voice or a native phoneme (see Chapter 3), it may be able to react faster and retrieve the representation automatically without the need to engage other mechanisms before memory retrieval.

Moreover, given that both kinds of information are indexed by physical dimensions which occasionally overlap, the quick availability of representations highlighted in Chapter 3 would allow sound features and, most importantly, physical variability to be reconducted to specific sources. This possibility is considered crucial for the models explaining the *language familiarity effect*. Perrachione, (2017), sketched the *phonetic familiarity hypothesis*, which states that talker identification is improved when listeners hear their native language because of the familiarity with the distributions of native phonetic features, which can finally be attributed to meaningful phonemic contrasts. Certainly, Italian participants hearing German talkers and phonemes employed in the study of Chapter 3 did not have detailed knowledge about the distribution of phonetic features of German language. Yet, exposure to a non-native language can enhance the

language familiarity effect (Orena et al., 2015). While we do not directly test this hypothesis, we cannot exclude that hearing small texts and words in the experiment could have facilitated the establishment of talker-identity representations of non-native speakers. Lastly, considering how early the cognitive system analyses phonetic aspects of speech (Pereira et al., 2018; Shahin et al., 2018), our result shows that already at 200 ms from stimulus onset listeners can associate acoustic features to representations of both talkers and phonemes. If listeners can identify which part of the signal are to be associated to phonemic properties, they could restrict the amount of variability in the signal that has to be attributed to talker-related information and vice-versa.

Among the mechanisms studied within this dissertation, whereas Chapter 3 might highlight a potential first step for the integration of linguistic and talker-related information, the talker-specific perceptual learning described in Chapter 4 represents the most evident connection between the two types of information. In fact, the aim of this mechanism is to gather information about how specific talkers produce specific linguistic structures which can be used in subsequent encounters with such talkers. The functional core of this mechanism is that listeners compute a talker-specific set of weights by which they can adjust their perception to correctly interpret acoustic cues during speech perception. However, we should not frame perceptual learning as a single mechanism, but rather consider it as a process in which different cognitive mechanisms work together to reach one common goal. Abstraction and memory retrieval might have a significant role in this workflow.

Considering abstraction, several studies showed that perceptual learning generalizes to previously unheard words (Bosker, 2021; Cutler et al., 2010; Eisner & McQueen, 2005). Note that in Chapter 4, we do not provide indication of a generalization of talker-specific cues across the lexicon, but only indications of generalization across word episodes given the physical differences between training and test stimuli. Nevertheless, generalizing talker-specific knowledge about production idiosyncrasies across the lexicon means that if listeners can learn the way by which one talker produces, e.g., the /s/ phoneme by hearing specific words, they can apply what they learned to new word tokens. Hence, perceptual learning by definition would require abstract representations of linguistic structures (Cutler et al., 2010).

Secondly, since perceptual learning is talker specific as it does not generalize across talkers (Eisner & McQueen, 2005), listeners need first to identify the talker to learn about its production idiosyncrasies and to apply what they learned in the future. Given that the talker, across different encounters, has to be identified by hearing different words, it is reasonable to assume that

perceptual learning also hinges upon abstract talker's voice representations. Considering that also face stimuli can trigger talker-specific perceptual learning (Zhang & Holt, 2018), listeners might use different multimodal cues to understand who is speaking. Thus, the proposal that listeners need abstract voice representations only applies to situations in which the voice is the only available cue to talker identity.

Additionally, both kinds of abstract representations must naturally be retrieved from memory, as well as the perceptual mapping between phonetic cues and representations. Given that during normal conversation the talker-cue mappings must be retrieved to recognize words, but also updated on the basis of new phonetic evidence, a fast and automatic memory retrieval mechanism would provide considerable support to talker-specific learning.

5.2 Theoretical implications for models of speech perception

The implications discussed in this section could provide initial insights for a model of speech perception that addresses the uptake of both linguistic and talker-related information as well as their interaction. The models of speech perception in which the talker information is used can be positioned on a continuum ranging from abstractionist accounts (Gaskell & Marslen-Wilson, 1998; McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008) – in which talker information is used to normalize the speech signal to access abstract phonological representations but is then discarded – to episodic models – in which listeners retain fine-grained phonetic information without needing additional signal transformations such as normalization or abstraction (Goldinger, 1996, 1998; K. A. Johnson & Mullennix, 1997; Pierrehumbert, 2001).

Normalization is often considered as a basic process by which listeners can deal with between-talker variability. By stripping away talker-related information from the signal (Pisoni, 1992; Sjerps et al., 2011a, 2019), listeners can access linguistic information encoded in abstract representations. The results we provide in Chapter 2 are not completely in line with the way in which abstractionist models would envisage normalization. In fact, these results suggest that talker-related information is not stripped away from the signal. Instead, listeners appear to automatically form an abstract representation of the talker's voice equally as they do with phonemes (Formisano et al., 2008; Jacobsen, Schröger, & Sussman, 2004). One way to reconcile our results with the notion of normalization in abstractionist accounts is that listeners may use talker normalization to enhance the access to linguistic information, but possibly retain abstract

talker-related information which can be used for other purposes such as talker-identification (Lavan et al., 2019).

These results would not be compatible with episodic models of speech perception (Goldinger, 1998) given the compelling evidence of the formation of abstract voice representations. Nonetheless, it is important to note that such results cannot account for the absence of episodic memory traces which are the information unit of this theoretical framework. Although, the results presented in Chapter 2 clearly advocate for the retainment of talker-related information, which is a fundamental feature of episodic accounts, in contrast with the abstractionist ones which advocate for its removal from the speech signal.

Additionally, the importance of talker-identity information is highlighted by the results of Chapter 3, in which familiarizing with the talker allowed listeners to automatically retrieve a talker-identity representation. This feature could be useful to explain the talker-familiarity effect (Johnsrude et al., 2013; Nygaard et al., 1994; Nygaard & Pisoni, 1998), which further stresses the importance of talker-identity for word recognition. As stated with respect to the relationship between normalization and retainment of talker-related information, talker-familiarity and normalization do not need to be mutually exclusive, but can coexist. Relatedly, Magnuson et al. (2021) showed that despite talker-familiarity provided a significant advantage for word processing in noisy conditions, it could not alter the processing cost attributed to normalization, which was triggered by changes in the talker-identity. To this regard, in Chapter 3 we show that memory retrieval processes for phonemes and voice appear as functionally dissociated.

It would be difficult to interpret our findings using episodic models of speech perception, as the notion of abstraction is in sharp contrast with the idea that listeners only retain fine-grained episodic instances of speech tokens. We could not exclude that some kind of episodic components may have contributed to the results of the study in Chapter 3, as we provided no evidence for the abstractness of the retrieved memory traces for phonemes and voices. In fact, we are not endorsing a perspective in which episodic information is impossible to grasp or useless. With respect to this aspect, the Ideal Adapter Framework (Kleinschmidt & Jaeger, 2015) was developed to integrate abstract and episodic information and authors state that "*some balance between complete abstraction and complete lack of abstraction is optimal"*. The core assumption of this framework is that listeners benefit from having a set of abstract linguistic categories, but also from being able to flexibly adapt such categories to new phonetic cue distributions, which often relate to the

idiosyncratic way by which specific talkers — or group of similarly accented talkers — produce speech. This framework is particularly fitting with the results of the study in Chapter 4, in which we report evidence for talker-specific learning of cues to lexical stress in Dutch listeners. In the General Discussion we previously argued that talker-specific perceptual learning might crucially rely on abstraction and memory retrieval. While theoretically fitting with the framework, direct evidence about an involvement of such processes in perceptual learning is still lacking.

5.3 Future perspectives and conclusion

Taken together, the results of all the three studies presented in this dissertation suggest that future models of speech perception should put substantial efforts in describing how linguistic and talker-specific information interact by providing precise computational and temporal specifications. First, we would like to stress the importance of establishing what kind of automatic processes are activated, considering both linguistic as well as talker-specific information as their target. It is likely that listeners often do pay attention to what is being said or to who is speaking. Critically though, listener cannot choose not to understand what is being said or to ignore who is speaking (Mullennix & Pisoni, 1990). Hence, the investigation of automatic processes might provide strong computational constraints for the development of new models of speech perception.

Second, once this feature is achieved, determining how listeners make an active and attentive use of the information which is automatically grasped by the cognitive system could elucidate the conditions by which the interaction between linguistic and talker-related information have positive or negative consequences for an accurate speech perception. While listeners can rely on several automatic mechanisms to deal with variability and to recognize words or talkers, their behaviour is often dictated by specific environmental requests. For instance, if listeners are requested to extract linguistic information, they surely need talker-related information as outlined in this dissertation but not *all* the information about the talker (e.g., social or biological characteristics). Therefore, they must orient their attention and deploy their cognitive resources in an optimal way, accordingly.

Future studies could investigate this specific issue and characterize the role of selective attention in guiding volitional processes during speech perception. In fact, the results presented in this dissertation would not allow to draw strong conclusions on attentive processes, and only provide initial insights about their involvement in the perception of voice changes in presence vs

absence of linguistic information. In particular, the study presented in Chapter 2 could be extended with a follow-up experiment in which participants perform a dichotic listening task with the exact same stimuli. In the new experiment, deviant stimuli would be presented either only at the right or left hear and participants would be asked to pay attention only to the stimulus delivered to one ear, while ignoring the other. In this way, participants would perform an active task where attention is engaged and automatically allocated to deviants across both streams of stimuli, but selective attention is directed only to one specific channel. This could allow to make a first differentiation between the automatic and volitional involvement of attention in the task of interest.

In conclusion, the experimental works presented in this dissertation highlight specific features of three cognitive mechanisms that allow listeners to benefit from the integration of linguistic and talker-identity information and to deal with their intrinsic variability. First, we have shown that listeners can form abstract representations of the talker's voice which are invariant to changes pertaining phonological information. Further, we have suggested that the abstraction mechanism works similarly across different auditory domains. Second, we have shown that the automatic memory retrieval processes for linguistic and talker-identity information appear as functionally dissociated. Third, we provided evidence for the ability of listeners to learn how specific talkers produce specific linguistic structures by adjusting their perceptual weights associated to particular acoustic cues. Finally, we have outlined specific temporal and computational features of the studied mechanisms which could be informative for the development of future models of speech perception that will adequately address the relationship between linguistic and talker-related information.

APPENDIX A

(Chapter 2)

APPENDIX A

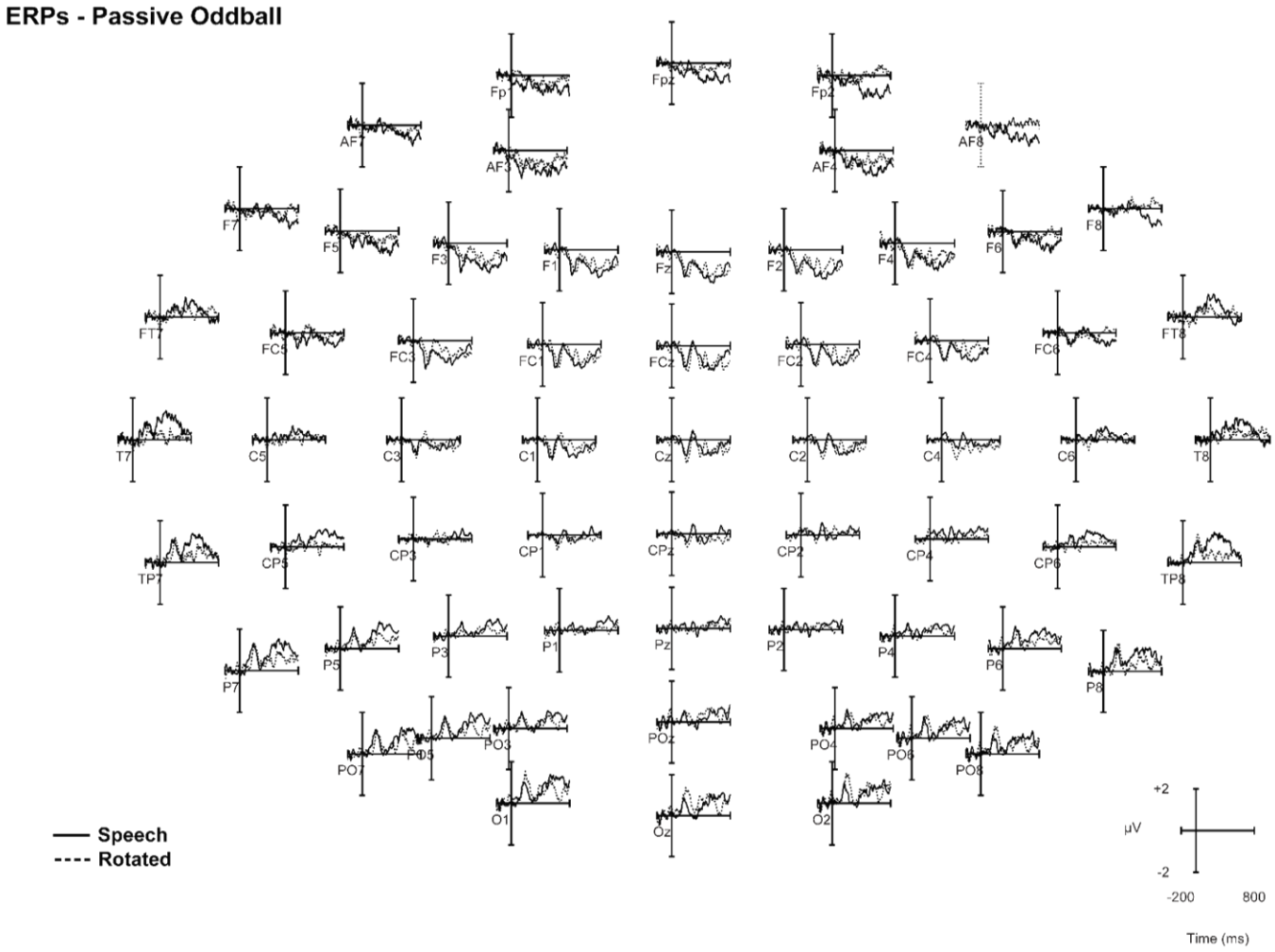


Figure S1. Differential waveforms of the ERPs for the passive oddball task for the speech (solid line) and the rotated speech (dashed line) calculated by subtracting deviant ERPs from control ERPs across all the electrode sites.

APPENDIX A

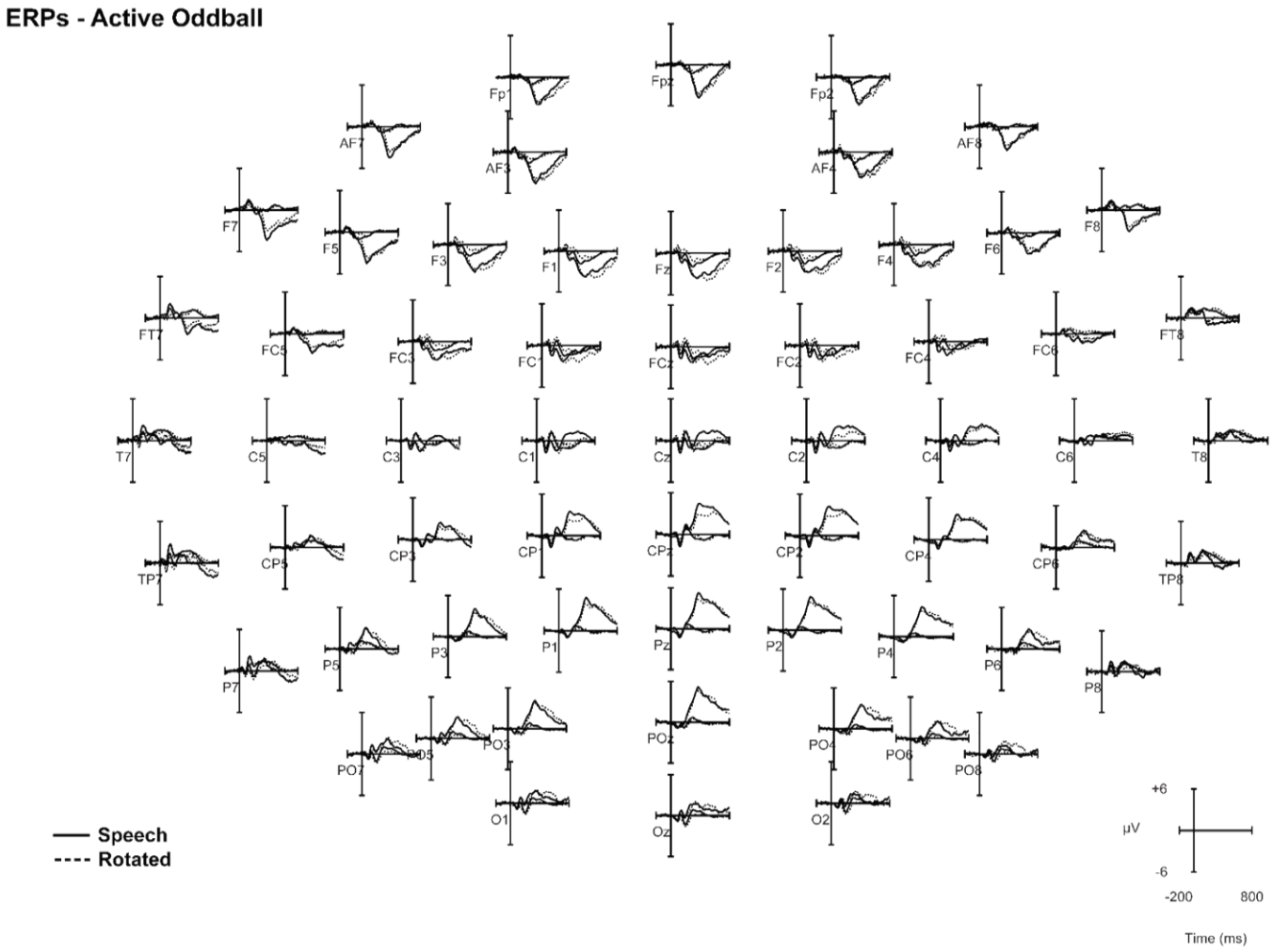


Figure S2. Standard and deviant ERPs for the active oddball task for the speech (solid line) and the rotated speech (dashed line) across all the electrode sites.

APPENDIX A

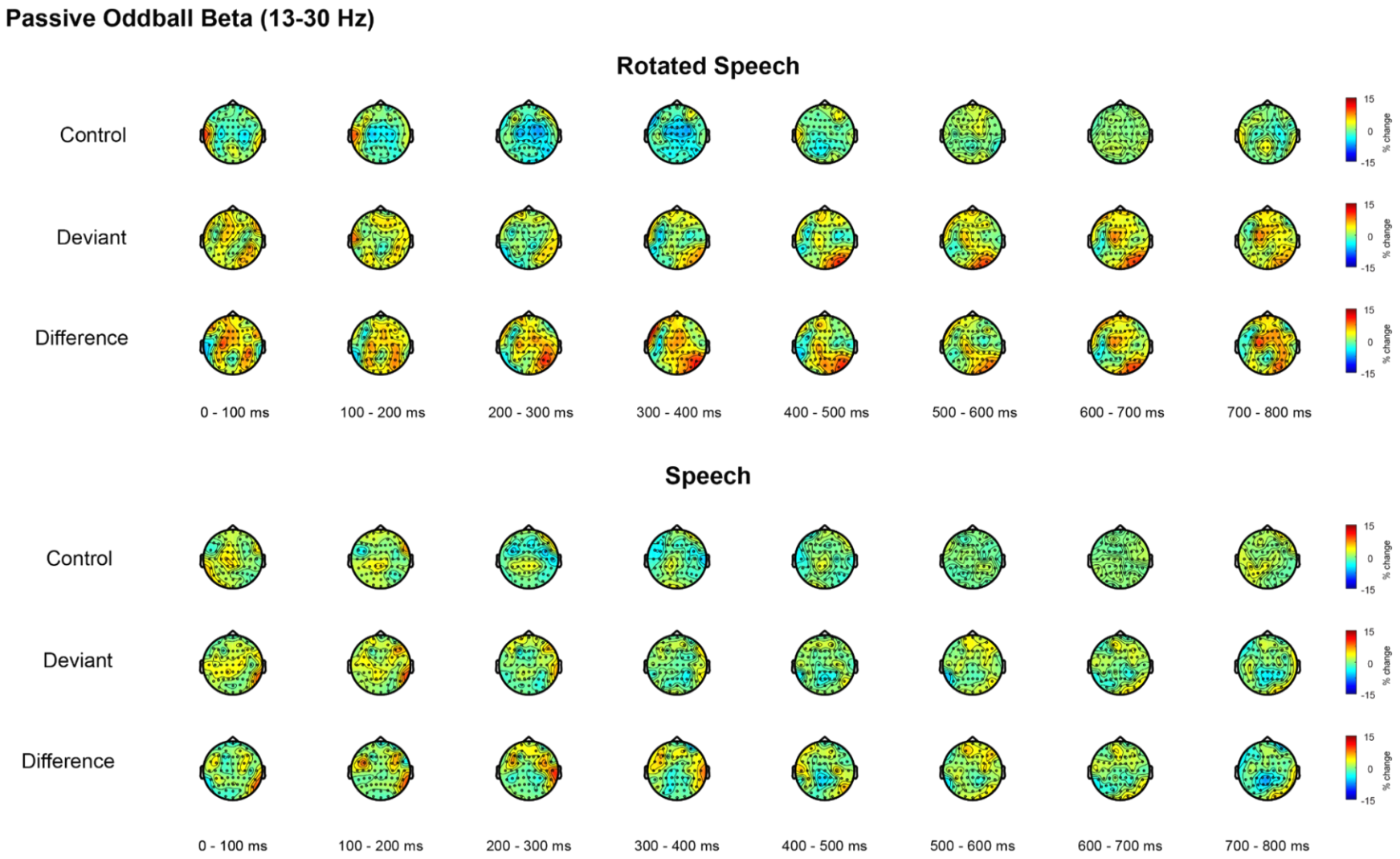


Figure S3. Topographies show the spatial distribution of beta ERDs/ERSs averaged in contiguous 100 ms time windows, characterizing the ERSPs for Control and Deviant events (1st,2nd,4th,5th row) as well as the difference between them (3rd, 6th row) for the Rotated Speech (1st-3rd row) and the Speech condition (4th-6th row) for the passive oddball task.

APPENDIX A

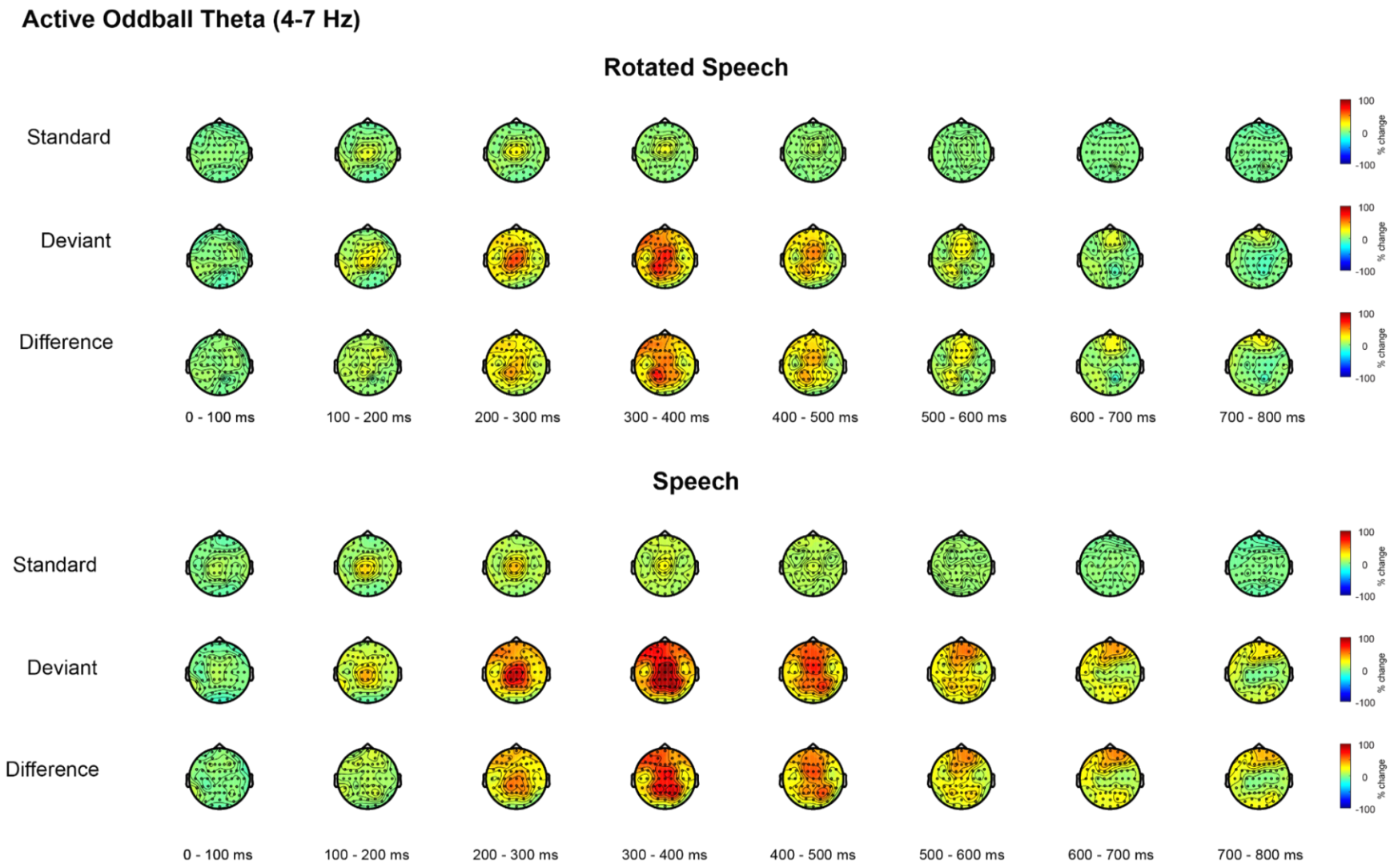


Figure S4. Topographies show the spatial distribution of theta ERDs/ERSs averaged in contiguous 100 ms time windows, characterizing the ERSPs for Standard and Deviant events (1st,2nd,4th,5th row) as well as the difference between them (3rd, 6th row) for the Rotated Speech (1st-3rd row) and the Speech condition (4th-6th row) for the active oddball task.

APPENDIX A

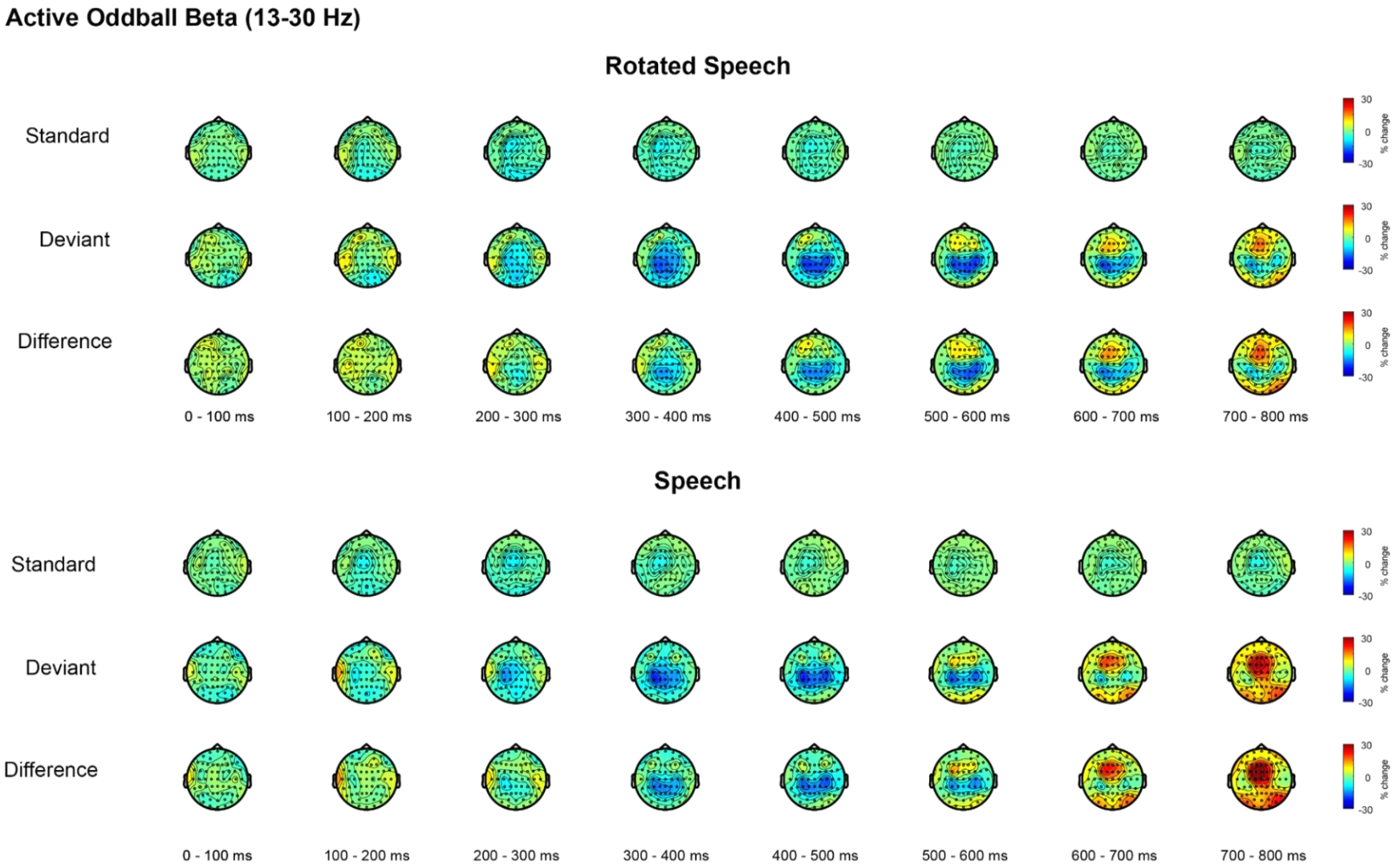


Figure S5 Topographies show the spatial distribution of beta ERDs/ERSs averaged in contiguous 100 ms time windows, characterizing the ERSPs for Standard and Deviant events (1st,2nd,4th,5th row) as well as the difference between them (3rd, 6th row) for the Rotated Speech (1st-3rd row) and the Speech condition (4th-6th row) for the active oddball task.

APPENDIX B

(Chapter 3)

APPENDIX B

Table S1. The table shows data of participants' age (years), sex (F = Female, M = Male), years of education, hours of language use and self-reported level of proficiency for L2 averaged across written and oral skills (1 = *really low*; 10 = *really high*). Standard deviations are in brackets.

	Groups		
	Talker-identification training	Syllable-identification training	Whole sample
Age	21.53 (2.69)	22.53 (2.06)	22.03 (2.41)
Sex	F = 13; M = 2	F = 13; M = 2	F = 26; M = 4
Years of Education	15.53 (2.47)	15.73 (1.98)	15.66 (2.20)
Daily use of L2 (hours)	4.63 (3.17)	3 (2.95)	3.89 (3.28)
L2 proficiency	6.85 (0.69)	7 (1.26)	6.92 (0.95)

APPENDIX B

Table S2. Duration values in milliseconds of the recorded syllables from talkers 1 and 4. Syllables were selected on the base of duration similarity. The selected syllables were then manipulated and used for the EEG experiment.

Syllable	Talker	Token	Duration	Selected
/pi:/	4	1	290 ms	Yes
		2	323 ms	No
		3	272 ms	No
/pi:/	1	1	283 ms	Yes
		2	455 ms	No
		3	425 ms	No
/py:/	4	1	293 ms	Yes
		2	327 ms	No
		3	275 ms	No

APPENDIX B

Table S3. Word stimuli in the talker-identification training

German	English
wann	when
Wahn	delusion
Seele	soul
Säle	halls
Bett	bed
Beet	vegetable patch
Mitte	centre
Miete	rent
Hülle	cover
Hölle	hell
losen	to draw lots
lösen	to solve
Nuss	hazelnut
nass	wet
jener	that (m)
jene	that (n)
Öhr	needle's eye
Ur	aurochs
(Ich) bäte	(I) prayed
beten	to pray
Bete	beetroot
Lamm	lamb
lahm	lame

APPENDIX B

Table S4. The mean MMN amplitude values calculated by group, by session by condition and by channel in the 215-255 ms time window for the voice-change condition and in the 199-239 ms time window for the phoneme-change condition. Standard deviations are in brackets. Asterisks show the level of significance (FDR corrected) of one sample t-test that compared Standard and Deviant events for every cell

Group	Session	Condition	Fz	FCz	Cz
Talker Identification Training	Pre-training	Voice-Change	-1.25 (1.14) ***	-1.50 (1.08) ***	-1.11 (0.95) ***
		Phoneme-Change	-0.90 (1.19) **	-0.90 (0.96) **	-0.57 (0.81) *
	Post-training	Voice-Change	-1.04 (0.68) ***	-1.19 (0.69) ***	-0.79 (0.64) ***
		Phoneme-Change	-1.08 (0.95) **	-0.99 (0.94) **	-0.74 (0.85) **
Syllable Identification Training	Pre-training	Voice-Change	-1.12 (0.73) ***	-1.17 (0.96) ***	-0.71 (0.59) ***
		Phoneme-Change	-1.20 (0.88) ***	-1.21 (0.75) ***	-0.86 (0.59) ***
	Post-training	Voice-Change	-1.40 (0.83) ***	-1.39 (0.72) ***	-0.93 (0.63) ***
		Phoneme-Change	-1.34 (0.70) ***	-1.26 (0.61) ***	-1.09 (0.65) ***

* p <0.05 ** p <0.01 *** p <0.001

APPENDIX C

(Chapter 4)

APPENDIX C

Pre-tests of experimental stimuli

Control items pretests and stimulus selection

In total, we ran two pretests on the control stimuli. Both pretests were built and hosted using the Gorilla Experiment Builder (www.gorilla.sc) and were two-alternative forced choice (2AFC) categorization experiment. Participants were presented with carrier sentences containing the target word in sentence-final position (e.g., *Het woord is VOORnaam*, 'The word is first name'), and presented with two options on the screen (e.g., *VOORnaam* and *voorNAAM*). They were instructed to respond with button presses (left or right) to which word they had heard. The task was identical across pre-tests, but the target stimuli differed.

In the first pretest, we tested the lexical stress continua (N = 18, M_{age} = 22.333, SD_{age} = 3.395). The aim of this pretest was to select the best steps along the continua following two criteria. First, we selected the steps that were most distinct from each other, serving as clear SW and WS tokens. Second, we selected the steps that were comparable across talkers and cues. For example, we ensured that the selected step that served as a SW token from Talker 1, using intensity would be similar to the SW token from Talker 2 using intensity, and similarly across cues.

We thus tested all seven steps of the F0 and intensity continuum, for both talkers, for all eight items. The pretest also tested another 16 words that were eventually not used in the experiment. The experimental list thus contained a total of 672 stimuli, which was divided over two lists. In each list, participants heard all possible steps, cues and talkers for half of the words. Stimuli were presented in randomized order.

To analyze the results, we calculated the proportion of SW responses on each step which provided an indication of how each step was categorized (Figure S6). This illustrated that across all words, there was a clear switch from SW to WS. Further, we calculated the proportion of SW responses divided by talker and cue, and divided by word. In contrast to the results across words, these results illustrated that not every word contained a clear switch. In addition, we found that the acoustically ambiguous step (step 4) was not always the perceptual ambiguous step for all the words.

Since we did require clear SW and WS items as well as clear ambiguous stimuli for all items, we decided to create new continua based on the following strategy. First, based on the results, we took step 5 as the new acoustic middle. Second, we created more extreme steps to create

clear SW and WS items. To achieve this, we increased the number of steps from 7 to 10. Using the same step size as in the 7-step continuum, we added steps to the extremes of the continua to obtain even more extreme versions of the items.

The second pretest tested this new 10-step continuum for all the words (N = 18, M_{age} = *22.222,* SD_{age} *= 6.734*). To reduce the number of stimuli in the Pilot 2, we only tested a subset of the items on the continuum, which were step 1, 3, 5, 6, 8 and 10. By testing these steps, we obtained information on whether (1) more extreme acoustic values would be perceived as clear SW and WS items (step 1 and 10) and (2) where the perceptual middle lied (step 5 or 6).

Similar to pretest 1, we calculated the proportion of SW responses on all the steps across all items and separately for each word, talker, and cue. Overall, we found that the more extreme steps did not result in different categorization responses (see Figure S8). In contrast, the 10-step continua did confirm that the perceptual middle was step 6 rather than step 5.

After extensive piloting, we thus obtained data for two lexical stress continua (7-step and 10-step) for each cue, word, and talker. Based on these data, we selected the steps that best suited the two criteria (most distinct patterns and comparable across talkers and cues). Given the large variability in percentage of SW responses between items, we could not select one single step for each pattern across all items. We thus made the selection on an item-specific basis. Furthermore, we observed better results (i.e., results that better fitted our criteria) in the 7-step continuum for some words, and in the 10-step continuum for others. Therefore, we selected the continuum with the best perceptual data (7-step or 10-step), for each word separately, and chose the steps that met our criteria within that continuum. To ensure that the selection indeed fulfilled our criteria, we calculated the percentage of SW responses for each pattern (SW or WS) split by talker and cue (see Tables S3 and S4). This confirms that there was a clear switch between patterns which was comparable across talkers and cues.

Mixed items manipulations, pretest, and stimulus selection

To create the mixed items, we needed to select the optimal combination of steps along the continua. The resulting items had to meet the criteria of overall ambiguity and clear stress patterns when focusing on one cue. Using the results from the first two pretests, and based on auditory evaluations by the first authors, we created these items.

APPENDIX C

Manipulations involved several different steps. First, as starting point, we took the selected steps that best signaled clear SW and WS items, as evident from the pretests, and combined those values. For example, for the mixed item *voornaam*, we took step 1 on the intensity continuum (syllable 1: 69 dB, syllable 2: 46 dB) and step 9 on the F0 continuum (syllable 1: 94 Hz, syllable 2: 130 Hz). In the mixed item, these values were thus combined (syllable 1: 69 dB, 94 Hz; syllable 2: 46 dB, 130 Hz). Second, the resulting stimuli were evaluated by the first authors. We noticed that in these items, a certain threshold of F0 was required to compete with intensity as a cue. Third, we developed a new strategy for finding the optimal combination. Using this first version of the mixed items, we chose an F0 step that resulted in an overall ambiguous word and still signalled a clear pattern when focusing on only F0 (again, based on auditory evaluations). To find the best combination in mixed items, we took this F0 step and combined it with two different intensity steps. More specifically, it was combined with the most extreme step based on the perception data and one less extreme step. For example, if intensity signalled a SW pattern, we combined F0 step 7 with step 1 and 2 for intensity. This allowed us to find, using one fixed F0 value, which combination with an amplitude value would serve as the best mixed item.

The stimulus set thus contained two combinations of mixed items (the most extreme and one less extreme intensity step) for each pattern (F0-Intensity vs. Intensity-F0), for two talkers and for all words. Importantly, these combinations were always made based on data within one of the two pilots mentioned before. That is, a step from the 7-step continuum was always combined with one from the same continuum, never with another step from the 10-step continuum. Also, recall that we only tested steps 1, 3, 5, 6, 8, 10 in pilot 2. However, for these mixed items, we did make combinations with the 'missing' steps. We did this based on the assumption that steps that lie acoustically within this continuum would also be perceptually within the continuum.

Next, we ran the final pretest in which we tested the different combinations of mixed items (N = 18, $M_{age} = 21.166$, $SD_{age} = 3.185$). In addition, we used this pretest to assess the selected steps of the control items. This allowed us to observe how those items would be categorized in absence of a continuum. Also, they would serve as perceptual anchors of clear SW and WS items. The pretest was built and hosted using the Gorilla Experiment Builder (www.gorilla.sc) and the task was identical to the first two pretests.

This pretest consisted of two separate blocks. In the first block, we tested only the control items in a randomized order. In the second block, we tested the mixed items. Additionally, we still

provided the control items as anchors (on 50% of the trials) in this second block. The items were presented in a pseudorandomized order, ensuring that a mixed item for one specific word pair was never preceded by the control item of that word pair.

Similar to the previous pilots, we calculated the proportion of SW responses for all items. That is, we calculated it for the control items (SW and WS tokens) and for all the combinations of mixed items (F0-Intensity, Intensity-F0; the most extreme intensity step, one less extreme step).

Concerning the control items, results illustrated that there was a clear switch between SW and WS items. More specifically, the proportion of SW responses was 81% for SW items while it was 38% for WS items. Notably, the control items did not reach complete unambiguity (i.e., close to 100% or 0%). A possible explanation is that in contrast to normal speech, in which multiple cues signal lexical stress, these items only contain one cue. Hence, complete unambiguity might not be reached with these items. Still, we did not expect that this would cause problems in the main experiment. That is, there is still a large difference between the SW and WS items. In addition, the main experiment included feedback on participants' responses, and we expected it to further increase the difference between SW and WS.

Concerning the mixed items, we found that perfect ambiguity could not be reached. As Table S11 and Figure S11 illustrate, the F0-Intensity items are perceived as more SW-like compared to chance-level while the Intensity-F0 items as more WS-like. This suggests that despite the presence of two cues, F0 is always perceived as the more dominant cue in these items. Still, the results did illustrate that the mixed items are perceived as less extreme than the control items, which confirms that the opposing cues conflict with correct perception of the stress patterns. Finally, we calculated the proportion of SW responses for the mixed items for each word separately and selected the ideal mixed item (i.e., the item that came closest to chance-level) on an item-specific basis.

In sum, our stimulus set consisted of eight different minimal stress pairs. For each pair and for both talkers, we had a clear SW and WS item (control items), produced using either F0 or intensity, and two mixed items (F0-Intensity, Intensity-F0).

APPENDIX C

Table S5. Dutch words (in Dutch orthography, English translations and IPA transcription) that were used as stimuli. Capitalization indicates lexical stress.

Dutch SW item	English SW translation	Dutch WS item	English WS translation	IPA transcription
AANvaart	(I) collide with	*aanVAARDT*	(I) accept	anvart
MISbruik	Abuse	*misBRUIK*	(I) abuse	mɪsbrœyk
SERvisch	Serbian	*serVIES*	Tableware	sɛrvis
VOORnaam	First name	*voorNAAM*	Respectable	vɔ:rnam
VOORkomen	To appear	*voorKOmen*	To prevent	vɔ:rkomən
DOORlopen	To move along	*doorLOpen*	To go through	dɔ:rlopən
VOORuitgang	Front exit	*voorUITgang*	Progress	vɔ:rœtgɑg
VOORuitzicht	Front view	*voorUITzicht*	Prospect	vɔ:rœtzɪχt

APPENDIX C

Table S6. Item-specific ambiguous values and step sizes of the three manipulated acoustic cues for Talker 1

Word-pair	Syllable	Ambiguous value			Step size	
		F0	Intensity	Duration	F0	Intensity
AANvaart	1	122	69	226	8	2
aanVAARDT	2	119	66	399	6	2
MISbruik	1	126	71	284	8	2
misBRUIK	2	116	64	339	3	2
SERvisch	1	126	71	284	16	2
serVIES	2	113	64	339	3	2
VOORnaam	1	126	71	255	8	2
voorNAAM	2	116	64	369	3	2
VOORkomen	1	126	71	236	8	2
voorKOmen	2	104	64	202	4.5	2
DOORlopen	1	122	71	224	8	5
doorLOpen	2	113	64	171	4.5	2
VOORuitgang	1	122	71	236	8	2
voorUITgang	2	110	64	171	6	2
VOORuitzicht	1	122	71	224	8	2
voorUITzicht	2	110	64	171	6	2

.

APPENDIX C

Table S7. Item-specific ambiguous values and step sizes of the three manipulated acoustic cues for Talker 2

Word-pair	Syllable	Ambiguous value			Step size	
		F0	Intensity	Duration	F0	Intensity
AANvaart	1	122	71	226	8	2
aanVAARDT	2	113	64	399	3	2
MISbruik	1	122	71	284	8	2
misBRUIK	2	116	64	339	3	2
SERvisch	1	122	71	284	8	2
serVIES	2	113	64	339	4.5	2
VOORnaam	1	122	71	255	8	2
voorNAAM	2	116	64	369	6	2
VOORkomen	1	122	71	236	8	2
voorKOmen	2	113	64	202	4.5	2
DOORlopen	1	126	73	224	8	4
doorLOpen	2	113	59	171	4.5	2
VOORuitgang	1	126	71	224	8	2
voorUITgang	2	110	64	171	4.5	2
VOORuitzicht	1	126	71	224	8	2
voorUITzicht	2	110	64	171	4.5	2

APPENDIX C

Table S8. Pre-test 1 - Proportion of SW responses for each word pair of the selected steps from the 1-7 continuum and the 1-10 continuum, divided by Talker, Cue and Pattern.

		Cue			
		Pitch		Intensity	
		Pattern		Pattern	
Word Pair	Talker	SW	WS	SW	WS
AANvaart	1	.8	.2	.77	.2
aanVAARDT	2	.9	.2	.8	.3
MISbruik	1	.81	.45	.81	.18
misBRUIK	2	.72	.18	.90	.09
SERvisch	1	.9	.2	.9	.4
serVIES	2	1.0	.5	.9	.5
VOORnaam	1	1.0	.09	.45	0
voorNAAM	2	.90	.09	.81	.18
VOORkomen	1	.7	0	.8	.3
voorKOmen	2	.8	.4	.9	.4
DOORlopen	1	1.0	.18	.63	.09
doorLOpen	2	.90	.45	.72	.36
VOORuitgang	1	1.0	.18	.72	.27
voorUITgang	2	.72	.54	.81	.36
VOORuitzicht	1	.72	.36	.72	.36
voorUITzicht	2	.90	.36	.90	.27

APPENDIX C

Table S9. Pre-test 1 – Mean proportion and Standard Deviation of SW responses averaged across word pairs of the selected steps from the 1-7 continuum and the 1-10 continuum, divided by Talker, Cue and Pattern.

		Pattern	
Talker	Cue	SW	WS
1	Pitch	.91 (SD = .11)	.18 (SD =.18)
2		.89 (SD = .07)	.33 (SD = .13)
1	Intensity	.76 (SD = .14)	.21 (SD = .12)
2		.88 (SD = .08)	.26 (SD = .10)

APPENDIX C

Table S10. Pre-test 2 - Proportion of SW responses for each word pair of the selected combinations of steps divided by Pattern

		Pattern	
Word Pair	Talker	F0-Intensity	Intensity-F0
AANvaart	1	.78	.57
aanVAARDT	2	.66	.47
MISbruik	1	.73	.57
misBRUIK	2	.63	.63
SERvisch	1	.68	.36
serVIES	2	.63	.36
VOORnaam	1	.52	.36
voorNAAM	2	.55	.47
VOORkomen	1	.63	.36
voorKOmen	2	.78	.57
DOORlopen	1	.68	.47
doorLOpen	2	.73	.57
VOORuitgang	1	.78	.42
voorUITgang	2	.68	.38
VOORuitzicht	1	.52	.57
voorUITzicht	2	.47	.42

APPENDIX C

Table S11. Pre-test 2 - Proportion of SW responses for mixed items averaged across word pairs of the selected combinations of steps divided by Pattern.

	Pattern	
Talker	F0-Intensity	Intensity-F0
1	.67 (SD = .24)	.46 (SD = .28)
2	.64 (SD = .23)	.49 (SD = .33)

APPENDIX C

Table S12. Output of the Model for Control Items with all words (left part) and without unknown words (right part)

Control items - Exposure and Test

	All words				No unknown words			
Predictors	*Log-Odds*	SE	*z*	*p*	*Log-Odds*	SE	*z*	*p*
(Intercept)	0.04	0.09	0.44	0.661	0.04	0.09	0.49	0.625
Pattern	-3.14	0.21	-15.00	**<0.001**	-3.12	0.21	-14.89	**<0.001**
Phase	0.47	0.08	5.68	**<0.001**	0.49	0.09	5.60	**<0.001**
Trial N (normalized)	-0.00	0.06	-0.05	0.958	-0.02	0.07	-0.29	0.774
Talker	0.12	0.18	0.67	0.500	0.15	0.18	0.80	0.424
Pattern * Phase	0.44	0.15	2.95	**0.003**	0.49	0.16	3.09	**0.002**
Pattern * Trial N	-0.05	0.13	-0.38	0.703	-0.08	0.14	-0.60	0.551
Phase * Trial N	-0.64	0.13	-4.95	**<0.001**	-0.64	0.14	-4.65	**<0.001**
Pattern * Talker	0.12	0.30	0.40	0.693	0.04	0.28	0.15	0.883
Pattern * Phase * Trial N	-0.95	0.26	-3.68	**<0.001**	-1.03	0.27	-3.76	**<0.001**

APPENDIX C

Table S13. Post-hoc tests for the Pattern x Phase interaction of the model for Control items with all words (left part) and without unknown words (right part).

Control Items - Pattern x Phase Post-hoc tests

	All words				No unknown words			
Contrast	*Log-Odds*	*SE*	*z*	*p*	*Log-Odds*	*SE*	*z*	*p*
SW training - WS training	3.18	0.20	15.91	**< .001**	3.18	0.20	15.95	**< .001**
SW test - SW training	-0.17	0.06	-2.64	**0.011**	-0.18	0.07	-2.73	**0.009**
WS test - WS training	-0.13	0.06	-2.11	**0.035**	-0.16	0.07	-2.38	**0.017**
SW test - WS test	3.15	0.20	15.37	**< .001**	3.16	0.20	15.40	**< .001**

APPENDIX C

Table S14. Post-hoc tests for the Pattern x Phase x Trial Number interaction of the model for Control items with all words (left part) and without unknown words (right part).

Control Items - Pattern x Phase x Trial Number Post-hoc tests

	All words				**No unknown words**			
Contrast	*Log-Odds*	*SE*	*z*	*p*	*Log-Odds*	*SE*	*z*	*p*
SW training - WS training	0.52	0.15	3.52	**< .001**	0.60	0.16	3.78	**< .001**
SW test - SW training	0.16	0.18	0.90	0.368	0.12	0.19	0.63	0.531
WS test - WS training	1.11	0.18	6.07	**< .001**	1.15	0.20	5.93	**< .001**
SW test - WS test	-0.42	0.21	-2.00	0.06	-0.43	0.23	-1.92	0.073

APPENDIX C

Table S15. Output of the Model for Mixed Items with all words (left part) and without unknown words (right part)

Mixed Items - Test

	All words				No unknown words			
Predictors	*Log-Odds*	*SE*	*z*	*p*	*Log-Odds*	*SE*	*z*	*p*
(Intercept)	0.18	0.11	1.68	0.092	0.17	0.11	1.58	0.115
Predicted Response	-0.74	0.14	-5.20	**<0.001**	-0.73	0.15	-4.77	**<0.001**
Talker	0.44	0.18	2.47	**0.014**	0.44	0.17	2.60	**0.009**
Pattern	1.42	0.27	5.34	**<0.001**	1.47	0.26	5.65	**<0.001**
Trial N (normalized)	0.04	0.11	0.37	0.712	0.03	0.12	0.27	0.787
Predicted Response * Talker	0.64	0.44	1.45	0.148	0.62	0.44	1.40	0.161
Predicted Response * Trial N	0.36	0.19	1.92	0.055	0.36	0.20	1.78	0.074

APPENDIX C

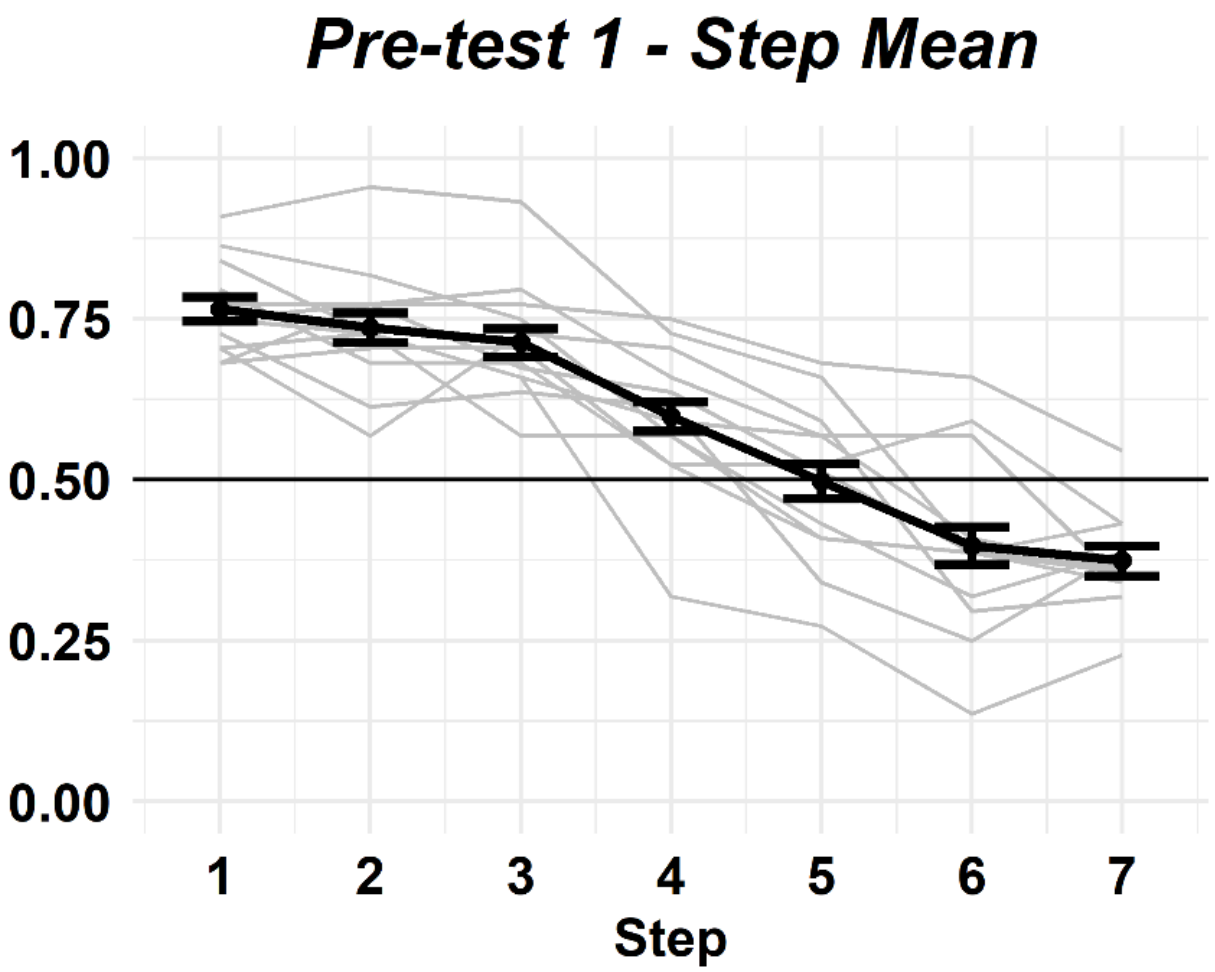


Figure S6. Pre-test 1 – Proportion of SW responses (y axis) along the 7-step continua averaged across participants, Talkers and Cues. Grey lines refer to individual word pairs while thick black lines represent the average across word pairs. Error bars represent the standard error.

APPENDIX C

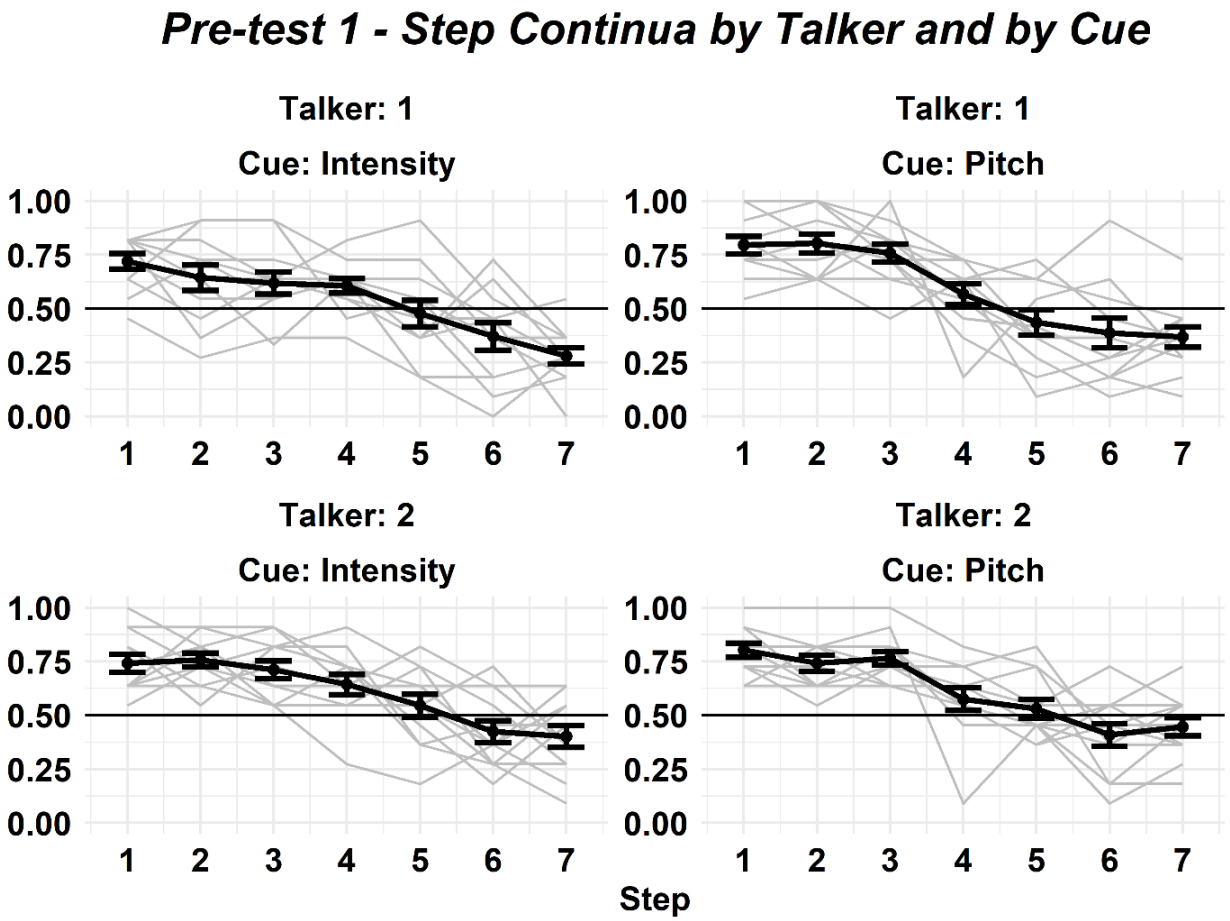


Figure S7. Pre-test 1 – Proportion of SW responses (y axis) along the 7-step continua averaged across participants divided by Talker and Cue. Grey lines refer to individual word pairs while thick black lines represent the average across word pairs. Error bars represent the standard error.

APPENDIX C

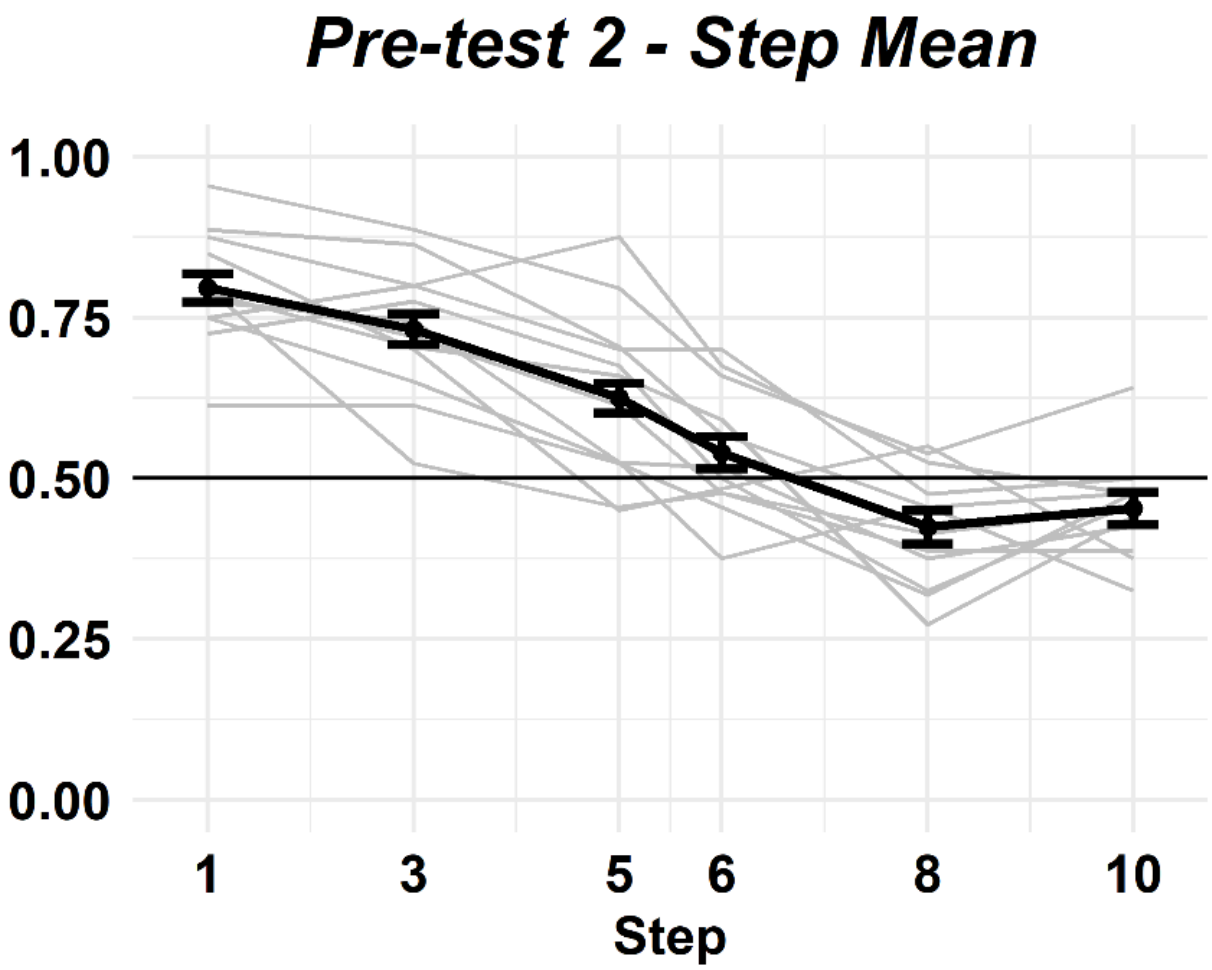


Figure S8. Pre-test 2 – Proportion of SW responses (y axis) along the 7-step continua averaged across participants, Talkers and Cues. Grey lines refer to individual word pairs while thick black lines represent the average across word pairs. Error bars represent the standard error.

APPENDIX C

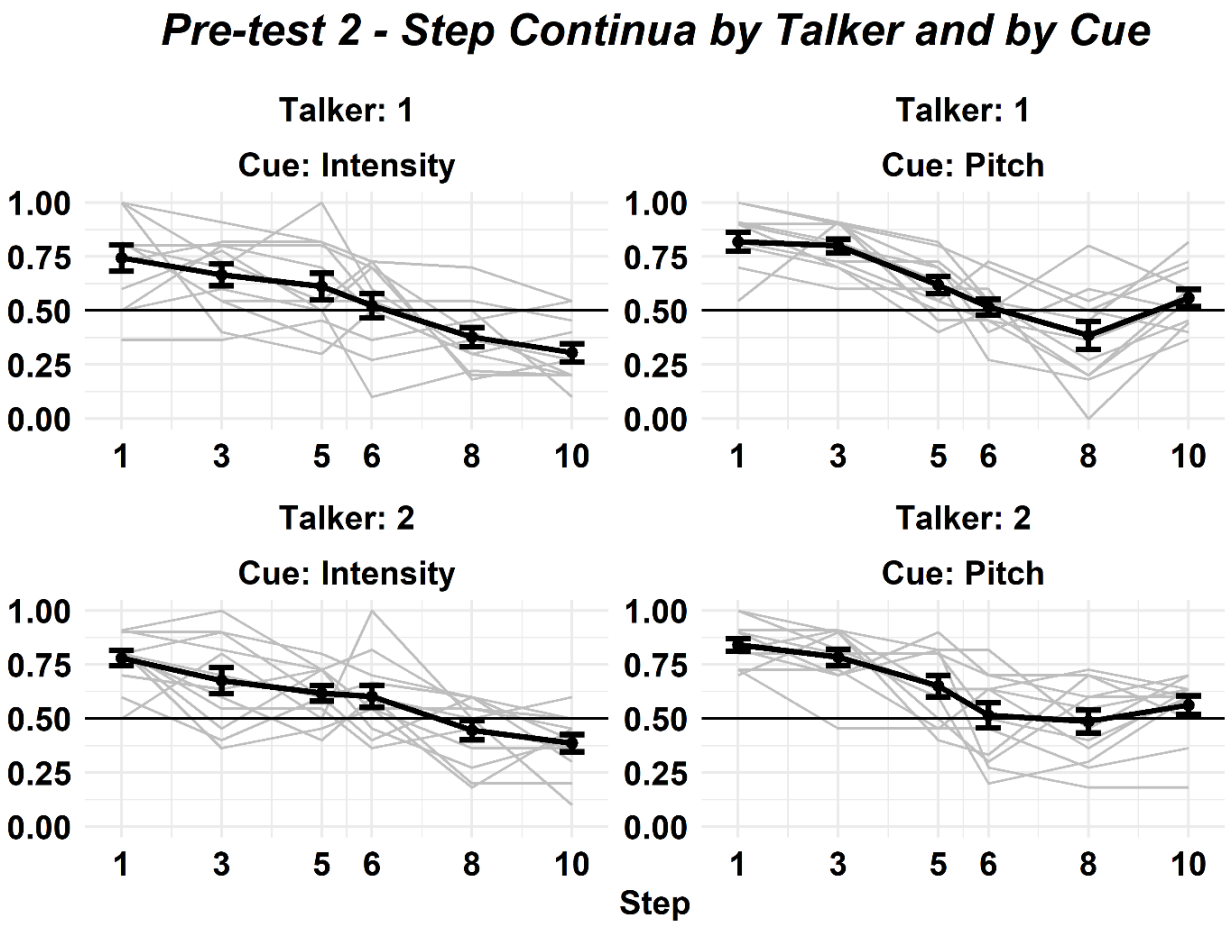


Figure S9. Pre-test 2 - Proportion of SW responses (y axis) along the 10-step continua (steps 1,3,5,6,8,10) averaged across participants divided by Talker and Cue. Grey lines refer to individual word pairs while thick black lines represent the average across word pairs. Error bars represent the standard error.

APPENDIX C

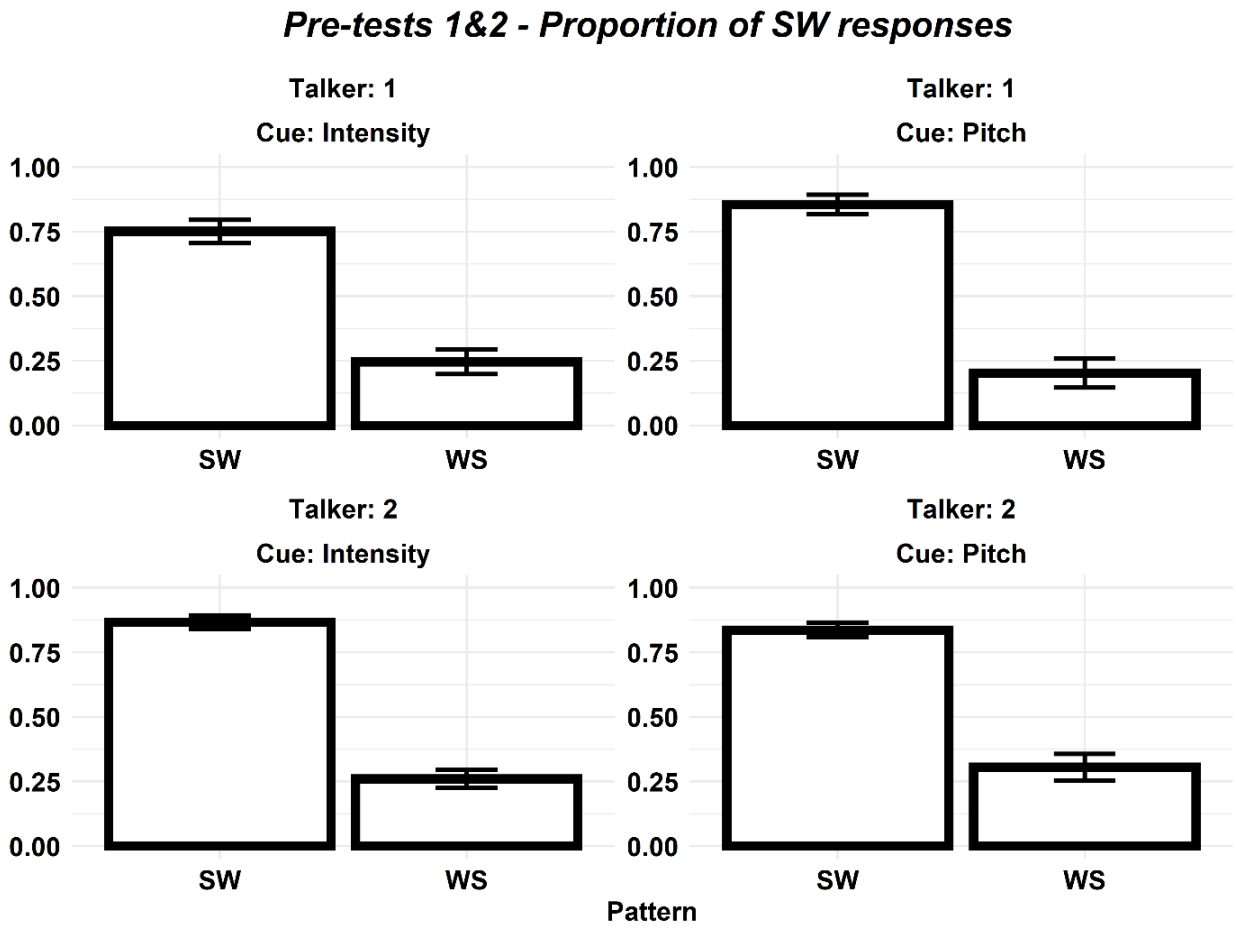


Figure S10. Pre-test 1&2 – Proportion of SW responses (y axis) averaged across word pairs of the selected steps from the 1-7 continuum and the 1-10 continuum, divided by Talker, Cue and Pattern. Error bars represent the Standard Error.

APPENDIX C

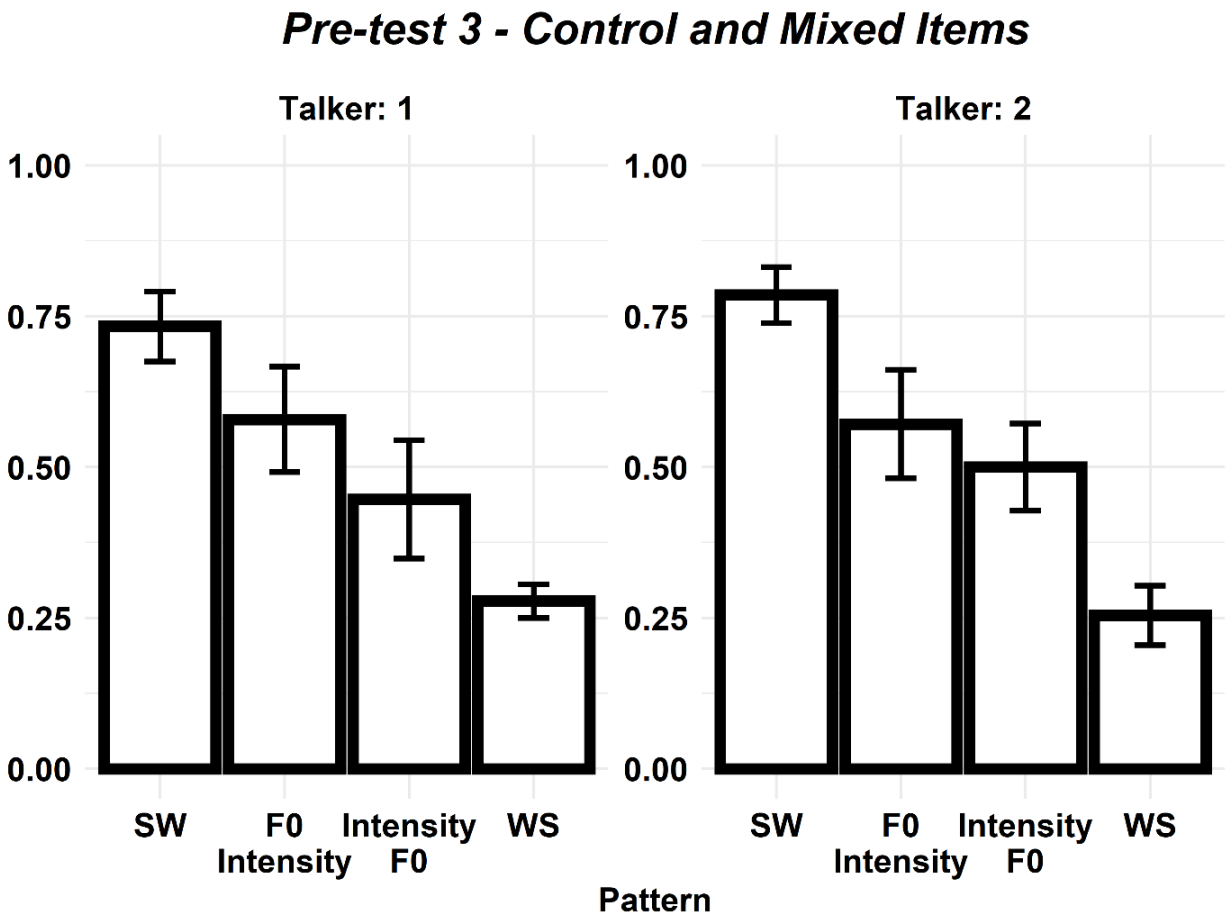


Figure S11. Pre-test 3 - Proportion of SW responses (y-axis) averaged across word pairs and participants of the selected combinations of steps divided b

References

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*(5), 3099–3107. https://doi.org/10.1121/1.1795335

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.

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*, *49*(1), 97–110. https://doi.org/10.3758/s13428-015-0689-6

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

Andics, A., McQueen, J. M., & Petersson, K. M. (2013). Mean-based neural coding of voices. *NeuroImage*, *79*, 351–360. https://doi.org/10.1016/j.neuroimage.2013.05.002

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage*, *52*(4), 1528–1540. https://doi.org/10.1016/j.neuroimage.2010.05.048

Assmann, P. F., & Nearey, T. M. (2007). Relationship between fundamental and formant frequencies in voice preference. *The Journal of the Acoustical Society of America*, *122*(2), EL35–EL43. https://doi.org/10.1121/1.2719045

Azadpour, M., & Balaban, E. (2008). Phonological Representations Are Unconsciously Used when Processing Complex, Non-Speech Signals. *PLoS ONE*, *3*(4), e1966. https://doi.org/10.1371/journal.pone.0001966

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

Backus, A. R., Schoffelen, J.-M., Szebényi, S., Hanslmayr, S., & Doeller, C. F. (2016). Hippocampal-Prefrontal Theta Oscillations Support Memory Integration. *Current Biology*, *26*(4), 450–457. https://doi.org/10.1016/j.cub.2015.12.048

Başar-Eroglu, C., Başar, E., Demiralp, T., & Schürmann, M. (1992). P300-response: Possible psychophysiological correlates in delta and theta frequency channels. A review. *International Journal of Psychophysiology*, *13*(2), 161–179. https://doi.org/10.1016/0167-8760(92)90055-g

Bastiaansen, M. C. M., Linden, M. van der, Keurs, M. ter, Dijkstra, T., & Hagoort, P. (2005). Theta Responses Are Involved in Lexical—Semantic Retrieval during Language Processing. *Journal of Cognitive Neuroscience*, *17*(3), 530–541. https://doi.org/10.1162/0898929053279469

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., & Bolker, M. B. (2015). Package 'lme4.' *Convergence*, *12*(1), 2.

Batterink, L. J., & Paller, K. A. (2019). Statistical learning of speech regularities can occur outside the focus of attention. *Cortex*, *115*, 56–71. https://doi.org/10.1016/j.cortex.2019.01.013

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research Psychologische Forschung*, *74*(1), 110–120. https://doi.org/10.1007/s00426-008-0185-z

Beauchemin, M., De Beaumont, L., Vannasing, P., Turcotte, A., Arcand, C., Belin, P., & Lassonde, M. (2006). Electrophysiological markers of voice familiarity. *European Journal of Neuroscience*, *23*(11), 3081–3086. https://doi.org/10.1111/j.1460-9568.2006.04856.x

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe: *NeuroReport*, *14*(16), 2105–2109. https://doi.org/10.1097/00001756-200311140-00019

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Bendixen, A., & Schröger, E. (2008). Memory trace formation for abstract auditory features and its consequences in different attentional contexts. *Biological Psychology*, *78*(3), 231–241. https://doi.org/10.1016/j.biopsycho.2008.03.005

Birkett, P. B., Hunter, M. D., Parks, R. W., Farrow, T. F., Lowe, H., Wilkinson, I. D., & Woodruff, P. W. (2007). Voice familiarity engages auditory cortex. *NeuroReport*, *18*(13), 1375–1378. https://doi.org/10.1097/WNR.0b013e3282aa43a3

Bishop, D. V. M., Hardiman, M. J., & Barry, J. G. (2011). Is auditory discrimination mature by middle childhood? A study using time-frequency analysis of mismatch responses from 7 years to adulthood: Is auditory discrimination mature? *Developmental Science*, *14*(2), 402–416. https://doi.org/10.1111/j.1467-7687.2010.00990.x

Blesser, B. (1972). Speech Perception Under Conditions of Spectral Transformation: I. Phonetic Characteristics. *Journal of Speech and Hearing Research*, *15*(1), 5–41. https://doi.org/10.1044/jshr.1501.05

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

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-Dependent Decoding of Speaker and Vowel Identity from Auditory Cortical Response Patterns. *Journal of Neuroscience*, *34*(13), 4548–4557. https://doi.org/10.1523/JNEUROSCI.4339-13.2014

Bonte, M., Valente, G., & Formisano, E. (2009). Dynamic and Task-Dependent Encoding of Speech and Voice by Phase Reorganization of Cortical Oscillations. *Journal of Neuroscience*, *29*(6), 1699–1706. https://doi.org/10.1523/JNEUROSCI.3694-08.2009

Bosker, H. R. (2021). Evidence for selective adaptation and recalibration in the perception of lexical stress. *Language and Speech*.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, *130*(1), 85–95.

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

Carral, V., Huotilainen, M., Ruusuvirta, T., Fellman, V., Näätänen, R., & Escera, C. (2005). A kind of auditory 'primitive intelligence'already present at birth. *European Journal of Neuroscience*, *21*(11), 3201–3204. https://doi.org/10.1111/j.1460-9568.2005.04144.x

Chandrasekaran, B., Krishnan, A., & Gandour, J. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain and Language*, *108*(1), 1–9. https://doi.org/10.1016/j.bandl.2008.02.001

Cheour, M., Korpilahti, P., Martynova, O., & Lang, A.-H. (2001). Mismatch Negativity and Late Discriminative Negativity in Investigating Speech Perception and Learning in Children and Infants. *Audiology and Neuro-Otology*, *6*(1), 2–11. https://doi.org/10.1159/000046804

Choudhury, N. A., Parascando, J. A., & Benasich, A. A. (2015). Effects of Presentation Rate and Attention on Auditory Discrimination: A Comparison of Long-Latency Auditory Evoked Potentials in School-Aged Children and Adults. *PLOS ONE*, *10*(9), e0138160. https://doi.org/10.1371/journal.pone.0138160

Christmann, C. A., Berti, S., Steinbrink, C., & Lachmann, T. (2014). Differences in sensory processing of German vowels and physically matched non-speech sounds as revealed by the mismatch negativity (MMN) of the human event-related brain potential (ERP). *Brain and Language*, *136*, 8–18. https://doi.org/10.1016/j.bandl.2014.07.004

Citherlet, D., Boucher, O., Tremblay, J., Robert, M., Gallagher, A., Bouthillier, A., Lepore, F., & Nguyen, D. K. (2020). Spatiotemporal dynamics of auditory information processing in the insular cortex: An intracranial EEG study using an oddball paradigm. *Brain Structure and Function*, *225*(5), 1537–1559. https://doi.org/10.1007/s00429-020-02072-z

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

Cohen, M. X. (2014). *Analyzing neural time series data: Theory and practice*. The MIT Press.

Comerchero, M. D., & Polich, J. (1999). P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology*, *110*(1), 24–30. https://doi.org/10.1016/S0168-5597(98)00033-1

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

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*(2), 633–664. https://doi.org/10.1016/j.cognition.2007.03.013

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., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In *Laboratory phonology 10* (pp. 91–112). De Gruyter Mouton.

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

Datta, H., Shafer, V. L., Morr, M. L., Kurtzberg, D., & Schwartz, R. G. (2010). Electrophysiological Indices of Discrimination of Long-Duration, Phonetically Similar Vowels in Children With Typical and Atypical Language Development. *Journal of*

Speech, Language, and Hearing Research, *53*(3), 757–777. https://doi.org/10.1044/1092-4388(2009/08-0123)

David, C., Roux, S., Bonnet-Brilhault, F., Ferré, S., & Gomot, M. (2020). Brain responses to change in phonological structures of varying complexity in children and adults. *Psychophysiology*, *57*(9). https://doi.org/10.1111/psyp.13621

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport*, *8*(4), 919–924.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage*, *24*(1), 21–33. https://doi.org/10.1016/j.neuroimage.2004.09.039

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Demiralp, T., Ademoglu, A., Comerchero, M., & Polich, J. (2001). Wavelet analysis of P3a and P3b. *Brain Topography*, *13*(4), 251–267. https://doi.org/10.1023/A:1011102628306

DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences*, *109*(8), E505–E514. https://doi.org/10.1073/pnas.1113427109

Donchin, E. (1981). Surprise!… surprise? *Psychophysiology*, *18*(5), 493–513. https://doi.org/10.1111/j.1469-8986.1981.tb01815.x

Dong, S., Reder, L. M., Yao, Y., Liu, Y., & Chen, F. (2015). Individual differences in working memory capacity are reflected in different ERP and EEG patterns to task difficulty. *Brain Research*, *1616*, 146–156. https://doi.org/10.1016/j.brainres.2015.05.003

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., Polich, J., Reinvang, I., & Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, *120*(11), 1883–1908. https://doi.org/10.1016/j.clinph.2009.07.045

Duncan, D., & Theeuwes, J. (2020). Statistical learning in the absence of explicit top-down attention. *Cortex*, *131*, 54–65. https://doi.org/10.1016/j.cortex.2020.07.006

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

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

Eisner, F., & McQueen, J. M. (2018). Speech perception. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *3*, 1–46.

Elmer, S., Hausheer, M., Albrecht, J., & Kühnis, J. (2017). Human Brainstem Exhibits higher Sensitivity and Specificity than Auditory-Related Cortex to Short-Term Phonetic Discrimination Learning. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-07426-y

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—Signalling the status quo? *Current Opinion in Neurobiology*, *20*(2), 156–165. https://doi.org/10.1016/j.conb.2010.02.015

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. In Nelson Morgan (Ed.), *Proceedings of Interspeech 2016* (pp. 1059–1063). The International Speech Communication Association (ISCA); DiVA. 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.

Ervast, L., Hämäläinen, J. A., Zachau, S., Lohvansuu, K., Heinänen, K., Veijola, M., Heikkinen, E., Suominen, K., Luotonen, M., & Lehtihalmes, M. (2015). Event-related brain potentials to change in the frequency and temporal structure of sounds in typically

developing 5–6-year-old children. *International Journal of Psychophysiology*, *98*(3), 413–425. https://doi.org/10.1016/j.ijpsycho.2015.08.007

Escera, C., Leung, S., & Grimm, S. (2014). Deviance Detection Based on Regularity Encoding Along the Auditory Hierarchy: Electrophysiological Evidence in Humans. *Brain Topography*, *27*(4), 527–538. https://doi.org/10.1007/s10548-013-0328-4

Escera, C., & Malmierca, M. S. (2014). The auditory novelty system: An attempt to integrate human and animal research: The auditory novelty system. *Psychophysiology*, *51*(2), 111–123. https://doi.org/10.1111/psyp.12156

Eulitz, C., & Lahiri, A. (2004). Neurobiological Evidence for Abstract Phonological Representations in the Mental Lexicon during Speech Recognition. *Journal of Cognitive Neuroscience*, *16*(4), 577–583. https://doi.org/10.1162/089892904323057308

Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and Voice Representation: From Acoustic-Based Representation to Voice Averages. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01180

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science*, *322*(5903), 970–973. https://doi.org/10.1126/science.1164318

Fuentemilla, Ll., Marco-Pallarés, J., Münte, T. F., & Grau, C. (2008). Theta EEG oscillatory activity and auditory change detection. *Brain Research*, *1220*, 93–101. https://doi.org/10.1016/j.brainres.2007.07.079

Garner, W. R. (2014). *The processing of information and structure*. Psychology Press.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, *12*(5–6), 613–656. https://doi.org/10.1080/016909697386646

Gaskell, M. G., & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 380. https://doi.org/10.1037/0096-1523.24.2.380

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166

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

Grabski, K., & Sato, M. (2020). Adaptive phonemic coding in the listening and speaking brain. *Neuropsychologia*, *136*, 107267. https://doi.org/10.1016/j.neuropsychologia.2019.107267

Gu, F., Zhang, C., Hu, A., & Zhao, G. (2013). Left hemisphere lateralization for lexical and acoustic pitch processing in Cantonese speakers as revealed by mismatch negativity. *NeuroImage*, *83*, 637–645. https://doi.org/10.1016/j.neuroimage.2013.02.080

Gustavsson, L., Kallioinen, P., Klintfors, E., & Lindh, J. (2013). Neural processing of voices—Familiarity. *Proceedings of Meetings on Acoustics*, 060204–060204. https://doi.org/10.1121/1.4800901

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.

Halle, M. (2013). Speculations about the representations of words in memory. In *From Memory to Speech and Back* (pp. 122–136). De Gruyter Mouton.

Hamburg. (2019). In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Hamburg&oldid=188211539

Hewlett, N., & Beck, J. M. (2013). *An introduction to the science of phonetics*. Routledge.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

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. https://doi.org/10.1121/1.411872

Hirahara, T., & Kato, H. (1992). The effect of F0 on vowel identification. *Speech Perception, Production and Linguistic Structure*, 89–112.

Honbolygó, F., Kóbor, A., German, B., & Csépe, V. (2020). Word stress representations are language-specific: Evidence from event-related brain potentials. *Psychophysiology*, *57*(5). https://doi.org/10.1111/psyp.13541

Hsu, C.-H., Evans, J. P., & Lee, C.-Y. (2015). Brain responses to spoken F 0 changes: Is H special? *Journal of Phonetics*, *51*, 82–92. https://doi.org/10.1016/j.wocn.2015.02.003

Huotilainen, M., llmoniemi, R. J., Lavikainen, J., Tiitinen, H., Alho, K., Sinkkonen, J., Knuutila, J., & Nä, R. (1993). Interaction between representations of different features of auditory sensory memory: *NeuroReport*, *4*(11), 1279. https://doi.org/10.1097/00001756-199309000-00018

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.

Jacobsen, T., Schröger, E., & Alter, K. (2004). Pre-attentive perception of vowel phonemes from variable speech stimuli. *Psychophysiology*, *41*(4), 654–659. https://doi.org/10.1111/1469-8986.2004.00175.x

Jacobsen, T., Schröger, E., & Sussman, E. (2004). Pre-attentive categorization of vowel formant structure in complex tones. *Cognitive Brain Research*, *20*(3), 473–479. https://doi.org/10.1016/j.cogbrainres.2004.03.021

Jensen, O., & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task: Frontal theta increases with memory load. *European Journal of Neuroscience*, *15*(8), 1395–1399. https://doi.org/10.1046/j.1460-9568.2002.01975.x

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

Jin, Y., Díaz, B., Colomer, M., & Sebastián-Gallés, N. (2014). Oscillation Encoding of Individual Differences in Speech Perception. *PLoS ONE*, *9*(7), e100901. https://doi.org/10.1371/journal.pone.0100901

Johnson, K. (1990). The role of perceived speaker identity in F 0 normalization of vowels. *The Journal of the Acoustical Society of America*, *88*(2), 642–654.

Johnson, K. A., & Mullennix, J. W. (1997). Speech perception without speaker normalization. In *Talker Variability in Speech Processing.* (pp. 45–166). San Diego: Academic Press.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a Cocktail Party: Voice Familiarity Aids Speech Perception in the Presence of a Competing Voice. *Psychological Science*, *24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Jongman, S. R., Roelofs, A., & Lewis, A. G. (2020). Attention for Speaking: Prestimulus Motor-cortical Alpha Power Predicts Picture Naming Latencies. *Journal of Cognitive Neuroscience*, *32*(5), 747–761. https://doi.org/10.1162/jocn_a_01513

Jongman, S. R., Roelofs, A., & Meyer, A. S. (2015). Sustained attention in language production: An individual differences investigation. *Quarterly Journal of Experimental Psychology*, *68*(4), 710–730. https://doi.org/10.1080/17470218.2014.964736

Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research*, *1114*(1), 161–172. https://doi.org/10.1016/j.brainres.2006.07.049

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

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3933–3936.

Kemmerer, D. (2015). Speech Perception. In *Cognitive neuroscience of language*. Psychology Press.

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

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Research Reviews*, *29*(2–3), 169–195. https://doi.org/10.1016/S0165-0173(98)00056-3

Klimesch, W., Doppelmayr, M., Stadler, W., Pöllhuber, D., Sauseng, P., & Röhm, D. (2001). Episodic retrieval is reflected by a process specific increase in human electroencephalographic theta activity. *Neuroscience Letters*, *302*(1), 49–52. https://doi.org/10.1016/S0304-3940(01)01656-1

Ko, D., Kwon, S., Lee, G.-T., Im, C. H., Kim, K. H., & Jung, K.-Y. (2012). Theta Oscillation Related to the Auditory Discrimination Process in Mismatch Negativity: Oddball versus Control Paradigm. *Journal of Clinical Neurology*, *8*(1), 35. https://doi.org/10.3988/jcn.2012.8.1.35

Koerner, T. K., Zhang, Y., Nelson, P. B., Wang, B., & Zou, H. (2016). Neural indices of phonemic discrimination and sentence-level speech intelligibility in quiet and noise: A mismatch negativity study. *Hearing Research*, *339*, 40–49. https://doi.org/10.1016/j.heares.2016.06.001

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557–577. https://doi.org/10.1017/S0048577201990559

Kolev, V., Demiralp, T., Yordanova, J., Ademoglu, A., & Isoglu-Alkaç, Ü. (1997). Time–frequency analysis reveals multiple functional components during oddball P300. *NeuroReport*, *8*(8), 2061–2065.

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. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

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

Kraus, N., McGee, T., Carrell, T. D., King, C., Tremblay, K., & Nicol, T. (1995). Central Auditory System Plasticity Associated with Speech Discrimination Training. *Journal of Cognitive Neuroscience*, *7*(1), 25–32. https://doi.org/10.1162/jocn.1995.7.1.25

Kuhl, P. K. (2011). Who's Talking? *Science*, *333*(6042), 529–530. https://doi.org/10.1126/science.1210277

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

Kurkela, J. L. O., Hämäläinen, J. A., Leppänen, P. H. T., Shu, H., & Astikainen, P. (2019). Passive exposure to speech sounds modifies change detection brain responses in adults. *NeuroImage*, *188*, 208–216. https://doi.org/10.1016/j.neuroimage.2018.12.010

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

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104.

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The Spectrotemporal Filter Mechanism of Auditory Selective Attention. *Neuron*, *77*(4), 750–761. https://doi.org/10.1016/j.neuron.2012.11.034

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*(4), R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-Induced Changes in the Cerebral Processing of Voice Identity. *Cerebral Cortex*, *21*(12), 2820–2828. https://doi.org/10.1093/cercor/bhr077

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology*, *23*(12), 1075–1080. https://doi.org/10.1016/j.cub.2013.04.055

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10295-w

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

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38.

Li, X., & Chen, Y. (2018). Unattended processing of hierarchical pitch variations in spoken sentences. *Brain and Language*, *183*, 21–31. https://doi.org/10.1016/j.bandl.2018.05.004

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431. https://doi.org/10.1037/h0020279

Linares, D., & López i Moliner, J. (2016). quickpsy: An R package to fit psychometric functions for multiple groups. *The R Journal, 2016, Vol. 8, Num. 1, p. 122-131*.

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

Lisman, J., & Buzsaki, G. (2008). A Neural Coding Scheme Formed by the Combined Function of Gamma and Theta Oscillations. *Schizophrenia Bulletin*, *34*(5), 974–980. https://doi.org/10.1093/schbul/sbn060

Liu, H.-M., Chen, Y., & Tsao, F.-M. (2014). Developmental Changes in Mismatch Responses to Mandarin Consonants and Lexical Tones from Early to Middle Childhood. *PLoS ONE*, *9*(4), e95587. https://doi.org/10.1371/journal.pone.0095587

Liu, L., Ong, J. H., Tuninetti, A., & Escudero, P. (2018). One Way or Another: Evidence for Perceptual Asymmetry in Pre-attentive Learning of Non-native Contrasts. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00162

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

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. https://doi.org/10.3389/fnhum.2014.00213

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics*, *83*(4), 1842–1860. https://doi.org/10.3758/s13414-020-02203-y

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179–193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 245–256. https://doi.org/10.1037/a0019952

Marklund, E., Gustavsson, L., Kallioinen, P., & Schwarz, I.-C. (2020). N1 Repetition-Attenuation for Acoustically Variable Speech and Spectrally Rotated Speech. *Frontiers in Human Neuroscience*, *14*, 534804. https://doi.org/10.3389/fnhum.2020.534804

Marklund, E., Lacerda, F., & Schwarz, I.-C. (2018). Using rotated speech to approximate the acoustic mismatch negativity response to speech. *Brain and Language*, *176*, 26–35. https://doi.org/10.1016/j.bandl.2017.10.006

Mazaheri, A., & Picton, T. W. (2005). EEG spectral dynamics during discrimination of auditory and visual targets. *Cognitive Brain Research*, *24*(1), 81–96. https://doi.org/10.1016/j.cogbrainres.2004.12.013

McAuliffe, M., & Babel, M. (2016). Stimulus-directed attention attenuates lexically-guided perceptual learning. *The Journal of the Acoustical Society of America*, *140*(3), 1727–1738. https://doi.org/10.1121/1.4962529

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McGuire, G. L., & Babel, M. (2020). Attention to Indexical Information Improves Voice Recall. *Interspeech 2020*, 1595–1599. https://doi.org/10.21437/Interspeech.2020-3042

McLaughlin, D., Dougherty, S., Lember, R., & Perrachione, T. K. (2015). Episodic memory for words enhances the language familiarity effect in talker identification. *ICPhS*.

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

Monte-Ordoño, J., & Toro, J. M. (2017). Early positivity signals changes in an abstract linguistic pattern. *PLOS ONE*, *12*(7), e0180727. https://doi.org/10.1371/journal.pone.0180727

Moran, R. J., Campo, P., Maestu, F., Reilly, R. B., Dolan, R. J., & Strange, B. A. (2010). Peak frequency in the theta and alpha bands correlates with human working memory capacity. *Frontiers in Human Neuroscience*, *4*, 200. https://doi.org/10.3389/fnhum.2010.00200

Mueller, V., Brehmer, Y., Von Oertzen, T., Li, S.-C., & Lindenberger, U. (2008). Electrophysiological correlates of selective attention: A lifespan comparison. *BMC Neuroscience*, *9*(1), 1–21. https://doi.org/10.1186/1471-2202-9-18

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. https://doi.org/10.3758/BF03210878

Muller, D., Widmann, A., & Schroger, E. (2005). Auditory streaming affects the processing of successive deviant and standard sounds. *Psychophysiology*, *42*(6), 668–676. https://doi.org/10.1111/j.1469-8986.2005.00355.x

Näätänen, R. (1995). The mismatch negativity: A powerful tool for cognitive neuroscience. *Ear and Hearing*, *16*(1), 6–18.

Näätänen, R., Jacobsen, T., & Winkler, I. (2005). Memory-based or afferent processes in mismatch negativity (MMN): A review of the evidence. *Psychophysiology*, *42*(1), 25–32. https://doi.org/10.1111/j.1469-8986.2005.00256.x

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*(6615), 432–434. https://doi.org/10.1038/385432a0

Näätänen, R., & Michie, P. T. (1979). Early selective-attention effects on the evoked potential: A critical review and reinterpretation. *Biological Psychology*, *8*(2), 81–136. https://doi.org/10.1016/0301-0511(79)90053-X

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*(12), 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026

Nager, W., Teder-Sälejärvi, W., Kunze, S., & Münte, T. F. (2003). Preattentive evaluation of multiple perceptual streams in human audition: *NeuroReport*, *14*(6), 871–874. https://doi.org/10.1097/00001756-200305060-00019

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

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

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. https://doi.org/10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Obleser, J., Elbert, T., Lahiri, A., & Eulitz, C. (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research*, *15*(3), 207–213. https://doi.org/10.1016/S0926-6410(02)00193-3

Obleser, J., & Weisz, N. (2012). Suppressed Alpha Oscillations Predict Intelligibility of Speech and its Acoustic Details. *Cerebral Cortex*, *22*(11), 2466–2477. https://doi.org/10.1093/cercor/bhr325

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Ollen, J. E. (2006). *A criterion-related validity test of selected indicators of musical sophistication using expert ratings*. The Ohio State University.

Öniz, A., & Başar, E. (2009). Prolongation of alpha oscillations in auditory oddball paradigm. *International Journal of Psychophysiology*, *71*(3), 235–241. https://doi.org/10.1016/j.ijpsycho.2008.10.003

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition*, *143*, 36–40. https://doi.org/10.1016/j.cognition.2015.06.002

Paavilainen, P. (2013). The mismatch-negativity (MMN) component of the auditory event-related potential to violations of abstract regularities: A review. *International Journal of Psychophysiology*, *88*(2), 109–123. https://doi.org/10.1016/j.ijpsycho.2013.03.015

Pakarinen, S., Huotilainen, M., & Näätänen, R. (2010). The mismatch negativity (MMN) with no standard stimulus. *Clinical Neurophysiology*, *121*(7), 1043–1050. https://doi.org/10.1016/j.clinph.2010.02.009

Paul Boersma & David Weenink. (2018). *Praat: Doing phonetics by computer [Computer program]* (Version 6.0.37) [Computer software]. http://www.praat.org/

Pereira, O., Gao, Y. A., & Toscano, J. C. (2018). Perceptual Encoding of Natural Speech Sounds Revealed by the N1 Event-Related Potential Response. *Auditory Perception & Cognition*, *1*(1–2), 112–130. https://doi.org/10.1080/25742442.2018.1545106

Perrachione, T. K. (2017). *Speaker recognition across languages*. Oxford University Press.

Perrachione, T. K., & Choi, J. Y. (2016). Extrinsic talker normalization via rapid accumulation of talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, *139*(4), 2019–2019. https://doi.org/10.1121/1.4949955

Perrachione, T. K., Dougherty, S., McLaughlin, D., & Lember, R. (2015). The effects of speech perception and speech comprehension on talker identification. *ICPhS*.

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910. https://doi.org/10.1016/j.neuropsychologia.2006.11.015

Pesnot Lerousseau, J., & Schön, D. (2021). Musical Expertise Is Associated with Improved Neural Statistical Learning in the Auditory Domain. *Cerebral Cortex*, bhab128. https://doi.org/10.1093/cercor/bhab128

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Petkov, C. I., & Vuong, Q. C. (2013). Neuronal coding: The value in having an average voice. *Current Biology*, *23*(12), R521–R523. https://doi.org/10.1016/j.cub.2013.04.077

Pierrehumbert, J. (2001). Lenition and contrast. *Frequency and the Emergence of Linguistic Structure*, *45*, 137.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Pisoni, D. B. (1992). Talker normalization in speech perception. *Speech Perception, Production and Linguistic Structure*, 143–151.

Plante-Hébert, J., Boucher, V. J., & Jemel, B. (2017). Electrophysiological Correlates of Familiar Voice Recognition. *Interspeech*, 3907–3910. https://doi.org/10.21437/Interspeech.2017-1392

Polich, J. (1987). Task difficulty, probability, and inter-stimulus interval as determinants of P300 from auditory stimuli. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *68*(4), 311–320. https://doi.org/10.1016/0168-5597(87)90052-9

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Prins, N. (2016). *Psychophysics: A practical introduction*. Academic Press.

Pulvermüller, F., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., Alho, K., Martinkauppi, S., Ilmoniemi, R. J., & Näätänen, R. (2001). Memory Traces for Words as Revealed by the Mismatch Negativity. *NeuroImage*, *14*(3), 607–616. https://doi.org/10.1006/nimg.2001.0864

Pulvermüller, F., & Shtyrov, Y. (2006). Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Progress in Neurobiology*, *79*(1), 49–71. https://doi.org/10.1016/j.pneurobio.2006.04.004

Pulvermüller, F., Shtyrov, Y., Kujala, T., & Näätänen, R. (2004). Word-specific cortical activity as revealed by the mismatch negativity. *Psychophysiology*, *41*(1), 106–112. https://doi.org/10.1111/j.1469-8986.2003.00135.x

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. (2013). *R: A language and environment for statistical computing*.

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. https://doi.org/10.1080/17470210903104412

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. *Categorical Perception: The Groundwork of Cognition*, 89–112.

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

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S., & von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology*, *24*(19), 2348–2353. https://doi.org/10.1016/j.cub.2014.08.048

Saarbrücken. (2019). In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Saarbr%C3%BCcken&oldid=188463051

Saarinen, J., Paavilainen, P., Schöger, E., Tervaniemi, M., & Näätänen, R. (1992). Representation of abstract attributes of auditory stimuli in the human brain. *NeuroReport*, *3*(12), 1149–1151. https://doi.org/10.1097/00001756-199212000-00030

Saloranta, A., Alku, P., & Peltola, M. S. (2020). Listen-and-repeat training improves perception of second language vowel duration: Evidence from mismatch negativity (MMN) and N1 responses and behavioral discrimination. *International Journal of Psychophysiology*, *147*, 72–82. https://doi.org/10.1016/j.ijpsycho.2019.11.005

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2015). Voice Identity Recognition: Functional Division of the Right STS and Its Behavioral Relevance. *Journal of Cognitive Neuroscience*, *27*(2), 280–291. https://doi.org/10.1162/jocn_a_00707

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2017). Comparison of the Working Memory Load in N-Back and Working Memory Span Tasks by Means of EEG Frequency Band Power and P300 Amplitude. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00006

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

Schneider, E., & Zuccoloto, A. (2007). E-prime 2.0 [Computer software]. *Pittsburg, PA: Psychological Software Tools*.

Scott, T. L., & Perrachione, T. K. (2019). Common cortical architectures for phonological working memory identified in individual brains. *NeuroImage*, *202*, 116096. https://doi.org/10.1016/j.neuroimage.2019.116096

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*.

Shahin, A. J., Backer, K. C., Rosenblum, L. D., & Kerlin, J. R. (2018). Neural Mechanisms Underlying Cross-Modal Phonetic Encoding. *The Journal of Neuroscience*, *38*(7), 1835–1849. https://doi.org/10.1523/JNEUROSCI.1566-17.2017

Sheehan, K. A., McArthur, G. M., & Bishop, D. V. M. (2005). Is discrimination training necessary to cause changes in the P2 auditory event-related brain potential to speech sounds? *Cognitive Brain Research*, *25*(2), 547–553. https://doi.org/10.1016/j.cogbrainres.2005.08.007

Shestakova, A., Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A., Sams, M., Ilmoniemi, R. J., & Näätänen, R. (2002). Abstract phoneme representations in the left temporal cortex: Magnetic mismatch negativity study. *NeuroReport*, *13*(14), 1813–1816. Scopus. https://doi.org/10.1097/00001756-200210070-00025

Shestakova, A., Huotilainen, M., Čeponien, R., & Cheour, M. (2003). Event-related potentials associated with second language learning in children. *Clinical Neurophysiology*, *114*(8), 1507–1512. https://doi.org/10.1016/S1388-2457(03)00134-2

Shtyrov, Y., Nikulin, V. V., & Pulvermüller, F. (2010). Rapid Cortical Plasticity Underlying Novel Word Learning. *Journal of Neuroscience*, *30*(50), 16864–16867. https://doi.org/10.1523/JNEUROSCI.1376-10.2010

Shtyrov, Y., & Pulvermüller, F. (2002). Neurophysiological evidence of memory traces for words in the human brain. *Neuroreport*, *13*(4), 521–525.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10365-z

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011a). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception, & Psychophysics*, *73*(4), 1195–1215. https://doi.org/10.3758/s13414-011-0096-8

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011b). 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

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

Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin*, *133*(5), 780–799. https://doi.org/10.1037/0033-2909.133.5.780

Spencer, K. M., & Polich, J. (1999). Poststimulus EEG spectral analysis and P300: Attention, task, and probability. *Psychophysiology*, *36*(2), 220–232. https://doi.org/10.1111/1469-8986.3620220

Steinberg, J., Truckenbrodt, H., & Jacobsen, T. (2011). Phonotactic constraint violations in German grammar are detected automatically in auditory speech processing: A human event-related potentials study: Preattentive phonotactic processing. *Psychophysiology*, *48*(9), 1208–1216. https://doi.org/10.1111/j.1469-8986.2011.01200.x

Steinmetzger, K., & Rosen, S. (2017). Effects of acoustic periodicity and intelligibility on the neural oscillations in response to speech. *Neuropsychologia*, *95*, 173–181. https://doi.org/10.1016/j.neuropsychologia.2016.12.003

Strange, W. (1989). Evolving theories of vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2081–2087. https://doi.org/10.1121/1.397860

Strauß, A., Kotz, S. A., Scharinger, M., & Obleser, J. (2014). Alpha and theta brain oscillations index dissociable processes in spoken word recognition. *Neuroimage*, *97*, 387–395. https://doi.org/10.1016/j.neuroimage.2014.04.005

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.

Sulpizio, S., Toti, M., Del Maschio, N., Costa, A., Fedeli, D., Job, R., & Abutalebi, J. (2019). Are you really cursing? Neural processing of taboo words in native and foreign language. *Brain and Language*, *194*, 84–92. https://doi.org/10.1016/j.bandl.2019.05.003

Sun, Y., Giavazzi, M., Adda-Decker, M., Barbosa, L. S., Kouider, S., Bachoud-Lévi, A.-C., Jacquemot, C., & Peperkamp, S. (2015). Complex linguistic rules modulate early auditory brain responses. *Brain and Language*, *149*, 55–65. https://doi.org/10.1016/j.bandl.2015.06.009

Sussman, E. S., Chen, S., Sussman-Fort, J., & Dinces, E. (2014). The Five Myths of MMN: Redefining How to Use MMN in Basic and Clinical Research. *Brain Topography*, *27*(4), 553–564. https://doi.org/10.1007/s10548-013-0326-6

Tamminen, H., Peltola, M. S., Kujala, T., & Näätänen, R. (2015). Phonetic training and non-native speech perception—New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures. *International Journal of Psychophysiology*, *97*(1), 23–29. https://doi.org/10.1016/j.ijpsycho.2015.04.020

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics*, *77*(5), 1674–1684. https://doi.org/10.3758/s13414-015-0854-0

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detaila). *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

Tremblay, K., Kraus, N., Carrell, T. D., & McGee, T. (1997). Central auditory system plasticity: Generalization to novel stimuli following listening training. *The Journal of the Acoustical Society of America*, *102*(6), 3762–3773. https://doi.org/10.1121/1.420139

Tuninetti, A., Chládková, K., Peter, V., Schiller, N. O., & Escudero, P. (2017). When speaker identity is unavoidable: Neural processing of speaker identity cues in natural speech. *Brain and Language*, *174*, 42–49. https://doi.org/10.1016/j.bandl.2017.07.001

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

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The Neural Integration of Speaker and Message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591. https://doi.org/10.1162/jocn.2008.20054

Virtala, P., Berg, V., Kivioja, M., Purhonen, J., Salmenkivi, M., Paavilainen, P., & Tervaniemi, M. (2011). The preattentive processing of major vs. Minor chords in the human brain: An event-related potential study. *Neuroscience Letters*, *487*(3), 406–410. https://doi.org/10.1016/j.neulet.2010.10.066

Wöstmann, M., Lim, S.-J., & Obleser, J. (2017). The Human Neural Alpha Response to Speech is a Proxy of Attentional Control. *Cerebral Cortex*, *27*(6), 3307–3317. https://doi.org/10.1093/cercor/bhx074

Wronka, E., Kaiser, J., & Coenen, A. M. L. (2012). Neural generators of the auditory evoked potential components P3a and P3b. *Acta Neurobiologiae Experimentalis*, *72*(1), 51–64.

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

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the Brain to Weight Speech Cues Differently: A Study of Finnish Second-language Users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, *15*(1), 88–99. https://doi.org/10.1037/0882-7974.15.1.88

Yordanova, J., Devrim, M., Kolev, V., Ademoglu, A., & Demiralp, T. (2000). Multiple time-frequency components account for the complex functional reactivity of P300: *NeuroReport*, *11*(5), 1097–1103. https://doi.org/10.1097/00001756-200004070-00038

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271. https://doi.org/10.1016/j.tics.2013.04.004

Yue, J., Bastiaanse, R., & Alter, K. (2014). Cortical plasticity induced by rapid Hebbian learning of novel tonal word-forms: Evidence from mismatch negativity. *Brain and Language*, *139*, 10–22. https://doi.org/10.1016/j.bandl.2014.09.007

Zachau, S., Rinker, T., Körner, B., Kohls, G., Maas, V., Hennighausen, K., & Schecker, M. (2005). Extracting rules: Early and late mismatch negativity to tone patterns. *NeuroReport*, *16*(18). https://doi.org/10.1097/00001756-200512190-00009

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1). https://doi.org/10.1038/srep11475

Zäske, R., Awwad Shiekh Hasan, B., & Belin, P. (2017). It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. *Cortex*, *94*, 100–112. https://doi.org/10.1016/j.cortex.2017.06.005

Zäske, R., Volberg, G., Kovacs, G., & Schweinberger, S. R. (2014). Electrophysiological Correlates of Voice Learning and Recognition. *Journal of Neuroscience*, *34*(33), 10821–10831. https://doi.org/10.1523/JNEUROSCI.0581-14.2014

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., Peng, G., & Wang, W. S.-Y. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage*, *124*, 536–549. https://doi.org/10.1016/j.neuroimage.2015.08.064

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

[i] The work presented in this dissertation has been co-financed by Fondazione Cassa di Risparmio di Trento e Rovereto (CARITRO).