Early differentiation of memory retrieval processes for newly learned voices and phonemes as indexed by the MMN

Giuseppe Di Dona[1], Michele Scaltritti[1], and Simone Sulpizio[2]

[1] Dipartimento di Psicologia e Scienze Cognitive, Università degli Studi di Trento, Corso Bettini 84, 38068 – Rovereto (TN), Italy. e-mail: giuseppe.didona@unitn.it; michele.scaltritti@unitn.it

[2] Dipartimento di Psicologia, Università degli Studi di Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 – Milano (MI), Italy. e-mail: simone.sulpizio@unimib.it

Author Note

Correspondence concerning the article should be addressed to: Giuseppe Di Dona, Dipartimento di Psicologia e Scienze Cognitive, Università degli Studi di Trento, Corso Bettini 84, 38068 – Rovereto (TN), Italy. e-mail: giuseppe.didona@unitn.it

Abstract

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.

Keywords: Speech perception; Voice perception; EEG; Mismatch Negativity (MMN); Memory retrieval; Enhancement effect.

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, Fecteau, & Bédard, 2004) consider phonemes and voices as the fundamental information units for speech perception and talker identification, respectively (Formisano, De Martino, Bonte, & Goebel, 2008). Phonemes can be described on the basis of their first and second formant frequencies (F1 and F2) (Obleser, Elbert, Lahiri, & Eulitz, 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, Awwad Shiekh Hasan, & Belin, 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, Francis, and Melara (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, Sommers, & Pisoni, 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, Tian, Woods, & Poeppel, 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 (Formisano et al., 2008), 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,

Paavilainen, Rinne, & Alho, 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, Jacobsen, & Winkler, 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. 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, Dougherty, McLaughlin, & Lember, 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 /piː/ 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 /pyː/ 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, Blumstein, & Luthra, 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, Hämäläinen, Leppänen, Shu, & Astikainen, 2019).

2. Method

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 (see section 2.3.2) 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 Supplementary Materials). 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.

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 (see section 2.3.2) 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 and 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 1.

Table 1. *Mean values and Standard Deviations (SD) of fundamental frequency (F0) and dispersion across the fourth and the fifth formant (F4/F5) for every talker*[1]

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)

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

[1] The data of the talker that was excluded for the high level of external noise is not reported in Table 1.

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 Supplementary Materials). 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 2. 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. *Physical characteristics of Standard and Deviant stimuli used in the EEG experiment.*[2]

	Standard stimulus	Deviant Stimulus	
		phoneme-change condition	voice-change condition

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

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

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.

2.3.1 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 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 /pi:/ produced by Talker 1 was used as deviant, while in the phoneme-change condition the syllable /py:/ 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.

2.3.2 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, Love, & Latinus, 2017; Latinus, Crabbe, & Belin, 2011). In this kind of trainings, the use of multiple talkers can provide an acoustic space in which voices can be physically represented (Andics, McQueen, & Petersson, 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 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).

2.3.3 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, Peltola, Kujala, and Näätänen (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 /y:/. The training started with a familiarization phase, during which, via headphones, participants could listen to the /pi:/ and /py:/ 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 (/pi:/ or /py:/) 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.

2.3.4 Post-training EEG session.

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

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.

2.5 Statistical Analyses

2.5.1 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.

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.

2.5.2 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, Nikulin, & Pulvermüller, 2010; Tamminen et al., 2015). The mean peak latency of MMN was separately measured for the phoneme-change and the voice change conditions (Gu, Zhang, Hu, & Zhao, 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, Truckenbrodt, & Jacobsen, 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. Results

3.1 Behavioural data

3.1.1 Talker-Identification Training

The pre-training and post-training accuracy scores are represented in Figure 1A, 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 (β = 0.27, *SE* = 0.10, *z* = 2.51, *p* = .011) and Talker 3 (. β = 0.90, SE = 0.10, z = 8.72, p < .001).

3.1.2 Syllable-Identification Training

The pre-training and post-training identification responses and goodness ratings are represented in Figure 1C, 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.

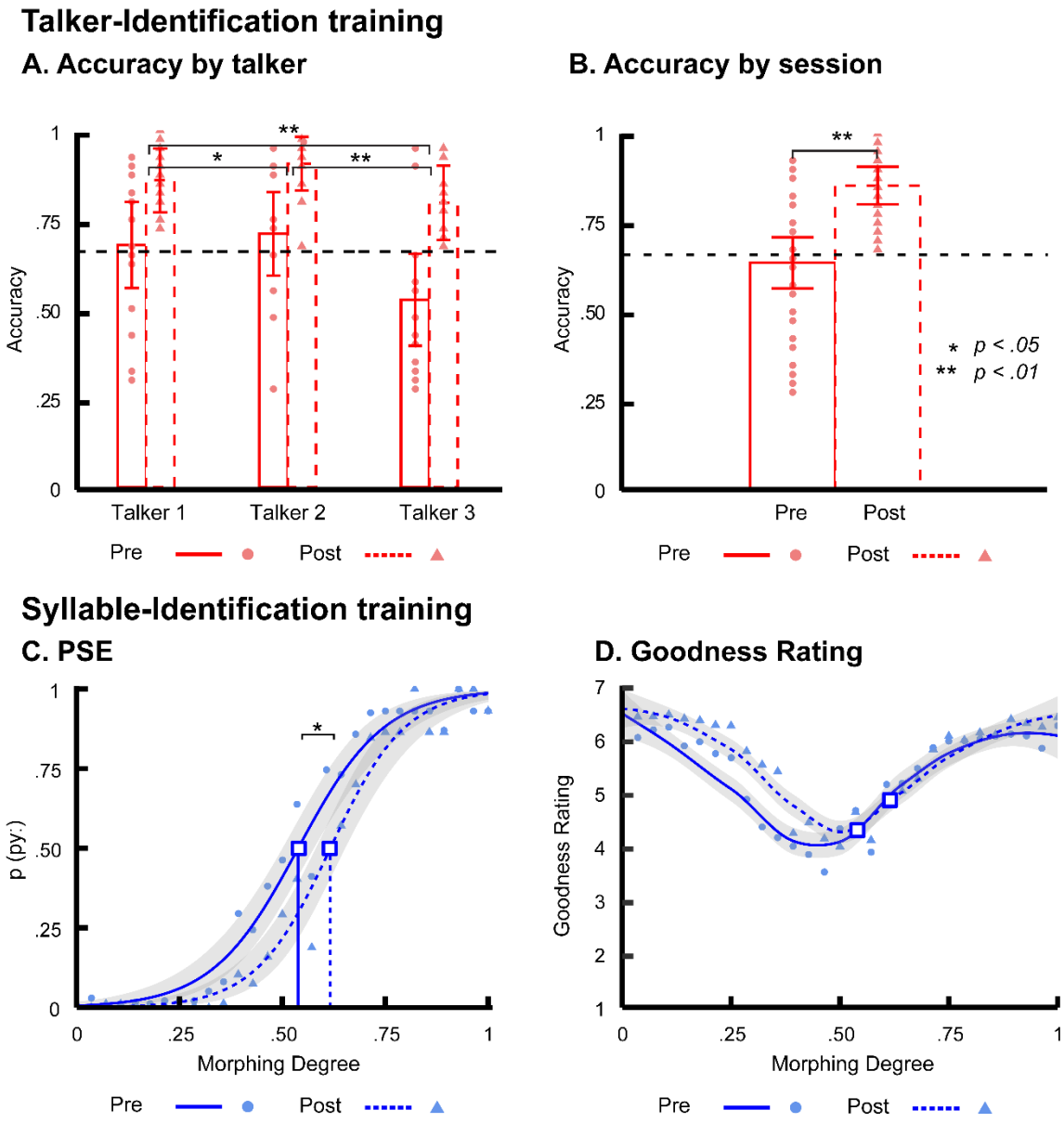

Figure 1. Behavioural results of the talker-identification (red) and the syllable-identification training (blue) in the pre- (continuous line) and in the post-training (dashed line) sessions. (A) and (B) show the

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.

3.2 EEG

3.2.1 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 Supplementary Materials) confirming that MMN was successfully elicited. MMN waveforms are displayed in Figure 2A, 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).

3.2.2 The enhancement effect

The analyses on the enhancement effect (Figure 2C) 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).

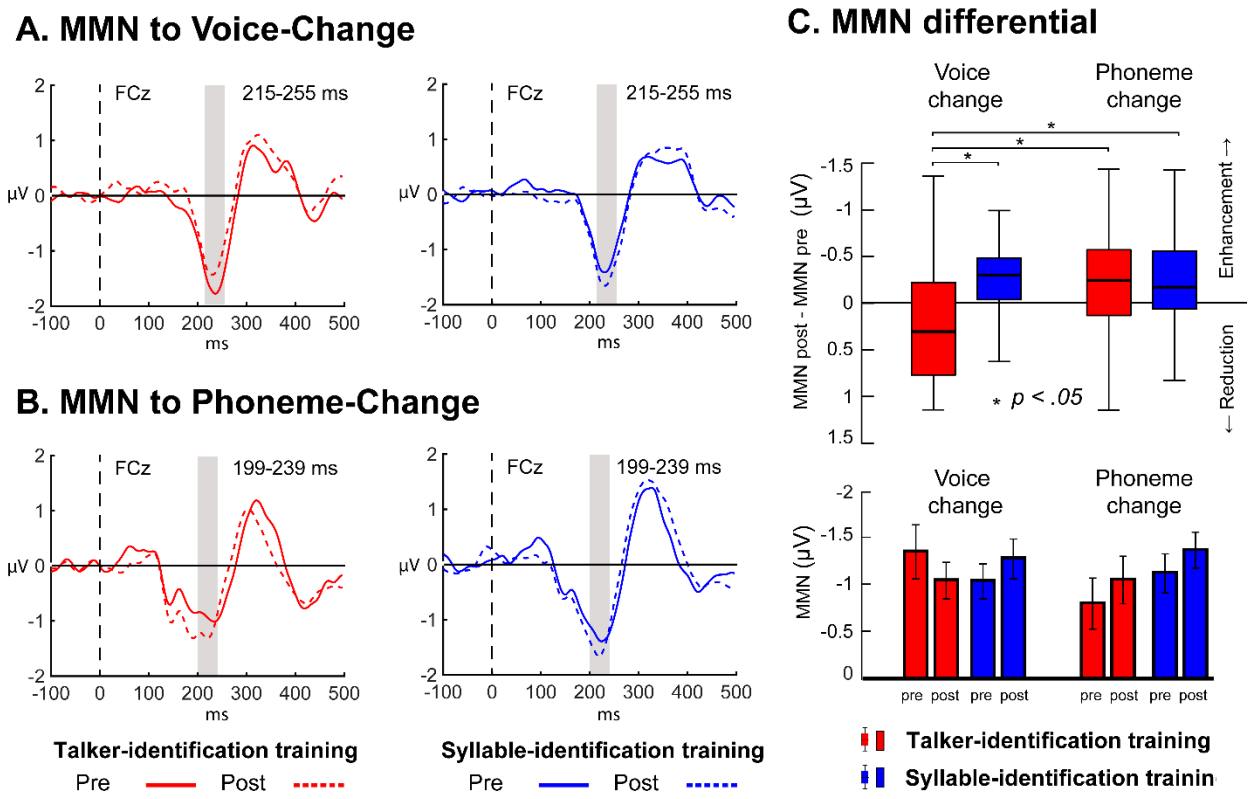

Figure 2. MMN for the different conditions in the group enrolled in the talker-identification (red) and the group enrolled in the syllable-identification training (blue). MMN was calculated in the pre- (continuous line) and in the post-training (dashed line) sessions at a representative channel (FCz) 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 (± SEM) 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).

3.2.3 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 3). 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).

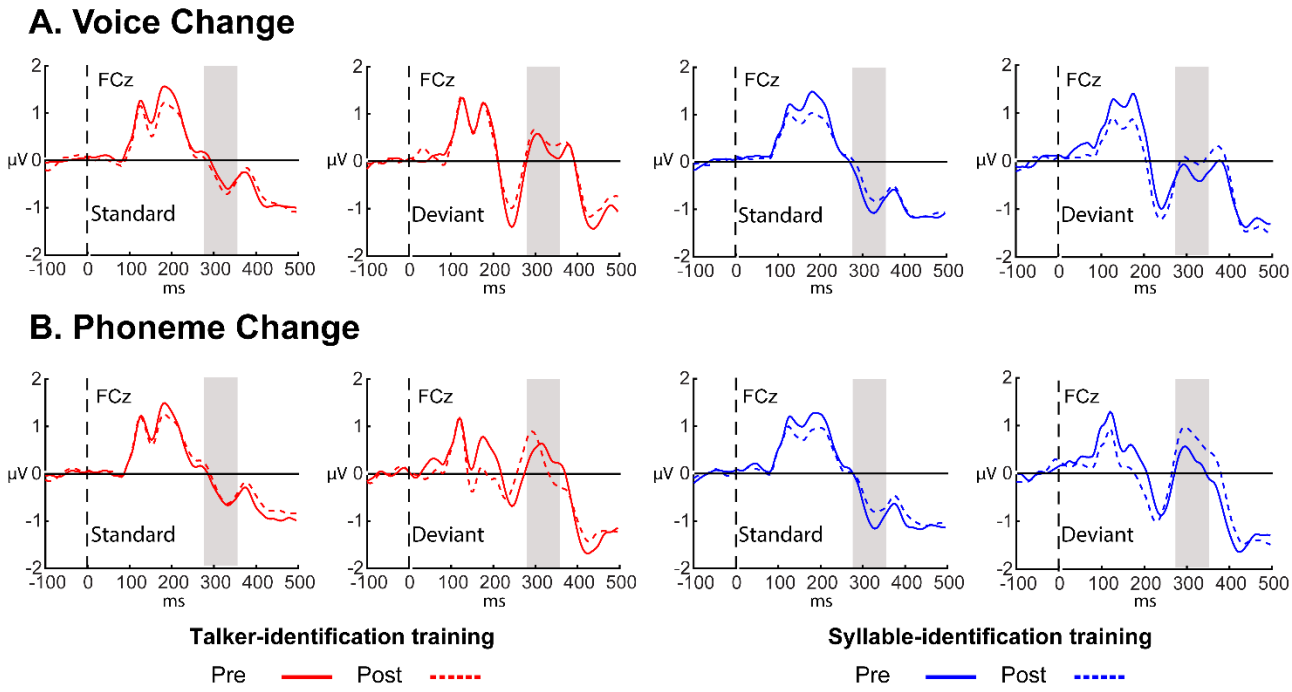

Figure 3: P3a for the different conditions in the group enrolled in the talker-identification (red) and the group enrolled in the syllable-identification training (blue). ERPs for standard and deviant events calculated in the pre- (continuous line) and the post training (dashed line) sessions at a representative 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.

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, Chládková, Peter, Schiller, & Escudero, 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.

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.[3]

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, Alku, and Peltola, 2020; Tamminen et al., 2015; Ylinen et al., 2010). Considering the replication of these findings, new phonemes appear to be learned

[3] 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).

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, Kaiser, & Coenen, 2012), and the target status – i.e., the P3a is larger for target than non-target stimuli (Comerchero & Polich, 1999).

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, Roelofs, & Meyer, 1999), phonetic encoding and articulation (Jongman, Roelofs, & Lewis, 2020; Jongman, Roelofs, & Meyer, 2015) – 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/).

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, McAleer, Bestelmeyer, & Belin, 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, Knight, & McGettigan, 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, McQueen, & Petersson, 2013; Andics et al., 2010). 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., 2014) 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 in 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, Dougherty, Lember, & Perrachione, 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, Kallioinen, Klintfors, & Lindh, 2013; Plante-Hébert, Boucher, & Jemel, 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.

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 (Liu, Ong, Tuninetti, & Escudero, 2018; Yue, Bastiaanse, & Alter, 2014). Contrastively, other studies showed that while different training tasks can modulate the amplitude of MMN (Kraus et al., 1995; Tremblay, Kraus, Carrell, & McGee, 1997), passive exposure alone is not sufficient to do so (Elmer, Hausheer, Albrecht, & Kühnis, 2017; Sheehan, McArthur, & Bishop, 2005). As described in Kurkela, Hämäläinen, Leppänen, Shu, & Astikainen, (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.

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, Kiebel, Maess, and von Kriegstein (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.

Bonte, Valente and Formisano (2009) suggested that the existence of a "default modality" for speech processing that mainly analyses linguistic information independently of the talker but does not do the opposite. As a matter of fact, the auditory system is able to automatically extract formant information from vowels while abstracting from continuously varying voice information (Jacobsen, Schroger, & Alter, 2004), whereas there is no evidence that it also automatically extracts phonemic-invariant vocal information. Listeners are indeed able to actively recognize voices notwithstanding the variability in the speech tokens. However, they might use different processes to store and retrieve voice-independent phoneme representations and phoneme-independent voice representations, but the retrieval processes of the latter may not be completely automatic.

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.

Acknowledgements: This study was conducted as part of G.D.D.'s PhD project, funded by the University of Trento. We are grateful to Giovanni Orlando for his help with data acquisition.

Declaration of Competing Interest: none.

Data Statement: The datasets generated and/or analyzed during the present study are not publicly available because we did not obtain consent for publication from the participants. Data are available from the corresponding author on reasonable request.

References

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

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Bolker, M. B. (2015). Package 'lme4.' *Convergence*, *12*(1), 2.

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.

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

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

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

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

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

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

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

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

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

Hamburg. (2019). In *Wikipedia*. Retrieved May 4, 2019, from https://de.wikipedia.org/w/index.php?title=Hamburg&oldid=188211539

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

Jacobsen, T., Schroger, 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

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

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

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

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. IEEE.

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

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

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

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*.

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

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

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

McGuire, G. L., & Babel, M. (2020). Attention to Indexical Information Improves Voice Recall. *Interspeech 2020*, 1595–1599. ISCA. 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*.

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

Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language*, *165*, 33–44. https://doi.org/10.1016/j.bandl.2016.11.001

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., … 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

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

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

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

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

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

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

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

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

Pulvermüller, F., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., … 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

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

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.

Saarbrücken. (2019). In *Wikipedia*. Retrieved May 11, 2019, from https://de.wikipedia.org/w/index.php?title=Saarbr%C3%BCcken&oldid=188463051

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

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

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

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., … Näätänen, R. (2002). Abstract phoneme representations in the left temporal cortex: Magnetic mismatch negativity study. *NeuroReport*, *13*(14), 1813–1816. https://doi.org/10.1097/00001756-200210070-00025

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.

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

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

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

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

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.

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

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

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

Supplementary Materials

Supplementary Table 1. 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 deviation are in brackets.

	Groups		Whole sample
	Talker-identification training	Syllable-identification training	
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)
Hours of daily use of L2	4.63 (3.17)	3 (2.95)	3.89 (3.28)
L2 proficiency	6.85 (0.69)	7 (1.26)	6.92 (0.95)

Supplementary Table 2. 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

Supplementary Table 3. 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

Supplementary Table 4. 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) ***

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