Behavioral/Cognitive

Electrophysiological Correlates of Voice Learning and Recognition

Romi Zäske,[1] Gregor Volberg,[2] Gyula Kovács,[3] and Stefan Robert Schweinberger[1]

[1]Department for General Psychology and Cognitive Neuroscience, Institute of Psychology, [2]Department for Psychology, Institute of Psychology, and [3]Institute of Psychology, Friedrich Schiller University of Jena, 07743 Jena, Germany

Listeners can recognize familiar human voices from variable utterances, suggesting the acquisition of speech-invariant voice representations during familiarization. However, the neurocognitive mechanisms mediating learning and recognition of voices from natural speech are currently unknown. Using electrophysiology, we investigated how representations are formed during intentional learning of initially unfamiliar voices that were later recognized among novel voices. To probe the acquisition of speech-invariant voice representations, we compared a "same sentence" condition, in which speakers repeated the study utterances at test, and a "different sentence" condition. Although recognition performance was higher for same compared with different sentences, substantial voice learning also occurred for different sentences, with recognition performance increasing across consecutive study-test-cycles. During study, event-related potentials elicited by voices subsequently remembered elicited a larger sustained parietal positivity (~250–1400 ms) compared with subsequently forgotten voices. This difference due to memory was unaffected by test sentence condition and may thus reflect the acquisition of speech-invariant voice representations. At test, voices correctly classified as "old" elicited a larger late positive component (300–700 ms) at Pz than voices correctly classified as "new." This event-related potential OLD/NEW effect was limited to the same sentence condition and may thus reflect speech-dependent retrieval of voices from episodic memory. Importantly, a speech-independent effect for learned compared with novel voices was found in beta band oscillations (16–17 Hz) between 290 and 370 ms at central and right temporal sites. Our results are a first step toward elucidating the electrophysiological correlates of voice learning and recognition.

Key words: ERPs; learning; memory; oscillations; speech; voice

Introduction

The ability to recognize voices is crucial for social interactions, particularly when visual information is absent. Whereas recognizing unfamiliar voices is error-prone (Clifford, 1980), familiar voice recognition is remarkably accurate across variable utterances (Skuk and Schweinberger, 2013). However, the neural mechanisms underlying the acquisition of representations of familiar voices remain poorly understood.

Understanding the transition from unfamiliar to familiar voices during learning is important, because processing familiar voices can be selectively impaired and relies on partially distinct cortical areas (Van Lancker and Kreiman, 1987; Kriegstein and Giraud, 2004). Temporal voice areas (Belin et al., 2000) were reported to process acoustical information regardless of voice familiarity for simple vowels (Latinus, Crabbe, and Belin, 2011). In contrast, right inferior frontal areas seemed sensitive to learned voices that were perceptually identical, suggesting their implication in voice identity processing (but see Andics et al., 2010). This notion of hierarchical processing is consistent with traditional person perception models, which provide a conceptual framework for the present research (Belin et al., 2011). Accordingly, after acoustic analysis, voices are compared with long-term voice representations. Crucially, for recognition across utterances, these representations are thought to respond to familiar voices independent of speech content. Therefore, voice learning entails acquiring representations independently of low-level information, similar to face learning (Yovel and Belin, 2013).

For faces, distinct event-related potentials (ERPs) marking these subprocesses include the occipitotemporal N170, which indexes face encoding preceding recognition (Eimer, 2011). The occipitotemporal N250 reflects the acquisition and activation of representations for individual faces (Kaufmann, Schweinberger, and Burton, 2009; Zimmermann and Eimer, 2013). A subsequent centroparietal positivity (from ~300 ms) may reflect retrieval of episodic and/or semantic person information (Schweinberger and Burton, 2003; Wiese, 2012); a similar positivity emerges for correctly recognized versus new nonface items (OLD/NEW effects; Friedman and Johnson, 2000). Finally, ERP "differences due to subsequent memory"

Received Feb. 11, 2014; revised May 27, 2014; accepted June 17, 2014.

Author contributions: R.Z. and S.R.S. designed research; R.Z. performed research; R.Z. and G.V. analyzed data; R.Z., G.V., G.K., and S.R.S. wrote the paper.

This work was funded by the Programm zur Förderung der Drittmittelfähigkeit (2012) of the University of Jena and by the Deutsche Forschungsgemeinschaft (Grant ZA 745/1-1, KO 3918/1-2; 2-1). We thank Leonie Fresz, Achim Hötzel, Christoph Klebl, Katrin Lehmann, Carolin Leistner, Constanze Mühl, Finn Pauls, Marie-Christin Perlich, Johannes Pfund, Mathias Riedel, Saskia Rudat, and Meike Wilken for stimulus acquisition and editing and Bettina Kamchen for help with data acquisition.

The authors declare no competing financial interests.

Correspondence should be addressed to Romi Zäske, Department for General Psychology and Cognitive Neuroscience, Institute of Psychology, Friedrich Schiller University of Jena, Am Steiger 3/1, 07743 Jena, Germany. E-mail: romi.zaeske@uni-jena.de.

DOI:10.1523/JNEUROSCI.0581-14.2014

0270-6474/14/3410821-11$15.00/0

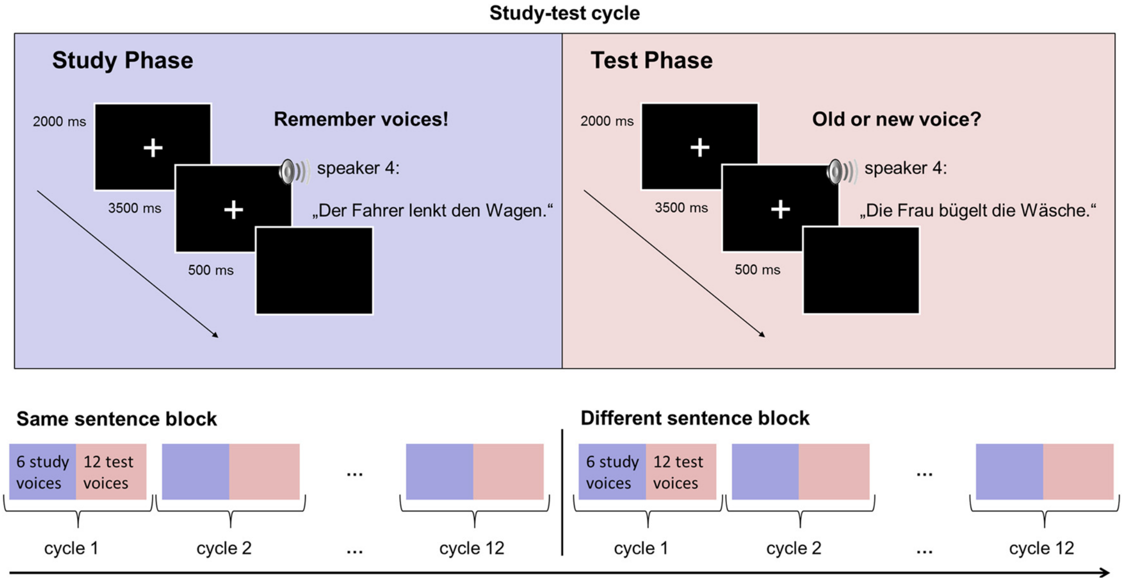

Figure 1. Top, Trial procedure for one study-test cycle. The example shows an "old" test voice presented with a different sentence than at study. Bottom, The 12 study-test cycles for same and different sentence blocks, respectively.

(Dms) indicate successful face encoding into memory (Sommer, Schweinberger, and Matt, 1991).

Auditory equivalents to these components are less well established. Whereas early auditory potentials (N1–P2) are sensitive to sound repetitions, including voices (Schweinberger, 2001; Schweinberger et al., 2011b), the "frontotemporal positivity to voice" (Charest et al., 2009) discriminates voices from other sounds from ~160 ms. Potential markers of voice recognition include the frontocentral mismatch negativity (~200 ms), which may index detection of familiar among unfamiliar voices (Beauchemin et al., 2006), and the parietal P3, which is sensitive to voice repetitions (from ~300 ms) despite phonetic changes (Schweinberger et al., 2011b). However, electrophysiological correlates of voice learning are currently unknown.

We investigated the acquisition of familiar voice representations in a recognition memory paradigm. Despite some interdependence of speech and speaker perception (Perrachione and Wong, 2007), familiar voices can be recognized across utterances (Skuk and Schweinberger, 2013). To disentangle speech-dependent and speech-invariant recognition, we tested voices with either the same or a different sentence than presented at study. Finally, because brain oscillations have been related to memory (Düzel, Penny, and Burgess, 2010), we analyzed performance, ERPs, and time-frequency data in electrophysiological responses to studied and novel voices.

Materials and Methods

Participants. Twenty-four student participants (12 female, all right-handed and native speakers of German, mean age 24.0 years, range 19–32) contributed data. None reported hearing difficulties, which was confirmed by a digital audio test by Cotral (Version 1.02B, available online), and none reported prior familiarity with any of the voices used in the experiment. Data from 8 additional participants were excluded due to extensive EEG artifacts ($n = 6$) or familiarity with any of the voices before the experiment ($n = 2$). Participants received a payment of 7.50 Euros or course credit, as well as an additional performance-based incentive of 1–3 Euros. All gave written informed consent. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Faculty Ethics Committee of the Friedrich Schiller University of Jena.

Stimuli. Stimuli were recordings from 48 adult native speakers of German (24 female, age 18–25 years, mean age 22.0 years). All 48 speakers uttered 12 German sentences (6 of which started with the article "der" and "die," respectively), resulting in 576 different stimuli. All sentences had the same syntactic structure and consisted of seven or eight syllables; for example, "Der Fahrer lenkt den Wagen" ("the driver steers the car") and "Die Kundin kennt den Laden" ("the customer knows the shop"). Utterances were chosen based on neutral valence ratings of written content as determined in a pilot questionnaire. Twelve raters (9 female, mean age 28 years, range 21–43) who did not participate in the main experiment judged 61 written sentences for their emotional valence as follows: negative (−1), neutral (0), or positive (+1). Twelve sentences with the most neutral ratings (mean ratings between −0.2 and 0.2) were chosen as stimuli. Moreover, speakers were instructed to intonate sentences as emotionally neutral as possible. To standardize intonation and sentence duration and to keep regional accents to a minimum, speakers were encouraged to mimic as closely as possible a prerecorded model speaker (first author) presented via loudspeakers. Each sentence was recorded 4–5 times in a quiet and semi-anechoic room by means of a Sennheiser MD 421-II microphone with pop protection and a Zoom H4n audio interface (16 bit resolution, 48 kHz sampling rate, stereo). The recordings with the least artifacts, the least background noise, and with clear pronunciation were chosen as stimuli. Using PRAAT software (Boersma and Weenink, 2001), voice recordings were cut to contain one sentence starting exactly at plosive onset of "Der"/"Die." Voice recordings were then resampled to 44.1 kHz, converted to mono, and RMS normalized to 70 dB. Voices were set into eight random sets of six speakers (three female) to serve as study and test voices in the experiment. Mean sentence duration was 1719 ms (SD = 114, range 1492–2048 ms) and did not differ significantly between speaker sets (univariate ANOVA with the factor set, $F < 1$). As practice stimuli, voices of four additional speakers (two female) uttering three sentences not used in the main experiment were used. Stimuli were presented diotically via Sennheiser HD 25–1 II headphones with an approximate peak intensity of 60 dB(A) as determined with a Brüel and Kjær Precision Sound Level Meter Type 2206.

Procedure. Participants were tested individually in a dimly lit, sound-attenuated booth. Instructions were presented in writing on a computer screen to minimize interference from the experimenter's voice. To maintain participants' motivation, they were told they could earn an additional performance-dependent bonus of 1–3 Euros.

The experiment was divided into two blocks, each comprising 12 study-test cycles (Fig. 1). Each cycle consisted of a study phase and a subsequent test phase. In each study phase of a given block, participants heard the same set of six voices in random order all uttering the same sentence once. Participants were instructed to remember these voices for a subsequent test. Study trials started with a white fixation cross on a black background (2000 ms) announcing a voice sample. The fixation cross remained on screen for another 3500 ms after voice onset and was followed by a blank screen (500 ms), after which the next study trial started. Total trial duration was thus 6000 ms at all times. No responses were required from the participants on study trials.

During the test phase, participants made old/new decisions to 12 voices all uttering the same sentence, which were presented in randomized order. Of those voices, six (old) voices had been presented during the study phase and six (new) voices were novel. The structure of test trials was identical to study trials except that responses were required starting from voice onset until 3500 ms thereafter. Depending on experimental block, test voices either repeated the sentence from the study phase or uttered a different sentence (in which case, sentences always started with a different article). Importantly, participants were instructed to respond only to voice identity (i.e., regardless of sentence content) as quickly and accurately as possible. Responses were entered via “d” and “l” keys of a standard PC keyboard using index fingers of both hands. Assignment of keys to “old” or “new” responses was counterbalanced between participants. Responses later than 3500 ms were followed by the feedback “Please respond faster” (500 ms), which replaced the blank screen.

To facilitate learning, the same set of six study voices was repeated in 12 consecutive cycles of each experimental block using a new study sentence on each cycle. After the first 12 cycles (first block), a new set of six study voices was used in the remaining 12 cycles (second block). With 12 sentences being available overall, the same 12 sentences were used during study phases of the same and different sentence blocks. With respect to test phases, sentence conditions (same/different) varied between blocks, whereas test voice conditions (old/new) varied randomly within blocks. The order of sentence conditions was counterbalanced between participants. Because two of eight speaker sets were used as study voices in the first and second block, a total of six speaker sets were available as “new” voices. Each set of “new” test voices was used in two consecutive cycles. Therefore, after 12 cycles (first block), all of the six speaker sets had been encountered twice. To minimize spurious recognition of these “new” voices in two consecutive cycles, we used different test sentences across cycles. For the remaining 12 cycles (second block), the same six “new” test speaker sets were used.

The order of study and test sentences was balanced across participants, as was the assignment of study and test speaker sets to experimental conditions (test voice condition and sentence condition). Furthermore, we reasoned that participants’ learning performance may depend on the phonetic variability of speech material. Therefore, we ensured that, for a given study voice, participants would hear exactly the same number of different sentences ($n = 12$) across study-test cycles of the same and different sentence blocks. In addition to differences in overall performance between blocks (i.e., the same sentence condition vs the different sentence condition), we were also interested in whether there would be performance improvements from the first half of each block to the second half. Therefore, we controlled for sentence variability not only across sentence condition blocks, but also across block halves. Accordingly, whereas all test sentences in the “same sentence block” had by definition occurred in the preceding study block, all test sentences in the “different sentence” block had occurred or would occur as study sentences in another cycle of the respective block and of the respective block half. In this way, we avoided a possible confound of sentence condition and phonetic variability.

Altogether, there were 144 study trials (two blocks × 12 cycles × six trials) and 288 test trials (two blocks × 12 cycles × 12 trials). Individual breaks were allowed after every cycle. In total, the experiment lasted ~60 min.

After the experiment, participants answered a short questionnaire assessing the amount of verbal contact to people on a 6-point Likert scale where 1 = “very often” and 6 = “very rarely.” We further asked how well participants thought they would recognize familiar and unfamiliar voices in everyday life situations without any visual cues, where 1 = “very well” and 6 = “very poorly,” and how well they thought they had recognized voices in the present experiment both compared with their own performance in real-life situations and compared with other participants of the experiment. This was to test whether actual recognition performance in the experiment would be correlated with the amount of verbal contact or with subjective voice recognition abilities.

EEG recordings. The electroencephalogram (EEG) was recorded using a 64-channel Biosemi Active II system at electrode positions Fp1, FT9, AF3, F1, F3, F5, F7, FT7, TP9, FC3, FC1, C1, C3, C5, T7, TP7, PO9, CP3, CP1, P1, P3, O9, P7, P9, PO7, PO3, O1, Iz, Oz, POz, Pz, CPz, Fpz, Fp2, FT10, AF4, Afz, Fz, F2, F4, F6, F8, FT8, TP10, FC4, FC2, FCz, Cz, C2, C4, C6, T8, TP8, PO10, CP4, CP2, P2, P4, O10, P8, P10, PO8, PO4, and O2 (according to the extended international 10/20 system). Note that the Biosemi system uses a combined ground/reference circuit (http://www.biosemi.com/faq/cms&drl.htm). In addition, the horizontal electrooculogram (EOG) was recorded from the outer canthi of the eyes and the vertical EOG was recorded bipolarly from above and below the left eye. The EEG was recorded continuously at a sampling rate of 512 Hz (bandwidth: DC to 120 Hz).

ERP analysis. For the ERP analysis, we created epochs offline (−200 to 1500 ms relative to voice onset) both for study and test trials. Ocular artifact correction was computed automatically with BESA 5.1.8.10 (MEGIS Software) and data were recalculated to average reference. Any remaining artifacts such as drifts were removed manually based on visual inspection. Thresholds for subsequent automated artifact rejection were 100 μV for amplitude, 75 μV for gradient, and 0.1 μV for low signal. Trials with missing responses were excluded as well. Trials were averaged according to four study conditions and four test conditions depending on voice recognition performance. Study conditions for the Dm analysis (subsequent hits − same test sentence; subsequent misses − same test sentence; subsequent hits − different test sentence; and subsequent misses − different test sentence) had average trial numbers (SEM) of 47.3 (1.3), 18.0 (1.1), 43.8 (1.9), and 22.6 (1.8), respectively. For the OLD/NEW analysis all test conditions (same test sentence − hits; same test sentence − correct rejections; different test sentence − hits; and different test sentence − correct rejections) had a minimum number of 16 trials per participant. Average trial numbers (SEM) were 47.2 (1.2), 45.3 (1.6), 44.9 (2.0), and 44.1 (1.8) for each test condition, respectively. Averaged ERPs were low-pass filtered at 20 Hz with a zero phase shift digital filter.

Time-frequency analyses. For the time-frequency analysis, the continuous data were segmented into epochs from −2500 to 2500 ms relative to the voice onset. Only trials with correct behavioral responses were used. Compared with the ERP preprocessing, the segments were longer and so contained more artifacts, which would have led to high rejection rates when using an automated trial rejection procedure. Therefore, the strategy for controlling artifacts was different from that used for ERPs. In a first step, trials with movement artifacts or electrode artifacts were removed. In a second step, an independent component analysis was computed on the precleaned data (Delorme and Makeig, 2004). Components that could be identified as artifactual were removed from the data. The main sources of artifacts were eye blinks, eye movements, and muscle activity. After back projecting the remaining components into EEG signal space, trials containing residual artifacts were identified by visual inspection and then removed from the data. Finally, the cleaned data were re-referenced to an average reference value.

Data were filtered by convolution in the time domain from 1 to 30 Hz in steps of 1 Hz and 10 ms. To that end, a filter containing seven cycles of the target frequency was multiplied with a data segment of the same length. Beforehand, the data segment was multiplied with a Hanning window to reduce the temporal smearing of the filter response. The resulting power estimates were baseline corrected by computing the percentage of signal change relative to the activity from −500 to 0 ms relative to the voice onset. The time-frequency decomposition and data analysis were accomplished with custom routines and the Fieldtrip toolbox (Oostenveld et al., 2011) for MATLAB environment (The Mathworks). The average number of trials (SEM) were 47.0 (1.2), 44.1 (1.7), 42.9 (2.0), and 42.1 (1.6) for the conditions old voice/same sentence, new voice/same sentence, old voice/different sentence, and new voice/different sentence, respectively.

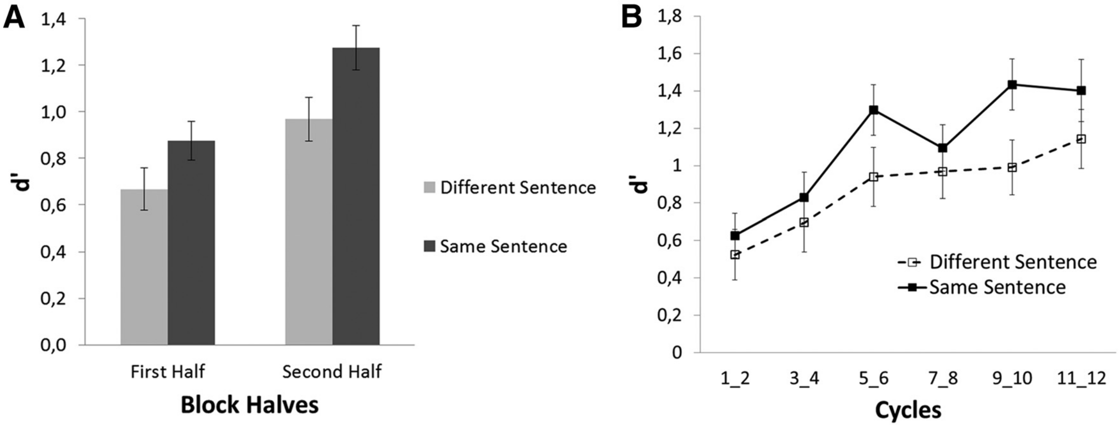

Figure 2. Mean d' for voice recognition in the same and different sentence conditions for both block halves (***A***) and collapsed over pairs of consecutive study-test cycles (***B***). Error bars indicate SEM.

Table 1. Accuracies for hits and CR, correct RT, d', and C for block halves, sentence conditions, and voice conditions

		Old voices		New voices			
Block halves	Sentence condition	Hits	RT (ms)	CR	RT (ms)	d'	C
First	Same	69.6% (1.7)	1574 (72)	63.1% (2.9)	1608 (75)	0.88 (0.08)	0.09 (0.05)
	Different	61.8% (2.6)	1494 (84)	62.8% (3.0)	1562 (73)	0.67 (0.09)	−0.02 (0.06)
Second	Same	75.3% (2.0)	1471 (71)	70.1% (2.5)	1447 (65)	1.28 (0.10)	0.07 (0.05)
	Different	69.8% (2.9)	1495 (79)	65.5% (2.5)	1509 (79)	0.97 (0.10)	0.07 (0.06)

SEMs are shown in parentheses.

Results

Behavioral and EEG data were submitted to ANOVA and t tests. Where appropriate, epsilon corrections for heterogeneity of covariances (Huynh and Feldt, 1976) were performed throughout.

Performance

Errors of omission and responses faster than 200 ms from voice onset were excluded (0.4% of responses). To test recognition performance as a function of voice conditions, we collapsed six consecutive cycles of each test sentence block to obtain two block halves for each sentence condition. We refrained from a more fine-grained statistical analysis of learning curves due to the low number of trials available in each cycle. However, d' as a function of sentence condition and cycle pairs (data collapsed across two consecutive cycles) is depicted in Fig. 2*B*. We then analyzed signal detection parameters d' and response bias (C) with repeated measures on block half (first/second) and sentence condition (same/different). Hit rates and false alarm rates equal to zero or one were adjusted according to MacMillan and Kaplan (1985). ANOVAs for accuracies and correct reaction times (RTs) were performed analogously, but with the additional within-subjects factor voice condition (old/new). Performance data are summarized in Table 1.

Sensitivity

We found higher d' in the second block half compared with the first block half ($F_{(1,23)} = 18.57$, $p < 0.001$, $\eta_p^2 = 0.447$; Fig. 2*A*). Moreover, d' was higher when voices were tested with the same sentence as in the study phase than with a different sentence ($F_{(1,23)} = 9.71$, $p = 0.005$, $\eta_p^2 = 0.297$). Recognition performance was substantially above-chance ($d' > 0$) in all four conditions, as suggested by one-sample t tests ($7.32 < t_{(23)} < 13.27$, $p < 0.001$, Bonferroni corrected).

Response criterion

For the criterion C, no significant main effects of block half or sentence condition were found ($p > 0.10$). The interaction approached significance ($F_{(1,23)} = 3.19$, $p = 0.087$, $\eta_p^2 = 0.122$).

Accuracies

Significant main effects of block half ($F_{(1,23)} = 18.07$, $p < 0.001$, $\eta_p^2 = 0.440$) and sentence condition ($F_{(1,23)} = 10.76$, $p = 0.003$, $\eta_p^2 = 0.319$) suggested an increase of accuracies from the first to the second block half, as well as higher accuracies in the same sentence condition than in the different sentence condition.

Correct RTs

A main effect of block half ($F_{(1,23)} = 13.35$, $p = 0.001$, $\eta_p^2 = 0.373$) suggested slower responses in the first compared with the second half. This effect seemed to be due to the same sentence condition, as indicated by the interaction of block half and sentence condition ($F_{(1,23)} = 6.08$, $p = 0.022$, $\eta_p^2 = 0.209$). A significant facilitation of response times from the first to the second block half occurred for the same sentence condition ($F_{(1,23)} = 13.64$, $p = 0.001$, $\eta_p^2 = 0.372$), but not for the different sentence condition ($F_{(1,23)} = 1.15$, $p > 0.05$, $\eta_p^2 = 0.047$). The interaction of block half and voice condition ($F_{(1,23)} = 7.74$, $p = 0.011$, $\eta_p^2 = 0.252$) reflected that hits were faster than correct rejections (CR) in the first half ($F_{(1,23)} = 6.49$, $p = 0.018$, $\eta_p^2 = 0.220$), with no difference in the second half ($F < 1$).

EEG results

ERPs in study phases

We performed ANOVAs on mean amplitudes of studied voices with subsequent recognition (subsequent hits/subsequent misses) and sentence condition (same/different) as within-subjects factors. We analyzed N1 and P2 at Cz where these components had their maximum repsonse. Based on findings that

Table 2. Statistical parameters for analyses of ERP effects in study phases

ERP	Effect	df	F	p	η_p^2
Dm I (250–400 ms)	Anteriority × subsequent recognition	2,46	5.54	0.011	0.194
Dm II (400–800 ms)	Anteriority	2,46	44.24	<0.001	0.658
	Subsequent recognition	1,23	4.45	0.046	0.162
	Laterality × anteriority	4,92	4.81	0.001	0.173
	Laterality × subsequent recognition	2,46	4.80	0.013	0.173
	Anteriority × subsequent recognition	2,46	12.89	<0.001	0.359
Dm III (800–1400 ms)	Laterality	2,46	4.03	0.024	0.149
	Anteriority	2,46	23.59	<0.001	0.506
	Laterality × anteriority	4,92	3.28	0.021	0.125
	Laterality × subsequent recognition	2,46	3.64	0.034	0.137
	Anteriority × subsequent recognition	2,46	4.01	0.032	0.148

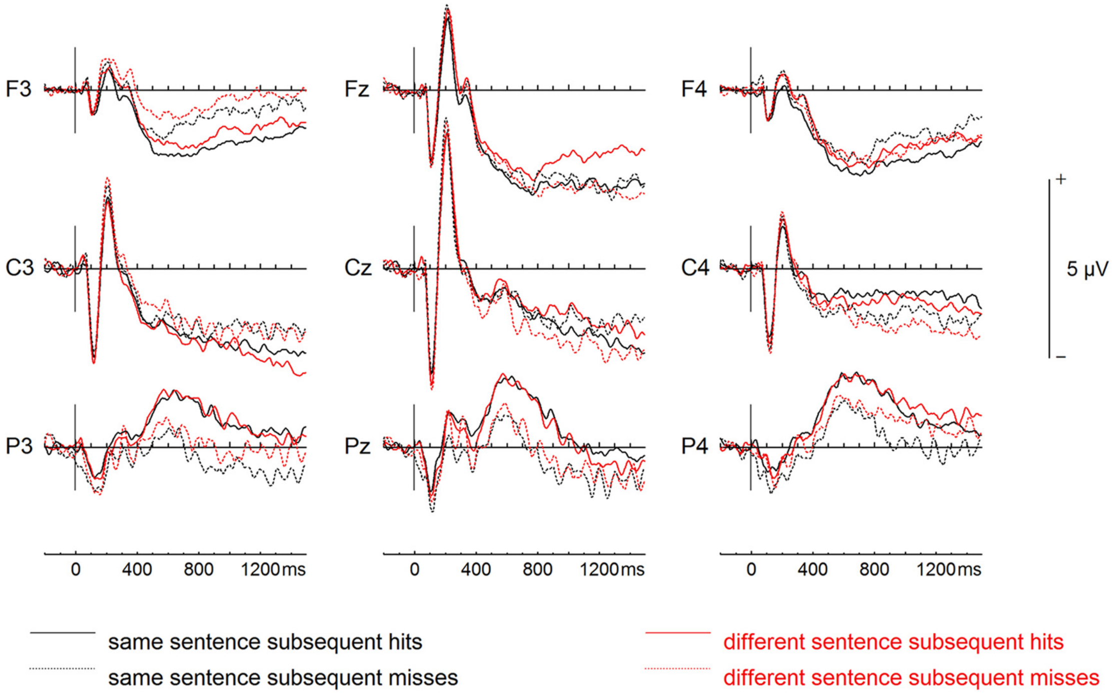

Figure 3. Grand mean ERPs during study phases for subsequently remembered voices (hits) and forgotten voices (misses) when tested with the same or different sentences than during study.

subsequent memory effects (Dms) can be broadly distributed both over time and across frontal-central-parietal electrodes (Paller et al., 1987) we analyzed Dm in three different time-windows between 250 and 1400 ms, and at F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4. Additional topographical factors were anteriority (frontal/central/parietal) and laterality (left/midline/right). Only effects involving the factors subsequent recognition and/or sentence condition are reported in the text as these are of particular interest. For a complete list of significant effects, please refer to Table 2.

N1 (90–130 ms) and P2 (190–230 ms). Anaysis of N1 and P2 mean amplitudes at Cz did not reveal any significant effects.

Dm I (250–400 ms). The earliest differences due to subsequent memory were seen at posterior sites from ~250 ms (Fig. 3). An ANOVA on mean amplitudes for Dm I (250–400 ms) revealed an interaction of anteriority and subsequent recognition ($F_{(2,46)} = 5.54$, $p = 0.011$, $\eta_p^2 = 0.194$), which was further explored with separate analyses for the three levels of anteriority. A main effect of subsequent recognition was only found at parietal sites ($F_{(1,23)} = 8.76$, $p = 0.007$, $\eta_p^2 = 0.276$), with larger positivity for subsequent hits compared with subsequent misses. This Dm was not significantly modulated by sentence condition ($F_{(1,23)} = 3.13$, $p = 0.09$, $\eta_p^2 = 0.120$) or laterality ($F < 1$). Please see Figure 4 for voltage maps.

Dm II (400–800 ms). The ANOVA revealed a main effect of subsequent recognition ($F_{(1,23)} = 4.45$, $p = 0.046$, $\eta_p^2 = 0.162$) with more positive responses for subsequent hits than for sebsequent misses. This Dm was unaffected by sentence condition, but interacted with anteriority ($F_{(2,46)} = 12.89$, $p < 0.001$, $\eta_p^2 = 0.359$) and laterality ($F_{(2,46)} = 4.80$, $p = 0.013$, $\eta_p^2 = 0.173$), in the absence of a three-way interaction ($F_{(4,92)} = 1.31$, $p = 0.272$, $\eta_p^2 = 0.054$). To examine the interaction of subsequent recognition and anteriority, separate analyses were performed for frontal, central and parietal sites. Whereas there were no significant Dm at frontal and central sites, analyses of parietal sites exhibited a prominent Dm ($F_{(1,23)} = 26.42$, $p < 0.001$, $\eta_p^2 = 0.535$) with more positive amplitudes for subsequently remembered (1.44 μV) compared with subsequently forgotten voices (0.54 μV). Finally, separate analyses for left, midline and right sites revealed a significant Dm over midline ($F_{(1,23)} = 5.41$, $p < 0.029$, $\eta_p^2 = 0.190$) and right ($F_{(1,23)} = 7.92$, $p = 0.010$, $\eta_p^2 = 0.256$), but not over left hemispheric sites ($F < 1$).

Dm III (800–1400 ms). Subsequent recognition interacted both with anteriority ($F_{(2,46)} = 4.0$, $p = 0.032$, $\eta_p^2 = 0.148$) and

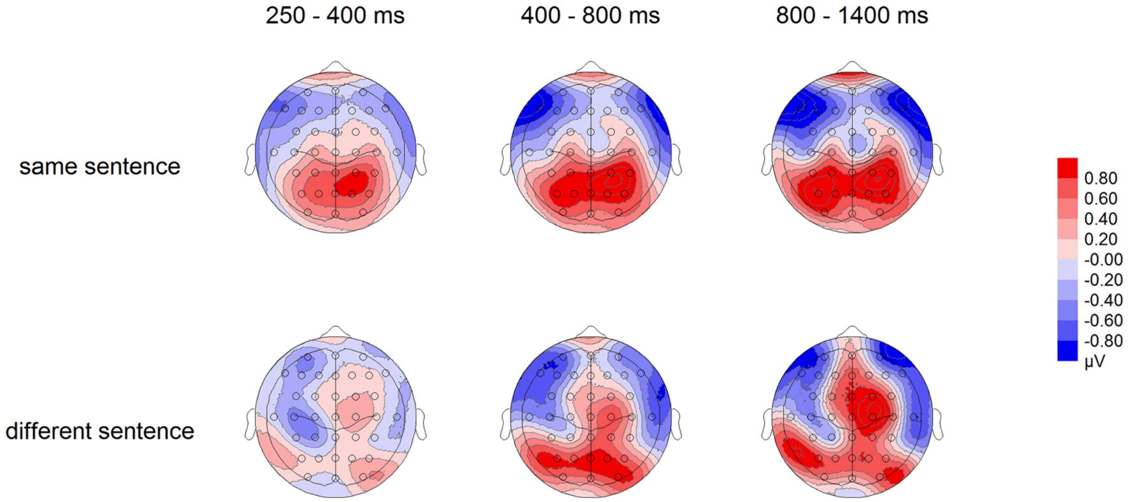

Figure 4. Voltage maps of ERP differences for the Dm (study voices subsequently remembered − subsequently forgotten), reflecting subsequent voice recognition, are depicted for time windows of the Dm I (250–400 ms), Dm II (400–800 ms), and Dm III (800–1400 ms). Note that maps are based on spherical spline interpolation with 90° equidistant projection.

with laterality ($F_{(2,46)} = 3.64$, $p = 0.034$, $\eta_p^2 = 0.137$). As before, a significant Dm was only present at parietal sites ($F_{(1,23)} = 8.82$, $p = 0.007$, $\eta_p^2 = 0.277$), but not over frontal and central sites ($F < 1$). Moreover, Dm was absent over left ($F < 1$) and midline ($F_{(1,23)} = 2.46$, $p > 0.05$) sites and reduced to a trend over the right hemisphere ($F_{(1,23)} = 3.41$, $p = 0.078$, $\eta_p^2 = 0.129$).

ERPs in test phases

We performed ANOVAs on mean amplitudes for correct responses (i.e., hits for old voices and CRs for new voices) with voice condition (old/new) and sentence condition (same/different) as within-subjects factors. We analyzed the early auditory components N1 and P2 at Cz, where these components had their maximum response, consistent with earlier studies using similar acoustic stimuli (Schweinberger, 2001). We assessed OLD/NEW effects in the parietal late positive component (LPC) at Pz, where the LPC is typically more pronounced for previously encountered compared with novel stimuli (Friedman and Johnson, 2000).

N1 (90–130 ms) and P2 (190–230 ms). Analysis of N1 and P2 mean amplitudes at Cz did not reveal any significant effects.

LPC (parietal OLD/NEW effect, 300–700 ms). An ANOVA gave rise to a significant interaction of voice condition and sentence condition ($F_{(1,23)} = 4.43$, $p = 0.046$, $\eta_p^2 = 0.162$). This interaction reflected an OLD/NEW effect for the same sentence condition, with more positivity for voices correctly recognized as "old" than for voices correctly rejected as "new" ($F_{(1,23)} = 4.78$, $p = 0.039$, $\eta_p^2 = 0.172$) and with no OLD/NEW effect for the different sentence condition ($F < 1$; Fig. 5).

700–1400 ms segment. An analogous analysis for mean amplitudes in a later segment (700–1400 ms) did not yield any significant effects involving the experimental variables.

Time-frequency analyses

In addition to the conventional analysis of ERP amplitudes, we performed a series of time-frequency analyses for test voices. To identify frequencies, time points, and electrodes of interest, in a first analysis step, a series of paired t tests were computed on the power differences between correctly classified old (hits) and new (CR) voices ($\alpha = 0.05$, two-sided). This was done at all 64 electrodes simultaneously and for each time and frequency bin within a subepoch from −400 to 800 ms relative to the voice onset. A randomization procedure was then conducted to adjust for multiple comparisons. In 5000 successive runs, the data were randomly exchanged between conditions within subjects and t tests were applied for all electrodes at random time points in the baseline interval. The number of significant electrodes was recorded after each run. This procedure was repeated for each frequency step, and then the fraction of runs was computed where the number of significant electrodes found in the randomized data exceeded the number of significant electrodes obtained during the first analysis step. The fraction can be interpreted as the probability p for finding a larger number of significant electrodes as those found in the actual data by chance. Results of the first-stage analysis were only accepted if $p < 0.05$. This strategy en-

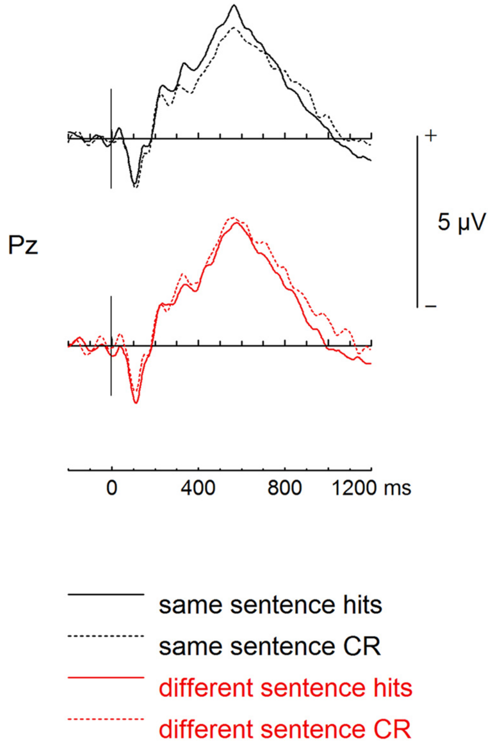

Figure 5. Grand mean ERPs at Pz during test phases for voices correctly recognized (hits) as old and correctly rejected as new.

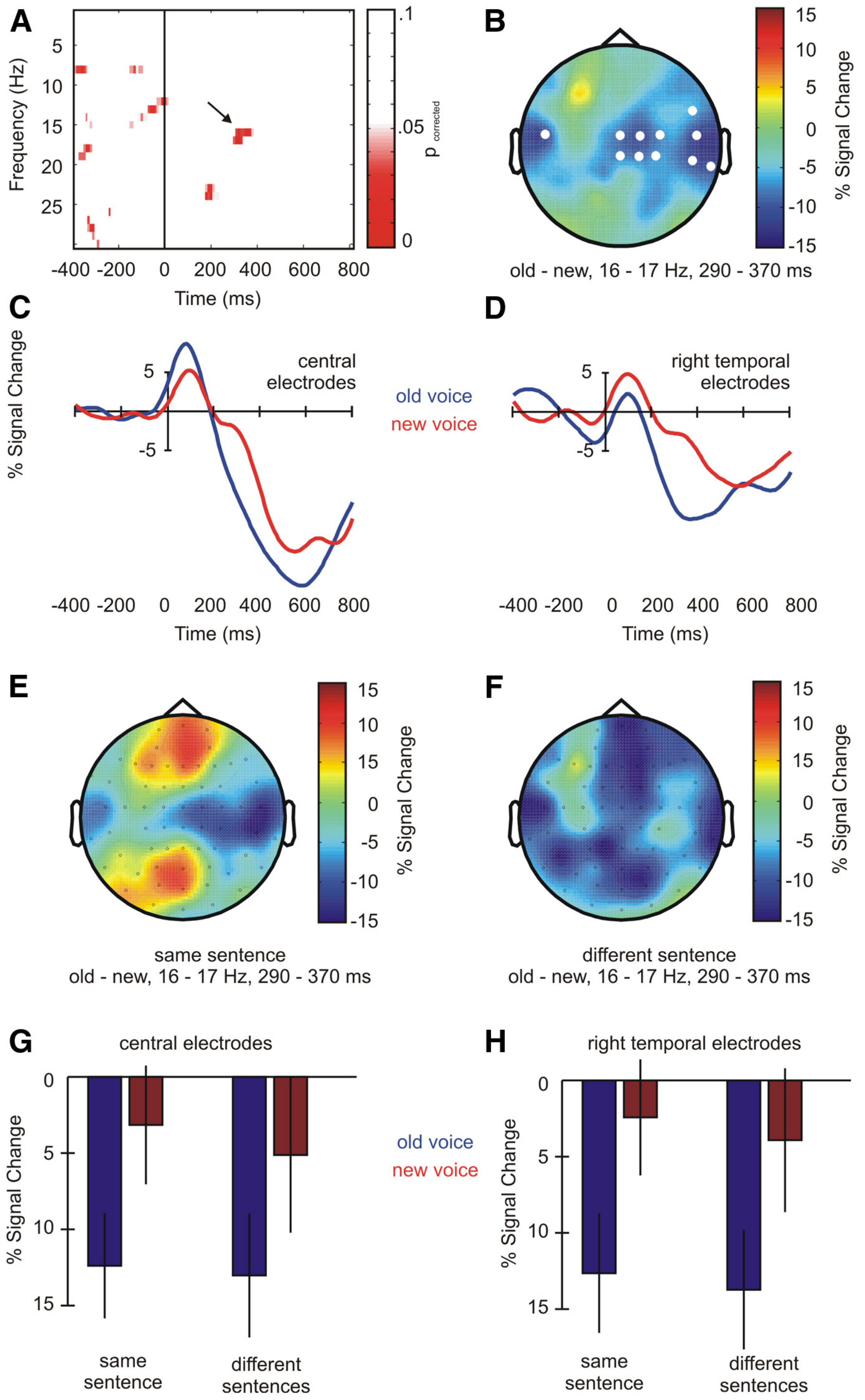

Figure 6. Results of the time-frequency analyses. ***A***, Significant time and frequency ranges (*p*-values) obtained from permutation tests. Only results with more than 10 adjacent significant bins were considered. The relevant time and frequency range is marked. ***B***, Head topography of the mean signal change, old versus new voices, at the relevant time and frequency range. Significant electrodes are marked. ***C***, Waveforms showing the mean signal chance (16–17 Hz) for old and new voices at the central electrode group. The waveforms show an amplitude drop after ~300 ms that was stronger for old compared with new voices. ***D***, Same as ***C***, but for the right temporal electrode group. ***E***, Head topography of the mean signal change, old versus new voices, for the "same sentence" condition. ***F***, Same as ***E***, but for the "different sentence" condition. ***G***, Bar graph showing the mean signal change (± SEM) in each condition at the central electrode group. ***H***, Same as ***G***, but at right temporal electrodes.

sured a correction for multiple comparisons across electrodes, with a two-sided familywise α level of 0.05 (Maris, Schoffelen, and Fries, 2007). To also correct for multiple comparisons across frequencies and sample points, we considered only effects that occurred in more than 10 adjacent time and frequency bins (Volberg et al., 2013).

The purpose of the randomization procedure was to select appropriate electrodes and time-frequency ranges for the further analysis. In a next step, we investigated whether the power at the regions that emerged from this procedure differed across conditions. To this end, the power for old and new voices was averaged over relevant time and frequency ranges as obtained from the permutation test. Electrodes in which the mean amplitude difference was significant were selected for further analysis. The mean amplitudes at the electrodes and within the time ranges of interest were subjected to a repeated-measures ANOVA with the factors voice condition (old/new) and sentence condition (same/different).

The results of the permutation procedure are depicted in Figure 6*A*. This revealed one significant time and frequency range at 16 and 17 Hz and between 290 and 370 ms. The difference had a maximum at 17 Hz, 300–310 ms, where 12 electrodes were significant ($p_{\mathrm{corr}} = 0.008$ by permutation test). Although spurious differences also appeared at ~200 ms and within the baseline interval, none of those exceeded the criterion of 10 adjacent significant time and frequency bins.

Figure 6*B* shows the mean power difference (percent signal change) for old minus new voices averaged within the relevant time (290–370 ms) and frequency range (16–17 Hz). There were two different scalp regions with significant differences. The first region had a central focus and comprised the electrodes CPz, Cz, C2, C4, CP4, and CP2. The second region had a right temporal focus and comprised electrodes FT8, TP10, T8, and TP8. The waveforms in Figure 6, *C* and *D*, show the mean power (16–17 Hz) for old and new voices at central and right temporal regions, respectively. After an initial power increase, the activity dropped below baseline at both regions. This decrease was more pronounced for old compared with new voices, so that a negative difference in the average power was observed.

Next, we investigated whether the difference between old and new voices was specific for the sentence conditions. Figure 6, *E* and *F*, show the difference topographies on the average power within the previously identified time and frequency ranges, separately for the "same sentence" (Fig. 6*E*) and "different sentence" (Fig. 6*F*) condition, old minus new voices, 16–17 Hz, 290–370 ms. Negative power differences at central and right temporal electrodes could be clearly identified in both cases.

The average amplitudes at the central and the right temporal electrode groups were then subjected to repeated-measures ANOVAs with the factors voice condition (old/new) and sentence condition (same/different). The main effect of voice condition was significant at both sites (central: $F_{(1,23)} = 8.74$, $p = 0.007$, $\eta_p^2 = 0.276$; right temporal: $F_{(1,23)} = 8.51$, $p = 0.008$, $\eta_p^2 = 0.270$). Because only those electrodes and time-frequency ranges were used for the ANOVA in which power differences between old and new items occurred, this main effect was expected. Importantly, no further effects were observed. Specifically, there was no interaction between the factors voice condition and sentence condition (central: $F_{(1,23)} = 0.04$, $p = 0.842$; right temporal: $F_{(1,23)} = 0.01$, $p = 0.935$).

Finally, separate *t* tests were applied to power differences between old and new voices in the "same sentence" and "different sentence" conditions. The results are depicted in Figure 6, *G* and *H* (mean ± SEM). They show more negative values for old compared with new voices at both regions and in both conditions. In addition, the power changes had a similar overall size across conditions. For the central electrode group, the difference between old and new voices was significant in the "same sentence" condition ($t_{(23)} = -2.36$, $p = 0.027$, $\eta_p^2 = 0.196$), but not in the "different sentence" condition ($t_{(23)} = -1.67$, $p = 0.108$, $\eta_p^2 = 0.109$). For the right temporal electrode group, both differences were significant (same sentence: $t_{(23)} = -2.44$, $p = 0.023$, $\eta_p^2 = 0.206$; different sentence: $t_{(23)} = -2.35$, $p = 0.028$, $\eta_p^2 = 0.194$).

Postexperimental questionnaire results

Note that scales for the questionnaire results were inverted for convenience of interpretation, such that positive correlations would reflect a higher performance with increasing contact or increasing subjective ratings of voice recognition abilities. There was a trend for a correlation of recognition scores (d') with the average amount of self-reported verbal contact to people ($r_{(22)} = 0.36$, $p = 0.09$). Because a previous study found a strong relationship between contact and recognition performance in female but not in male listeners (Skuk and Schweinberger, 2013), we performed the same correlation separately for participant sexes. The correlation was significant for female participants ($r_{(10)} = 0.75$, $p = 0.005$), indicating that recognition scores increased with the amount of verbal contact to people for female participants, but not for male participants ($r_{(10)} = 0.10$, $p > 0.05$). There was also a trend toward women reporting more verbal contact (4.4) than men (3.8) ($t_{(22)} = -1.76$, $p = 0.093$).

We further correlated recognition scores (d') with three measures of subjective voice recognition abilities: in everyday life in the absence of visual cues for once-heard unfamiliar people, for familiar people, and performance in the experiment compared with other participants. None of these correlations reached significance ($-0.16 \leq r_{(22)} \leq 0.37$, all $p \geq 0.08$).

Discussion

We show here that learning unfamiliar voices from short sentences results in substantial voice recognition, which generalizes well to new speech samples. During encoding in the study phase, subsequent voice recognition was predicted by an ERP Dm, emerging as a sustained positivity that started ~250 ms and showed a right parietal maximum. At the time of testing, we observed a parietal ERP OLD/NEW effect (300–700 ms), which was confined to instances when voices were tested with the same sentences as heard during study, and thus appears to reflect retrieval of the study episode rather than memory for voice identity per se. Voice identity recognition independent of test sentence was reflected in power differences in beta-band oscillations (16–17 Hz) at central and right temporal sites (290–370 ms).

The present study shows that listeners can recognize newly learned voices even after relatively limited exposure to only six sentences (first block half). Importantly, although recognition performance was somewhat superior for identical study and test sentences, substantial recognition was also established for voices tested with previously unheard sentences. This indicates the acquisition of representations that store idiosyncratic voice properties, allowing voice identity recognition independent of speech content. This pattern of results is reminiscent of findings from the visual domain that face learning transfers to unseen images of a person (Kaufmann et al., 2009; Zimmermann and Eimer, 2013). Our results are also consistent with reports that voice identification generalizes to unheard speech samples (Legge, Grosmann, and Pieper, 1984; Nygaard and Pisoni, 1998; Sheffert et al., 2002; Sheffert and Olson, 2004). At the same time, and also consistent with earlier research (Read and Craik, 1995; Schweinberger, Herholz, and Stief, 1997), voices were recognized somewhat more accurately from the same than from different sentences. We propose that this reflects a degree of interdependence between speech and voice in speaker perception (Perrachione and Wong, 2007; Remez, Fellowes, and Nagel, 2007; Perrachione, Del Tufo, and Gabrieli, 2011; Schweinberger et al., 2014), which complements similar findings in word recognition (Craik and Kirsner, 1974; Nygaard, Sommers, and Pisoni, 1994; Campeanu, Craik, and Alain, 2013). Note that this processing interdependence between linguistic and nonlinguistic information could also facilitate voice recognition from previously unheard sentences. Specifically, speakers may be recognized by individual patterns of articulation or pronunciation that persist across different utterances and represent important markers of identity (Belin, Fecteau, and Bédard, 2004). A prediction derived from that notion is that the amount of phonetic overlap between study and test sentences should influence voice recognition.

The correlation between individual voice recognition performance and self-reported verbal contact to people generally fits the notion that perceptual experience with people improves person recognition memory, as demonstrated for faces (Valentine and Endo, 1992). As a qualification, we saw this relationship for female listeners only. At present, we are unable to offer a straightforward account for why women, but not men, would benefit from voice experience. This issue deserves further investigation, particularly when considering that our results replicate findings for personally familiar voices (Skuk and Schweinberger, 2013). Finally, the lack of a correlation between confidence and performance in voice recognition concurs with findings from forensic research (Clifford, 1980).

Most importantly, our findings provide the first electrophysiological evidence for brain processes during voice encoding predictive of subsequent recognition. Study voices subsequently remembered (vs voices subsequently forgotten) elicited a larger sustained parietal positivity (from ~250 ms), which was most pronounced over right hemispheric electrodes. These novel Dm effects for voices extend similar findings for other stimulus domains (Friedman and Johnson, 2000; Paller and Wagner, 2002). Dm effects have been reported previously for written words (Paller, Kutas, and Mayes, 1987), faces (Sommer et al., 1991) and musical stimuli (Klostermann, Loui, and Shimamura, 2009) and are generally thought to reflect successful encoding of stimuli from the respective categories into memory. Whereas the medial temporal lobe is believed to act as a common hub that binds representations together, promoting episodic memory formation (Paller and Wagner, 2002), stimulus-specific activation patterns of Dm likely reflect the formation of category-specific representations in long-term memory (Sommer, Komoss, and Schweinberger, 1997). In that sense, the encoding of words may rely more on left-hemispheric regions (Wagner et al., 1998), whereas the encoding of pictorial information (Brewer et al., 1998), including faces (Sommer et al., 1991) and musical stimuli (Klostermann et al., 2009), appears to involve predominantly right-lateralized regions.

Previous studies (Belin et al., 2000; Belin, Zatorre, and Ahad, 2002; Belin and Zatorre, 2003; von Kriegstein et al., 2003; Formisano et al., 2008) suggested a prominent role of the right hemisphere for human voice processing regardless of speech content. Relative to those findings, the lateralization of the present voice Dm establishes a specific role of the right hemisphere for the encoding of voices into long-term memory. Importantly, the similarity between the same and different sentence conditions suggests that the present voice Dm reflects a correlate of the acquisition of speech-invariant representations of voice identity. Concerning timing, several studies reported an onset of voice identity processing at ∼250–300 ms in both recognition and matching tasks (Spreckelmeyer et al., 2009; Schweinberger et al., 2011b; Schweinberger, Kloth, and Robertson, 2011a). Although the voice Dm reflects encoding rather than recognition, its onset at ∼250 ms could provide an interesting parallel to those findings.

At the time of testing, voices correctly classified as "old" elicited a more pronounced parietal LPC (300–700 ms) than voices correctly classified as "new." Importantly, this OLD/NEW effect was confined to the same sentence condition, suggesting that it relates to the detailed explicit retrieval of the stimulus from episodic memory. This interpretation both fits and extends similar OLD/NEW effects for other categories, such as words (Paller et al., 1987), faces (Guerin and Miller, 2009), pictures and natural sounds (Shannon and Buckner, 2004), and musical sounds (Klostermann et al., 2009). Note that the functional interpretation of the parietal OLD/NEW effect differs from that of earlier perceptual ERP repetition effects, which typically start at ∼200 ms and are independent of explicit retrieval (Rugg and Allan, 2000). The parietal OLD/NEW effect is believed to reflect source memory for the study episode and thus relatively slow explicit recollection (Rugg and Curran, 2007). This entails retrieval of specific stimulus features (which here define the specific utterance during study).

Whereas an electrophysiological correlate reflecting voice (as opposed to item) recognition was absent in the conventional ERP analyses, time-frequency analyses revealed such an effect. Regardless of sentence condition at the time of testing, we observed a stronger decrease of power in beta-band oscillations (16–17 Hz) for old compared with new voices. This effect occurred in a central and a right temporal electrode cluster at a latency of 290–370 ms. Whereas current functional understanding of beta-band activity (BBA) is poor compared with other frequency bands, BBA has been related to various cognitive and motor processes (Engel and Fries, 2010). Regarding audition, BBA has been linked to the detection of novel sounds (Haenschel et al., 2000). Of potential relevance, Haenschel et al. (2000) found smaller BBA (12–20 Hz) in repeated versus novel sounds (sinusoidal tones) between 200 and 400 ms. We propose that the present BBA attenuation for learned compared with novel voices is a correlate of the access to speech-invariant voice identity representations established during learning. The topography of the effect over central and right temporal sites could reflect the respective involvement of bilateral auditory cortices in superior temporal gyri (Pantev et al., 1991) and in right temporal cortex areas selectively responsive to (familiar) voices (Belin et al., 2000, 2011; Belin and Zatorre, 2003).

We propose that our results from the test phase reflect two types of memory. The first is memory for a specific study item (episodic memory). This type of memory depends on speech content and is reflected in the parietal OLD/NEW effect. The second is memory for voice identity, reflecting the activation of speech-invariant representations for learned voices. Note that we investigated recognition after relatively limited prior exposure. Whereas this has the advantage of allowing precise experimental control over perceptual exposure during learning, the implications of our results for understanding learning and recognition of personally well known people's voices remain to be determined. Our results may capture the initial stages of voice learning, but not necessarily the acquisition of robust voice representations for highly familiar people. Moreover, whereas our results demonstrate the acquisition of speech-invariant representations for voice identity, our study did not require identification of individual speakers via naming or providing unique semantic information. It remains for future research to establish how identification, compared with recognition of voices as familiar, is reflected in electrophysiological brain responses.

The present study has established electrophysiological correlates of both voice encoding and recognition in a learning paradigm. Whereas the observed Dm reflects a correlate of encoding of voice identity into memory, the factors that determine whether a particular voice will be subsequently recognized remain to be explored. For faces, there is consistent evidence both that distinctive faces are recognized more easily than typical ones (Schulz et al., 2012) and that facial distinctiveness is intimately related to Dm during face encoding (Sommer et al., 1995). Although the role of distinctiveness for voice recognition is less well explored, there is evidence that distinctive personally familiar voices are recognized more easily (Skuk and Schweinberger, 2013). Moreover, perceived voice distinctiveness is parametrically related to activation in voice-sensitive temporal cortex areas (Latinus et al., 2013), perhaps suggesting that voice identity is represented in a norm-based fashion. Exploring the relationships between (norm-based) voice distinctiveness and electrophysiological correlates of voice encoding and recognition will be of considerable future interest.

References

Andics A, McQueen JM, Petersson KM, Gál V, Rudas G, Vidnyánszky Z (2010) Neural mechanisms for voice recognition. Neuroimage 52:1528–1540. CrossRef Medline

Beauchemin M, De Beaumont L, Vannasing P, Turcotte A, Arcand C, Belin P, Lassonde M (2006) Electrophysiological markers of voice familiarity. Eur J Neurosci 23:3081–3086. CrossRef Medline

Belin P, Zatorre RJ (2003) Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14:2105–2109. CrossRef Medline

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403:309–312. CrossRef Medline

Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 13:17–26. CrossRef Medline

Belin P, Fecteau S, Bédard C (2004) Thinking the voice: neural correlates of voice perception. Trends Cogn Sci 8:129–135. CrossRef Medline

Belin P, Bestelmeyer PE, Latinus M, Watson R (2011) Understanding voice perception. Br J Psychol 102:711–725. CrossRef Medline

Boersma P, Weenink D (2001) PRAAT, a system for doing phonetics by computer. Glot International 5:341–345.

Brewer JB, Zhao Z, Desmond JE, Glover GH, Gabrieli JD (1998) Making memories: brain activity that predicts how well visual experience will be remembered. Science 281:1185–1187. CrossRef Medline

Campeanu S, Craik FI, Alain C (2013) Voice congruency facilitates word recognition. PLoS One 8:e58778. CrossRef Medline

Charest I, Pernet CR, Rousselet GA, Quiñones I, Latinus M, Fillion-Bilodeau S, Chartrand JP, Belin P (2009) Electrophysiological evidence for an early processing of human voices. BMC Neurosci 10:127. CrossRef Medline

Clifford BR (1980) Voice identification by human listeners: on earwitness reliability. Law and Human Behavior 4:373–394. CrossRef

Craik FIM, Kirsner K (1974) Effect of speakers voice on word recognition. Quarterly Journal of Experimental Psychology 26:274–284. CrossRef

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21. CrossRef Medline

Düzel E, Penny WD, Burgess N (2010) Brain oscillations and memory. Curr Opin Neurobiol 20:143–149. CrossRef Medline

Eimer M (2011) The face-sensitivity of the N170 component. Front Hum Neurosci 5.

Engel AK, Fries P (2010) Beta-band oscillations–signalling the status quo? Curr Opin Neurobiol 20:156–165. CrossRef Medline

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322: 970–973. CrossRef Medline

Friedman D, Johnson R Jr (2000) Event-related potential (ERP) studies of memory encoding and retrieval: a selective review. Microsc Rec Tech 51:6–28. CrossRef Medline

Guerin SA, Miller MB (2009) Lateralization of the parietal old/new effect: An event-related fMRI study comparing recognition memory for words and faces. Neuroimage 44:232–242. CrossRef Medline

Haenschel C, Baldeweg T, Croft RJ, Whittington M, Gruzelier J (2000) Gamma and beta frequency oscillations in response to novel auditory stimuli: A comparison of human electroencephalogram (EEG) data with in vitro models. Proc Natl Acad Sci U S A 97:7645–7650. CrossRef Medline

Huynh H, Feldt LS (1976) Estimation of the box correction for degrees of freedom from sample data in randomized block and split block designs. Journal of Educational Statistics 1:69–82.

Kaufmann JM, Schweinberger SR, Burton AM (2009) N250 ERP correlates of the acquisition of face representations across different images. J Cogn Neurosci 21:625–641. CrossRef Medline

Klostermann EC, Loui P, Shimamura AP (2009) Activation of right parietal cortex during memory retrieval of nonlinguistic auditory stimuli. Cogn Affect Behav Neurosci 9:242–248. CrossRef Medline

Kriegstein KV, Giraud AL (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22:948–955. CrossRef Medline

Latinus M, Crabbe F, Belin P (2011) Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex 21:2820–2828. CrossRef Medline

Latinus M, McAleer P, Bestelmeyer PE, Belin P (2013) Norm-based coding of voice identity in human auditory cortex. Curr Biol 23:1075–1080. CrossRef Medline

Legge GE, Grosmann C, Pieper CM (1984) Learning unfamiliar voices. Journal of Experimental Psychology-Learning Memory and Cognition 10:298–303. CrossRef

Macmillan NA, Kaplan HL (1985) Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. Psychol Bull 98:185–199. CrossRef Medline

Maris E, Schoffelen JM, Fries P (2007) Nonparametric statistical testing of coherence differences. J Neurosci Methods 163:161–175. CrossRef Medline

Nygaard LC, Pisoni DB (1998) Talker-specific learning in speech perception. Percept Psychophys 60:355–376. CrossRef Medline

Nygaard LC, Sommers MS, Pisoni DB (1994) Speech-perception as a talker-contingent process. Psychol Sci 5:42–46. CrossRef Medline

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 156869.

Paller KA, Wagner AD (2002) Observing the transformation of experience into memory. Trends Cogn Sci 6:93–102. CrossRef Medline

Paller KA, Kutas M, Mayes AR (1987) Neural correlates of encoding in an incidental-learning paradigm. Electroencephalogr Clin Neurophysiol 67: 360–371. CrossRef Medline

Pantev C, Makeig S, Hoke M, Galambos R, Hampson S, Gallen C (1991) Human auditory evoked gamma-band magnetic-fields. Proc Natl Acad Sci U S A 88:8996–9000. CrossRef Medline

Perrachione TK, Wong PC (2007) Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. Neuropsychologia 45:1899–1910. CrossRef Medline

Perrachione TK, Del Tufo SN, Gabrieli JD (2011) Human voice recognition depends on language ability. Science 333:595. CrossRef Medline

Read D, Craik FIM (1995) Earwitness identification–some influences on voice recognition. Journal of Experimental Psychology-Applied 1:6–18. CrossRef

Remez RE, Fellowes JM, Nagel DS (2007) On the perception of similarity among talkers. Journal of the Acoustical Society of America 122:3688–3696. CrossRef Medline

Rugg MD, Allan K (2000) Event-related potential studies of memory. In: The Oxford handbook of memory (Tulving E, Craik FIM, Eds), pp. 521–537. Oxford: OUP.

Rugg MD, Curran T (2007) Event-related potentials and recognition memory. Trends Cogn Sci 11:251–257. CrossRef Medline

Schulz C, Kaufmann JM, Kurt A, Schweinberger SR (2012) Faces forming traces: Neurophysiological correlates of learning naturally distinctive and caricatured faces. Neuroimage 63:491–500. CrossRef Medline

Schweinberger SR (2001) Human brain potential correlates of voice priming and voice recognition. Neuropsychologia 39:921–936. CrossRef Medline

Schweinberger SR, Burton AM (2003) Covert recognition and the neural system for face processing. Cortex 39:9–30. CrossRef Medline

Schweinberger SR, Herholz A, Stief V (1997) Auditory long-term memory: repetition priming of voice recognition. Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology 50:498–517.

Schweinberger SR, Kloth N, Robertson DM (2011a) Hearing facial identities: Brain correlates of face-voice integration in person identification. Cortex 47:1026–1037. CrossRef Medline

Schweinberger SR, Walther C, Zäske R, Kovács G (2011b) Neural correlates of adaptation to voice identity. Br J Psychol 102:748–764. CrossRef Medline

Schweinberger SR, Kawahara H, Simpson AP, Skuk VG, Zaeske R (2014) Speaker perception. Wiley Interdisciplinary Reviews–Cognitive Science 5:15–25. CrossRef

Shannon BJ, Buckner RL (2004) Functional-anatomic correlates of memory retrieval that suggest nontraditional processing roles for multiple distinct regions within posterior parietal cortex. J Neurosci 24:10084–10092. CrossRef Medline

Sheffert SM, Olson E (2004) Audiovisual speech facilitates voice learning. Percept Psychophys 66:352–362. CrossRef Medline

Sheffert SM, Pisoni DB, Fellowes JM, Remez RE (2002) Learning to recognize talkers from natural, sinewave, and reversed speech samples. Journal of Experimental Psychology-Human Perception and Performance 28: 1447–1469. CrossRef Medline

Skuk VG, Schweinberger SR (2013) Gender differences in familiar voice identification. Hear Res 296:131–140. CrossRef Medline

Sommer W, Schweinberger SR, Matt J (1991) Human brain potential correlates of face encoding into memory. Electroencephalogr Clin Neurophysiol 79:457–463. CrossRef Medline

Sommer W, Heinz A, Leuthold H, Matt J, Schweinberger SR (1995) Metamemory, distinctiveness, and event-related potentials in recognition memory for faces. Mem Cognit 23:1–11. CrossRef Medline

Sommer W, Komoss E, Schweinberger SR (1997) Differential localization of brain systems subserving memory for names and faces in normal subjects with event-related potentials. Electroencephalogr Clin Neurophysiol 102: 192–199. CrossRef Medline

Spreckelmeyer KN, Kutas M, Urbach T, Altenmüller E, Müller TF (2009) Neural processing of vocal emotion and identity. Brain Cogn 69:121–126. CrossRef Medline

Valentine T, Endo M (1992) Towards an exemplar model of face processing–the effects of race and distinctiveness. Quarterly Journal of Experimental Psychology Section A–Human Experimental Psychology 44:671–703. CrossRef

Van Lancker D, Kreiman J (1987) Voice discrimination and recognition are separate abilities. Neuropsychologia 25:829–834. CrossRef Medline

Volberg G, Karmann A, Birkner S, Greenlee MW (2013) Short- and long-range neural synchrony in grapheme-color synesthesia. J Cogn Neurosci 25:1148–1162. CrossRef Medline

von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res 17:48–55. CrossRef Medline

Wagner AD, Poldrack RA, Eldridge LL, Desmond JE, Glover GH, Gabrieli JD (1998) Material-specific lateralization of prefrontal activation during episodic encoding and retrieval. Neuroreport 9:3711–3717. CrossRef Medline

Wiese H (2012) The role of age and ethnic group in face recognition memory: ERP evidence from a combined own-age and own-race bias study. Biol Psychol 89:137–147. CrossRef Medline

Yovel G, Belin P (2013) A unified coding strategy for processing faces and voices. Trends Cogn Sci 17:263–271. CrossRef Medline

Zimmermann FG, Eimer M (2013) Face learning and the emergence of view-independent face recognition: an event-related brain potential study. Neuropsychologia 51:1320–1329. CrossRef Medline