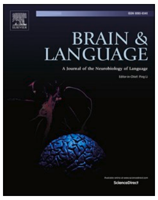
Brain & Language 240 (2023) 105264

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Using TMS to evaluate a causal role for right posterior temporal cortex in talker-specific phonetic processing[☆]

Sahil Luthra[a,*], Hannah Mechtenberg[a], Cristal Giorio[a], Rachel M. Theodore[a], James S. Magnuson[a,b,c], Emily B. Myers[a]

[a] *University of Connecticut, United States*
[b] *BCBL. Basque Center on Cognition Brain and Language, Donostia-San Sebastián, Spain*
[c] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*

ARTICLE INFO

Keywords:
Speech perception
Talker variability
Phonetic processing
Right hemisphere
TMS

ABSTRACT

Theories suggest that speech perception is informed by listeners' beliefs of what phonetic variation is typical of a talker. A previous fMRI study found right middle temporal gyrus (RMTG) sensitivity to whether a phonetic variant was typical of a talker, consistent with literature suggesting that the right hemisphere may play a key role in conditioning phonetic identity on talker information. The current work used transcranial magnetic stimulation (TMS) to test whether the RMTG plays a causal role in processing talker-specific phonetic variation. Listeners were exposed to talkers who differed in how they produced voiceless stop consonants while TMS was applied to RMTG, left MTG, or scalp vertex. Listeners subsequently showed near-ceiling performance in indicating which of two variants was typical of a trained talker, regardless of previous stimulation site. Thus, even though the RMTG is recruited for talker-specific phonetic processing, modulation of its function may have only modest consequences.

1. Introduction

The speech signal simultaneously conveys linguistic information, including phonetic information about which particular consonants and vowels are being produced and talker information about the person producing those speech sounds (e.g., Abercrombie, 1967). Classic neuropsychological data suggest at least partial separability between phonetic processing and talker processing, as patients with left hemisphere damage often exhibit selective impairments in speech perception but not vocal identity processing, whereas patients with right hemisphere damage often exhibit impairments in impairments in vocal identity processing but not speech perception (Van Lancker & Canter, 1982; Wernicke, 1874). More generally, contemporary neurobiological accounts hold that phonetic processing is largely supported by the left hemisphere while vocal identity information is largely processed by the right hemisphere, though early acoustic–phonetic analysis has been shown to recruit the temporal cortex bilaterally (Hickok & Poeppel, 2000; Maguinness, Roswandowitz, & von Kriegstein, 2018).

While the distinction between phonetic processing and talker processing can be useful, it is also an oversimplification. Individual talkers differ in how they produce their speech sounds (Allen, Miller, & DeSteno, 2003; Hillenbrand, Getty, Clark, & Wheeler, 1995; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Newman, Clouse, & Burnham, 2001; Peterson & Barney, 1952), and listeners appear to capitalize on the structure in this variability, forming talker-specific *generative models* – that is, sets of beliefs for how different talkers tend to produce their speech sounds (Kleinschmidt, 2019). A large body of evidence indicates that listeners can optimally capitalize on their knowledge of a talker's idiosyncrasies to guide speech perception (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Theodore & Monto, 2019), and familiarity with a talker's idiolect can facilitate speech perception (e.g., recognizing speech in noise; Nygaard et al., 1994; Souza, Gehani, Wright, & McCloy, 2013) as well as vocal identity recognition (Ganugapati & Theodore, 2019).

To illustrate this phenomenon, it is useful to consider a set of studies on listener sensitivity to talker-specific differences in voice-onset-time (VOT). VOT is an acoustic–phonetic property defined as the amount of time between the release of a stop consonant and the onset of vocal fold

[☆] This paper is based on the Dissertation Award at the 13th Society for the Neurobiology of Language (SNL) sponsored by Brain and Language.

[*] Corresponding author at: Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213-3815, United States.
E-mail address: sahilluthra@cmu.edu (S. Luthra).

https://doi.org/10.1016/j.bandl.2023.105264
Received 2 December 2021; Received in revised form 6 April 2023; Accepted 8 April 2023
Available online 21 April 2023

Table 1
Voice-onset-time (VOT) values for the stimuli used in this study.

Talkers	Continuum	Voice-Onset-Time (ms)		
		Voiced	*Short-VOT*	*Long-VOT*
Alvin/Carol	bowl/pole	20	Train: 60, 70 Test: 65	Train: 150, 160 Test: 155
Don/Joanne	dime/time	15	Train: 70, 80 Test: 75	Train: 160, 170 Test: 165
Peter/Sheila	gain/cane	20	Train: 80, 90 Test: 85	Train: 170, 180 Test: 175

vibration, and it is a primary cue for distinguishing voiced stop consonants (/b/, /d/ and /g/) from their voiceless counterparts (/p/, /t/ and /k/, respectively). Talkers can differ substantially in the precise VOTs they use to cue voiceless stop consonants (Allen et al., 2003), even after accounting for other factors that can affect VOT, such as speaking rate (Kessinger & Blumstein, 1997; Miller, 1989). Allen and Miller (2004) demonstrated that listeners are sensitive to these talker-specific differences. In their study, listeners were exposed to two talkers, one of whom produced the sound /t/ with a relatively long VOT and one of whom produced it with a relatively short VOT; notably, both variants were still unambiguously identified as /t/. After exposure to these two talkers, listeners were able to explicitly indicate which of two variants (long- or short-VOT) was typical of each talker. Additional work in this domain has shown that these judgments can generalize across place of articulation (i.e., that a talker who produces /k/ with a long VOT is likely to produce other voiceless stops with a long VOT; Theodore & Miller, 2010). Thus, exposure to a talker's idiosyncratic style of speaking leads listeners to make adjustments to a talker-specific generative model, allowing them to make explicit judgments about whether a production is typical or atypical of a particular talker.

The neural systems that support this talker-specific phonetic processing remain relatively underspecified; however, the right posterior temporal cortex is a promising candidate region that may support a listener's ability to contact talker-specific generative models (Luthra, 2021). Functional neuroimaging studies have implicated the right posterior temporal cortex in both phonetic processing and talker processing (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Kennedy-Higgins, Devlin, Nuttall, & Adank, 2020; Turkeltaub & Branch Coslett, 2010; von Kriegstein & Giraud, 2004), and strikingly, neural decoding studies indicate that portions of the right posterior temporal cortex support the classification of speech stimuli along both phonetic and talker dimensions (Formisano, De Martino, Bonte, & Goebel, 2008; Luthra, Magnuson, & Myers, 2023).

Further evidence of a potential role for right temporal cortex in talker-specific phonetic processing comes from an fMRI study by Myers and Theodore (2017), who investigated the neural mechanisms through which familiarity with a talker's idiolect can influence speech perception. Prior to scanning, listeners were exposed to two talkers who produced the words *gain* and *cane*. Following the Allen and Miller (2004) study described above, one talker produced the sound /k/ in *cane* ([keɪn]) with a short VOT and one produced it with a long VOT, and listeners showed high accuracy when asked to explicitly indicate which of two variants was typical of each talker. In the scanner, listeners completed a phonetic categorization task with the *gain* and *cane* stimuli. Critically, listeners heard both long-VOT and short-VOT variants of *cane* for each talker, meaning that they heard both typical and atypical variants. Myers and Theodore found that the response of the right posterior temporoparietal cortex – specifically, a cluster in the right posterior middle temporal gyrus (MTG) extending into the right superior temporal gyrus (STG) and right angular gyrus (AG) – depended on whether the variant was typical of the talker, even though the scanner task did not require making talker typicality judgments.

An open question, however, is whether recruitment of the right posterior temporal cortex is *necessary* for contacting a listener's beliefs about how a talker typically produces their speech sounds. In the current study, we leveraged transcranial magnetic stimulation (TMS) to test this question directly. Specifically, we tested whether magnetic stimulation applied during exposure to a talker's voice impacted listeners' ability to judge phonetic variants as typical or atypical of the talker during a subsequent test phase. Of interest was how performance would be impacted by stimulation to right posterior temporal cortex as compared to stimulation of the corresponding region in the left hemisphere and stimulation of a control site (the vertex of the scalp). Importantly, our goal is not for TMS to disrupt the encoding of phonetic information or listeners' ability to identify a talker — instead, we hypothesize that stimulation to the right posterior temporal cortex may impact a listener's ability to *link* phonetic information and talker information, as measured through a talker typicality judgment posttest (collected after the exposure phase).

The current study comprises two experiments. In Experiment 1, we sought to validate our behavioral paradigm for the TMS experiment, specifically testing whether listeners were able to show talker-specific learning for three pairs of talkers. In Experiment 2, we applied TMS to three different sites (RMTG, LMTG, vertex) over the course of the experiment, using a different pair of talkers for each stimulation site, and assessed the consequences of stimulation for determining which phonetic variants were typical of a talker. Stimuli, data, and analysis code for all experiments are available at https://osf.io/cf9t8/.

2. Experiment 1

Prior to conducting an experiment with TMS, we conducted an online experiment to verify that listeners could show talker-specific learning with a task design closely based on previous studies (Allen & Miller, 2004; Myers & Theodore, 2017; Theodore & Miller, 2010). In Experiment 1, listeners were exposed to three pairs of talkers (i.e., six different talkers), with each pair consisting of a male talker and a female talker. Within each pair, one talker produced their voiceless stop consonants (/p/, /t/ or /k/) with a relatively short VOT and the other talker produced the same consonant with a relatively long VOT. Because we would ultimately administer TMS at three stimulation sites for each participant (with listeners hearing a different pair of talkers for each stimulation site), we specifically aimed to establish that listeners could show talker-specific learning for each pair of talkers, with minimal generalization from one pair of talkers to the next. During training, listeners would hear three pairs of talkers, each pair consisting of a male and female talker. At test, listeners would indicate for the female talker only whether tokens with long vs short VOT were more typical of that talker. This modification from prior designs (which have used only two same-sex talkers) would enable us to ensure that learning effects in the TMS experiment (Experiment 2) reflected only the influence of the site being stimulated, rather than an aftereffect from stimulation at a previous site.

2.1. Methods

2.1.1. Stimulus construction

We first selected three minimal pair continua differing in VOT that had been used in previous studies. Specifically, we selected (1) a *bowl-pole* continuum originally constructed for Theodore and Miller (2010), (2) a *dime-time* continuum from Allen and Miller (2004), and (3) a *gain-cane* continuum from Theodore and Miller (2010). The talkers for these continua were all women and for the sake of this study are referred to as Carol, Joanne, and Sheila, respectively. Note that the continua differ in the place of articulation of the initial consonant (labial, alveolar, velar, respectively) as well as in the following vowel; by choosing continua with phonologically dissimilar words, we aimed to discourage generalization from talker to talker.

The voiced endpoint for each continuum (i.e., *bowl*, *dime*, *gain*) was synthesized through an LPC analysis of natural tokens produced by a different female native speaker of English. Each successive continuum step was created by iteratively modifying parameters of the LPC analysis to turn voiced frames into voiceless frames so as to increase VOT across successive steps. For additional details on the construction of these stimuli, the reader is referred to the studies for which they were originally constructed (Allen & Miller, 2004; Theodore & Miller, 2010).

We selected several tokens from each continuum for use in the current study, choosing a voiced token, three voiceless tokens with relatively short VOTs, and three voiceless tokens with relatively long VOTs. The specific VOT values of the stimuli (Table 1) were chosen based on VOT values used in the studies from which the stimuli were selected; note that the more posterior the place of articulation, the longer the VOT of the voiceless stimuli we chose, consistent with how these stimuli are produced naturally (Lisker & Abramson, 1964). Following previous studies using this paradigm (e.g., Allen & Miller, 2004), we selected two short-VOT variants (e.g., for the *pole* continuum, we selected one 60 ms VOT variant and one 70 ms variant) and two long-VOT steps for use during training; the VOTs of training variants differed by 10 ms and allowed us to simulate within-talker variability. At test, listeners did not hear the same tokens as had been presented at training but instead heard an intermediate one (e.g., *pole* with a 65 ms VOT).

We then made several modifications to the selected continuum steps using Praat (Boersma & Weenik, 2017). First, to decrease the perceptual similarity between the female talkers, we shifted the pitch contour of Carol's stimuli down by 15 Hz and the pitch contour of Sheila's stimuli up by 30 Hz. Subsequently, we created three male talkers by applying the "Change vocal tract size, pitch, and duration" function in the Praat Vocal Toolkit. A male talker named Alvin was synthesized by transforming Carol's stimuli; specifically, we applied a formant shift ratio of 0.85, set a new pitch median of 100 Hz, and set the pitch variation of Alvin's voice to be 80 % of Carol's. A male talker named Don was derived from Joanne's voice by applying a formant shift ratio of 0.80 and a median pitch value of 85 Hz; no change was made to the pitch variation. Finally, a male talker named Peter was created by applying a formant shift ratio of 0.70 to Sheila's speech and setting a median pitch value of 126 Hz. Note that no changes in stimulus duration were introduced during this step. The pitch and formant shift ratio manipulations resulted in a plausibly male voice for each transformation.

Owing to the particular way in which the VOT continua were constructed, stimuli with shorter VOTs (and therefore longer vowels) were associated with higher overall amplitude than stimuli with longer VOTs (see Allen & Miller, 2004). As a result, short-VOT tokens had a mean root-mean-square (RMS) amplitude of 0.070 Pa, whereas long-VOT stimuli had a mean RMS amplitude of 0.055 Pa. We followed the same approach as Allen and Miller to ensure that VOT was not confounded with amplitude; specifically, we created both a high-amplitude version (RMS amplitude set to 0.070 Pa) and low-amplitude version (RMS amplitude was set to 0.055 Pa) for each token, and both amplitude variants were presented throughout the experiment.

2.1.2. Stimulus pretest

Prior to conducting Experiment 1, we pretested our stimuli to ensure that our six talkers had perceptually distinct voices. For the pretest, we recruited 15 English-speaking monolinguals via the online participant recruitment platform Prolific (https://www.prolific.co/); all participants self-reported that they were currently residing in the United States, had normal or corrected-to-normal vision, and had no hearing difficulties and no language-related disorders. Participants completed a short screening test to ensure that they were wearing headphones (Woods, Siegel, Traer, & McDermott, 2017). In this test, participants must decide which of three tones is quietest; critically, one tone is presented 180 degrees out of phase across stereo channels, such that it is judged to be relatively quiet when presented over loudspeakers but not when presented via headphones. Thus, performance on this task differs depending on whether listeners are wearing headphones or listening over their computer speakers. Three participants failed the headphone screener twice and so were excluded from analyses, yielding a final sample of 12 (9 female, 3 male; mean age: 29 years, age range: 19–44 years). The experiment was programmed using the Gorilla experiment builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020), which is well-suited for online experiments. All procedures were approved by the University of Connecticut Institutional Review Board. Each participant provided informed consent prior to participating and received monetary compensation for their time.

During the pretest, listeners were first familiarized with each talker, then trained to associate each talker's voice to their name, and finally tested on their ability to identify each talker from their voice. The pretest was blocked by talker sex, with half the participants completing all three phases with the male talkers before hearing the female talkers, and half the participants completing the pretest with the female talkers first.

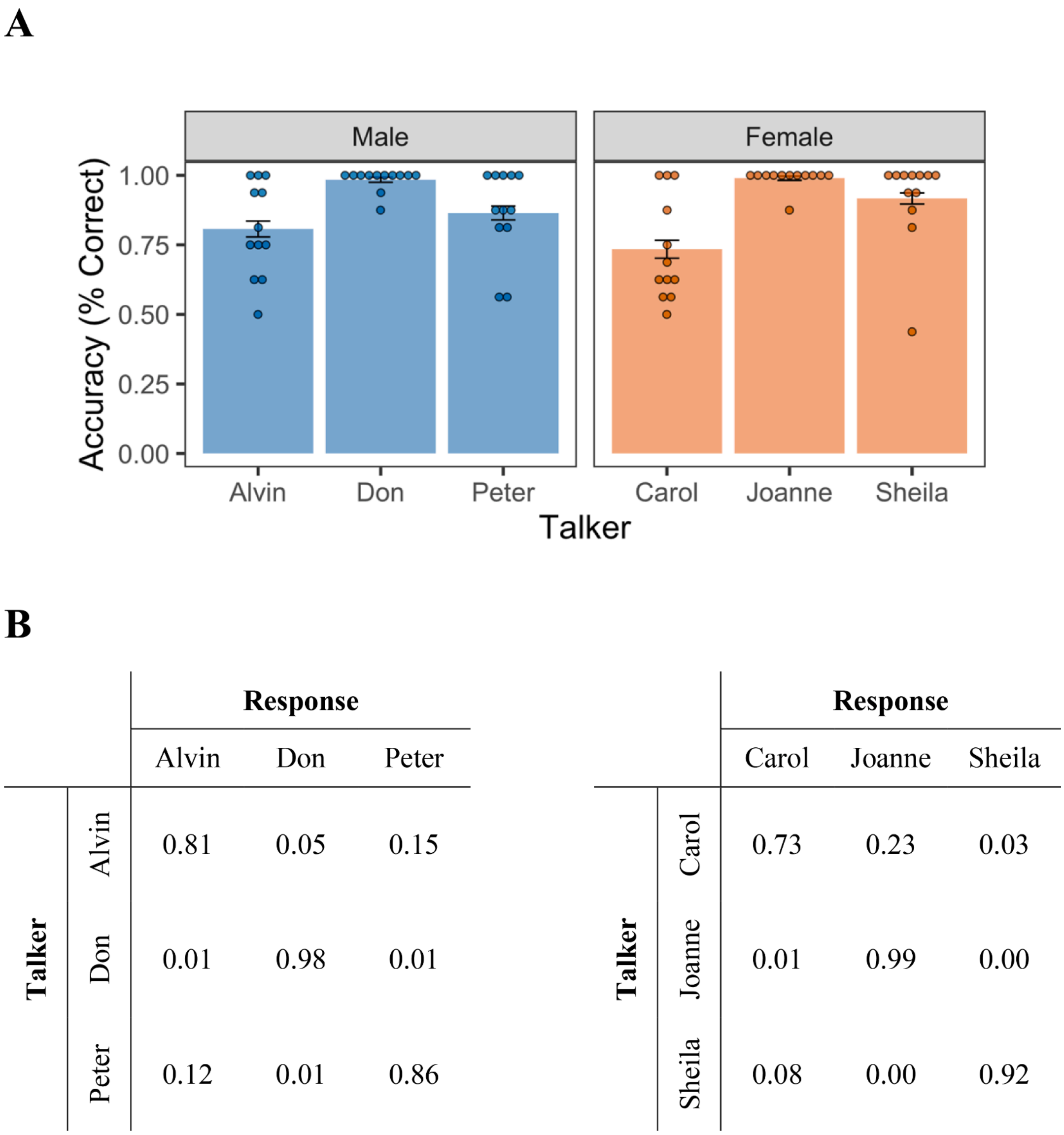
During the initial familiarization period, listeners heard four productions from each talker, with each talker saying a different word. For half the listeners, each male talker produced an item with word-initial voicing (i.e., Alvin said *bowl*, Don said *dime*, and Peter said *gain*), and each female talker produced an item that began with a voiceless consonant (i.e., Carol said *pole*, Joanne said *time*, and Sheila said *cane*). For the other half of the listeners, the male talkers produced the items beginning with voiceless consonants and the female talkers produced the items with word-initial voiced consonants. Familiarization was blocked by talker and the order of the talkers was fixed. The familiarization period had an inter-stimulus interval (ISI) of 1000 ms.

The training task was a 3-alternative forced choice task in which listeners heard a single word on each trial and were asked to identify the talker from the set of talkers of that sex. Listeners were told if they were correct or incorrect, and if they were incorrect, they were told what the correct answer was. To ensure that listeners did not learn to distinguish the talkers based solely on which word they were producing, we also included some stimuli that were not used in Experiment 1. Specifically, we also included productions of *gain* and *cane* spoken by Alvin and Carol as well as productions of *bowl* and *pole* produced by Peter and Sheila; these stimuli were constructed following the same approach described above, and for each talker, listeners heard an equal number of productions of the possible words (e.g., they heard Alvin saying *gain* just as often as they heard him saying *bowl*). Listeners completed a total of 96 trials (16 per talker), split between the two training blocks (with only talkers of the same sex presented in a given training block). Each talker produced an equal number of voiced and voiceless words, and of the voiceless tokens, half had a short VOT and half had a long VOT. There was an ISI of 1000 ms.

The test phase was identical to the training phase, except listeners did not receive feedback on the accuracy of their responses. As shown in Fig. 1A, participants had high accuracy during the test phase, with mean accuracy greater than 70 % for all talkers; note that chance-level accuracy is 33 %. Confusion matrices, shown in Fig. 1B, also indicate that participants were generally accurate in identifying the talkers, though there were some asymmetries in their errors. For instance, Carol was sometimes (23 %) misidentified as Joanne, but Joanne was rarely (1 %) identified as Carol.

2.1.3. Participants

For Experiment 1, we recruited 36 participants via the online system Prolific. These participants did not participate in the stimulus pretest described in Section 2.1.2. All Experiment 1 participants self-reported being English-speaking monolinguals residing in the United States



		Response		
		Alvin	Don	Peter
Talker	Alvin	0.81	0.05	0.15
	Don	0.01	0.98	0.01
	Peter	0.12	0.01	0.86

		Response		
		Carol	Joanne	Sheila
Talker	Carol	0.73	0.23	0.03
	Joanne	0.01	0.99	0.00
	Sheila	0.08	0.00	0.92

Fig. 1. We conducted a pretest to ensure that our six talkers had perceptually distinct voices. On each test trial, listeners had to identify who was speaking from among the set of same-sex talkers. (A) Accuracy on talker identification pretest. Talker names are shown on the x-axis and percent accuracy on the y-axis. Dots represent individual subject data. Error bars indicate standard error of the mean. (B) Confusion matrices for the talker identification pretest. Rows indicate which talker was speaking, and columns indicate participants' responses. Proportions in a row may not sum to 1 due to rounding.

with normal or corrected-to-normal vision. Participants reported that they did not have any hearing difficulties or language-related disorders. Five participants failed the headphone screening test twice and so were excluded. Data from one additional participant were excluded to equate the number of participants in each counterbalancing condition. Thus, 30 participants (12 female, 18 male; mean age: 32, age range: 20–64)[1] were included in the analysis; this sample size was based on previous studies using this paradigm (Allen & Miller, 2004; Myers & Theodore, 2017; Theodore & Miller, 2010), which observed the behavioral effect of interest with smaller samples (range: 17–20). In selecting this sample size, we were also guided by a set of previous speech perception studies (Bestelmeyer, Belin, & Grosbras, 2011; Heimrath, Spröggel, Repplinger, Heinze, & Zaehle, 2019; Kennedy-Higgins et al., 2020; Meyer, Elsner, Turker, Kuhnke, & Hartwigsen, 2018; Nixon, Lazarova, Hodinott-Hill, Gough, & Passingham, 2004; Romero, Walsh, & Papagno, 2006; Smalle, Rogers, & Möttönen, 2015) that observed TMS effects with a mean sample size of 18 (range: 6–48).

All procedures were approved by University of Connecticut Institutional Review Board. Participants provided informed consent prior to beginning the experiment and received monetary compensation for their time.

2.1.4. Procedure

Experiment 1 consisted of three blocks, and listeners heard a different pair of talkers (and thus a different continuum) in each block. Block order (Alvin/Carol, Don/Joanne, Peter/Sheila) was counterbalanced using a Latin square. Critically, the talkers in a single block had a different characteristic VOT for their voiceless stop consonants; the specific VOTs are provided in Table 1. To discourage generalization across blocks, the characteristic VOT for talkers of the same sex

[1] Experiment 1 used a larger age range (20–64) compared to Experiment 2 (19–35). To ensure that differences in age range did not drive the results seen in Experiment 1, we also conducted an analysis of the Experiment 1 data that only included participants aged 35 and younger. This analysis (N=22) yielded the same patterns of significance as compared to the full sample and is reported in Supplementary Materials.

alternated across blocks (always short-long-short for the female talkers and long-short-long for the male talkers).

At the start of each block, listeners were told that they would be exposed to two talkers who differed in how they produced a particular speech sound (e.g., the /p/ sound in *pole*). They were told that their job was to learn the unique way that each talker produced this sound. During an initial familiarization period, listeners heard eight tokens from each talker (four voiced and four voiceless). For the voiceless tokens, listeners only heard the variants that were typical for the talker. The talker's name was shown on screen as each stimulus played, and there was an ISI of 1000 ms. Stimuli were blocked by talker, with listeners always hearing the male talker first, and the order of items produced by each talker was randomized.

Following the familiarization phase, listeners completed a training phase and test phase (schematized in Fig. 2A). During the training phase of each block, listeners performed a 4-alternative forced choice task. For each stimulus, listeners made a keyboard response to indicate both who was talking and what word they said. Feedback (a green check for correct responses, a red x for incorrect responses) was shown on screen for 500 ms after listeners made their response, and there was a 1000 ms interval between trials. In total, training consisted of 96 trials (48 per talker), with trials presented in random order. Listeners heard an equal number of voiced and voiceless tokens from each talker, and they heard an equal number of high-amplitude and low-amplitude versions of each token.

During the test portion of the block, listeners heard only the female talker. We opted to test on only one talker's voice per block for two reasons, both related to potential effects of TMS being investigated in Experiment 2. First, a protracted test phase would give participants additional exposure to talker-atypical variants, potentially attenuating learning effects and possibly also encouraging generalization across sets of talkers; while this issue could in theory be ameliorated by a longer training phase, practical considerations related to the number of TMS pulses that can safely be delivered in a single session made this impractical (Rossi et al., 2009). Secondly, testing on only one voice allowed us to reduce the number of variables that would need to be counterbalanced (e.g., the order in which voices were tested), thereby removing a potential source of between-subject variability and improving our ability to observe potential effects of stimulation site (which we manipulated within participants, as described below) with a relatively small number of trials, as necessitated by safety considerations. On each trial, listeners heard a short-VOT and a long-VOT variant (with the order of variants counterbalanced) and were asked to indicate which was more typical of the talker. As noted in Table 1, the VOT heard during test was not exactly the same as the ones heard during training; for instance, if listeners had heard Sheila producing 80 ms and 90 ms variants of /k/ during training, the test phase would involve deciding whether an 85 ms or 175 ms variant was more typical of Sheila. The amplitude of the tokens was held constant within each trial. Listeners completed 32 trials during each test phase.

2.2. Results

Performance on the training task is visualized in Fig. 2B. In analyzing the training data, we separately assessed listeners' ability to identify the talker (regardless of whether they were correct in identifying which word was said) as well as their ability to determine which word was said (regardless of whether they were correct in identifying the talker). Listeners were highly accurate in the talker decision and the phonetic decision, regardless of the talker or their typical VOT (mean accuracy >91 % in all cases).

Because a short-VOT variant might be more easily confused with a voiced stimulus (compared to a long-VOT variant), we statistically assessed the influence of the talker's typical VOT (i.e., whether the talker produced voiceless stops with a short or long VOT) on the training task; separate models were conducted for talker identification performance and phonetic identification performance. These models were implemented in R (R Core Team, 2019) using the "mixed" function in the *afex* package (Singmann, Bolker, Westfall, & Aust, 2018). This function fits a mixed-effects model to the data (using the *lme4* package; Bates, Maechler, Bolker, & Walker, 2015) and evaluates the significance of each fixed effect by comparing the full model to a reduced model without that fixed effect. Here, each model included a fixed factor for Typical VOT (long/short, sum-coded) and random intercepts for each subject and for each talker. We specified a binomial family with a logit link, and we used likelihood ratio tests to evaluate significance. The talker's typical VOT did not influence performance on the talker identification component of the task, $\chi^2(1) = 0.00$, $p = 0.97$, but did have a significant effect on phonetic identification, $\chi^2(1) = 22.51$, $p < 0.0001$; this latter effect was driven by slightly less accurate responses when the talker had a short VOT (mean: 0.95, SD: 0.22) compared to when the talker had a long VOT (mean: 0.97, SD: 0.16). That is, short-VOT variants were more likely to be confused with voiced tokens, but long-VOT variants were mislabeled relatively less often.

Mean overall accuracy in the test phase was 68.9 % (SD: 16.0 %), and results from the test phase are plotted in Fig. 2C. Visually, it is clear that participants were more likely to select the long-VOT variant as the more typical one when the talker had previously produced long-VOT variants during training. To evaluate this statistically, test data were submitted to a linear mixed effects regression that assessed how fixed factors of Talker (Carol, Joanne, Sheila; sum-coded) and Typical VOT (long/short; sum-coded) influenced whether participants selected the long-VOT variant. To select our random effect structure, we began with the maximal random effect structure that converged (Barr, Levy, Scheepers, & Tily, 2013) and used a backward-stepping procedure to identify whether we could use a simpler model structure without significantly compromising model fit (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). In this way, we selected a random effects structure with random by-subject slopes for Typical VOT as well as random by-subject intercepts. We observed a significant effect of Typical VOT, $\chi^2(1) = 15.12$, $p = 0.0001$, driven by more long-VOT responses if the talker's characteristic VOT was long (mean: 0.56, SD: 0.50) than if it was short (mean: 0.24, SD: 0.43). No other effects were significant ($p > 0.19$).

2.3. Discussion

In Experiment 1, listeners were exposed to pairs of talkers that differed in how they produced their voiceless stop consonants. During training, participants demonstrated near-ceiling performance in their ability to identify who was talking and what word they were saying. Recall that for the TMS experiment, we are particularly interested in whether stimulation at *training* influences the extent of learning what variation is typical of a talker at *test*, rather than whether stimulation influences a listener's ability to perform phonetic identification and/or talker identification (as assessed during training). As described in the introduction, we hypothesized that TMS to right superior temporal cortex should impact listeners' ability to determine what phonetic variation is typical (or atypical) of each talker (i.e., to *link* talker information with phonetic detail), not that TMS should influence listeners' ability to encode talker or phonetic detail. Note that if accuracy on the Experiment 1 training task had been below ceiling, then any potential effects of TMS observed in Experiment 2 could be driven (at least in part) by disruptions to the earlier processes of encoding talker information or phonetic detail, rather than being driven by specific disruptions to the process of learning which phonetic variant is typical or atypical of a given talker's idiolect. For this reason, the near-ceiling performance on the training task in Experiment 1 is not a cause for concern.

Furthermore, results from the test phase indicate that participants were able to learn the phonetic idiosyncrasies of multiple talkers, as measured by their ability to explicitly identify whether a short VOT or a long VOT was typical for each talker's productions of voiceless stop consonants. Specifically, listeners were significantly more likely to select

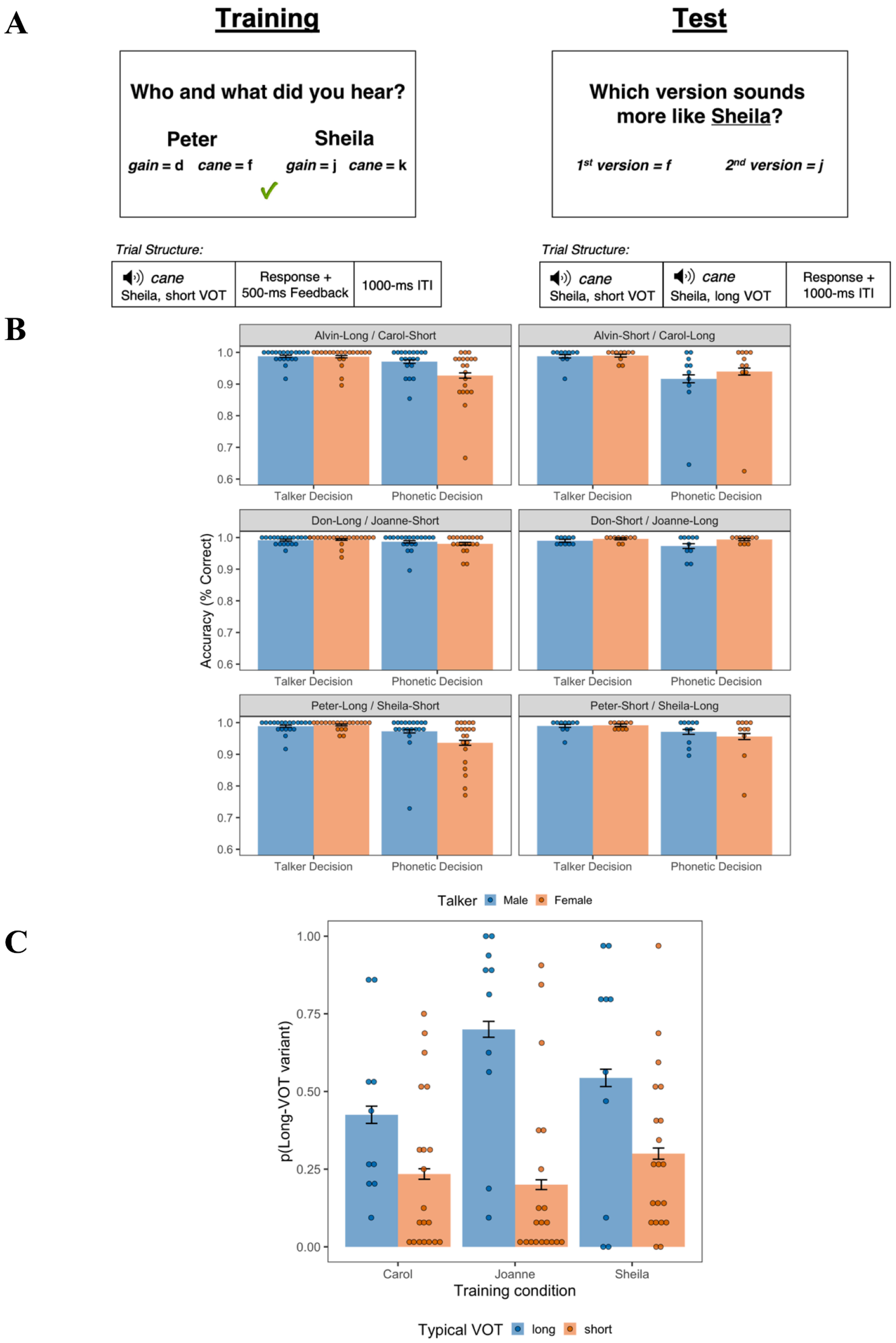


Fig. 2. Experiment 1 design and results. (A) Participants completed a 4-AFC training task that involved simultaneous talker and phonetic identification, followed by a test phase where participants were queried on which phonetic variant was typical of a talker. (B) Performance on the training task, separately considering whether listeners were accurate in identifying who was talking ("Talker Decision") and which word they said ("Phonetic Decision"). Accuracy values are shown on the y-axis. Each row shows performance on a different block. In plots on the left, the female talker produced voiceless stop consonants with a short VOT, and in plots on the right, she produced these consonants with a long VOT. Dots represent individual subject data. Error bars indicate standard error of the mean. (C) Results from the test phase of Experiment 1, showing the probability that a listener selected the long-VOT variant (y-axis) as a function of the talker (x-axis) and whether the talker produced voiceless stops with long (blue bars) or short (orange bars) VOTs during training. Dots represent individual subject data. Error bars indicate standard error of the mean.

a long-VOT variant as typical of the talker if that talker had previously produced voiceless stops with long VOTs than if she had produced them with short VOTs. Critically, participants did not exhibit ceiling- or floor-level performance on the typicality judgment task during Experiment 1, with a mean overall accuracy level of 69 % on the typicality judgments; this accuracy level allows us to measure both TMS-related enhancements or disruptions in talker-specific phonetic learning. Finally, the fact that we observed robust talker-specific learning for all talkers we tested suggests that this is a valid paradigm for our TMS experiment (Experiment 2).

However, it is worth noting that when the talker had previously produced long-VOT voiceless stop consonants, listeners appeared to be close to chance in their tendency to select the long-VOT variant, as illustrated in Fig. 2C. We verified this through a one-sample *t*-test (vs chance) conducted on the subject-by-subject proportions of long-VOT responses in the long-VOT condition, $t(29) = 0.90$, $p = 0.37$. We suggest that this result reflects listeners' general preference for short VOTs, as these are more typical of voiceless stop consonants in general. That is, short-VOT variants are a better fit to the English /p/, /t/ and /k/ phonetic categories than are long-VOT variants. Nonetheless, the fact that

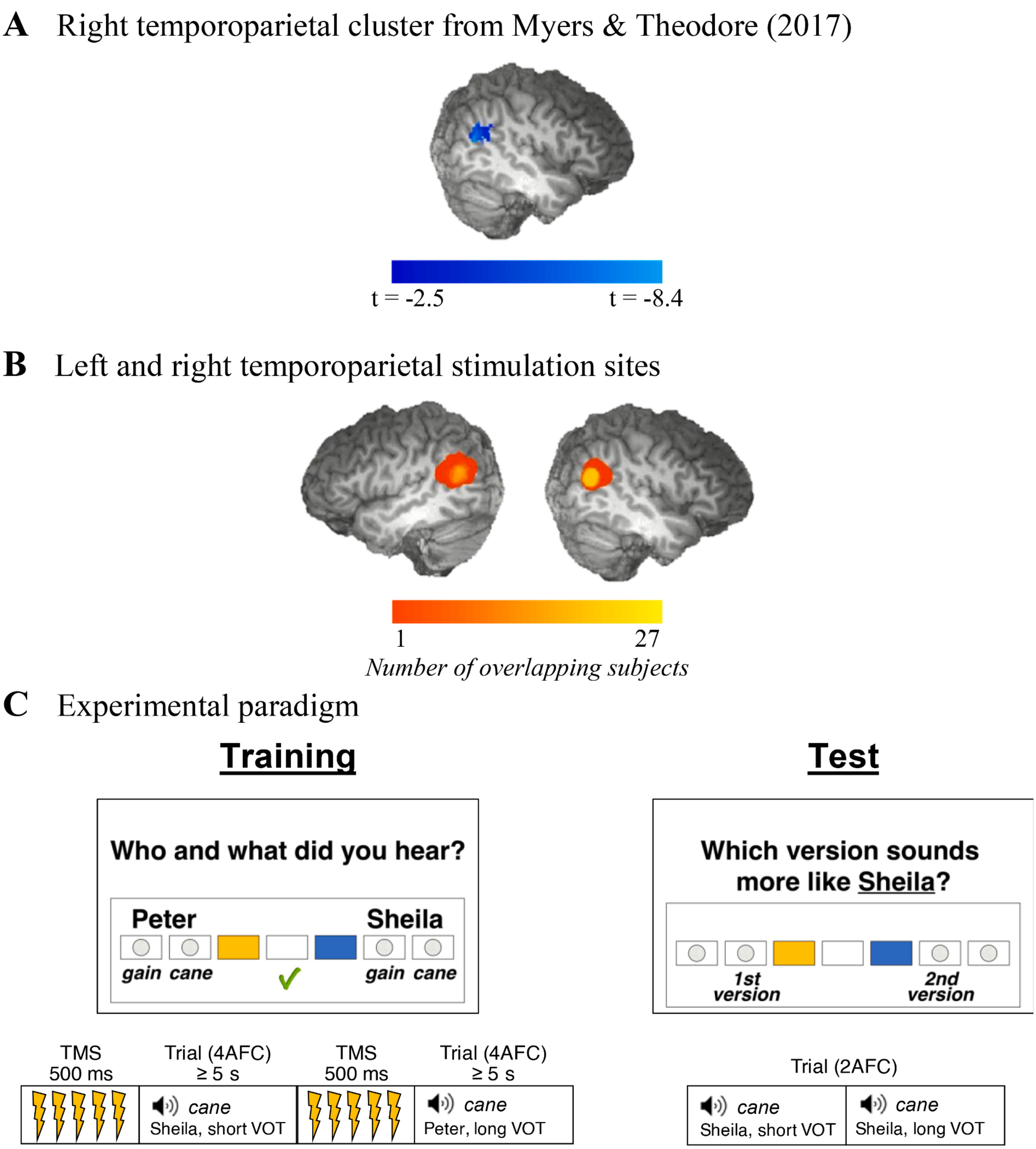


Fig. 3. (A) Myers and Theodore (2017) found that the activation of right posterior temporal cortex during an in-scanner phonetic categorization task was modulated by whether a phonetic variant was typical or atypical of the talker. (B) The current TMS experiment involved three stimulation sites: the right temporal region identified by Myers and Theodore, the homologous left hemisphere region, and scalp vertex (control site; not visualized here). (C) For each stimulation site, participants completed a training (4AFC) task and a test (2AFC) task. Sample screen displays are shown, illustrating how responses mapped onto the seven-button response box used in the experiment. (Participants responded by pressing the buttons with foam circles.) Stimulation was administered prior to each training trial, with no TMS administered during the test phase.

listeners made more long-VOT responses in the long-VOT condition than in the short-VOT condition suggests that they did indeed learn what was typical of the talker (even if they were hesitant to make long-VOT responses in general because of the low base rate of long-VOT voiceless stops in English). Based on this result, however, we decided that in Experiment 2, we would orally emphasize to participants that they should make their response with regard to which test variant was more typical of the talker rather than more typical of the phonetic category. This would be easier to do in Experiment 2 as it was conducted in person, unlike Experiment 1.

Additionally, we note that there was no significant interaction between Talker and Typical VOT in Experiment 1, suggesting that the degree of learning was comparable across all three pairs of talkers. Visually, however, Fig. 2C suggests that the degree of learning may have been larger for some talkers than for others; when listeners heard Joanne, for instance, they made long-VOT responses 70 % of the time to long-VOT variants and 20 % of the time to short-VOT variants, but when listeners heard Carol, they made long-VOT responses 43 % of the time to long-VOT variants and 23 % of the time to short-VOT variants. Despite the lack of a significant statistical interaction between these factors, we decided to err on the side of caution and opted to counterbalance which set of talkers (Alvin/Carol, Don/Joanne, Peter/Sheila) was associated with which particular stimulation site (RMTG, LMTG, vertex) in Experiment 2.

3. Experiment 2

In Experiment 2, participants were tasked with learning the phonetic signatures of three pairs of talkers, as in Experiment 1. In previous work, Myers and Theodore (2017) found that a posterior RMTG cluster was sensitive to whether the phonetic variant heard during a phonetic categorization task was typical or atypical of a talker; this cluster is shown in Fig. 3A. In the current study, rapid TMS was delivered to a different stimulation site (RMTG, LMTG, vertex) during each training block; these clusters are visualized in Fig. 3B. Each subject received stimulation at all three sites over the course of the experiment, with site order counterbalanced across participants. Of interest was how TMS at each site would affect performance during the subsequent test phase; we hypothesized that TMS to the right MTG would influence a participant's ability to determine what phonetic variation was typical of each talker, as measured during test. Note that we opted to compare performance to a control (vertex) site instead of applying sham stimulation; the application of TMS to the specific temporal sites used in this study can result in participants experiencing mild jaw twitches due to direct stimulation of facial muscles, making it difficult to apply a convincing sham stimulation.

3.1. Methods

3.1.1. Stimuli

We used the same stimuli as in Experiment 1.

3.1.2. Participants

Thirty-one right-handed native speakers of American English were recruited from the University of Connecticut community. Participants reported having normal or corrected-to-normal vision, no hearing loss and no history of neurological impairment. Each participant was screened for MRI and TMS contraindications following established safety protocols (Rossi et al., 2009). Data from four participants had to be excluded due to a programming error. Analyses therefore represent data from 27 participants (20 female, 7 male, mean age: 24, age range: 19–35). All participants provided informed consent prior to participating and received monetary compensation for their time. No participants who participated in Experiment 1 participated in Experiment 2. All procedures were approved by the University of Connecticut Institutional Review Board.

3.1.3. Procedure

Experiment 2 was conducted over two sessions, both of which took place at the Brain Imaging Research Center at the University of Connecticut. During the first session, we first acquired a T1-weighted structural magnetic resonance image (unless we already had such an image on file for the participant from a previous study). Anatomical images were acquired on a 3-T Siemens Prisma scanner with a 64-channel head coil using a T1-weighted magnetisation-prepared rapid acquisition gradient echo (MP-RAGE) sequence (TR = 2400 ms, TE = 2.15 ms, FOV = 256 mm, flip angle = 8 degrees, 1 mm sagittal slices). These images were used in conjunction with the Localite TMS Navigator (Localite, St. Augustin, Germany) to monitor the TMS coil position relative to each stimulation site.

To identify the appropriate level of stimulation for each participant, we determined each person's resting motor threshold – that is, the minimal amount of stimulation that must be applied to the motor hand area to elicit a reliable muscle response in the hand. The motor hand area in the left hemisphere was identified through visual inspection of the participant's brain anatomy (Yousry et al., 1997). The muscle activity of the contralateral thumb was recorded while we stimulated the motor hand area and nearby brain regions; the location at which we elicited the strongest response was identified as the motor hotspot (Ahdab, Ayache, Brugières, Farhat, & Lefaucheur, 2016). The motor threshold at the hotspot was then determined using the Motor Threshold Assessment Tool (Awiszus & Borckardt, 2011), which includes an adaptive Parameter Estimation by Sequential Testing (PEST) procedure for determining motor thresholds. Motor-evoked potentials were recorded using a Biopac MP160 system (Biopac Systems Inc., Goleta, CA), and stimulation was delivered using a MagPro X100 TMS device with a dynamically cooled butterfly double coil in combined active and sham (Cool B-65 A/P) configuration (MagVenture, Inc., Atlanta, GA). Motor thresholding was performed during the first session for most participants, though for three participants, it was performed at the start of the second session. These motor thresholds have been shown to be reliable within a participant as well as across sessions (Varnava, Stokes, & Chambers, 2011).

Participants completed the experimental task at their second session, following a similar procedure to that of Experiment 1. The experimental paradigm is summarized in Fig. 3C. On training trials, participants made a 4-alternative forced-choice decision, indicating both who was talking (e.g. "Alvin" or "Carol") and what they were saying (e.g. "bowl" or "pole"); note that during training, listeners heard both the male and female talkers. On test trials, participants indicated which of two variants was more typical of the talker; note that during test, listeners only heard the female talkers, as in Experiment 1. On test trials, participants responded with their left hand to indicate that the first variant was more typical of the talker and with the right hand if the second variant was more typical; any response on the left side of the button box was coded as a first-variant response, and any right-side response was coded as a second-variant response.

TMS was administered online during the training portion of each block using a stimulation protocol that was consistent with established safety recommendations (Rossi et al., 2009). Specifically, prior to each training trial, we administered five biphasic burst TMS pulses at a 10 Hz frequency; this stimulation rate was based on previous studies in which 10 Hz stimulation of the temporal cortex led to impairments in vocal identity processing (Bestelmeyer et al., 2011) and speech perception (Kennedy-Higgins et al., 2020). We used a 5000 ms ITI in between training trials, in contrast to the 1000 ms ITI used in Experiment 1. Stimulation intensity was set to 90 % of the participant's resting motor threshold; this corresponded to a mean of 49 % of the maximum stimulation output (MSO), with a range of 37–65 % MSO. Occasionally, TMS to the temporal lobes resulted in participants experiencing jaw twitches, due to direct stimulation of facial muscles; though considered a negligible safety risk (Rossi et al., 2009), we checked in with any participant who experienced these twitches to ensure they were not experiencing severe discomfort and still wished to continue.

During each experimental block, stimulation was applied to a different site (RMTG, LMTG, vertex), with stimulation site order counterbalanced using a Latin square. For each participant, the RMTG stimulation site was defined by projecting the functionally defined RMTG cluster from Myers and Theodore (2017) from Talairach and Tournoux (1988) space onto each subject's individual anatomy using the *3dFractionize* command in AFNI (Cox, 1996); recall that Myers and Theodore found that this RMTG cluster was sensitive to whether phonetic variants were typical or atypical of a talker in their phonetic categorization task. Note that we could not ask participants in our study to complete a functional localizer, as asking participants to do the Myers and Theodore (2017) task would have required them to learn what phonetic variation was typical of each talker — exactly the process we hoped to disrupt with TMS. However, other studies have successfully observed modulatory effects of TMS after localizing stimulation sites from anatomical MRI scans (Kennedy-Higgins et al., 2020; Meyer et al., 2018; Nixon et al., 2004; Romero et al., 2006). The LMTG site was defined by identifying the homologous site in the left hemisphere, and the scalp vertex was identified visually using the Localite navigation software. To visualize the left and right stimulation sites (Fig. 3B), we drew a sphere with an 8-mm radius around each subject's stimulation site and projected each sphere into Talairach and Tournoux space; we then overlaid the different subject-specific stimulation sites.

For the experimental task, stimuli were delivered through a Focusrite Scarlett 2i2 digital audio interface (High Wycombe, England) coupled to a pair of ER-3C insert headphones with foam eartips (Etymotic Research, Elk Grove Village, IL). This setup allowed the participants to hear the stimuli while also providing hearing protection against the acoustic clicks of the TMS coil. The experiment was programmed in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012), and participants made their responses via a handheld button box.

3.2. Results

Overall performance on the training task is visualized in Fig. 4. From Fig. 4A, it is clear that performance was high across both the talker identification and phonetic identification components of the task, but strikingly, phonetic identification performance appears to have been modestly impaired when participants received stimulation to RMTG. Specifically, listeners were slightly less accurate in deciding which word they heard after RMTG stimulation (mean: 0.96, SD: 0.19) compared to LMTG (mean: 0.99, SD: 0.11) and control (mean: 0.98, SD: 0.14) stimulation.

To assess this statistically, trial-level data from the training task were submitted to logistic mixed effects regression analyses. As in Experiment 1, separate analyses considered the likelihood of correctly identifying the talker versus the likelihood of making the correct phonetic decision. For each analysis, we first fit the data using a model that included fixed factors of Stimulation Site (left/right/vertex; sum-coded) and Typical VOT (long/short). The fit of this model was compared to that of a simpler model, which just tested for a fixed effect of Stimulation Site; the simpler model was preferred only if it did not entail a significant loss in the goodness-of-fit between the model and the data (Matuschek et al., 2017). Both models included random intercepts for each subject as well as for each talker. This procedure led us to select the simpler model for the talker identification analysis and the more complex model for the phonetic identification analysis. For all models, we specified a binomial family with a logit link, and we used likelihood ratio tests to evaluate significance.

Talker identification ability was not significantly affected by stimulation site, $\chi^2(1) = 2.30$, $p = 0.32$; for all stimulation sites, mean accuracy was greater than 99 %.

While accuracy on the phonetic identification component of the task was also high, performance was influenced by our factors of interest. Specifically, a marginal effect of Stimulation Site, $\chi^2(1) = 5.60$, $p = 0.06$, as well as a significant effect of Typical VOT (long or short), $\chi^2(1) = 32.35$, $p < 0.0001$, driven by higher accuracy when the talker produced their voiceless stops with a long VOT (mean: 0.98, SD: 0.12) compared to when they produced them with a short VOT (mean: 0.97, SD: 0.18). We also observed a significant interaction between the two factors, $\chi^2(1) = 9.30$, $p = 0.01$, visualized in Fig. 4B.

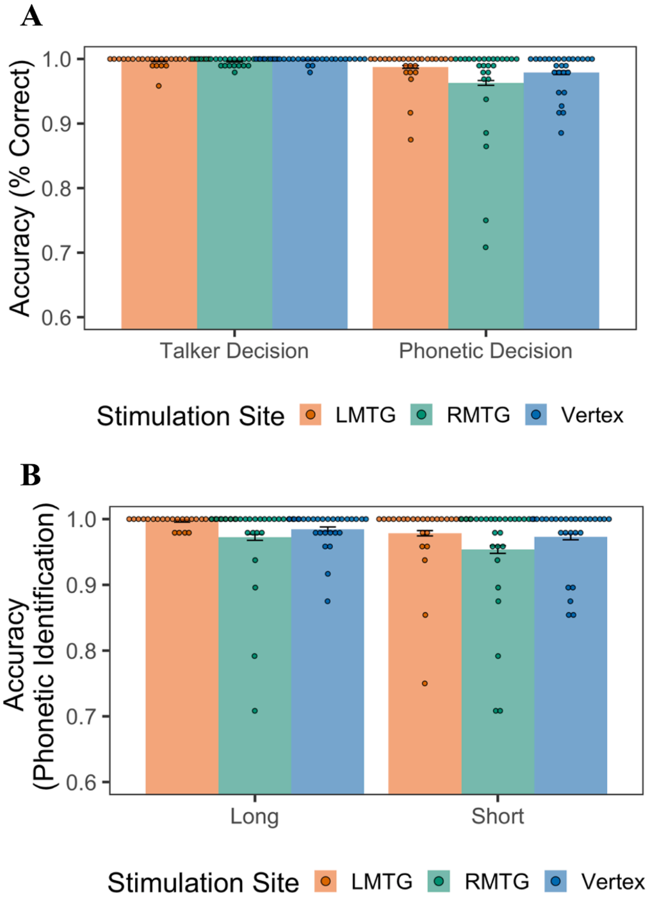


Fig. 4. Performance on the training task (during which TMS was applied) in Experiment 2. In panel (A), we separately show performance on the talker decision and phonetic decision components of the task. Panel (B) shows only the phonetic identification component of the task, separately considering whether the talker produced long-VOT voiceless stops (left bars) or short-VOT voiceless stops (right bars). For both panels, accuracy values are shown on the y-axis. Bar color indicates whether stimulation was applied to left MTG (orange bars), right MTG (green bars) or scalp vertex (blue bars) for that block. Dots represent individual subject data. Error bars indicate standard error of the mean.

To further probe the marginal effect of Stimulation Site, we conducted follow-up pairwise comparisons for each of our stimulation sites; this was implemented using the *emmeans* package (Lenth, 2021), and p values were Tukey-adjusted to correct for multiple comparisons. This analysis suggested that there was a marginal difference between the phonetic identification accuracy for RMTG stimulation compared to LMTG stimulation, $p = 0.05$, but nonsignificant differences for the other two pairs (LMTG vs vertex: $p = 0.11$; RMTG vs vertex: $p = 0.93$).

To follow-up on the significant interaction between Stimulation Site and Typical VOT, we used the *emmeans* package to evaluate the effect of Stimulation Site for each level of Typical VOT. We found that for talkers with long-VOT voiceless stops, participants were most accurate when receiving LMTG stimulation (LMTG vs RMTG: p = 0.01; LMTG vs vertex: $p = 0.02$; RMTG vs vertex: $p = 0.96$), but no pairwise differences were observed for talkers with short-VOT voiceless stops (LMTG vs RMTG: p = 1.00; LMTG vs vertex: $p = 0.95$; RMTG vs vertex: $p = 0.93$).

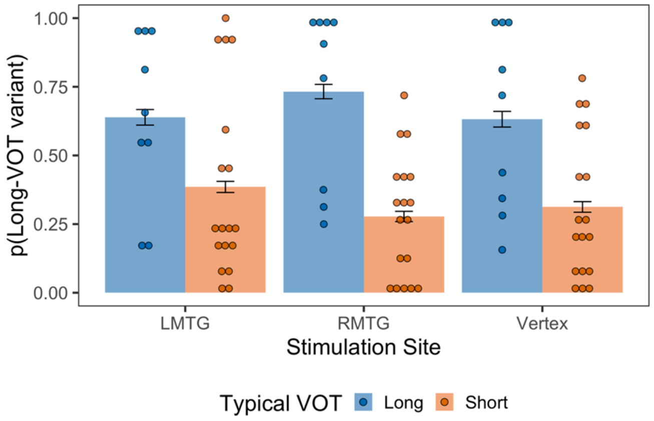


Fig. 5. Results from the test phase of Experiment 2, showing the probability that a listener selected the long-VOT variant (y-axis) as a function of the stimulation site (x-axis) and whether the talker produced voiceless stops with long (blue bars) or short (orange bars) VOTs during training. Dots represent individual subject data. Error bars indicate standard error of the mean.

Data from the test phase of each block are visualized in Fig. 5; recall that TMS was not applied on these trials. From the plot, it is clear that participants tended to learn which variant was typical, as participants were more likely to identify the long-VOT variant as more typical if the talker had previously produced their voiceless stops with a long VOT.

To assess test performance statistically, test data were submitted to a mixed effects regression that assessed how fixed factors of Stimulation Site (LMTG/RMTG/Vertex) and Typical VOT (long/short) influenced whether participants selected the long-VOT variant. The maximal random effect structure (Barr et al., 2013) included both random by-subject intercepts and random slopes for Stimulation Site. This was also the most parsimonious model structure, as a simpler random effect structure led to a significant reduction in model fit (Matuschek et al., 2017). We observed a non-significant effect of Stimulation Site, $\chi^2(1) = 0.43$, $p = 0.81$ and a significant effect of Typical VOT, $\chi^2(1) = 18.60$, $p = 0.0001$, driven by more long-VOT responses if the talker's characteristic VOT was long (mean: 0.67, SD: 0.47) than if it was short (mean: 0.33, SD: 0.47). The interaction between Stimulation Site and Typical VOT did not reach significance, $\chi^2(1) = 4.22$, $p = 0.12$.

3.3. Discussion

In Experiment 2, listeners were tasked with learning the phonetic signatures of multiple talkers – specifically, whether they produced voiceless stops with a long VOT or a short VOT. On each training trial, listeners simultaneously indicated which talker they had heard (e.g., Peter or Sheila) and which word the talker had produced (e.g., *gain* or *cane*); prior to each training trial, TMS was delivered to the RMTG, LMTG, or scalp vertex, with a different stimulation site for each block of training. On test trials, no TMS was applied, and participants had to indicate which of two variants (a long-VOT and a short-VOT variant) was typical of the talker. Based on previous literature implicating the right posterior temporal cortex in talker-specific phonetic processing, we hypothesized that TMS applied during training would influence participants' ability to learn what variation was typical of each talker, as assessed at test.

During training, listeners were near-ceiling in indicating which talker they heard, regardless of stimulation site. Though listeners were also highly accurate in their phonetic decisions, analyses indicated a modest, marginally significant effect of stimulation, such that for long-VOT stimuli, listeners were slightly more accurate in deciding which word they heard after LMTG stimulation compared to RMTG and vertex stimulation. Notably, these long-VOT stimuli are atypical of voiceless phonetic categories, and previous studies have found that the left MTG is sensitive to the phonetic typicality of a stimulus during phonetic categorization (Blumstein, Myers, & Rissman, 2005; Myers, 2007). We speculate that the effect of LMTG stimulation for long-VOT stimuli during the phonetic categorization task may be related to the LMTG's sensitivity to the goodness of fit between the acoustic–phonetic details of a production and its phonetic category.

For Experiment 2, we were principally interested in whether TMS applied during training would affect listeners' ability to learn what phonetic variation was typical of each talker, hypothesizing that stimulation to the RMTG during training might interfere with learning. However, stimulation did not strongly influence test performance, as talker-specific phonetic learning was observed regardless of stimulation site.

4. General discussion

A burgeoning literature has implicated the right posterior temporal cortex in integrating talker detail and phonetic information (Evans & Davis, 2015; Formisano et al., 2008; von Kriegstein et al., 2010). Thus, the right posterior temporal cortex might be particularly important for adapting to the idiosyncratic ways that different talkers produce their speech sounds. Some recent evidence for this view comes from Myers and Theodore (2017), who exposed listeners to two talkers who differed in how they produced a voiceless stop consonant; one produced it with a relatively short VOT (though the sound was still unambiguously voiceless) and one with a relatively long VOT. The authors found that when listeners performed a phonetic categorization task after training, activation in the right temporoparietal cortex varied as a function of whether the phonetic variant heard was typical or atypical of that talker. In the current study, listeners were trained on talkers who differed in how they produced their voiceless stops, with TMS applied prior to each training trial. Strikingly, we observed only modest influences of TMS. Stimulation of the LMTG led to a modest improvement in listeners' ability to perform phonetic identification of long-VOT productions during training, consistent with previous studies showing LMTG sensitivity to the "goodness of fit" between a production and a phonetic category (Blumstein et al., 2005; Myers, 2007). However, there were no significant long-term consequences for a listener's ability to recognize which variant was typical of the talker following TMS to any of our stimulation sites.

One possibility is that the absence of a strong TMS effect on learning in the present work is due to the particular stimulation parameters chosen for this experiment. We note, however, that the stimulation site used in the current study was well-aligned with the region identified by Myers and Theodore (see Fig. 3). Furthermore, our decision to stimulate at a frequency of 10 Hz prior to each training trial was consistent with the rate used in relevant previous studies. For instance, Bestelmeyer et al. (2011) found that 10 Hz stimulation of the right anterior superior temporal sulcus led to impaired performance in discriminating between vocal and non-vocal sounds, while Kennedy-Higgins et al. (2020) found that 10 Hz stimulation to either the left STG or right STG impaired listeners' ability to identify words spoken against a noise background. In the present study, stimulation intensity was calibrated to 90 % of each individual's resting motor threshold, leading to stimulation intensities ranging from 37 to 65 % (mean: 49 %) of the maximum stimulation output (MSO). This is comparable to the stimulation intensities used by Bestelmeyer et al. (range of 55–60 % MSO) and Kennedy-Higgins et al. (who used an intensity of 40 % MSO for all participants).

We suggest that the relatively modest effects of TMS in the current study may therefore be due not to the specific stimulation parameters but rather to the task itself. Here, we had hypothesized that stimulation of the RMTG might impair a listener's ability to explicitly indicate whether a phonetic variant was typical or atypical of a talker; notably, however, the RMTG cluster identified by Myers and Theodore (2017) showed differential activation during a phonetic categorization task, not an explicit talker typicality judgments in the scanner. Indeed, a number of other studies have also identified regions within the RMTG that are sensitive to talker-specific phonetic detail in tasks when listeners must make explicit phonetic judgments (Myers & Mesite, 2014). Thus, the RMTG may play a causal role in phonetic identification, especially when phonetic details differ across talkers, rather than in explicitly judging whether a particular phonetic variant is typical of a given talker. It would therefore be informative to test for effects of TMS to the RMTG with a slightly different paradigm, such as one in which listeners learn what phonetic variation is typical of a talker and then hear both talker-typical and talker-atypical variants during a phonetic categorization task. More generally, it is clear that in order to fully assess a causal role for right temporal regions in talker-specific phonetic processing, it will be necessary to examine the impact of TMS in other listening paradigms as well.

An additional consideration is the timing of stimulation relative to the process of interest. In the current study, TMS was applied prior to every training trial, though we were primarily interested in effects during the test portion of each block (when listeners made talker typicality judgments). This was by design; our goal was to test whether recruitment of the RMTG is necessary for updating a listener's beliefs about how a talker produces their speech sounds. Because this belief-updating process takes place during training, we decided to apply TMS immediately prior to each training trial. Future work might examine the impact of TMS applied immediately prior to each test trial, instead of (or in addition to) prior to each training trial. However, it might be that the relative timing of TMS (i.e., whether it is applied during training or test) does not strongly influence performance. During both the training and test portions of each block, listeners must access their beliefs of how each talker produces their speech sounds, whether to update these beliefs (during training) or to use them to guide a typicality judgment (during test); if the same cognitive process is at play during both training and test, then the specific decision of when to apply TMS may not be hugely consequential.

In any study where only modest effects are observed, it is important to address the issue of statistical power; it may be the case that stimulation to the RMTG could in theory influence how well listeners learn what phonetic variants are typical of a talker but that we simply did not have the appropriate number of participants to detect the effect. The lack of prior literature on TMS effects in similar paradigms made it difficult to do a principled power analysis; in designing the present study, our strategy instead was to exceed the mean sample size of prior studies that used TMS to affect speech perception (Bestelmeyer et al., 2011; Heimrath et al., 2019; Kennedy-Higgins et al., 2020; Meyer et al., 2018; Nixon et al., 2004; Romero et al., 2006; Smalle et al., 2015). It is certainly possible that a lack of statistical power may underlie the lack of TMS effects in the present study, and future work would be needed to more precisely determine the statistical power of our study. Nevertheless, we believe that the lack of a strong TMS effect in the present study is striking in and of itself, as such a finding suggests that if the RMTG does play a role in adapting to talker-specific phonetic idiosyncrasies, its role may be relatively small.

More generally, we believe the most likely explanation of the current results is that talker-specific phonetic processing is largely accomplished by both the left and right hemisphere. Such a view is consistent with previous fMRI data showing that phonetic information and talker information are simultaneously represented by both the left hemisphere and the right hemisphere (Evans & Davis, 2015; Formisano et al., 2008; von Kriegstein et al., 2010). As such, if the recruitment of one hemisphere is impaired (e.g., by TMS), a listener can still use the other hemisphere to perform talker-specific phonetic processing. Thus, even though current neurobiological accounts posit that phonetic processing principally involves the left hemisphere (Hickok & Poeppel, 2007) and that the system for processing vocal identity is largely right-lateralized (Maguinness et al., 2018), it is not the case that phonetic processing falls solely within the purview of the left hemisphere and that talker processing is solely a matter for the right hemisphere. Under some circumstances, vocal identity processing can entail recruitment of the left hemisphere (Perrachione, Pierrehumbert, & Wong, 2009; Roswandowitz, Kappes, Obrig, & Von Kriegstein, 2018; Salvata, Blumstein, & Myers, 2012; von Kriegstein et al., 2010), and phonetic processing often involves recruitment of the right temporal cortex in phonetic processing (Leonard, Baud, Sjerps, & Chang, 2016; Luthra, Guediche, Blumstein, & Myers, 2019; Myers, 2007).

In summary, the present study found that temporarily interfering with the recruitment of the right posterior temporal cortex did not influence talker-specific phonetic learning. These results are consistent with the view that talker-specific phonetic processing is achieved through the coordinated activity of both the left and right hemispheres (Luthra, Magnuson, & Myers, 2023; von Kriegstein & Giraud, 2004; von Kriegstein et al., 2010). That is, even though the left and right hemispheres may have preferences for different aspects of the speech signal, with the left hemisphere favoring phonetic detail and the right hemisphere favoring talker information, these are not hard-and-fast rules. Rather, the considerable degree of redundancy in what information is represented in the left and right temporal cortices allows for a remarkable degree of flexibility in the recruitment of the two hemispheres, thereby promoting robust processing of talker-specific phonetic variation.

Author Note

This research was supported by NSF 1554810 (PI: EBM), NIH NIDCD 2R01 DC013064 (PI: EBM) and NSF NRT 1747486 (PI: JSM). This research was supported in part by the Basque Government through the BERC 2022–2025 program, by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and award PID2020-119131 GB-I000. SL was supported by an NSF Graduate Research Fellowship. All stimuli, data and analysis scripts are publicly available at https://osf.io/cf9t8/.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

As noted in the manuscript, all stimuli, data and analysis scripts are publicly available at https://osf.io/cf9t8/

Acknowledgements

We thank Roeland Hancock for technical assistance with the TMS experiment as well as for helpful discussion on a previous version of this manuscript. We also thank Christian Brodbeck and Dave Kleinschmidt for their comments on a previous version of this document. Finally, we thank members of the Language and Brain Lab and the Computational Cognitive Neuroscience of Language Lab for helpful feedback throughout this project.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2023.105264.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh, Scotland: Edinburgh University Press.

Ahdab, R., Ayache, S. S., Brugières, P., Farhat, W. H., & Lefaucheur, J. P. (2016). The hand motor hotspot is not always located in the hand knob: A neuronavigated transcranial magnetic stimulation study. *Brain Topography, 29*(4), 590–597.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America, 115*(6), 3171–3183.

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

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*, 388–407.

Awiszus, F., & Borckardt, J. J. (2011). TMS Motor Threshold Assessment Tool (MTAT 2.0).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309–312.

Bestelmeyer, P. E. G., Belin, P., & Grosbras, M.-H. (2011). Right temporal TMS impairs voice detection. *Current Biology, 21*(20), R838–R839.

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: An fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience, 17*(9), 1353–1366.

Boersma, P., & Weenik, D. (2017). Praat: Doing phonetics by computer.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162–173.

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex, 25*(12), 4772–4788.

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.

Ganugapati, D., & Theodore, R. M. (2019). Structured phonetic variation facilitates talker identification. *The Journal of the Acoustical Society of America*, *145*(6), EL469–EL475.

Heimrath, K., Spröggel, A., Repplinger, S., Heinze, H., & Zaehle, T. (2019). Transcranial static magnetic field stimulation over the temporal cortex modulating the right ear advantage in dichotic listening. *Neuromodulation: Technology at the Neural Interface*, *2019*.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*(4), 131–138.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402.

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

Kennedy-Higgins, D., Devlin, J. T., Nuttall, H. E., & Adank, P. (2020). The causal role of left and right superior temporal gyri in speech perception in noise: A transcranial magnetic stimulation study. *Journal of Cognitive Neuroscience, 32*(6), 1092–1103.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics, 25*(2), 143–168.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience, 34*(1), 43–68.

Lenth, R. V. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.6.0. https://CRAN.R-project.org/package=emmeans.

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications, 7*, 1–9.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431–461.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *WORD, 20*(3), 384–422.

Luthra, S. (2021). The role of the right hemisphere in processing phonetic variability between talkers. *Neurobiology of Language, 2*(1), 138–151.

Luthra, S., Guediche, S., Blumstein, S. E., & Myers, E. B. (2019). Neural substrates of subphonemic variation and lexical competition in spoken word recognition. *Language, Cognition and Neuroscience, 34*(2), 141–169.

Luthra, S., Magnuson, J. S., & Myers, E. B. (2023). Right posterior temporal cortex supports integration of phonetic and talker information. *Neurobiology of Language, 4* (1), 145–177.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia, 116*, 179–193.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

Meyer, L., Elsner, A., Turker, S., Kuhnke, P., & Hartwigsen, G. (2018). Perturbation of left posterior prefrontal cortex modulates top-down processing in sentence comprehension. *NeuroImage, 181*, 598–604.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America, 85*(5), 2114–2134.

Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation. *Neuropsychologia, 45*(7), 1463–1473.

Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language, 76*, 80–93.

Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language, 165*, 33–44.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America, 109*(3), 1181–1196.

Nixon, P., Lazarova, J., Hodinott-Hill, I., Gough, P., & Passingham, R. (2004). The inferior frontal gyrus and phonological processing: An investigation using rTMS. *Journal of Cognitive Neuroscience, 16*(2), 289–300.

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

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1950–1960.

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

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Romero, L., Walsh, V., & Papagno, C. (2006). The neural correlates of phonological short-term memory: A repetitive transcranial magnetic stimulation study. *Journal of Cognitive Neuroscience, 18*(7), 1147–1155.

Rossi, S., Hallett, M., Rossini, P. M., Pascual-Leone, A., Avanzini, G., Bestmann, S., … Ziemann, U. (2009). Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clinical Neurophysiology, 120*(12), 2008–2039.

Roswandowitz, C., Kappes, C., Obrig, H., & Von Kriegstein, K. (2018). Obligatory and facultative brain regions for voice-identity recognition. *Brain, 141*(1), 234–247.

Salvata, C., Blumstein, S. E., & Myers, E. B. (2012). Speaker invariance for phonetic information: An fMRI investigation. *Language and Cognitive Processes, 27*(2), 210–230.

Singmann, H., Bolker, B., Westfall, J., & Aust, F.. afex: Analysis of Factorial Experiments. R package version 0.21-2. https://CRAN.R-project.org/package=afex.

Smalle, E. H. M., Rogers, J., & Möttönen, R. (2015). Dissociating contributions of the motor cortex to speech perception and response bias by using transcranial magnetic stimulation. *Cerebral Cortex, 25*(10), 3690–3698.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology, 24*(8), 689–700.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: An approach to cerebral imaging.*

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

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin and Review, 26*(3), 985–992.

Turkeltaub, P. E., & Branch Coslett, H. (2010). Localization of sublexical speech perception components. *Brain and Language, 114*(1), 1–15.

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition, 1*(2), 185–195.

Varnava, A., Stokes, M. G., & Chambers, C. D. (2011). Reliability of the "observation of movement" method for determining motor threshold using transcranial magnetic stimulation. *Journal of Neuroscience Methods, 201*(2), 327–332.

von Kriegstein, K., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*(2), 948–955.

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience, 30*(2), 629–638.

Wernicke, C. (1874). Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis. Cohn.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics, 79*(7), 2064–2072.

Yousry, T. A., Schmid, U. D., Alkadhi, H., Schmidt, D., Peraud, A., Buettner, A., & Winkler, P. (1997). Localization of the motor hand area to a knob on the precentral gyrus: A new landmark. *Brain, 120*(1), 141–157.