PNAS Nexus

PNAS Nexus, 2024, **3**, pgae354
https://doi.org/10.1093/pnasnexus/pgae354
Advance access publication 23 August 2024
Research Report

When Jack isn't Jacques: Simultaneous opposite language-specific speech perceptual learning in French–English bilinguals

Tiphaine Caudrelier[a,b], Lucie Ménard[c,d], Marie-Michèle Beausoleil[c,d], Clara D. Martin[b,e] and Arthur G. Samuel[b,e,f,*]

[a]Laboratoire d'Etude des Mécanismes Cognitifs, Université Lumière Lyon 2, 5 avenue Pierre Mendès France, 69676 BRON Cedex, Lyon, France
[b]Basque Center on Cognition Brain and Language (BCBL), Paseo Mikeletegi 69, Gipuzkoa, San Sebastian 20009, Spain
[c]Département de Linguistique, Pavillon Hubert-Aquin, A-3405, 400 rue Sainte-Catherine Est, Université du Québec à Montréal, Montréal, QC H2L 2C5, Canada
[d]Center for Research on Brain, Language, and Music (CRBLM), 2001 av McGill College, 6th Floor, Montréal, QC H3A 1G1, Canada
[e]Ikerbasque, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Bizkaia, Spain
[f]Department of Psychology, Stony Brook University, 100 Nicolls Road, Stony Brook, NY 11794, USA
*To whom correspondence should be addressed: BCBL, Paseo Mikeletegi 69, Gipuzkoa, San Sebastian 20009, Spain. Email: a.samuel@bcbl.eu
Edited By Michael Muthukrishna

Abstract

Humans are remarkably good at understanding spoken language, despite the huge variability of the signal as a function of the talker, the situation, and the environment. This success relies on having access to stable representations based on years of speech input, coupled with the ability to adapt to short-term deviations from these norms, e.g. accented speech or speech altered by ambient noise. In the last two decades, there has been a robust research effort focused on a possible mechanism for adjusting to accented speech. In these studies, listeners typically hear 15 – 20 words in which a speech sound has been altered, creating a short-term deviation from its longer-term representation. After exposure to these items, listeners demonstrate "lexically driven phonetic recalibration"—they alter their categorization of speech sounds, expanding a speech category to take into account the recently heard deviations from their long-term representations. In the current study, we investigate such adjustments by bilingual listeners. French–English bilinguals were first exposed to nonstandard pronunciations of a sound (/s/ or /f/) in one language and tested for recalibration in both languages. Then, the exposure continued with both the original type of mispronunciation in the same language, plus mispronunciations in the other language, in the opposite direction. In a final test, we found simultaneous recalibration in opposite directions for the two languages—listeners shifted their French perception in one direction and their English in the other: Bilinguals can maintain separate adjustments, for the same sounds, when a talker's speech differs across two languages.

Significance Statement

More people are multilingual than monolingual, yet most research on spoken language has focused on monolinguals. For example, almost all research on how listeners adapt to nonstandard (e.g. accented) speech has been done with monolinguals. In the current study, we investigate how bilinguals adjust to "accented" speech. In particular, we test whether bilingual listeners can simultaneously adjust their perception one way for one language that a bilingual talker speaks, but adjust it in an opposite way if the talker's nonstandard productions in their other language are different. We find that listeners can indeed do this. The results provide a much more nuanced understanding of how humans understand spoken language.

Introduction

Humans are remarkably good at understanding spoken language, despite the huge variability of the signal as a function of the talker (e.g. native vs nonnative; child vs adult), the situation (e.g. a casual conversation vs a formal talk), and the environment (e.g. in a quiet library vs in a train station). This success relies on having access to stable representations based on years of speech input, coupled with the ability to adapt to short-term deviations from these norms, e.g. accented speech or speech altered by ambient noise.

To appreciate the potential problem of an accent, imagine a situation in which Marie must understand Jacques, if Marie grew up in France and spent summers in England with British cousins, while Jacques grew up in a bilingual French–English home in Montreal. Jacques' French and his English will be notably different than Marie's, potentially imposing difficulties in understanding, but typically these initial difficulties decrease as Marie listens to Jacques. In the current study, we essentially ask whether Marie's experience with Jacques' accented French will help her to

Competing Interest: The authors declare no competing interests.
Received: May 31, 2024. **Accepted:** August 2, 2024

understand Jacques' accented English (or vice versa). This question is bound up with the question of how tightly linked the languages are in a bilingual's mind. In recent years, researchers studying bilingualism have pushed against the notion that a bilingual is simply a combination of two monolinguals. Understanding how processes in one language are connected to or are independent of processes in the other language, can clarify the nature of bilingual language use, and cognition more broadly.

In the last two decades, there has been a robust research effort focused on a possible mechanism for adjusting to accented speech, using a technique pioneered by Norris et al. (1). In these studies, listeners hear a modest number of words (typically 15–20) in which a speech sound has been altered, creating a short-term deviation from its longer-term representation. For example, words that should have /s/ (e.g. "episode") instead have a sound that is designed to be ambiguous, usually made by mixing the original sound with another (e.g. /f/). After exposure to these items, listeners demonstrate "lexically driven phonetic recalibration." In this example, recalibration would entail expanding the range of sounds that are heard as /s/ to now include sounds that previously were ambiguously /f/.

Adjusting phonemic category boundaries to match the local statistics can potentially improve word recognition. If a talker systematically produces nonstandard versions of a sound (e.g. due to an accent), recalibrated categories should improve the mapping between the input and the intended words. However, the perceptual system faces a fundamental problem: What is the best scope for the recalibration? Should the adjustment only apply to sounds in the same position (e.g. word-initial vs word-final) as the "odd" input sounds, or should the adjustment generalize over position? Should the adjustment be limited to the particular talker, or should it generalize to others who share some potentially relevant property (e.g. same native language; same sex)? Should the adjustment be limited to the particular "odd" phoneme, or should it generalize to related sounds (e.g. from one voiced stop consonant to another)? For Marie listening to Jacques, should an adjustment in one language affect the other?

The scope of recalibration critically depends upon a tradeoff between the system's stability and its flexibility; the former favors a limited scope, while the latter favors broader generalization. For our purposes, there are two domains of potential generalization that are most relevant—generalization from one language to another language, and generalization from one talker to another talker. We briefly review the relevant literatures:

Does recalibration transfer from one language to another?

We know of three studies that employed exposure to recalibration-inducing words in one language and tests for generalization to another language. All three studies used the contrast between /f/ and /s/; these sounds are acoustically quite similar across the testing languages. Reinisch et al. (2) exposed Dutch–English bilinguals to recalibration-inducing English words produced by a Dutch–English bilingual speaker. They found strong recalibration on Dutch test stimuli, demonstrating clear between-language generalization. Schuhmann (3) tested English–German and German–English listeners, using English exposure stimuli and test stimuli in both English and German. She found strong between-language generalization of recalibration for both listener groups. One published study found no transfer between languages (4). With English exposure stimuli and test stimuli in Dutch and in English, within-language (English–English) recalibration was robust, but no between-language generalization (English–Dutch) was found. As the authors noted, the subjects had an unusual language profile because as Dutch emigrants to Australia they had few Dutch interlocutors. In addition, the sample size was quite small. Collectively, the available literature indicates that recalibration in one language can apply to the same contrast in a closely related language (English, German, and Dutch are all Germanic languages). It remains to be determined how well recalibration might transfer between less closely related languages. In the current study, the participants are (Canadian) French–English bilinguals; these two languages are not as closely related as the pairs in previous studies, as French is not a Germanic language. These two languages are of course still not dramatically far apart (both are Western European languages, with some common Romance roots), but they are further apart than the Germanic pairs tested previously.

Does recalibration transfer from one talker to another?

There are many demonstrations of this type of generalization, but there are also relatively well-established boundaries. To a first approximation, recalibration driven by exposure to one talker's speech will transfer to a "similar" talker; generalization is also affected by the type of speech sound being tested. Most recalibration studies have used voiceless fricatives, with a smaller literature using other sounds. Among the latter, Kraljic and Samuel (5, 6) included a cross-talker recalibration test using stop consonants (/d/-/t/) and found robust transfer between the two talkers. Notably, one talker was female and the other male, a relatively rare demonstration of transfer between voices that are not similar.

With the more commonly tested fricative stimuli, most investigations have found significant transfer for female–female or male–male cases, but weaker or no transfer for female–male or male–female tests. For example, Kraljic and Samuel (6) exposed listeners to a male talker's sounds that should push categorization of /s/ versus /ʃ/ ("sh") in one direction and to a female talker's sounds that should push categorization in the other direction, and found talker-dependent shifts (i.e. categorization moved in opposite directions, depending on the test talker). Luthra et al. (7) found the same result, using a slightly different design.

Cases of transfer across male and female talkers have been observed for fricatives in a few cases, but these generally involved unusual conditions. For example, Kraljic and Samuel (8) found no transfer from a male talker to a female talker, but did find transfer in the opposite direction, apparently driven by overlap in the acoustics for the female exposure sounds and the male test sounds. Similarly, Eisner and McQueen (9) found such transfer when they spliced fricative sounds across male and female talkers (with talker identity subjectively driven by vowel portions that were not swapped). Reinisch and Holt (10) found no transfer from a female to a male talker in one experiment, but did find transfer when they restricted the range of tokens for the male test tokens.

The most extensive tests in this domain were reported by Cummings and Theodore (11). In these experiments, listeners heard exposure stimuli in two different voices that were designed to push categorization in opposite directions, as in Kraljic and Samuel (6) and Luthra et al. (7). In multiple experiments using one voice perceived as male and one perceived as female, they found robust shifts in opposite directions, consistent with talker-specific recalibration (i.e. no generalization). In experiments using two voices perceived as female, the talker specificity was reduced, suggesting some transfer. These results are consistent with those of Tamminga et al. (12), who found robust transfer from a female voice to another female voice, but no transfer from a female voice to a male voice.

Collectively, the literature indicates that for fricative sounds, listeners make adjustments with a relatively narrow scope—the recalibration applies to the exposure talker and to voices that are similar to it. To date, the literature has mostly operationalized similarity via male vs female voices; it remains to be seen what other dimensions may count toward vocal similarity. The frequency distribution of fricative energy reflects properties of the talker's vocal tract, potentially providing the perceptual system with a basis to make talker-specific adjustments. For the voicing distinction tested with stop consonants (5, 6), the timing cue does not provide such talker-related information, potentially giving any adjustments a broader scope.

The current study

The literature indicates that recalibration tends to generalize to "similar" sources: Speakers who are similar (e.g. sharing gender) to the exposure voice are affected, and contrasts that are acoustically very similar across languages are also affected. This pattern suggests that in maintaining a balance between stability and flexibility, the perceptual system sorts inputs along some similarity metric, and recalibrates the categorization of speech only for inputs that are within a relatively narrow similarity space.

In the current study, we ask whether the language presenting nonstandard sounds is itself a potential basis for sorting. In the studies showing cross-language transfer, there was no reason for the perceptual system to distinguish between /s/ or /f/ sounds in one language versus the other, as all of the exposure stimuli were presented in only one language. If, instead, the statistical properties of sounds systematically differ across two exposure languages, then perhaps recalibration would operate separately for the two: If Jacques' production of /f/ or /s/ is shifted from the norm in French in one way, but shifted from the norm in English in a different way, is it possible for Marie to make adjustments separately for the two languages? We test whether the perceptual system of a bilingual can simultaneously shift categorization in one direction in one language but in the opposite direction in the other language. Such a result would indicate that "similarity" can be defined by the language itself, for the purpose of phonetic recalibration.

Bilingual participants underwent two exposure and test phases in the current study. The first phase was comparable to prior studies testing cross-language transfer of recalibration—exposure was with words in one language, and tests for recalibration were in both of the listeners' languages. In the second phase, listeners continued to hear items like those in the first phase, but in addition they heard words in the other language that had sounds designed to push the phonetic categories in the opposite direction. The critical test in the second phase is whether test stimuli in the two languages show recalibration in opposite directions: Can listeners simultaneously shift the /f/-/s/ distinction for a bilingual talker in one direction for one language, while shifting it in the opposite direction for the other language?

Method

Human subjects protections

All participants provided informed consent before undertaking the experiment. The project was approved by the Comité d'éthique de la recherche avec des êtres humains (CIEREH) of the Université du Québec à Montréal, Montréal, Québec Canada (approval number 2021-3385).

Participants

Ninety-three French–English bilingual speakers (age: 18–40 years; French AoA = 0; English AoA range = 0–12 years) from Montreal, Canada, were tested online. The sample size was based on the large existing literature on lexically driven recalibration. A typical study in this literature has a sample size of ~ 50, usually about 25 people for each of the two sides of a contrast. We used a larger sample here, due to a critical test being based on an interaction rather than a main effect, and anticipating attrition due to multiple factors. Of the original 93 participants, seven failed to complete Phase 1 due to hardware/software problems. Eight of the remaining 86 participants were unable or unwilling to identify members of the test continua. More specifically, the participants included in the analyses had a spread (i.e. the difference in "f" report for the endpoint /f/ item versus the endpoint /s/ item) of about 88% in English and about 90% in French. The experimental test requires this sort of clean identification in both languages. The corresponding spreads for the excluded participants were as follows: (i) English 38%, French 25%; (ii) E 19%, F 38%; (iii) E 0%, F 94%; (iv) E 25%, F 88%; (v) E 63%, F 38%; (vi) E 44%, F 100; (vii) E 94, F 38; and (viii) E 25%, F 69%. With the exclusions, there was a final sample size of 78 for Phase 1. Ten participants failed to complete Phase 2, and three were not included in Phase 2 due to corrupted data files, leaving a final sample size of 65 in Phase 2.

Listeners were paid CA$20 for their participation. Their language experience and proficiency were assessed with a subset of the Language Experience and Proficiency Questionnaire (LEAP-Q (13)). They reported using French 77% (SD = 18%) of the time and English 20% (SD = 16%) of the time on average. See Table S1 in the Supplementary Material for more information from the LEAP-Q. Fifty-three of them also reported having an L3 but using it <10% of the time. There was a headphone screening at the beginning of the online experiment to make sure they were wearing headphones or earbuds (14). Participants who did not pass the headphone check could not take part in the study.

Design

Participants were randomly divided into four groups (see Fig. 1), defined by crossing the Phase 1 exposure language (English vs French) and recalibration direction (/f/ vs /s/). In Phase 2, exposure blocks *in the other language*, *in the opposite direction*, were interleaved with smaller blocks designed to maintain the recalibration induced in the first phase. In each phase, exposure was followed by an identification test to assess categorization of /f/ and /s/ in French words and in English words.

Stimuli

All stimuli were recorded by a French–English bilingual talker, from Montreal, Canada.

Exposure stimuli

In each language, the (lexical decision) exposure stimuli consisted of 36 critical items (18 /f/-words and 18 /s/-words), 36 filler words, and 72 pseudowords. There were thus 144 stimuli per language. For /f/ recalibration exposure conditions, the /f/-words were presented with ambiguous critical segments and the /s/-words were presented in their original form; the reverse was the case for the /s/-recalibration conditions.

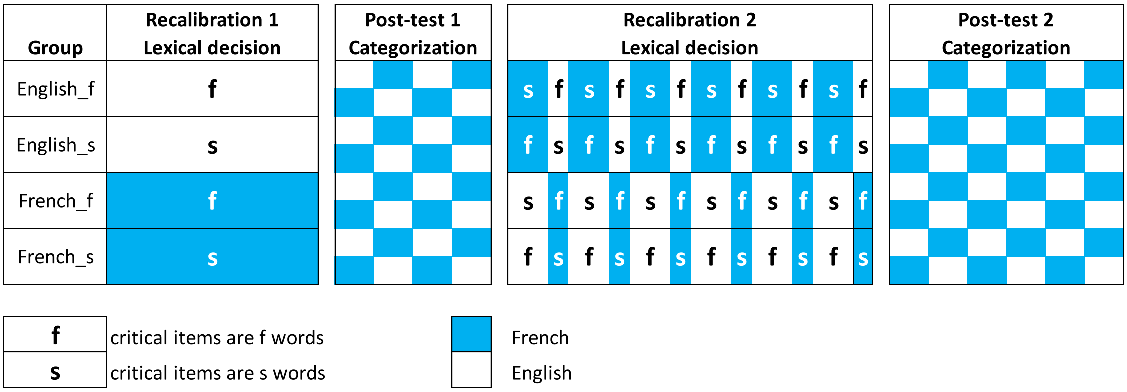

Fig. 1. General procedure by group. Each phase included an exposure task followed by a posttest. The letters "f" and "s" represent the direction of the induced recalibration. Colors (blue vs white) represent the language in which stimuli were presented.

Critical exposure words were 2 to 4 syllables long, with /f/ or /s/ appearing at the onset of the last or second-to-last syllable (e.g. "beautiful", "democracy"). Fillers were real words matched in length and frequency with critical items. Pseudowords were matched in length with real words using the pseudoword generator Wuggy (15). There were no fricative consonants (/f/, /v/, /s/, /z/, /ʃ/ or /ʒ/) in the exposure items other than the /f/ or /s/ in each critical item. An /f/-version and an /s/-version of each critical item were recorded and then mixed to obtain a 10-step continuum between them. The most ambiguous stimulus in each continuum was then selected through a pilot study with six native speakers.

Posttest stimuli

In each phase, the exposure task was followed by a posttest. For the posttest, four minimal pairs were selected in each language: *feel/seal*, *for/soar*, *fine/sign*, and *found/sound* in English and *fil/cil*, *fort/sort*, *feuille/seuil*, and *fond/son* in French. The two words of a minimal pair phonetically differed only on the /f/-/s/ contrast. A 20-step continuum was created for each pair using the same procedure as for critical items (16). Five adjacent steps near the middle of the continuum were selected based on pilot testing, chosen to include the most ambiguous region, while still including end points that were relatively unambiguous. The pilot subjects for the English stimuli were six native speakers of American English (from either the United States or Canada) with limited to no exposure to French. The pilot subjects for the French stimuli were six native speakers of Canadian French, with significant exposure to English.

For more details, stimulus construction/selection is reported in detail in Caudrelier et al. (17).

Procedure

The experiment was run online. It was coded using JsPsych, a JavaScript library for the development of Web-based experiments (18) and executed on the Jatos platform (19).

Phase 1

The first phase included an auditory lexical decision task that exposed listeners to 18 critical (recalibration-inducing) words in either English (Groups English_f and English_s) or French (Groups French_f and French_s), followed by a categorization posttest that measured /f/-/s/ perception in both languages. On the lexical decision task, participants pushed one of two keys on each trial to indicate whether an utterance was a real word or not (in the language of the task). On the posttest, for each token, participants made a forced choice between /f/ and /s/. Tokens were members of /f/-/s/ continua made with minimal pairs (see "Stimuli"). The posttest in Phase 1 consisted of four blocks of 40 trials each (4 minimal pairs × 5 steps on the continuum × 2 repetitions), alternating between French and English. The language of the first block was counterbalanced across participants. There is a tension between the need to collect enough responses to have stable results (favoring more test items) and the erosion of the recalibration effect due to the test stimuli themselves (favoring fewer test items) (20, 21); we discuss this further below.

Phase 2

The lexical decision task consisted of six blocks designed to induce an opposite shift, in the other language, than Phase 1. Each block contained three critical items in their ambiguous form and three critical items in their original form (e.g. three ambiguous s-words and three unambiguous f-words), as well as five filler words, and 11 pseudowords. These blocks were interleaved with smaller blocks in the other language designed to maintain the recalibration induced in Phase 1. These miniblocks contained one ambiguous critical item and one unambiguous critical item each (e.g. one ambiguous f-word and one unambiguous s-word), along with one filler word and three pseudowords; the items in these miniblocks had been presented during Phase 1 ((22) have shown that repeated items can be used in this paradigm). The posttest in Phase 2 had two more blocks than the posttest in Phase 1 because there was less concern about eroding the effect at that point, with no further testing after then.

Data preprocessing and analysis

The data were preprocessed and analyzed using methods that our laboratory and others have used many times. One aspect of the preprocessing relates to the erosion of the recalibration effect caused by the test items themselves that was mentioned above. A second aspect relates to the typical finding that with properly designed test continua, the categorization functions have floor and ceiling effects at their endpoints, with the effect of interest localized in the middle of each continuum. Finally, participants must be identified who either did not complete all parts of the study, or who were unable/unwilling to identify the test items.

Erosion

Recent recalibration studies ((20, 21), followed by multiple others) have shown that the recalibration effect is strongest early in testing, eroding as listeners hear more and more test items. Unlike these recent studies, the current study included test items from two languages and four different test continua within each language. A priori, the erosion could either be tied to hearing test items from a particular continuum (as in the prior studies), or it might apply across all of the different test continua, regardless of language or minimal pair. We therefore included longer posttests (in terms of the total number of test stimuli, spread across the eight test continua) than would be used if there was just a single test continuum. An initial preprocessing step was to determine, for our complex set of test stimuli, how quickly erosion occurred. We determined that in each Phase, the categorization functions based on the first two blocks (recall that there were 40 test items per block in each language, with order counterbalanced) produced the expected recalibration effects; after that, the erosion was quite substantial. This pattern indicates that the erosion is grounded in the number of test stimuli, regardless of whether they are from a single test continuum or from many. Therefore, our statistical analyses are reported for these first two blocks; in the Supplementary Material, we show the (eroded) effects from the next two blocks. In Phase 1 (Fig. S1), this means that the analyses are based on the first half of the data collected, and in Phase 2 (Fig. S2), the first third.

Floor/ceiling endpoints

Our standard data analysis procedure involves using the middle few continuum items for all of the statistical analyses. For five-step continua, as we used here, this means that analyses were based on how people identified steps 2, 3, and 4; identification of step 1 was at floor, and step 5 was at ceiling, by design (see (23), for a discussion of the strengths of this design/analysis approach).

Results

Phase 1: does recalibration transfer between a bilingual's two languages?

The 78 participants who successfully completed Phase 1 included 36 whose exposure was to critical /f/ words and 42 with exposure to critical /s/ words. On the Phase 1 posttest, each participant contributed 48 data points (2 languages × 4 minimal pairs per language × 3 steps × 2 repetitions). These data points were analyzed in a three-factor mixed ANOVA. One within-subject factor was step (2, 3, or 4), a between-subject factor was the exposure's Direction of "push" (toward /f/ or toward /s/), and a within-subject factor was Match (test items that matched the exposure language, or mismatched it). The Direction factor is the index of recalibration (i.e. is categorization different after an /f/ push than after an /s/ push), and its interaction with Match assesses whether recalibration transfers from the exposure language to the other language (a significant interaction would indicate incomplete transfer).

The identification functions collected after the Phase 1 exposure task are shown in Fig. 2, broken down by the language of the exposure stimuli (French in the top two panels, English in the bottom two), and whether the minimal pair test stimuli matched (left two panels) or mismatched (right two panels) the exposure items' language. As is clear in the figure, robust recalibration occurred overall, yielding a significant main effect of Direction, $F(1,76) = 9.149$, $P = 0.003$. The shift was numerically larger for the matched cases (14.0%) than for the mismatched cases (6.5%), but the interaction was not significant, $F(1,76) = 2.346$, $P = 0.130$. Nevertheless, the 6.5% shift for the mismatched cases by itself did not reach significance, $F(1,76) = 2.250$, $P = 0.138$, whereas for the matched cases alone the shift was reliable, $F(1,76) = 12.000$, $P = 0.001$. The effect of step was of course significant, $F(2,152) = 505.141$, $P < 0.001$. No other main effects or interactions were significant.

Given that our bilingual listeners were L1 French and L2 English, we can examine whether recalibration effects differed for French versus English exposure. In fact, effects were more robust for listeners whose exposure items were in their L1 French (top two panels), though the overall advantage for matching the exposure and test languages held for both sets of listeners. For French exposure, there was an 18% recalibration effect for matched-language (French) minimal pairs (top-left panel), versus an 11% shift for mismatched-language, English (top-right panel). For English exposure, matched-language (English) test items yielded an 11% shift (bottom-left panel), while mismatched-language (French) test items only showed a 2% shift (bottom-right panel). Overall, the French exposure groups had a robust recalibration effect ($F(1,31) = 9.669$, $P = 0.004$) and language-matching effect ($F(1,31) = 9.496$, $P = 0.004$), whereas these effects were weaker for the L2 English exposure groups (overall recalibration: $F(1,43) = 2.224$, $P = 0.143$; language-match: $F(1,43) = 3.131$, $P = 0.084$).

The analyses leave the question of cross-language transfer in a gray area. The numerical difference for the mismatched case and the lack of an interaction between Direction and Match suggest that there was transfer. However, the nonsignificant effect for the mismatched case alone undercuts any strong claim for transfer. It may be that there is weaker transfer between a Germanic language (English) and a non-Germanic one (French) than the transfer within Germanic languages (English–Dutch (2); English–German (3)). As Fig. 2 shows, for the listeners' dominant language (French), we did see smaller but reasonably robust transfer to English; for the nondominant language, the within-language recalibration was weak enough that the remaining effect with the language change was close to zero. We turn now to the key question of the current study—can a bilingual listener simultaneously shift categorization in opposite directions for the same talker, if the input provides evidence for such a split?

Phase 2: can recalibration operate in opposite directions simultaneously for a bilingual talker?

Of the 78 participants who successfully completed Phase 1, 65 successfully completed Phase 2. Of these, 35 were exposed to critical /f/ words in English and critical /s/ words in French; 30 were exposed to critical /s/ words in English and critical /f/ words in French. As in the Phase 1 PostTest, in the Phase 2 PostTest each participant contributed 48 data points (2 languages × 4 minimal pairs per language × 3 steps × 2 repetitions) to the analyses (see Fig. S2 in the Supplementary Material for the results from the corresponding 48 data points after erosion).

The data were analyzed in a mixed ANOVA. One within-subject factor was step (2, 3, or 4), a between-subject factor was the exposure's (complex) Direction of "push" (toward /f/ in English and toward /s/ in French, or toward /s/ in English and toward /f/ in French), and Language of the test items (English or French) was a within-subject factor. We also included the theoretically neutral counterbalancing factor of Order (i.e. whether a participant's exposure/test blocks were first in French or first in English). The critical statistical test is the interaction of Direction and Language: If listeners can simultaneously recalibrate their categorization in opposite directions in English and in French then the complex Direction manipulation will increase /f/ report relative to /s/ report for one language while decreasing it for the other language.

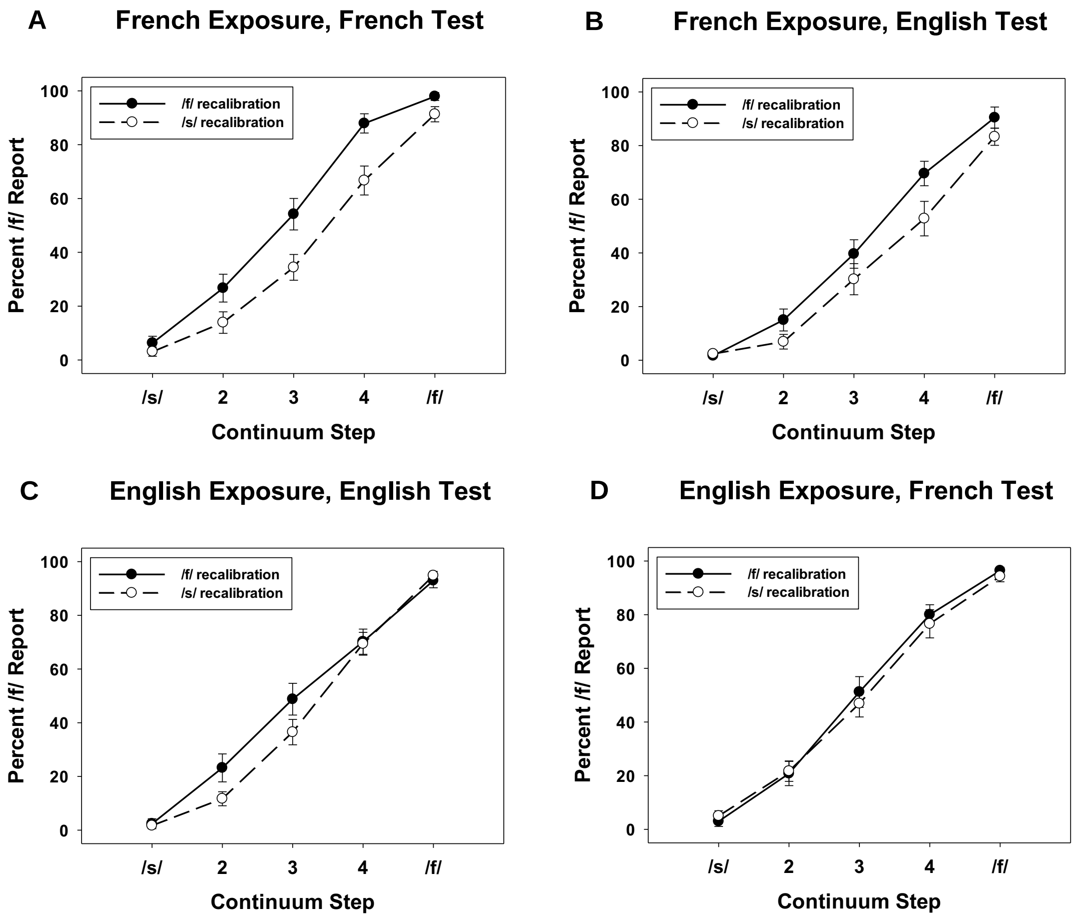

Fig. 2. Categorization functions in Phase 1. Panels A, B, C, and D show the four combinations of Exposure and Test languages. The error bars are SEs.

The categorization functions collected after the Phase 2 exposure task are shown in Fig. 3, broken down by whether the language of the test stimuli was French (left panel) or English (right panel). In each panel, the solid curve is based on the two subject groups for whom English exposure items favored /s/ and French exposure items favored /f/ (the second and third rows in Fig. 1); the dashed curve comes from the two subject groups with the opposite directions (the first and fourth rows in Fig. 1). The interaction shown in the figure (dashed curve to the right of solid curve for French test items, but dashed curve to the left of solid curve for English test items) shows that recalibration of the /f/-/s/ category distinction was simultaneously moved in opposite directions: Listeners treated the distinction one way in English, but the opposite way in French, for the same individual (i.e. the talker who

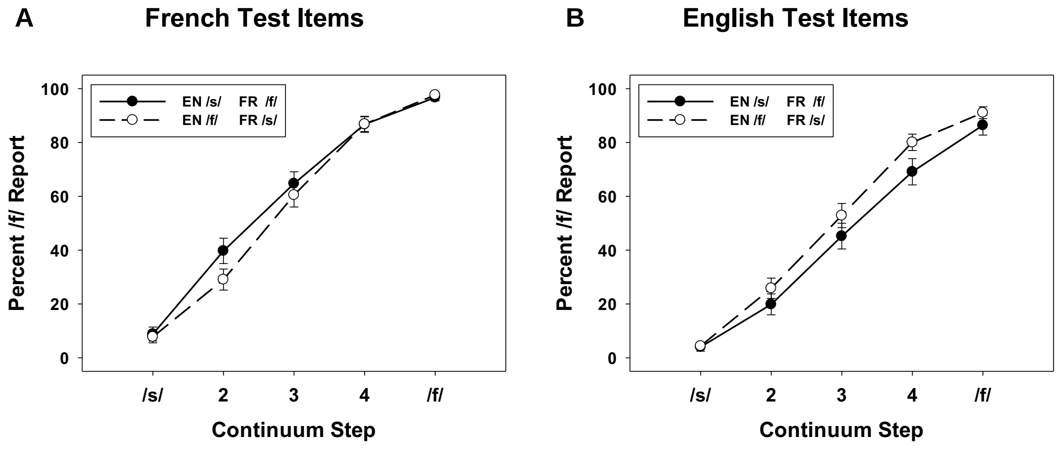

Fig. 3. Identification functions in Phase 2 for French test items (Panel A) and for English test items (Panel B). Solid symbols/lines plot results for listeners who received /s/-recalibration exposure in English, and /f/-recalibration exposure in French. Open symbols/dashed-lines plot results for listeners who received /f/-recalibration exposure in English, and /s/-recalibration exposure in French. The error bars are SEs.

produced the stimuli). This critical Language and Direction interaction was statistically reliable, $F(1,61) = 5.211$, $P = 0.026$. This interaction was not affected by Order (Language × Direction × Order: $F(1,61) < 1$), nor was there a main effect of Order, $F(1,61) < 1$. As in Phase 1, step was also highly reliable, $F(2,122) = 269.846$, $P < 0.001$. There was also a main effect of Language, with listeners categorizing more French test items as /f/ than English test items, $F(1,61) = 16.171$, $P < 0.001$. There were no other significant main effects or interactions among the three factors.

Discussion

In the Introduction, we noted that listeners rely on a combination of long-term knowledge and short-term adjustments to decode spoken language despite its substantial variability. Each phase of the current study tested how well the system uses the available information. In Phase 1, we found robust recalibration when the exposure stimuli matched the language of the test stimuli. Recalibration was notably stronger when the exposure stimuli were in the listeners' dominant language (French) than when they were in the less-dominant language. When the test items' language mismatched the language of the exposure items, we observed some recalibration, but the effect was about 8% smaller than the within-language case. Collapsing across the two languages, this size difference was not statistically reliable (i.e. the interaction was not significant), but the mismatching case by itself was also not reliable. These results suggest that across-language transfer may occur for languages that are not in the same family, but the effect is not as robust as those found in previous studies showing cross-language transfer between pairs of Germanic languages (Dutch–English (2); German–English (3)).

Our primary question was tested in Phase 2. After the listeners' Phase 1 experience with "odd" sounds in only one language, there was a change: Those odd sounds continued to be heard in the original language, but now other odd sounds were heard, in the other language. Critically, the combination now gave listeners evidence that adjustments to each language should be made in opposite directions. In terms of how a listener (Marie) adjusts to a talker (Jacques), when the only input from Jacques is in one language, Marie's adjustment applies to everything that Jacques says, even potentially to a different language. But, once Jacques has provided input in both languages, Marie is able to make adjustments for the languages separately. This pattern is consistent with an ideal adapter model (24) in which listeners update their long-term beliefs (categorization) based on the distribution of new information.

Two prior studies provide some context for the Phase 2 results. Saltzman and Myers (25) alternated exposure blocks between the two sides of an /s/-/ʃ/ contrast in English and found that they could move categorization in opposite directions from block to block. So, as in Phase 2, listeners were sensitive to input that pushed in opposite directions, within one talker, but the effects were found successively, not simultaneously; there was no need to maintain two different models for a single individual. Tzeng et al. (26) also presented listeners with opposing exposure input, again from a single talker in English. However, they blocked direction—one set of items pushed one way, followed by other items pushing the other. As in our Phase 2, they measured recalibration at the end of this mixed exposure. Given that the opposing input was all from one talker, in one language, the end result was no shift; the system had no basis to sort the conflicting inputs, whereas in our study the languages themselves offered a basis.

In the Introduction, we reviewed the literature that has examined whether exposure to recalibration-inducing stimuli from one talker will generate recalibration for stimuli produced by a different talker. Overall, the literature indicates that transfer depends on the similarity between two voices—typically there is transfer from one female talker to another, or from one male talker to another, but not from female to male or vice versa. The talker in the current study was the same female voice, regardless of whether she produced French words or English words. Thus, the acoustic match–mismatch that can explain the between-talker transfer pattern cannot explain the pattern in the current study. Instead, our results demonstrate that the language of the speech itself can be used to sort the exposure input when bilingual listeners hear bilingual talkers.

To the extent that this is true, it indicates that listeners maintain two separable models for the same bilingual talker. Individual talkers like Jacques often use two different "modes" of speaking in different languages; they may even take on different names in different language contexts, with Jacques preferring to go by Jack amongst his English-speaking friends. Listeners can tune into this, maintaining two separable (and possibly even conflicting) models for the same bilingual talker. In this sense, Jacques and Jack are not the same person talking—the listener has a model of French for Jacques, and a model of English for Jack. This allows the listener to use the details of the input speech more completely than would be possible if Jacques is always Jacques, regardless of whether he is talking in French or English. This result provides an important new constraint on models of bilingualism: The two languages are linked sufficiently to allow input in one language to affect processing in the other (Phase 1), but the organization is sophisticated enough to shift to language-specific modifications when the speech input provides a basis for doing so (Phase 2). This kind of highly adaptive system allows a bilingual to flexibly utilize the two languages, e.g. in code switching, to optimize communication.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

Support for this project was provided by the Spanish Ministry of Science and Innovation through Grant # PID2020-113348GB-I00 (A.G.S.), PID2020-113926GB-I00 (C.D.M.), and FJC2019-039005-I (T.C.). This work was supported by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S. The research was also supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Grant # 819093 (C.D.M.) and by the Social Sciences and Humanities Research Council of Canada (Insight Grant # 435-2020-1072) (L.L.M.) and the Natural Sciences and Engineering Research Council of Canada (Discovery grant #RGPIN-2020-05439) (L.L.M.).

Author Contributions

T.C.: conceptualization, data curation, investigation, methodology, and writing—review and editing; L.M.: conceptualization, supervision, funding acquisition, and writing—review and editing; M.-M.B.: investigation and writing—review and editing; C.D.M.: conceptualization, supervision, funding acquisition, methodology, and writing—review and editing; and A.G.S.: conceptualization, writing—original draft, and writing—review and editing.

Data Availability

The data files for each participant and the input files used for the ANOVAs are available via: Caudrelier (29).

References

1 Norris D, McQueen JM, Cutler A. 2003. Perceptual learning in speech. *Cogn Psychol*. 47:204–238.

2 Reinisch E, Weber A, Mitterer H. 2013. Listeners retune phoneme categories across languages. *J Exp Psychol Hum Percept Perform*. 39: 75–86.

3 Schuhmann KS. 2016. Cross-linguistic perceptual learning in advanced second language listeners. *Proc Ling Soc Am*. 1:31.

4 Bruggeman L, Cutler A. 2020. No L1 privilege in talker adaptation. *Biling Lang Cogn*. 23:681–693.

5 Kraljic T, Samuel AG. 2006. Generalization in perceptual learning for speech. *Psychon Bull Rev*. 13:262–268.

6 Kraljic T, Samuel AG. 2007. Perceptual adjustments to multiple speakers. *J Mem Lang*. 56:1–15.

7 Luthra S, Mechtenberg H, Myers EB. 2021. Perceptual learning of multiple talkers requires additional exposure. *Atten Percept Psychophys*. 83:2217–2228.

8 Kraljic T, Samuel AG. 2005. Perceptual learning for speech: is there a return to normal? *Cogn Psychol*. 51:141–178.

9 Eisner F, McQueen JM. 2005. The specificity of perceptual learning in speech processing. *Percept Psychophys*. 67:224–238.

10 Reinisch E, Holt LL. 2014. Lexically guided phonetic retuning of foreign-accented speech and its generalization. *J Exp Psychol Hum Percept Perform*. 40:539–555.

11 Cummings SN, Theodore RM. 2023. Hearing is believing: lexically guided perceptual learning is graded to reflect the quantity of evidence in speech input. *Cognition*. 235:105404.

12 Tamminga M, Wilder R, Lai W, Wade L. 2020. Perceptual learning, talker specificity, and sound change. *Papers Historical Phonol*. 5:90–122.

13 Marian V, Blumenfeld HK, Kaushanskaya M. 2007. The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J Speech Lang Hear Res*. 50:940–967.

14 Woods KJP, Siegel MH, Traer J, McDermott JH. 2017. Headphone screening to facilitate web-based auditory experiments. *Atten Percept Psychophys*. 79:2064–2072.

15 Keuleers E, Brysbaert M. 2010. Wuggy: a multilingual pseudoword generator. *Behav Res Methods*. 42:627–633.

16 McAllister Byun T, Tiede M. 2017. Perception-production relations in later development of American English rhotics. *PLoS One*. 12:e0172022.

17 Caudrelier T *et al*. 2023. Lexically-guided phonetic recalibration transfers across languages in French-English bilinguals. Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS); Prague, Czech Republic.

18 de Leeuw JR. 2015. jsPsych: a JavaScript library for creating behavioral experiments in a web browser. *Behav Res Methods*. 47(1):1–12.

19 Lange K, Kühn S, Filevich E. 2015. "Just another tool for online studies" (JATOS): an easy solution for setup and management of web servers supporting online studies. *PLoS One*. 10:e0130834.

20 Liu L, Jaeger TF. 2018. Inferring causes during speech perception. *Cognition*. 174:55–70.

21 Liu L, Jaeger TF. 2019. Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *J Exp Psychol Hum Percept Perform*. 45:1562–1588.

22 Leach L, Samuel AG. 2007. Lexical configuration and lexical engagement: when adults learn new words. *Cogn Psychol*. 55: 306–353.

23 Samuel AG, Dumay N. 2021. Auditory selective adaptation moment by moment, at multiple timescales. *J Exp Psychol Hum Percept Perform*. 47:596–615.

24 Kleinschmidt DF, Jaeger TF. 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol Rev*. 122:148–203.

25 Saltzman D, Myers E. 2021. Listeners are initially flexible in updating phonetic beliefs over time. *Psychon Bull Rev*. 28:1354–1364.

26 Tzeng CY, Nygaard LC, Theodore RM. 2021. A second chance for a first impression: sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychon Bull Rev*. 28:1003–1014.

27 Charoy J, Samuel AG. 2023. Bad maps may not always get you lost: lexically driven perceptual recalibration for substituted phonemes. *Atten Percept Psychophys*. 85:2437–2458.

28 Zheng Y, Samuel AG. 2023. Flexibility and stability of speech sounds: the time course of lexically-driven recalibration. *J Phon*. 97:101222.

29 Caudrelier T. 2024. Dataset of Jack and Jacques study [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11397999